STA108 : Enquêtes et sondages

Le sondage indirect, la méthode du partage des poids

- Sondages en plusieurs degrés, sondage en deux phase et sondages indirects
- Exemple : les enquêtes par téléphone, bases de sondages fixes et mobiles
- En résume

le cnam

Sondages en plusieurs degrés, sondage en deux phase et sondages indirects

Sondage indirect, motivation

Jusqu'à présent, nous avons majoritairement présenté des méthodes de sondage qui se basent sur la sélection directe d'individus dans une liste donnant un accès ***direct*** aux unités de la population.

Les différents plans de sondage abordés se différentient par la façon dont ils intègrent de l'information auxiliaire dans leurs constructions : les deux principaux exemples sont les sondages à probabilités inégales ou stratifiés. Mais dans ces deux cas, l'accès aux unités est ***direct*** et ***bijectif***.

Sondage indirect, motivation

Il arrive bien souvent qu'au lieu de posséder une liste contenant les unités de collecte souhaitées, on ne dispose que d'une liste d'unités reliées d'une certaine façon à celle des unités de collecte. On a donc deux populations U_A et U_B liées l'une à l'autre, et on souhaite produire une estimation pour U_B. Mais on en possède de base de sondage que pour U_A .

On peut alors envisager de sélectionner un échantillon s_A dans U_A fin de produire une estimation pour U_B en s'appuyant sur la correspondance qui existe entre les deux populations.

On parle alors de sondage ***indirect***

Nous avons déjà vu certains cas de sondage similaires avec les sondages en plusieurs degrés qui répondaient déjà à des problèmes de constitution de base de sondage

Les sondages en plusieurs degrés, motivation

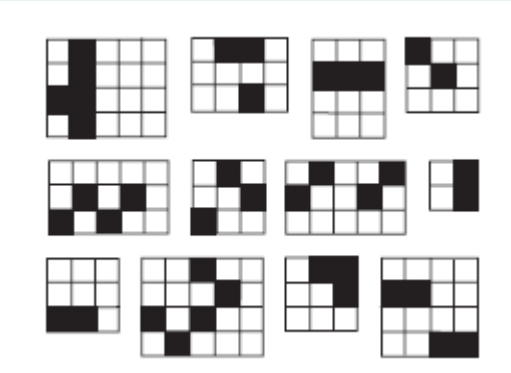
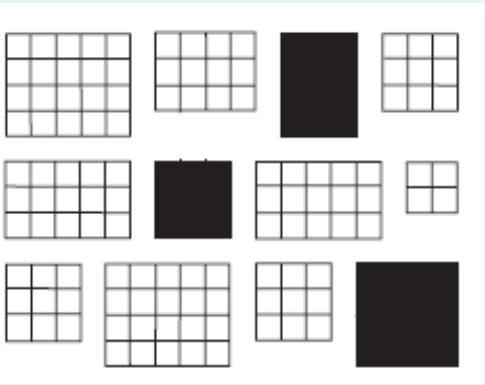
1 - Difficultés et mise à jour : une base de sondage pose le problème du coût de sa création et de sa mise à jour. Dans le cas d'individus physiques qui sont des unités mobiles, elle n'ont pas d'obligation légale de signaler leur lieu de résidence. Les logements sont des unités plus stables, mais qui évoluent également dans le temps. Les entreprises ont par contre une obligation légale de se signaler et de s'enregistrer, mais il reste la fréquence de la mise à jour de la base de donnée et de la mise à disposition

2 - Impossibilité pratique de création d'une base de sondage : par exemple, pour une enquête auprès de patients atteints d'une certaine pathologie, il est plus facile de procéder en deux phases. On commence par la sélection d'un échantillon de praticiens dans un registre de la ou des spécialité(s) concernées, puis la selection des unités parmi les patients. Ceci évite de créer une base de sondage exhaustive de tous les patients atteints de cette pathologie (ce qui est impossible pratiquement d'ailleurs !)

Les sondages en plusieurs degrés, motivation

3 - Coûts de mise en place : même si une base de sondage est disponible (fichiers de la taxe d'habitation) une enquête en face à face auprès des ménage réalisée par sondage aléatoire simple conduirait à un échantillon avec des coûts de déplacement prohibitifs si le territoire est grand, peu densément peuplé ou accidenté

➔ *Les motivations principales des sondages à plusieurs degrés sont des motivations de réduction de coûts ou de simplification de mise en place*

Sondage quelconque à plusieurs degrés	**Sondage en grappes**
	

Echantillonnage en deux phases, motivation

Dans le cadre d'une enquête, on peut être **intéressé** par **une sous population très spécifique**, ou une sous-population peu représentée dans la population générale (personnes illettrées, personnes soufrant de handicap,…). Sélectionner un échantillon de ces individus ne pose ***pas de difficulté technique*** particulière.

On peut par exemple tirer un échantillon dans une population plus large, puis n'utiliser que les individus échantillonnés qui appartiennent à la sous-population qui nous intéresse. On est alors dans un cas d'estimation sur domaine.

Cette méthode a plusieurs désavantages :

- on ne contrôle pas la taille de l'échantillon final,
- elle a un coût prohibitif si la prévalence est faible,
- elle n'utilise pas forcément bien de l'information auxiliaire qui serait disponible auprès de cette population
- …

Echantillonnage en deux phases, motivation

Une **solution possible** consiste à **sélectionner l'échantillon en deux temps** :

- On sélectionne tout d'abord un gros sur-échantillon (sélectionné pour l'occasion, ou utilisé dans une autre enquête).
- On collecte de l'information sur cet échantillon.
- On tire un sous-échantillon dans ce gros échantillon en utilisant l'information collectée (par exemple, pour définir une stratification).

A la différence d'un tirage à deux degrés, **ce sont les mêmes unités** qui sont **sélectionnées ici lors des deux phases** de tirage.

C'est une solution qui est fréquemment utilisée pour adosser à une enquête ménage (ou à une enquête entreprise) une enquête sur un sujet spécifique, ou sur une population spécifique.

Elle est utilisée par les instituts d'études privés, aussi pour cibler des population très rares (acheteurs de marques de luxe, gros consommateurs de chocolat, …)

Exemple de sondage en deux phase, l'enquête HID

Enquête Handicaps-Incapacités-Dépendances, réalisée en 1998 en institutions et à domicile (Joinville, 2002).

C'est une enquête auprès de ménages.

Deux étapes :

1. Utilisation d'une enquête de filtrage adossée au recensement (Vie Quotidienne et Santé), portant sur 360 000 personnes environ, afin de repérer les individus dans le champ de l'enquête.

2. Tirage parmi les répondants de l'enquête VQS en stratifiant selon un indicateur de handicap, âge zone géographique)

Certains cas avec des liens complexes entre populations

Certain cas de liens entre les deux populations U_A et U_B sont encore plus complexes et dans ces cas, l'estimation des caractéristiques de la population U_B peut poser des problème de taille si les liens entre les deux populations ne sont pas ***bijectifs***.

C'est pour ce types de cas que la méthode généralisée du partage des poids a été proposée

Un exemple courant de liens non bijectifs est celui des cas pours lesquels aucun registre ne couvre à lui seul la population, il faut en considérer une réunion. On parle de base de sondage multiples chevauchantes (Multiple frame Survey).

Exemple : les enquêtes par téléphone, bases de sondages fixes et mobiles

Au canada …

http://www.statcan.gc.ca/pub/12-001-x/2011002/article/11608-fra.pdf

Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes

Sharon L. Lohr [1]

Résumé

Les plans de sondage et les estimateurs des enquêtes à base de sondage unique utilisés à l'heure actuelle par les organismes gouvernementaux américains ont été élaborés en réponse à des problèmes pratiques. Les programmes d'enquêtes-ménages fédéraux doivent faire face aujourd'hui à la diminution des taux de réponse et de la couverture des bases de sondage, à la hausse des coûts de collecte des données et à l'accroissement de la demande de statistiques pour des petits domaines. Les enquêtes à bases de sondage multiples, dans lesquelles des échantillons indépendants sont tirés de bases de sondage distinctes, peuvent être utilisées en vue de relever certains de ces défis. La combinaison d'une liste et d'une base de sondage aréolaire ou l'utilisation de deux bases de sondage pour échantillonner les ménages ayant une ligne de téléphone fixe et ceux ayant une ligne de téléphone mobile en sont des exemples. Nous passons en revue les estimateurs ponctuels et les ajustements de la pondération qui peuvent être utilisés pour analyser les données d'enquête à bases de sondage multiples au moyen de logiciels standard et nous résumons la construction des poids de rééchantillonnage pour l'estimation de la variance. Étant donné leur complexité croissante, les enquêtes à bases de sondage multiples obligent à résoudre des difficultés qui ne se posent pas dans le cas des enquêtes à base de sondage simple. Nous étudions le biais dû à l'erreur de classification dans les enquêtes à bases de sondage multiples et proposons une méthode pour corriger ce biais quand les probabilités d'erreur de classification sont connues. Enfin, nous discutons des travaux de recherche nécessaires en ce qui concerne les erreurs non dues à l'échantillonnage dans les enquêtes à bases de sondage multiples.

En France : la téléphonie, un univers complexe

L'absence d'un annuaire universel,

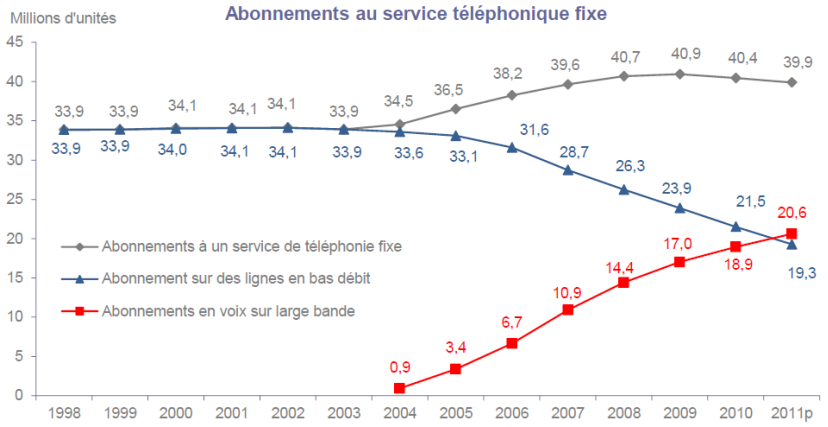
14% des individus qui ne sont pas joignables par téléphone mobile, 11% des adultes qui ne sont pas joignables par téléphone fixe, plus de 39% des abonnés en téléphonie fixe qui ne sont pas inscrits sur l'annuaire,

La substitution RTC et maintenant Bistream au profit de la VLB qui se fait sur un rythme très dynamique avec de nombreux facteurs qui influent sur le numéro d'appel (conservation du numéro/attribution d'un numéro en 09, deux numéros, racines spécifiques à certains opérateurs, …) [cf. ARCEP mars 2012]

…

Dégroupage				
Millions	2007	2008	2009	201
Nombre de lignes partiellement dégroupées	1,613	1,393	1,309	1,1
Nombre de lignes totalement dégroupées	3,625	4,939	6,414	7,6
Nombre de lignes dégroupées	**5,238**	**6,332**	**7,723**	**8,8**

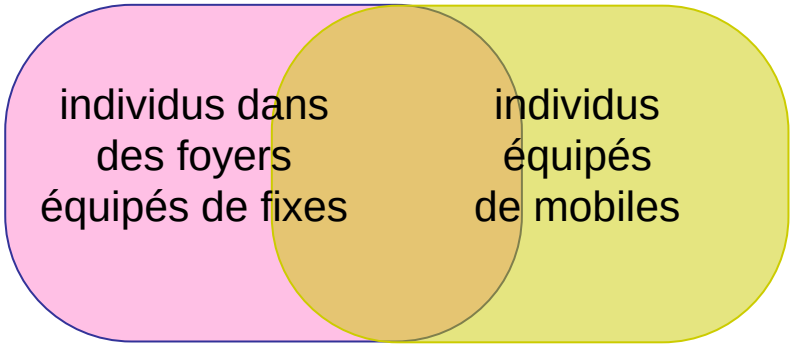
Bitstream (ATM et IP régional) et IP national				
Millions	2007	2008	2009	201
Nombre de lignes en "bitstream nu"	0,942	1,186	1,245	1,2
Nombre de lignes en "bitstream classique" et IP national	1,291	1,010	0,647	0,4
Nombre total de lignes	**2,233**	**2,196**	**1,892**	**1,7**



Bases de sondage multiples entre telephones fixes et mobiles, en France

Le téléphone fixe ***ou*** mobile couvre plus de 99% de la population en France. Le taux d'équipement des individus en mobile est de 86%, le taux d'équipement en fixe est de 89%.

On peut facilement constituer des bases de sondage par génération aléatoire de numéros fixes ou mobile, mais **seule la réunion de ces bases couvre bien la population**, mais les relations entre la liste ainsi constituée et les unités ne sont pas bijectives. Certains individus (la majorité ..) sont multi équipés



Il y a 3 Groupes:

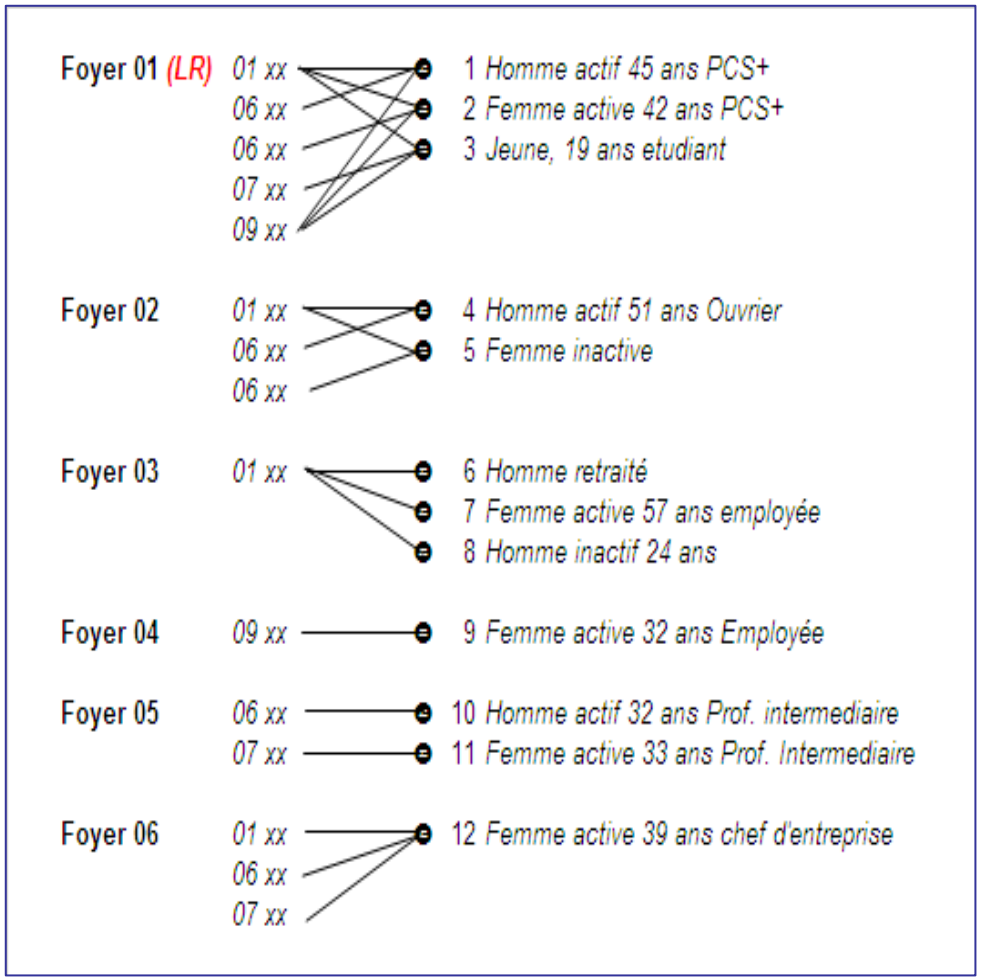
1. *Individus dans des foyers équipés de fixes, mais non possesseurs de mobiles*
2. *Individus dans des foyers équipés de fixes, possesseurs de mobiles*
3. *Individus dans des foyers non équipés de fixes, mais possesseurs de mobiles (mobiles only)*

Un exemple introductif : la population A est de taille 15, B est de taille 12 : 12 individus pour 15 numéros de téléphone

Soit une base de sondage téléphonique générée aléatoirement (A) et une population d'individus (B), les doubles liens ne sont pas des problèmes,

… si on connait les liens entre les deux populations, et les probabilités sur la population (A) on peut **calculer formellement les probabilités d'inclusion de chaque individu de (B) par la méthode de partage des poids** *[Deville, Lavallée 1995 - 2002]*

Plutôt que de reprendre le formalisme matriciel un peu abstrait des auteurs, nous partons d'un exemple simple pour montrer ce qui n'est qu'une répartition des poids



Point de départ, la matrice des liens : 15x12

On peut formaliser l'exemple précédent avec la matrice suivante qui décrit les liens entre les individus des deux populations. (p.ex, l'individu1, dans le foyer 1 est accessible avec les numéros 01 xx, 06 xx et 09 xx qui ont les probabilités de tirage π1, π2, π5).

		Liste des foyers et des individus de la population B (population cible)											
Probabilités		Foyer01			Foyer02		Foyer03			Foyer04	Foyer05		Foyer06
πi (A)	**Individu de la population A (numéro de téléphone)**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
π1	*01 xx*	1	1	1	0	0	0	0	0	0	0	0	0
π2	*06 xx*	1	0	1	0	0	0	0	0	0	0	0	0
π3	*06 xx*	0	1	0	0	0	0	0	0	0	0	0	0
π4	*07 xx*	0	0	1	0	0	0	0	0	0	0	0	0
π5	*09 xx*	1	1	1	0	0	0	0	0	0	0	0	0
π6	*01 xx*	0	0	0	1	1	0	0	0	0	0	0	0
π7	*06 xx*	0	0	0	1	0	0	0	0	0	0	0	0
π8	*06 xx*	0	0	0	0	1	0	0	0	0	0	0	0
π9	*01 xx*	0	0	0	0	0	1	1	1	0	0	0	0
π10	*09 xx*	0	0	0	0	0	0	0	0	1	0	0	0
π11	*06 xx*	0	0	0	0	0	0	0	0	0	1	0	0
π12	*07 xx*	0	0	0	0	0	0	0	0	0	0	1	0
π13	*01 xx*	0	0	0	0	0	0	0	0	0	0	0	1
π14	*06 xx*	0	0	0	0	0	0	0	0	0	0	0	1
π15	*09 xx*	0	0	0	0	0	0	0	0	0	0	0	1

Matrice des liens normalisée

La matrice des liens normalisée s'obtient en divisant chaque indicatrice par le nombre de liens que chaque unité de A possède avec B

		Liste des foyers et des individus de la population B (population cible)											
Probabilités		Foyer01			Foyer02		Foyer03			Foyer04	Foyer05		Foyer06
πi (A)	**Individu de la population A (numéro de téléphone)**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
π1	*01 xx*	0.33	0.33	0.25	0	0	0	0	0	0	0	0	0
π2	*06 xx*	0.33	0	0.25	0	0	0	0	0	0	0	0	0
π3	*06 xx*	0	0.33	0	0	0	0	0	0	0	0	0	0
π4	*07 xx*	0	0	0.25	0	0	0	0	0	0	0	0	0
π5	*09 xx*	0.33	0.33	0.25	0	0	0	0	0	0	0	0	0
π6	*01 xx*	0	0	0	0.5	0.5	0	0	0	0	0	0	0
π7	*06 xx*	0	0	0	0.5	0	0	0	0	0	0	0	0
π8	*06 xx*	0	0	0	0	0.5	0	0	0	0	0	0	0
π9	*01 xx*	0	0	0	0	0	1	1	1	0	0	0	0
π10	*09 xx*	0	0	0	0	0	0	0	0	1	0	0	0
π11	*06 xx*	0	0	0	0	0	0	0	0	0	1	0	0
π12	*07 xx*	0	0	0	0	0	0	0	0	0	0	1	0
π13	*01 xx*	0	0	0	0	0	0	0	0	0	0	0	0.33
π14	*06 xx*	0	0	0	0	0	0	0	0	0	0	0	0.33
π15	*09 xx*	0	0	0	0	0	0	0	0	0	0	0	0.33

Matrice des poids sur la population A (Horvitz Thompson)

Supposons que l'on a tiré les individus {1,3,5,6,9,12,14,15} de la population A

Alors on peut former la matrice des poids de Horvitz Thompson de terme général 1/πi si l'unité est choisie, 0 sinon :

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
		01 xx	06 xx	06 xx	07 xx	09 xx	01 xx	06 xx	06 xx	01 xx	09 xx	06 xx	07 xx	01 xx	06 xx	09 xx
1	01 xx	1/π1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	06 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	06 xx	0	0	1/π3	0	0	0	0	0	0	0	0	0	0	0	0
4	07 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	09 xx	0	0	0	0	1/π5	0	0	0	0	0	0	0	0	0	0
6	01 xx	0	0	0	0	0	1/π6	0	0	0	0	0	0	0	0	0
7	06 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	06 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	01 xx	0	0	0	0	0	0	0	0	1/π9	0	0	0	0	0	0
10	09 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	06 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	07 xx	0	0	0	0	0	0	0	0	0	0	0	1/π12	0	0	0
13	01 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	06 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	1/π14	0
15	09 xx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/π15

Calcul des poids sur la population B (Horvitz Thompson)

Le tirage des unités {1,3,5,6,9,12,14,15} de la population A, 'touche' les unités {1,2,3,4,5,6,7,8,11,12} de la population B. On peut former un estimateur de Horvitz Thomson pour une variable d'intérêt de la population B en appliquant les poids donnés par les sommes sur les lignes

		Liste des foyers et des individus de la population B (population cible)											
Probabilités		**Foyer01**			**Foyer02**		**Foyer03**			**Foyer04**	**Foyer05**		**Foyer06**
πi (A)	**Individu de la population A (numéro de téléphone)**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
π1	01 xx	0.33/π1	0.33/π1	0.25/π1	0	0	0	0	0	0	0	0	0
π2	06 xx	0.33	0	0.25	0	0	0	0	0	0	0	0	0
π3	06 xx	0	0.33/π3	0	0	0	0	0	0	0	0	0	0
π4	07 xx	0	0	0.25	0	0	0	0	0	0	0	0	0
π5	09 xx	0.33/π5	0.33/π5	0.25/π5	0	0	0	0	0	0	0	0	0
π6	01 xx	0	0	0	0.5/π6	0.5/π6	0	0	0	0	0	0	0
π7	06 xx	0	0	0	0.50	0	0	0	0	0	0	0	0
π8	06 xx	0	0	0	0	0.50	0	0	0	0	0	0	0
π9	01 xx	0	0	0	0	0	1/π9	1/π9	1/π9	0	0	0	0
π10	09 xx	0	0	0	0	0	0	0	0	1.00	0	0	0
π11	06 xx	0	0	0	0	0	0	0	0	0	1.00	0	0
π12	07 xx	0	0	0	0	0	0	0	0	0	0	1/π12	0
π13	01 xx	0	0	0	0	0	0	0	0	0	0	0	0.33
π14	06 xx	0	0	0	0	0	0	0	0	0	0	0	0.33/π14
π15	09 xx	0	0	0	0	0	0	0	0	0	0	0	0.33/π15

Estimateur de Horvitz Thomspon

On peut alors former un estimateur de Horvitz Thompson pour une moyenne ou un total en appliquant à chaque unité i de la population B échantillonnée le poids wi donné par la méthode MGPP (méthode généralisée du partage des poids)

Le poids de i est une somme des inverses de probabilités de sélection divisée par le nombre de liens B et A pour i

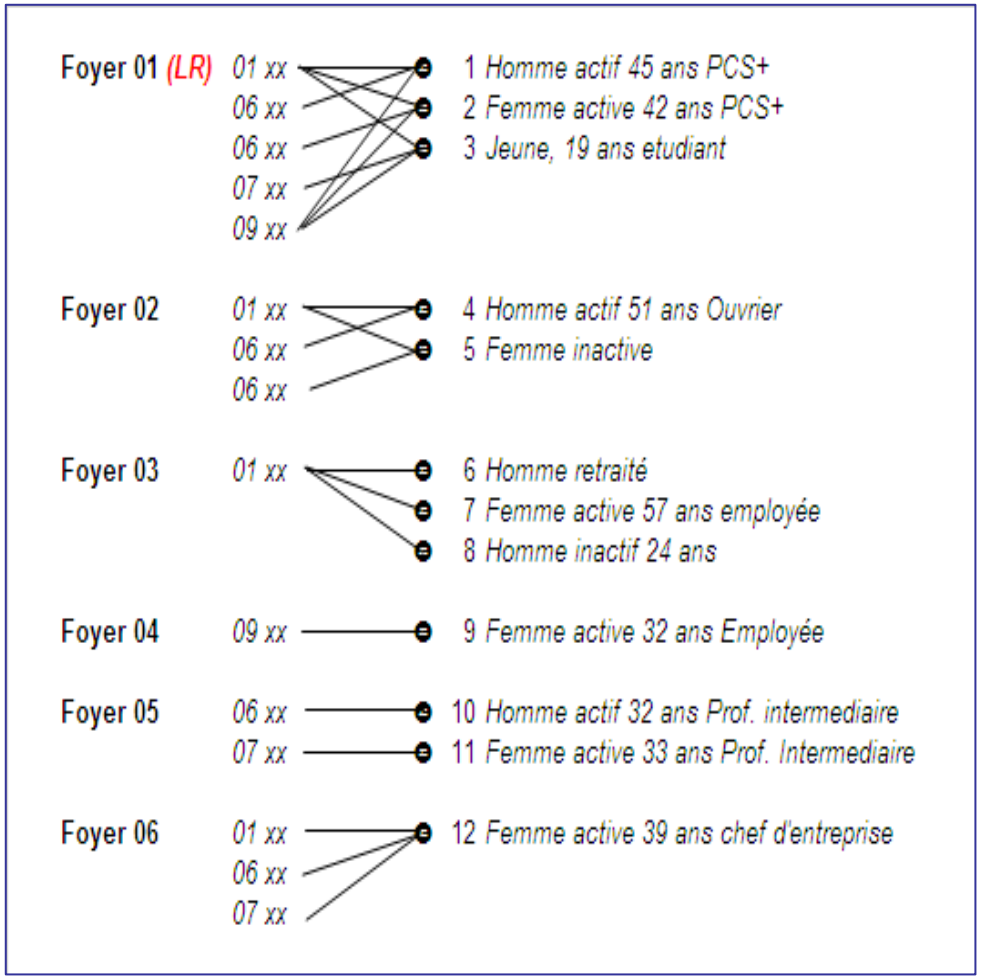
On retrouve alors tous les résultats classiques déjà présentés dans le cours sur l'estimateur de Hortvitz Thompson

Pour l'estimation d'un total sur Y par exemple : $\widehat{T_B} = \sum_{i=1}^{N_B} y_i$

Individus dans l'échantillon	Wi (poids)
1	w1 =0.33/π1+0.33/π5 = 1/3(1/π1+1/π5)
2	w2 =0.33/π1+0.33/π3+0.33/π5= 1/3(1/π1+1/π3+1/π5)
3	w3 =0.25/π1+0.25/π5= 1/4(1/π1+1/π5)
4	w4 =0.5/π6= 1/2(1/π6)
5	w5 =0.5/π6= 1/2(1/π6)
6	w6 =1/π9
7	w7 =1/π9
8	w8 =1/π9
9	
10	
11	w11 =1/π12
12	w12 =0.33/π14+0.33/π15= 1/3(1/π14+1/π15)

Bases de sondage multiples entre telephones fixes et mobiles, en France

Mise en œuvre : dans le cas d'une enquête au téléphone, il faut bien penser à poser à chaque individu interrogé une question sur son équipement téléphonique et l'utilisation de de celui-ci (mobile, fixes, boxes …), afin de pouvoir calculer le partage des poids



le cnam

En résumé …

En résumé (Vincent Loonis)

Le « partage des poids » est une méthode d'estimation adaptée aux situations d'échantillonnage plus complexes : sondages indirects, bases multiples, panels rotatifs...

À l'instar de l'échantillonnage équilibré, si les idées sous-jacentes de ces techniques étaient déjà présentes au début des années quatre-vingt, leur application a été systématisée grâce à des travaux pionniers (Ernst, 1989 ; Deville, 1998) généralisés par la suite (Lavallée, 2002).

De manière simplifiée, le partage des poids intervient dès lors que les individus peuvent potentiellement être présents plusieurs fois dans l'échantillon final. Cette multiplicité peut être le fait des sondages indirects ou de l'existence de bases de sondages multiples.

http://www.insee.fr/fr/ffc/docs_ffc/cs126e.pdf

Applications à l'INSEE

Les domaines d'application et la théorie sous-jacente sont également trop vastes pour être exposés ici ; on se bornera à constater que le partage des poids intervient fréquemment et dans des domaines variés :

– Pondérations des enquêtes par panel (SRCV) ;

– Gestion de la charge de travail des enquêteurs (VQS 2006) ;

– Possibilité de cibler des sous-populations très spécifiques (Sans-domicile 2001, Logement 2006) ;

– Dans les études préliminaires pour la construction de l'échantillon de l'enquête Famille 2011 ;

– Dans la gestion au quotidien de toutes les enquêtes pour le cas de logements éclatés ou « fusionnés » ;

– Dans les enquêtes couplées (Famille-employeur 2004)

Applications en Instituts privés

Principalement sur les études par téléphone, avec du fait de l'absence d'annuaire universel et de l'évolution rapide de l'environnement (portabilité du numéro, mobilité entre opérateurs favorisée par l'arrivée de Free, usage majoritaire du téléphone mobile sur le fixe depuis 4 ans)

Un champ d'application récent : les études par téléphone mobile en Afrique, pour lesquelles la multiplicité de cartes SIM, d'opérateurs et de terminaux sont la règle pour les foyers.

le cnam

Bibliographie

Bibliographie

Ouvrage :

Lavallée, Pierre (2002), Le sondage indirect, ou la Méthode généralisée du partage des poids, Éditions de l' Université de Bruxelles , Belgique, Éditions Ellipses, France.

INSEE :

Loonis, Vincent : L'échantillonnage de la théorie à la pratique

http://www.insee.fr/fr/ffc/docs_ffc/cs126e.pdf

Deville, J.-C. (1998), Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? Suivi de: Comment attraper une population en se servant d' une autreI, NSEE Méthodes, No. 84-85-86, pp. 63-82.

Pascal Ardilly et Pierre Lavallée La méthode du partage des poids et l'enquête européenne sur le revenu et les condition de vie

http://www.agro-montpellier.fr/sfds/CD/textes/ardilly1.pdf

Statistique Canada :

Sondage indirect : les fondements de la méthode généralisée du partage des poids

http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-001-X20060029551&lang=fra

Fondements statistiques des enquêtes par téléphone mobile

http://www.statcan.gc.ca/pub/12-001-x/2010002/article/11382-fra.pdf

Autres plans de sondage :échantillonnage avec bases de sondage multiples chevauchantes

http://www.statcan.gc.ca/pub/12-001-x/2011002/article/11608-fra.pdf