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

Jean-Claude Deville[1] Myriam Maumy[2]

[1]CREST, ENSAI
Paris, France

[2]IRMA, Université Louis Pasteur
Strasbourg, France

Journées Méthodologie Statistique, 2005

Sommaire

Sommaire

Sommaire

Sommaire

Sommaire

Sommaire

- Une **enquête aux frontières** sur la fréquentation touristique extra-régionale en Bretagne (hormis celle des Bretons) réalisée d'avril à septembre 1997.
- **But :** Recommencer ce type d'enquête.
- **Le problème :** Impossibilité de recueillir cette masse d'informations récoltées aux frontières régionales ou intra-régionales.
- **Pourquoi ?** La Gendarmerie ne désire plus intervenir dans la réalisation des enquêtes au bord des routes.

- **La solution :** Mise en place par l'ORTB avec l'aide d'un comité technique d'une nouvelle méthodologie.

- Les 2 objectifs principaux de cette enquête :
  - remplacer **l'enquête aux frontières de 1997**
  - obtenir les informations précises recherchées par l'ORTB.

- **Remarque :** Par exemple, si l'ORTB désire obtenir des renseignements sur les dépenses du séjour, il faut introduire **la notion de voyage** comme u.s. pour produire l'information.

- **Un des problèmes majeurs :** Absence d'une base de sondage permettant d'interroger directement les touristes.
- **L'idée principale :** Échantillonner des services destinés principalement aux touristes sur les différents lieux de ces nombreuses prestations touristiques.

- **Le problème :** Un touriste peut utiliser **une ou plusieurs fois un ou plusieurs services** de la base de sondage pendant l'enquête.
- **Un autre problème :** Pour estimer des paramètres d'intérêt, il faut relier le jeu de poids des services échantillonnés à celui des touristes qui ont fréquenté ces services.
- **Une solution :** la *méthode généralisée du partage des poids* (MGPP) mise au point par Lavallée (1995, 2002).

Sommaire

- Le principe de l'enquête : **Atteindre les touristes (français ou étrangers) par le biais de services destinés à satisfaire leurs besoins élémentaires,** comme l'hébergement, les transports, les activités de loisirs, la nourriture...

- Le principe a déjà été utilisé dans l'enquête des sans domicile réalisée par l'INSEE.

- Le principe est intéressant lorsqu'il y a absence de base de sondage.

- **La période de référence D :** de février 2005 à décembre 2005.
- **Le champ géographique G :** les 4 départements bretons.
- **L'unité statistique :** le voyage défini par le ménage touristique (l'ensemble du groupe en voyage) et par un intervalle de temps.

- **Les lieux d'interrogation :**
  - des boulangeries,
  - 16 lieux de passage sur des sites célèbres comme la pointe du Ratz, le Cap Fréhel, l'Île de Batz, l'Aquarium de Vannes, Saint-Malo..
  - le péage autoroutier de La Gravelle.

- **La base de sondage** est constituée de 3 strates représentant les services de l'enquête :
  1. les achats en boulangerie;
  2. les passages sur les 16 sites.
  3. les passages au péage autoroutier de La Gravelle

  Dans *la première strate :* échantillon à 3 degrés :
  - échantillon de boulangeries;
  - échantillon de jours d'enquête;
  - échantillon de clients dans la boulangerie à un jour donné.

Dans *la deuxième strate :* un échantillon à 2 degrés :

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

Dans *la troisième strate :* un échantillon à 2 degrés :

- un échantillon de jours d'enquête;
- un échantillon de voitures qui franchissent le péage de La Gravelle à un jour donné.

Sommaire

Les paramètres d'intérêt :

- des totaux,
- des effectifs,
- des ratios.

Intéressons nous, par exemple, à l'estimation d'un total T^B relatif à la population des voyages notée U^B, défini par :

$$T^B = \sum_{i \in U^B} y_i.$$

Exemples : T^B peut-être

- le nombre de personnes ayant pratiqué une certaine activité,
- le budget total dépensé,
- la provenance géographique,
- le nombre de jours passés en Bretagne...

Pour beaucoup de variables, T^B dépend de

- la taille du nombre de personnes en voyage
- et de la longueur du séjour : uniquement les jours passés en Bretagne.

Introduisons quelques notations

- A_1 : l'ensemble des boulangeries repéré par l'indice a_1
- A_2 : les 16 lieux de passage 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.

- Définissons une application F par,

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

- Notons U^B, la population des voyages de la période D. U^B est l'image par F de l'ensemble des services durant D dans les 3 lieux de l'enquête.

- Pour tout $i \in U^B$, notons

$$R_i(B) = \text{card}(F^{-1}(i)),$$

le nombre de services j utilisés par le voyage i.

- Le total T^B devient alors

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

où

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

et

$$R_i(B) = \text{card}(F^{-1}(i)).$$

Sommaire

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

- s^B : l'ensemble des voyages 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 T^B par

$$\widehat{T}^B = \sum_{i \in s^B} w_i y_i$$

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 voyages U^B.
- Cette formule provient de la **MGPP** (Lavallée, 2002).
- $\delta_j = \dfrac{1}{\pi_j^A}$, où $U^A = \bigcup_{l=1}^{3} U^{A_l}$.

Sommaire

- Sur certains sites parmi les 16 : le nombre de visiteurs est inconnu.
- Impossibilité donc d'avoir $\pi_j^{A_2}$ et donc δ_j pour $j \in A_2$.
- Par conséquent obtenir une **estimation du nombre de visiteurs** afin d'en déduire une estimation de $\pi_j^{A_2}$ définie par

$$\widetilde{\pi}_j^{A_2} = \frac{n_{A_2}}{\widetilde{N}_{A_2}}.$$

2 approches possibles pour estimer le nombre de visiteurs :

- une approche qui utilise un échantillon de voitures destiné à estimer le nombre de visiteurs
- une approche qui utilise un échantillon de visiteurs destiné à estimer le nombre de visiteurs

Sommaire

Dans cette première approche, un enquêteur relève

- le nombre de personnes dans une voiture qui franchit l'endroit où un compteur électronique a été placé pour relever le nombre de voitures qui entrent sur un parking d'un site en rase campagne.

- T_V le nombre total de voitures défini par

$$T_V = \sum_{k,\ldots} t_k,$$

où t_k désigne le nombre de voitures transportant k personnes.

- T_V défini également par :

$$T_V = \sum_{k \in U_V} \mathbf{1}, \tag{5.1}$$

où U_V désigne l'univers des voitures.

- T_P le nombre total de personnes visitant le site

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

- Par analogie avec (5.1), on a

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

où U_P désigne l'univers des personnes.

- T_P défini également par

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

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

- Définissons $\widehat{T}_P$ le π-estimateur

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

où s_V est un échantillon de voitures de taille n et $w_l = T_V / n$.

- Donc

$$\widehat{T}_P = \frac{T_V}{n} \sum_{i \in s_V} v_i = T_V \, \overline{v},$$

où $\overline{v} = 1/n \sum_{l \in s_V} v_l$.

- On veut estimer une variable d'intérêt Y définie par

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

où y_i est une variable d'intérêt mesurée dans le questionnaire.

- $\widehat{Y}$ est le π-estimateur

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

où $w_i = \widehat{T}_P / m$.

- $\widehat{Y}$ s'écrit alors :

$$\widehat{Y} = \frac{\widehat{T}_P}{m} \sum_{i \in s_P} y_j = \widehat{T}_P \, \overline{y}.$$

- Calculons la variance de $\mathrm{Var}\left[\widehat{Y}\right]$

$$V_Y = \mathrm{Var}\left[\widehat{Y}\right] = \overline{Y}^2 \mathrm{Var}\left[\widehat{T}_P\right] + T_P^2 \mathrm{Var}\left[\overline{y}\right] + \mathrm{Var}\left[\widehat{T}_P\right] \mathrm{Var}\left[\overline{y}\right]. \tag{5.2}$$

- On assimile l'échantillon de voitures à un SAS sans remise, (5.2) devient :

$$\begin{aligned}\mathrm{Var}\left[\widehat{Y}\right] &= \left(\overline{Y}^2 - \frac{1}{T_P} S_{\overline{y}}^2\right) T_V^2 S_V^2 \frac{1}{n} + \left(T_P^2 - T_V S_V^2\right) S_{\overline{y}}^2 \frac{1}{m} \\ &\quad + T_V^2 S_V^2 S_{\overline{y}}^2 \frac{1}{nm} + \frac{T_V}{T_P} S_V^2 S_{\overline{y}}^2 - \overline{Y}^2 T_V^2 S_V^2 - T_P S_{\overline{y}}^2\end{aligned} \tag{5.3}$$

- Cherchons l'allocation des tailles des échantillons s_P et s_V qui minimise en n, m $\mathrm{Var}\left[\widehat{Y}\right]$ pour des tailles de population T_P et T_V fixées,

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

- sous la contrainte

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

- En calculant le lagrangien, en annulant les dérivées partielles par rapport à n, m, λ et en organisant les termes, on a une équation du troisième degré en n à résoudre

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

- **Problème :** cette équation admet une solution à déterminer avec des méthodes numériques.

- Raisonnement analogue : en organisant les termes, on a une équation du troisième degré en m à résoudre

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

- **Problème :** cette équation admet une solution à déterminer avec des méthodes numériques.

- Pour remédier à ce problème, on fait une approximation dans (5.3).
- Supposons que $1/nm$ est négligeable devant $1/n$ et $1/m$.
- Cette hypothèse n'est pas absurde puisque n et m peuvent prendre des grandes valeurs.

- On a alors

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

- L'étape est de chercher l'allocation des tailles des échantillons s_P et s_V qui minimise $\mathrm{Var}\left[\widehat{Y}\right]$ pour des tailles de population T_P et T_V fixées.

- En simplifiant (5.3), en calculant le lagrangien, en annulant les dérivées partielles par rapport aux n, m, λ, et après calculs, on obtient

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

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

Sommaire

- 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.

- Rappelons l'égalité

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

où v_l est 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 qu'il y a dans une voiture défini par

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

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

- Cette dernière définition permet de donner une dernière écriture de T_P

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

- Par conséquent un estimateur de T_P s'écrit

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

où T_V est donné par le compteur.

- Pour connaître $\widehat{\widehat{T}}_P$, déterminons $\widehat{\overline{v}}$

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

où m_k est le le nombre de personnes de l'échantillon voyageant dans une voiture à k passagers.

- $\widehat{\overline{v}}$ s'écrit également :

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

- Cette dernière égalité permet d'écrire

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

ce qui représente la moyenne empirique des $1/u_j$.

- Calculons sa variance. On a

$$\text{Var}\left[\frac{1}{\widehat{\overline{v}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (5.4)$$

- Il reste à calculer la variance de $\widehat{\overline{v}}$ sachant (5.4).

- Remarquons que

$$\frac{1}{\widehat{\overline{v}}} = \frac{1}{\overline{v}}\left(1 - \frac{\widehat{\overline{v}} - \overline{v}}{\overline{v}} + o\left(\frac{\widehat{\overline{v}} - \overline{v}}{\overline{v}}\right)\right).$$

- 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 avec (5.4)

$$\mathrm{Var}\left[\widehat{\overline{v}}\right] \simeq \overline{v}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (5.5)$$

- Or comme T_P est inconnu, $S_{1/u}$ est estimée

$$\frac{1}{m-1} \sum_{j \in s_P} \left(\frac{1}{u_j} - \frac{1}{\overline{v}}\right)^2. \qquad (5.6)$$

- Grâce à (5.5) et (5.6) on peut donc connaître la variance de $\widehat{\overline{v}}$ et par conséquent celle de $\widehat{\widehat{T}}_P$ et celle de $\widehat{Y}$.

Bibliographie I


J.C. Deville.
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, 1999.


P. Lavallée.
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, p.27-35, 1995.

Bibliographie II


P. Lavallée
Le Sondage Indirect, ou la méthode généralisée du partage des poids.
Éditions de l'Université de bruxelles, Bruxelles, 2002.

Merci de votre attention

Cette présentation est désormais terminée.

Avez-vous des questions ?

Au delà de cette présentation

Pour tous commentaires, questions, ou suggestions à venir, n'hésitez pas :

- Jean-Claude.Deville@ensae.fr
- mmaumy@math.u-strasbg.fr