# Extensions de la méthode d'échantillonnage indirect et son application à l'enquête dans le tourisme : M.O.R.G.O.A.T

*Jean-Claude DEVILLE[(*)] et Myriam MAUMY[(**)]*

*[(*)] ENSAI/CREST, Laboratoire de Statistique d'Enquête*
*[(**)] Universités de Strasbourg et de Rennes 2, Laboratoire de Statistique*

## 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 (O.R.T.B.) 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 (2001) en Espagne 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 son séjour. 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 cette note est de présenter une méthode qui permet de faire ce calcul.

Cette méthode va s'appuyer principalement sur la *Méthode Généralisée du Partage des Poids* (MGPP) mise au point par Lavallée (1995, 2002) et Deville (1999).

# 1. Les autres enquêtes régionales

Dans le cadre du tourisme, quelques méthodes d'enquête existent et nous allons, dans cette section, en présenter certaines très rapidement.

Une première méthode d'enquête en France est celle utilisée par la région Riviera Côte d'Azur. Cette région est principalement fréquentée par des touristes utilisant l'avion comme moyen de transport pour entrer et sortir de cette région. L'aéroport de Nice Côte d'Azur accueille la plupart des touristes de cette région et les comptages de touristes dans un aéroport sont aisés à réaliser puisque l'aéroport est un espace fermé.

Une seconde méthode correspond au cas où l'avion n'est pas utilisé comme moyen de transport. Le touriste a alors recours le plus souvent à l'autoroute du Sud qui dessert toute la région. On ne peut pas reproduire totalement ce type d'enquête pour la Bretagne. En effet, il n'existe pas d'équivalent de l'aéroport Nice Côte d'Azur, ni d'équivalent de l'autoroute du Sud avec des postes de péage qui permettent un comptage exhaustif des entrées sorties en région. Les régions PACA et Aquitaine utilisent ce type de méthode uniquement aux postes de péage autoroutier.

Une troisième méthode d'enquête en France a été utilisée en région Centre. Cette région mène des enquêtes téléphoniques auprès des offices de tourisme afin de récolter le plus d'informations possibles sur les touristes. Ce principe présente un inconvénient majeur : si la région cherche une information très particulière comme par exemple le nombre de touristes pratiquant telle activité sportive dans la région, l'office du tourisme n'est pas toujours en mesure de lui fournir. Cette enquête est complétée également par des enquêtes sur les sites touristiques de la région Centre. Cette méthode peut atteindre ses limites lorsqu'il s'agit de produire des statistiques, puisque le problème majeur est l'absence d'une base de sondage de touristes.

Un des problèmes majeurs dans le domaine du tourisme, comme nous venons de le constater avec la région Centre est l'absence d'une base complète de sondage permettant d'interroger directement les touristes. Pour contourner ce problème, une solution acceptable et réalisable est d'échantillonner des services destinés principalement aux touristes et de les interroger sur les différents lieux de ces nombreuses prestations touristiques.

Cette idée a été réalisée dans la région des Asturies. Depuis décembre 1997, le Système d'Information Touristique des Asturies (SITA) fournit un bulletin mensuel à l'état espagnol sur les statistiques du tourisme dans la région des Asturies. Un point de cette méthode a été dégagé et reste inexpliqué :

**Comment relier le poids des services échantillonnés au poids des touristes ?**

En effet, il est bien évident qu'un touriste peut utiliser une ou plusieurs fois un ou plusieurs services de la base de sondage pendant leur séjour. Une méthode permettant de faire des calculs rigoureusement existe et s'appuie principalement sur la méthode généralisée de partage des poids mise au point par Lavallée (1995, 2002) et Deville (1999). Cette dernière a été en autre appliquée dans l'enquête des sans-domiciles réalisée par l'INSEE (Ardilly et Le Blanc) en 2001.

# 2. La méthode généralisée du partage des poids

On va rappeler très brièvement le principe de la *méthode généralisée du partage des poids*. 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 définie par

$$\Theta_{AB} = [\theta_{ji}^{AB}],$$

de taille $N^A \times N^B$ où chaque élément est noté $\theta_{ji}^{AB} \geq 0$. Autrement dit, l'unité $j$ de la population $U^A$ est reliée à l'unité $i$ de la population $U^B$ à condition que l'élément $\theta_{ji}^{AB} > 0$ ; 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 la population $U^A$ 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 l'échantillon $s^A$, on identifie les unités $i$ de la population $U^B$ pour lesquelles $\theta_{ji}^{AB} > 0$. Soit $s^B$, l'ensemble des $n^B$ unités de la population $U^B$ 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 l'échantillon $s^B$, une variable d'intérêt $y_i$ est mesurée à partir de la population $U^B$. On suppose que, pour toute unité $j$ de l'échantillon $s^A$, on peut obtenir les valeurs de $\theta_{ji}^{AB}$ pour $i = 1,\ldots,N^B$ par entrevue directe ou à partir d'une source administrative. Pour toute unité $i$ identifiée de la population $U^B$, on suppose que l'on peut obtenir les valeurs de $\theta_{ji}^{AB}$ pour $j = 1,\ldots N^A$. Par conséquent, il n'est pas nécessaire de connaître les valeurs de $\theta_{ji}^{AB}$ pour la totalité de la matrice de liens $\Theta_{AB}$. En fait, on ne doit connaître les valeurs de $\theta_{ji}^{AB}$ que pour les lignes $j$ de $\Theta_{AB}$, où $j \in s^A$, ainsi que pour les colonnes $i$ de $\Theta_{AB}$ où $i \in s^B$.

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

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

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

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \qquad (2.2)$$

où $w_i$ est le poids d'estimation de l'unité $i$ de $s^B$, avec $w_i = 0$ pour $i \in s^B$. 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 $i$ 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 méthode généralisée de partage des poids. Dans celle-ci les poids sont donnés par

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A}, \qquad (2.3)$$

où

$$\tilde{\theta}_{ji}^{AB} = \frac{\theta_{ji}^{AB}}{\sum_{j=1}^{N^A} \theta_{ji}^{AB}}. \qquad (2.4)$$

De cette construction, l'estimateur de $Y^B$ est sans biais. De même, la variance de cet estimateur peut-être calculée et estimée car elle est identique à celle de

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A}, \qquad (2.5)$$

où

$$z_j = \sum_{i \in N^B} \tilde{\theta}_{ji}^{AB} y_i. \qquad (2.6)$$

# 3. L'enquête tourisme en milieu ouvert

## 3.1. Principe de l'enquête et hypothèse principale à la réalisation de celle-ci

Le principe de l'enquête est le suivant :

*« Atteindre les touristes (étrangers ou français résidant ou non en Bretagne) par le biais de services destinés à satisfaire leurs besoins élémentaires »,*
comme l'hébergement, la nourriture, les activités de loisirs, les transports.

L'hypothèse principale à la réalisation de l'enquête est la suivante :

« *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 touristiques ne consomment aucun d'entre eux.* »

## 3.2. La population d'intérêt

- Soit $G$ **un champ géographique**. Ici dans l'enquête $G$ représentera les quatre départements bretons, à savoir :
  1. les Côtes d'Armor,
  2. le Finistère,
  3. l'Ille-et-Vilaine,
  4. et le Morbihan.

- Soit $P$ **une période de référence**. Pour nous, la période de référence est 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 le champ géographique $G$ hors de sa résidence principale (nuitée).

- Pour un touriste $t$, **un séjour** est un intervalle $s$ de la période $P$ de durée le cardinal du séjour $s$ noté $|s|$, au cours duquel le touriste passe toutes ses nuits dans le champ géographique $G$ hors de sa résidence principale et, les nuits immédiatement avant ou après le séjour $s$ étant passées hors du champ géographique $G$. Par exemple il peut avoir passé ces nuits-là à sa résidence principale ou dans tout autre hébergement qui ne se situe pas dans le champ géographique $G$.

- **Un voyage** ou encore **un ménage touristique** est un ensemble de touristes partageant le même séjour et avec le même hébergement au cours du séjour**.**

- **L'unité statistique** $i$ de l'enquête est le ménage touristique.

- **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 $|s| n_v$ nuitées. Ici la population $U^B$ est donc l'ensemble des voyages dans le champ géographique $G$ au cours de la période $P$.

- **Le champ de l'enquête** : la méthodologie permet de décrire principalement le tourisme d'agrément puisque les principaux points d'enquête sont des sites de loisirs et des boulangeries. En effet, il y aura peu de touristes d'affaires attrapés car nous n'enquêterons pas dans les modes d'hébergement marchand, en particulier les hôtels qui nous permettaient d'atteindre cette cible.

## 3.3. Le plan de sondage de l'enquête

Pour utiliser la méthode généralisée de partage des poids, la population $U^A$ est constituée par un ensemble de « services ». Dans cette enquête, ceux-ci sont :

- les achats en boulangerie, constituant une première sous population de la population $U^A$ ;

- les visites de 16 sites emblématiques de la Bretagne qui se divisent en trois familles :

    a. les sites naturels,
    b. les sites patrimoniaux,
    c. les sites familiaux.

    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 sous population de la population $U^A$ ;

- les passages sortant de Bretagne au péage autoroutier de La Gravelle. La voiture caractérise 80% des séjours de non-résidents bretons. Ces passages constituent la troisième sous population de la population $U^A$.

Dans *la première sous population*, on réalise un échantillon à 3 degrés :

1. un échantillon de boulangeries : 19 boulangeries;
2. un échantillon de jours d'enquête ;
3. un échantillon de clients dans les 19 boulangeries sélectionnées à un jour donné.

Dans *la deuxième sous population*, on réalise un échantillon à 2 degrés :

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

Enfin dans *la troisième sous population*, on réalise un échantillon à 2 degrés :

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

**Remarques :** 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 catastrophique liée à l'intervention des hôteliers dans le processus de sélection des enquêtés. Cette partie du dispositif de l'enquête sera donc abandonnée et remplacée par le passage au péage autoroutier de La Gravelle.

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 acceptent les enquêtes dans leurs établissements uniquement si le groupe I en est exclu ! En revanche, l'adhésion des artisans boulangers au concept de l'enquête a été excellente.

### 3.4. Les spécificités de M.O.R.G.O.A.T.

**La première spécificité : le bâtonnage**. Le système de bâtonnage est une mini enquête mise en œuvre préalablement à la passation du questionnaire. Ce système a essentiellement **deux objectifs** :

1. déterminer les proportions des touristes étrangers, des touristes français non bretons et des touristes intra-régionaux. C'est l'objectif fondamental du bâtonnage puisque ces proportions seront utilisées lors du calcul des poids. Il est donc impossible de supprimer le bâtonnage du dispositif de l'enquête. Ces proportions sont inconnues quel que soit le lieu enquête. Par exemple, on peut obtenir le nombre de visiteurs journaliers d'un zoo, d'un château ou d'un musée mais la part de visiteurs étrangers est inconnue. De même, on peut obtenir le nombre de clients journaliers venus effectuer un achat en boulangerie et même le nombre d'acheteurs pendant les heures d'enquête dans la boulangerie sélectionnée ; mais la proportion d'achats des locaux dans la boulangerie reste inconnue pour les artisans boulangers.
2. Filtrer les populations hors champs, c'est-à-dire les excursionnistes, les autocaristes et les locaux. C'est un objectif d'ordre pratique.

**La seconde spécificité : les sites en rase campagne**. Sur certains sites de loisirs, en particulier ceux qui ne sont pas dotés de systèmes de billetteries, nous ne connaissons pas le nombre de visiteurs journaliers. Ce nombre est nécessaire pour appliquer la méthode généralisée de partage des poids. Comme ce total est inconnu, il va falloir obtenir une estimation de ce nombre de visiteurs journaliers. Nous avons donc mis en place des compteurs, généralement placés à l'entrée des parkings et qui comptent le nombre de voitures qui franchissent l'entrée du parking. Ces compteurs nous donnent un nombre de véhicules et non un nombre d'individus. Or sur les autres sites, notre référentiel est le nombre d'individus visitant le site. Pour disposer du même référentiel, un enquêteur placé sur le parking ou sur le site (nous discuterons par la suite de quel dispositif nous devons choisir afin d'obtenir le dispositif le plus précis en terme de variance et le plus économique) demande aux visiteurs si la voiture qui les a conduit est garée sur le parking du site et dans le cas échéant combien de voyageurs étaient présents dans ce véhicule. Il ne reste plus qu'à relier le nombre total de voitures noté $T_v$ à l'estimateur du nombre total de visiteurs noté $\hat{T}_P$ .

**Remarque** : Le nombre total de voitures stationnées sur le parking du site en rase campagne $T_v$ est parfaitement connu aux erreurs de mesure près puisqu'il est fourni par un compteur installé par l'Observatoire Régional du Tourisme de Bretagne spécialement pour les besoins de la méthodologie de l'enquête.

## 4. Les paramètres d'intérêt

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 voyage *i* utilisateur de ce service.

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

Soit $U^B$, la population des voyages *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. La population $U^A$ est l'image par $F^{-1}$ de l'ensemble des voyages durant la période de référence *D*. 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 de l'enquête** peuvent être des totaux, des effectifs et 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 $U^B$

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

Dans l'égalité (4.1), $y_i$ est une variable mesurée auprès de l'individu $i$. Dans le cadre de l'enquête, ici, l'individu $i$ est un des membres du ménage touristique $i$.

Un cas particulier de ces totaux est le nombre total de touristes venus visiter la Bretagne.

Le total $T^B$ peut aussi 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
- …

**Remarque importante :** Pour beaucoup de variables, le total 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

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

où

- $A_1$ désigne l'ensemble des boulangeries du champ de l'enquête repéré par l'indice $a_1$,
- $A_2$ désigne les 16 lieux de passage du champ de l'enquête repérés par l'indice $a_2$,
- $A_3$ désigne le péage de La Gravelle repéré par l'indice $a_3$,
- $D_l$ désigne 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, $C_{dl}$ désigne l'ensemble des services dans un établissement $a_l$ de l'ensemble des établissements $A_l$ de la journée $d_l$ de l'ensemble des jours $D_l$ repérés par l'indice $j$.

# 5. Estimation sans biais d'un total

Dans le paragraphe précédent, nous avons montré que le total s'écrit comme un total sur l'ensemble des services du champ de l'enquête.

Supposons que l'on dispose d'un échantillon de services répondants $j$, auxquels on peut associer des poids de sondage $\delta_j$ . Ces poids seront supposés sans biais comme on l'a démontré dans la section 2. Disposant d'un jeu de poids de sondage $\delta_j$ pour les services répondants, et si on connaît le nombre de services $R_i(B)$ utilisés par le ménage touristique $i$, on estime alors le total $T^B$ sans biais par

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

où

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

Pour alléger les notations dans les formules ci-dessus, on ne fait pas apparaître tous les degrés de tirage de l'échantillon en fonction de l'établissement $a_l$.

Dans les formules ci-dessus, on désigne par :

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

On est ramené à une estimation sur la population des ménages touristiques. Cette formule n'est autre que celle donnée par la méthode généralisée de partage des poids.

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 $i$
- $\delta_j = 1 / \pi_j^A$.

L'estimation de la variance de $\hat{T}_B$ est possible selon les mêmes principes (Lavallée (2002)). Elle ne sera pas détaillée ici car elle n'est qu'une application assez lourde en calcul des principes généraux.

# 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 visiter le site ou se promener sur le site.

En effet, dans l'ensemble des 16 sites emblématiques de la Bretagne noté $A_2$ on ne connaît pas tous les services (ici le nombre de visites) de la population car comme mentionné au paragraphe 3.4, certains sites parmi les 16 sélectionnés ne disposent pas de systèmes de billetterie ou de système équivalent capable de fournir le nombre exact de visiteurs. C'est ce nombre qui sert de dénominateur dans le calcul de la probabilité. On ne peut donc pas avoir directement la probabilité $\pi_j^{A2}$ qui est un rapport de cardinaux (nombre de visiteurs interrogés par nombre total de visiteurs du site) et donc $\delta_j$ pour $j \in A_2$.

Pour contourner ce problème, on estime alors le nombre de visiteurs journaliers afin de déduire une estimation de la probabilité $\pi_j^{A2}$, qui se définit par

$$\tilde{\pi}_j^{A_2} = \frac{n_{A_2}}{\hat{T}_P^{A_2}}.$$

Dans la suite, nous allons développer deux 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 interrogé 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 échantillonnage 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 $T_V$ est connu à des erreurs de mesure négligeables

### 6.1.1. Définition de l'estimateur du nombre de visiteurs

Soit $T_V$ le nombre total de voitures défini par

$$T_V = \sum_{k=1,\dots} t_k, \qquad (6.1)$$

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

$$T_V = \sum_{k \in U_V} 1, \qquad (6.2)$$

où $U_V$ désigne l'univers des voitures.

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

$$T_P = \sum_{k=1,\dots} k t_k. \qquad (6.3)$$

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

$$T_P = \sum_{l \in U_P} 1, \qquad (6.4)$$

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

$$T_P = \sum_{l \in U_V} v_l \qquad (6.5)$$

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 le $\pi$ -estimateur de $T_P$ défini par :

$$\widehat{T}_P = \sum_{l \in s_V} w_l^V v_l, \qquad (6.6)$$

où $s_V$ est un échantillon aléatoire simple de voitures de taille $n$ et le poids $w_l^V$ est égal à $T_V/n$, ce qui permet d'écrire l'estimateur de $T_P$ sous la forme suivante

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

en posant

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

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

### 6.1.2. Calcul de la variance de l'estimateur

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

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

où $S^2_V$ désigne la variance corrigée de la population $U_V$.

## 6.2. 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 $j$ le nombre $u_j$ de passagers de la voiture $i$ qui l'a transportée. Ce nombre $u_j$ est ici égal à $v_l$. Cette méthode a l'avantage d'obtenir avec précision le nombre de passagers.

### 6.2.1. Définition de l'estimateur du nombre de visiteurs

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

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

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

$$\overline{v} = \frac{\sum_{k \in U_V} k t_k}{\sum_{k \in U_V} t_k} = \frac{\sum_{k \in U_P} M_k}{\sum_{k \in U_P} M_k / k}, \qquad (6.9)$$

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 $T_P$

$$T_P = T_V \overline{v}. \qquad (6.10)$$

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

$$\hat{\hat{T}}_P = T_V \hat{\overline{v}}, \qquad (6.11)$$

où le nombre total de voitures $T_V$ est parfaitement connu. En observant cette expression, on constate que pour connaître l'estimateur de $T_P$ défini par l'équation (6.11), il suffit de déterminer un estimateur de $\overline{v}$. 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. L'estimateur de $\overline{v}$ peut s'écrire également de la façon suivante

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

ou encore

$$\widehat{\overline{v}} = \frac{m}{\sum_{j \in s_P} 1/u_j}. \qquad (6.12)$$

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

$$\frac{1}{\widehat{\overline{v}}} = \frac{1}{m} \sum_{j \in s_P} \frac{1}{u_j}. \qquad (6.13)$$

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

$$\mathrm{Var}\left[\frac{1}{\widehat{\overline{v}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (6.14)$$

### 6.2.2. Calcul de la variance de l'estimateur

Reste à calculer la variance de l'estimateur de $\overline{v}$ sachant (6.14).

Pour cela, remarquons que l'on peut écrire

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

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 (6.14)

$$\mathrm{Var}\left[\widehat{\overline{v}}\right] \simeq \overline{v}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (6.15)$$

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

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{j \in U_P} \left(\frac{1}{u_j} - \frac{1}{\overline{v}}\right)^2. \qquad (6.16)$$

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

$$\frac{1}{m-1} \sum_{j \in s_P} \left(\frac{1}{u_j} - \frac{1}{\overline{v}}\right)^2. \qquad (6.17)$$

Grâce à (6.15) et à (6.17), on peut donc connaître facilement la variance de l'estimateur de $\overline{v}$ et par conséquent celle de l'estimateur de $T_P$ défini par l'équation (6.11).

**Remarque :** L'estimateur de $T_P$ défini par l'équation (6.11) est biaisé et asymptotiquement sans biais.

## 6.3. Illustration numérique

Un compteur mécanique d'un site en rase campagne donne $T_V = 100$ voitures. On suppose 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 $S^2_V$ 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 $S^2_{1/u}$. Après simplifications de (6.17) et en supposant que $T_P$ est suffisamment grand devant 1, on a :

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

Ainsi, on a :

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

Puisque nous connaissons $S^2_{1/u}$, nous pouvons calculer la variance de l'estimateur de $\overline{v}$ . Ainsi on a :

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

Enfin, on peut calculer la variance de l'estimateur du total :

$$\begin{aligned} \mathrm{Var}\left[\widehat{\widehat{T}}_P\right] &= T_V^2 \mathrm{Var}\left[\widehat{\overline{v}}\right] \\ &\simeq 10^4 \times 3^4 \times \frac{37}{30^2} \times \frac{1}{m} \end{aligned}$$

Donc, afin que l'estimateur défini par l'équation (6.11) ait la même variance que celui défini par l'équation (6.6), il suffit que la taille $m$ de l'échantillon $s_P$ soit égale à :

$$m \simeq 1.66\, n.$$

**Conclusion :** La seconde approche est moins coûteuse en termes d'individus et peut aboutir à des résultats aussi performants que la première approche à condition d'imposer à la taille $m$ de l'échantillon de personnes $s_P$ la condition suivante :

$$m \simeq 1.66\, n.$$

De ce fait, l'estimateur de la seconde approche serait tout aussi bon que l'estimateur de la première approche.

# Bibliographie

[1] Ardilly P., Le Blanc D., « Echantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français », *Techniques d'enquête*, 27, pp. 109-118, 2001

[2] Deville J.C., « 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*, n° 84-85-86, pp 63-82, 1999.

[3] Lavallée P., « 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* vol. 21, pp.27-35, 1995.

[4] Lavallée P., *Le sondage indirect, ou la méthode généralisée du partage des poids*, Editions de l'Université de Bruxelles, Editions Ellipses, Bruxelles, 2002.

[5] Lavallée P., Caron P., « Estimation Using the Generalized Weight Share Method: The Case of Record Linkage », *Survey Methodology*, vol. 27, No. 2, pp. 155-169, 2001.

[6] Torres Manzanera E., Sustacha Melijosa I., Menéndez Estébanez J.M., Valdés Pelaàez L., « A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places », Akos Probàld (Ed.), *Proceedings Of The Sixth International Forum On Tourism Statistics. Hungarian Central Statistical Office*, Budapest, 2002

[7] Valdés Pelavaàez L., et al., « A methodology to measure tourism expenditure and total tourism production at the region level», Lennon, J. (Editor), *Tourism Statistics*. Continuum, London.