le cnam

Sondages à probabilités inégales

STA108 Enquêtes et sondages

Octobre 2013

Philippe Périé, Gilbert Saporta

Sondages à probabilités inégales

- Les sondages aléatoires simples ne sont utilisés qu'en l'absence de toute autre information.
- D'autres méthodes permettent d'attribuer à chaque individu de la population une probabilité identique d'appartenir à l'échantillon, tout en gagnant en précision, mais en perdant en simplicité d'exécution. Nous le verrons plus tard, le sondage stratifié à allocation proportionnelle est dans ce cas.
- Les tirages à probabilités inégales sont une manière d'utiliser de l'information auxiliaire, c'est une approche assez complexe formellement et toujours très active en recherche.
- Il existe une infinité de plans à probabilités inégales et sans remise respectant des probabilités d'inclusion d'ordre 1 fixés a priori. Nous en présentons deux en fin de chapitre.
- Un autre point de complexité est que les probabilités d'inclusion d'ordre 2 sont liées aux algorithmes eux-mêmes, donc avec elles le calcul de la variance.

Résumé sur les plans de sondage classiques (source Pascal Ardilly 'Les techniques de sondage',1999)

	Plan de sondage	Réalisation du tirage et estimation	précision	coût terrain
	Sondage aléatoire simple	=	=	=
	Sondage stratifié allocation quelconque	-	+	=
	Sondage stratifié alloc proportionnelle	-	+	=
	Sondage stratifié allocation optimale	--	+++	=
	Sondage 'quelconque' à plusieurs degrés	--	-	+
	Sondage en grappes	-	--	++
	Sondages à probabilité inégales	-	Si Yi proportionnel à Xi ++, sinon --	=
	Sondage équilibré	---	+++	=
	Sondage par quota	-	?	++

Par rapport au SAS, les '+ 'et les '–' indiquent un gain/perte en termes de facilité de réalisation du tirage et estimation, précision, et coût terrain à taille égale

le cnam

Perception par le public du sondage à probabilités inégales

- Beaucoup de personnes ont l'impression qu'avec les probabilités inégales, le sondeur 'triche' en favorisant certaines unités. Le sondage ne serait plus 'représentatif', c'est faux bien sûr,
- Voici ce qu'en dit Yves Tillé :

3. Représentativité

L'objectif d'un sondage est de fournir un certain nombre d'informations sur une population en n'examinant qu'une partie de celle-ci, appelée l'échantillon. On dit souvent qu'un échantillon est représentatif d'une population s'il en constitue le modèle réduit. La ``représentativité'' est ainsi invoquée en tant qu'argument de validité : un bon échantillon devrait ``ressembler'' autant que possible à la population à étudier de sorte que certaines catégories apparaissent en même proportion dans l'échantillon et dans la population. Pourtant cette théorie, couramment véhiculée par les médias et même par certains ouvrages de méthodologie, est erronée (voir par exemple Ghiglione et Matalon, 1991, pp. 29-30, ou Javeau, 1988, pp. 42-46) : pour être valide, un échantillon ne doit pas être représentatif (au sens où nous venons de le définir). Il est, en effet, souvent souhaitable d'effectuer des tirages à probabilités inégales ou de sur-représenter certaines fractions de la population. Pour estimer avec précision un paramètre, il faut aller chercher l'information de manière judicieuse plutôt que d'accorder la même importance à chaque unité.

Exemple : EAE INSEE

- L'EAE est un enquête annuelle sur les entreprises

- Le plan de sondage est ici stratifié (sujet abordé plus loin dans le cours), les strates sont tirées avec des probabilités différentes les unes des autres

Institut national de la statistique et des études économiques

Échantillon de l'EAE Commerce de gros 2007 : poids de sondage par secteur d'activité (NAF 700) et par tranche de taille

Les entreprises du commerce de gros et les intermédiaires du commerce en 2007 - Résultats de l'enquête annuelle d'entreprise

Insee Résultats N° 44 Economie - septembre 2009

Synthèse des résultats

Données détaillées

Retour au sommaire de la publication

Échantillon de l'EAE Commerce de gros 2007 : poids de sondage par secteur d'activité (NAF 700) et par tranche de taille

En règle générale, **les entreprises occupant 20 salariés et plus sont interrogées exhaustivement. De même, sont interrogées exhaustivement les entreprises qui, quel que soit l'effectif, avaient un chiffre d'affaires supérieur à 38 millions d'euros lors de l'enquête précédente.**

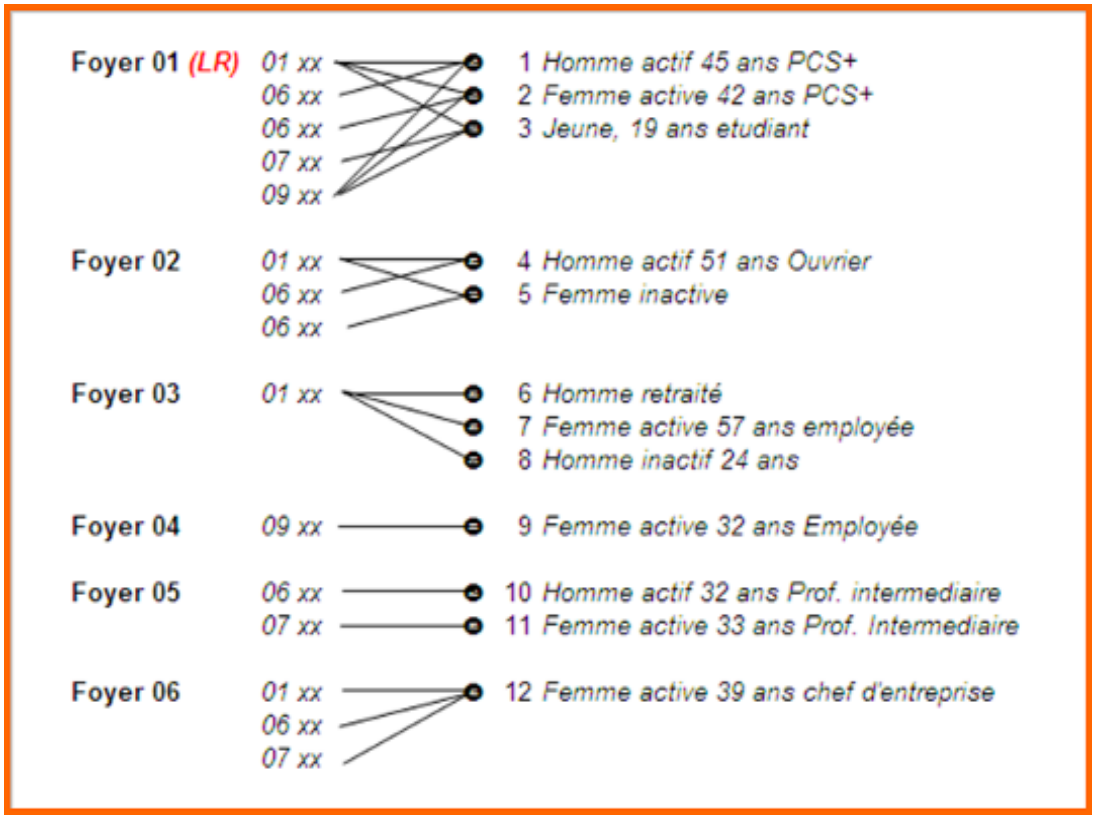
	Tranche de taille (nombre de salariés)				
Secteurs	**0 salarié**	**1 à 5 salariés**	**6 à 9 salariés**	**10 à 19 salariés**	**20 à 49 salariés**
46.11Z	25	3	2	2	1
46.12A	1	1	1	1	1
46.12B	3	3	1	2	1
46.13Z	9	7	3	1	1
46.14Z	14	6	4	3	1
46.15Z	14	9	2	1	1
46.16Z	20	7	2	1	1
46.17A	5	3	1	2	1
46.17B	10	4	1	1	1
46.18Z	33	9	5	3	1
46.19A	16	4	2	1	1
46.19B	23	6	1	1	1

Note : le poids est l'inverse du taux de sondage ; pour les entreprises de 50 salariés et plus, le tirage est exhaustif.

le cnam

Exemple : Enquête par téléphone sur numéros générés aléatoirement

- 99.7% de la population est équipée de téléphones fixes ou mobiles : la base de numéros est donc une base de sondage donc (presque) exhaustive
- En générant aléatoirement à probabilité égales des numéros fixes (01 à 05, 09) et mobiles (06 et 07), on réalise dans les faits un sondages à probabilités inégales : même si chaque numéro est généré avec la même probabilité, un individu aura une probabilité de sélection fonction du nombre et de la nature des lignes téléphoniques auxquelles il a accès.
- L'exemple ci contre est un exemple de **sondage dit indirect** entre une base de numéros de téléphone et une liste d'individus



Estimation

- Soit π_i la probabilité d'inclusion de l'individu i. On a : $\sum_{i=1}^{N} \pi_i = n$ pour un tirage de taille fixe n.
- Pour estimer le total, on peut utiliser l'estimateur : $\widehat{T} = \sum_{i \in s} \frac{Y_i}{\pi_i}$
- C'est l'estimateur de **Horvitz-Thompson** ou des valeurs dilatées. C'est une somme pondérée où chaque terme est l'inverse de la probabilité d'inclusion.
- On montre que $\hat{T}$ est le seul estimateur linéaire et sans biais du total : $E(\hat{T}) = \mathrm{T}$.
- Ceci est vrai si tous les individus ont une probabilité non nulle d'appartenir à l'échantillon : pas de défaut de couverture.
- Pour estimer la moyenne on divise simplement $\hat{T}$ par N

Un exemple

- Exemple (Ardilly) : une population de 5 communes, nombre d'habitants Y inconnu, nombre de logements X connu.
- Estimation du nombre moyen d'habitants par tirage à probabilités proportionnelles aux nombre de logements.
- On a les moyens d'échantillonner 2 communes

Communes	Nombre de logements = X	Nombre d'habitants = Y	Probabilité d'inclusion
(1) Antibes	48 812	70 688	0,99
(2) Cagnes	23 227	41 303	0,47
(3) St Laurent du Var	12 383	24 475	0,25
(4) Vence	9 341	15 364	0,19
(5) Villefranche/Mer	4 915	8 123	0,10
Moyenne	19 736	31 991	–

- On vérifie bien : $\sum_{i=1}^{N} \pi_i = n$: 0.99+0.47+0.25+0.19+0.1=2

Un exemple

- Il y a 10 échantillons possibles de deux communes: en voici la liste avec les estimateurs de Horvitz Thomson associés et l'estimateur que l'on obtiendrait avec un sondage aléatoire simple (SAS)
- Par exemple pour {1,2} : $\widehat{T} = \sum_{i \in S} \frac{Y_i}{\pi_i}$ =(1/0.99*70688+1/0.47*41303)=318561.5
- La moyenne est 31856.15

➔ Le calcul formel de la variance de l'estimateur dépend de l'algorithme de tirage utilisé, nous ne le développerons pas, mais au vu des données il est clair que le tirage à probabilités inégales est meilleur que le SAS

Échantillon s	$\widehat{\bar{Y}}$ (s)	$\bar{y}$ (s) (SAS)
1,2	31 856	55 996
1,3	33 860	47 582
1,4	30 453	43 026
1,5	30 526	39 406
2,3	37 156	32 889
2,4	33 748	28 334
2,5	33 822	24 713
3,4	35 753	19 920
3,5	35 826	16 299
4,5	32 419	11 744
Espérance	31 991	31 991

Précision

- Variance:

$$V(\hat{T}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) + \sum_{i \neq j}^{N} \sum \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j)$$

si n fixe formule de Yates-Grundy :

$$V(\hat{T}) = \frac{1}{2} \sum_{i \neq j}^{N} \sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 (\pi_i \pi_j - \pi_{ij})$$

- Estimation de la variance (par Horvitz-Thomson):

Première formule:

$$\widehat{V(\hat{T})} = \sum_{i \in s} y_i^2 \frac{1-\pi_i}{\pi_i^2} + \sum_{i \neq j \ \in s} \sum y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}}$$ peut être < 0

Deuxième formule:

$$\widehat{V(\hat{T})} = \frac{1}{2} \sum_{i,j \in s} \sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}$$

Précision

$$V(\hat{T}) = \frac{1}{2}\sum_{i\neq j}^{N}\sum\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \left(\pi_i\pi_j - \pi_{ij}\right)$$

- Pour avoir une variance nulle, il suffit que le terme $\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)$ soit égal à zéro pour tous les couples (i,j), donc que $\frac{Y_i}{\pi_i}$ soit constant.

- Supposons que l'on ait $Y_i = \alpha\pi_i$ avec α le coefficient de proportionnalité
- Alors on a : $T = \sum_{i=1}^{N} Y_i = \alpha \sum_{i=1}^{N} \pi_i = \alpha n$ où n est la taille fixe de l'échantillon
- Quel que soit l'échantillon tiré on a : $\hat{T} = \sum_{i\in s} \frac{Y_i}{\pi_i} = \sum_{i\in s} \frac{\alpha\pi_i}{\pi_i} = \alpha n = T$

- Ce résultat est somme toute logique : connaitre une variable parfaitement proportionnelle à Y sur toute la population revient à connaitre Y …

Précision

- En pratique , la formule de Yates Grundy montre que l'on a intérêt à tirer proportionnellement aux valeurs d'une variable auxiliaire X corrélée (positivement!) à Y.

- Intéressant en cas d'effet taille (CA, nb d'employés, bénéfice…)

Calcul des probabilités d'inclusion

- Calcul des probabilités d'inclusion : $\pi_i = \frac{nx_i}{\sum_{i=1}^{N} x_i}$
- Exemple: tirage de 3 individus parmi 6 proportionnellement à : x1=300 x2=90 x3=70 x4=50 x5=20 x6=20

	Taille	Probabilité d'inclusion		Probabilité d'inclusion (2)
x1	300	1.67		
x2	90	0.50	x2	0.75
x3	70	0.39	x3	0.59
x4	40	0.22	x4	0.33
x5	20	0.11	x5	0.17
x6	20	0.11	x6	0.17

- On sélectionne d'office l'unité X1, puis on recommence la sélection de 2 unités parmi 5

Une infinité de méthodes respectant les probabilités d'inclusion d'ordre 1

- Comme nous l'avons dit en introduction, le problème est encore très ouvert théoriquement
- Yves Tillé (dans son ouvrage sur la théorie des sondages) liste 4 critères pour ce qu'il appelle une 'bonne' procédure
  a. Exactitude : les probabilité d'inclusion d'ordre 1 sont respectées
  b. Taille fixe : la taille de l'échantillon est de taille fixe n
  c. Généralité : la méthode est applicable à toutes les probabilités et les tailles d'échantillon relativement à la population
  d. Sans remise : une unité ne peut être sélectionnée qu'une seule fois
- Il cite aussi :
  - ⇨ Les probabilités d'inclusion d'ordre 2 doivent être strictement positives (toutes les combinaisons sont possibles)
  - ⇨ Les probabilités doivent respecter les conditions de Sen-Yates-Grundy $\left(\pi_{ij} \leq \pi_i \pi_j\right)$
  - ⇨ La méthode doit fournir des probabilités d'inclusion d'ordre 2 telles que la variance de l'estimateur doivent être aisée à mettre en œuvre
  - ⇨ L'algorithme doit être rapide et en particulier éviter l'énumération de tous les échantillons possibles de taille n dans N
  - ⇨ L'algorithme doit être séquentiel, c'est-à-dire qu'il doit pouvoir s'appliquer en une passe sur le fichier de donnée

Un exemple de plan : Plan de Poisson

- Une généralisation du plan de Bernoulli pour les probabilités inégales
- Chaque unité est sélectionnée de manière indépendante avec une probabilité π_i
- Une population de taille N, i un entier, u un réel
- Répéter pour k = 1, ..N
  - ⇨ Tirer u dans une loi uniforme U[0,1]
  - ⇨ Si u < π_i, sélectionner l'unité i, sinon passer l'unité i

Un exemple de plan : Plan de Poisson

Avantages :

- Les unités sont sélectionnées indépendamment, donc on a $\pi_{ij} = \pi_i \pi_j$. Le calcul de la variance est très facile
- Le plan est à entropie maximale pour des probabilités d'inclusion données (c'est-à-dire que compte tenu des probabilités d'inclusion, c'est le plan qui est le plus 'aléatoire' possible)

Inconvénient :

- La méthode a un gros désavantage, la taille de l'échantillon est aléatoire, il y a donc une chance (petite …) d'avoir un échantillon vide ou de sélectionner toute la population

Dans Excel avec la fonction ALEA()

Individu	probabilité	selection ?	formule
1	0.8	1	*=(ALEA()<E5)*1*
2	0.3	0	*=(ALEA()<E6)*1*
3	0.4	1	*=(ALEA()<E7)*1*
4	0.2	0	*=(ALEA()<E8)*1*
5	0.6	0	*=(ALEA()<E9)*1*
6	0.7	1	*=(ALEA()<E10)*1*

Un exemple de plan : Sondage systématique à probabilités inégales

- On cumule pour tous les individus les probabilités d'inclusion

  ⇨ $Vk = \pi 1 + \pi 2 + ... + \pi k$

  ⇨ On génère une seule réalisation u de la loi U[0,1[

  ⇨ On sélectionne k tel que $Vk-1 \leq u < V\,k$

  ⇨ puis i tel que $Vi-1 \leq u + 1 < Vi$

  ⇨ puis j tel que $Vj-1 \leq u + 2 < Vj$

- etc ... on obtient in fine n individus

Un exemple de plan : Sondage systématique à probabilités inégales

Avantages :

- Simplicité
- Echantillon de taille fixe

Inconvénient :

- certaines probabilités d'inclusion d'ordre 2 peuvent être nulles
- Dépend de l'ordre du fichier
- Tri aléatoire avant tirage?

Un exemple de plan : Sondage systématique à probabilités inégales

- Exemple : N = 6, n = 3

  $\pi 1=0.2$

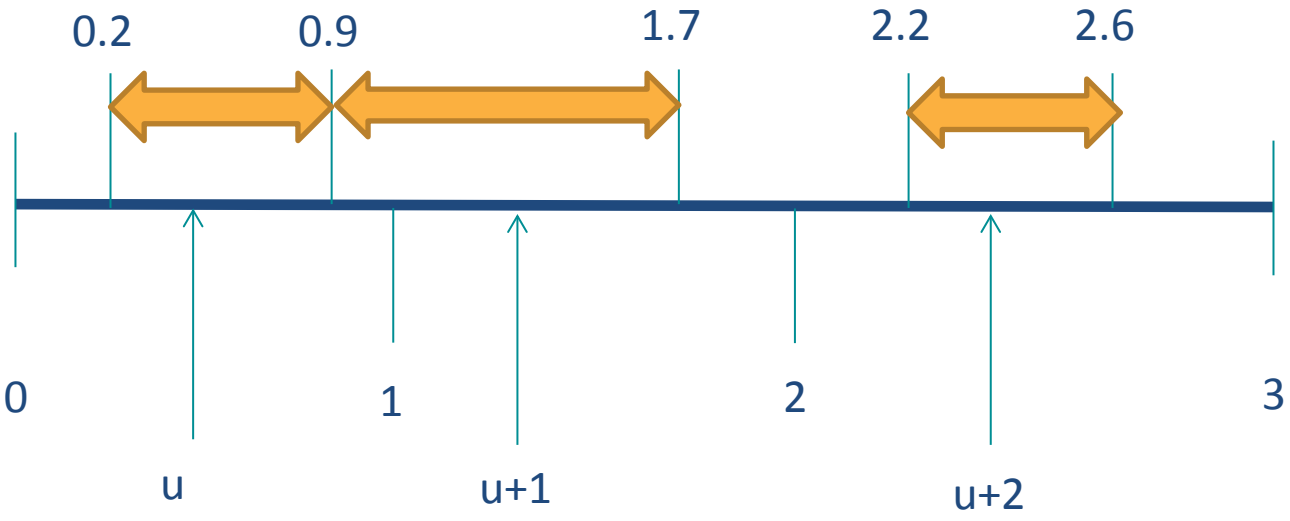
  $\pi 2=0.7$

  $\pi 3=0.8$

  $\pi 4=0.5$

  $\pi 5=P6=0.4$

- Supposons u = 0.3658 : les unités 2,3 et 5 sont sélectionnées : entre 0.2 et 0.9 c'est l'unité 2, entre 0.9 et 1.7 c'est l'unité 3, entre 2.2 et 2.6 c'est l'unité 5 …



Sondage systématique à probabilités inégales

- La méthode a des probabilités d'inclusion d'ordre 2 nulles. Par exemple, si 1 est sélectionné, il est impossible de sélectionner les unités 2, 5 et 6
- (En fait deux unités séparées par une distance inférieure au 'pas' de sondage ne peuvent pas être sélectionnées simultanément)

- La matrice des probabilités d'inclusion d'ordre 2 est :

-	0	0.2	0.2	0	0
0	-	0.5	0.2	0.4	0.3
0.2	0.5	-	0.3	0.4	0.2
0.2	0.2	0.3	-	0	0.3
0	0.4	0.4	0	-	0
0	0.3	0.2	0.3	0	-