Théorie des sondages : cours 3
Sondages par grappes et à deux degrés

Camelia Goga

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Master Besançon-2010

Préliminaires

Dans le chapitre précédant, on a étudié les plans de sondage "directs" des individus lors d'une seule étape. En pratique, ce type de sondage est souvent impossible à cause des plusieurs facteurs :

1. l'absence d'une base de sondage ou le prix pour en fabriquer une est trop élevé ;
2. les éléments de la population sont trop dispersés sur une aréa géographique et le coût d'interroger les individus serait élevé aussi. Dans ce type de situations, les taux de nonréponse sont aussi importants.

Exemple : On veut mesurer l'usage de drogue parmi les élèves de terminale. On sélectionne pour cela un échantillon de lycées et dans les lycées choisis, on sélectionne un échantillon de classes terminales. Tous les élèves de classes terminales ainsi choisies seront interrogés.

Chapitre 3.1 : Sondages en grappes

1. U est composée de N_I **grappes** $U_1, \ldots, U_i, \ldots, U_{N_I}$ de tailles $|U_i| = N_i$ for $i = 1, \ldots, N_I$ avec N_i souvent inconnu :

$$U_I = \{U_1, \ldots, U_i, \ldots, U_{N_I}\} = \{1, \ldots, i, \ldots, N_I\}.$$

2. **Premier degré** : Un échantillon $s_I \subset U_I$ de **grappes** de taille n_{s_I} est obtenu selon un plan $p_I(\cdot)$.
   Probabilités d'inclusion des grappes :

$$\pi_{Ii} = P(i \in s_I),$$

$$\pi_{Iij}, \quad i \neq j \in s_I$$

3. Pour chaque $i \in s_I$, on observe toutes les unités de U_i.
4. L'échantillon d'unités observées est

$$s = \cup_{i \in s_I} U_i$$

Probabilités d'inclusion des éléments

$$\pi_k = P(k \in s) = P(i \in s_I) = \pi_{Ii} \quad \text{si} \quad k \in U_i$$

$$\pi_{kl} = \begin{cases} \pi_{Ii} & \text{si} \quad k \neq l \in U_i \\ \pi_{Iij} & \text{si} \quad k \in U_i, l \in U_j \end{cases}$$

Objectif : estimer le total

$$t_y = \sum_U y_k = \sum_{U_I} t_i,$$

$$t_i = \sum_{U_i} y_k \quad \text{le total du grappe i}$$

Résultat : Pour un sondage par grappes, le π-estimateur est

- $\hat{t}_\pi = \sum_{s_I} \frac{t_i}{\pi_{Ii}}$;
- $V(\hat{t}_\pi) = \sum_{U_I} \sum_{U_I} \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}$;
- $\widehat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}$.

Remarques :

• comme la taille N est inconnue, la moyenne $\overline{y}_U$ ne peut pas être estimée par simple division par N dans $\hat{t}_\pi$ de dessus. Cette situation sera traitée dans le chapitre 4 ;

• le plan SYS peut être considéré comme un sondage par grappe avec $n_I = 1$ et la population U est constituée de N_I grappes correspondants aux a échantillons possibles.

Résultat

Pour un sondage en grappe de taille fixe n_I, nous avons

1. $V(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_{U_I} \Delta_{Iij}\left(\frac{t_i}{\pi_{Ii}} - \frac{t_j}{\pi_{Ij}}\right)^2$ *estimé par*
2. $\widehat{V}(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_{s_I} \frac{\Delta_{Iij}}{\pi_{Iij}}\left(\frac{t_i}{\pi_{Ii}} - \frac{t_j}{\pi_{Ij}}\right)^2$.

Remarques

• si tous les t_i/π_{Ii} sont égaux alors, $V(\hat{t}_\pi) = 0$; il résulte que si on prend $\pi_{Ii} \propto t_i$ alors, le sondage par grappe sera efficace ;

• si N_i est connu, on peut prendre $\pi_{Ii} \propto N_i$ qui sera un bon choix s'il y a peu de variation entre les moyennes de grappes.

Sondage SAS de grappes (SASG)

Nous considérons un échantillon SAS de grappes de taille n_I parmi N_I et dans chaque cluster sélectionné, on fait un recensement.

$$\pi_{Ii} = \frac{n_I}{N_I}, \quad \pi_{Iij} = \frac{n_I(n_I-1)}{N_I(N_I-1)}.$$

Résultat : Pour un sondage SAS de grappes, le π-estimateur est

- $\hat{t}_\pi = N_I \sum_{s_I} \frac{t_i}{n_I} = N_I \bar{t}_{s_I}$ *avec* $\bar{t}_{s_I} = \sum_{s_I} \frac{t_i}{n_I}$;
- $V_{SASG}(\hat{t}_\pi) = N_I \frac{1-f_I}{n_I} S^2_{tU_I}$;
- $\widehat{V}_{SASG}(\hat{t}_\pi) = N_I \frac{1-f_I}{n_I} S^2_{ts_I}$.

Exemple

Efficacité de SAS de grappes : 1

- On utilise $\delta = 1 - \frac{S_{yW}^2}{S_{yU}^2}$ coefficient d'homogénéité pour

$$S_{yW}^2 = \frac{1}{N - N_I} \sum_{i \in U_I} \sum_{k \in U_i} (y_k - \overline{y}_{U_i})^2$$ la variance intra-grappes.

- On a pour $S_{yU_i}^2 = \sum_{U_i} \sum_{k \in U_i} (y_k - \overline{y}_{U_i})^2 / (N_i - 1)$ que

$$S_{yW}^2 = \frac{\sum_{i \in U_I} (N_i - 1) S_{yU_i}^2}{\sum_{i \in U_I} (N_i - 1)}$$

ou S_{yW}^2 est moyenne pondérée des variances des N_I grappes.

- ANOVA : $\underbrace{(N-1)S_{yU}^2}_{SST} = \underbrace{(N - N_I)S_{yW}^2}_{SSW} + \underbrace{\sum_{i \in U_I} N_i (\overline{y}_{U_i} - \overline{y}_U)^2}_{SSB}$

Efficacité de SAS de grappes : 2

Propriété : $-\dfrac{N_I - 1}{N - N_I} \leq \delta \leq 1.$

Conséquence :

- si δ petit, alors les éléments du même grappe ne sont pas homogènes et δ est grand dans le cas contraire ;
- le cas extrême $\delta = 1$ correspond à une variation nulle à l'intérieur de chaque grappe (homogénéité complète) ;
- le cas extrême $\delta = -\dfrac{N_I - 1}{N - N_I}$ implique que les moyennes de grappes sont égales ou une variation nulle entre les grappes ;

Comparaison de SAS et SAS de grappes :1

Considérons un sondage SAS de taille $n = n_I \overline{N}$ pour $\overline{N} = N/N_I$ le nombre moyenne d'éléments par grappe.
On calcule $deff(SASG, \hat{t}_\pi)$.

- Si $N_i = N$ pour tous les $i \in U_I$, alors

$$deff(SASG, \hat{t}_\pi) = \frac{V_{SASG}(\hat{t}_\pi)}{V_{SAS}(\hat{t}_\pi)} = 1 + \frac{N - N_I}{N_I - 1}\delta \simeq 1 + (\overline{N}_1 - 1)\delta$$

ce qui implique que $V_{SASG} < V_{SAS}$ si est seulement si $\delta < 0$ ou si on a une variation assez grande à l'intérieur des grappes. En pratique, on a plutôt le contraire puisque les éléments de la même grappe se rassemblent beaucoup entraînant une variance du plan SASG supérieure à celle du SAS.

Comparaison de SAS et SAS de grappes : 2

• Si N_i varient, la variance V_{SASG} accroît encore plus par rapport au cas précédant ce qui fait que dans cette situation, le plan SASG est encore pire que le plan SAS.

Conclusion : le plan SASG est dans la plupart des cas moins efficace que le plan SAS surtout si les grappes sont homogènes et/ou des tailles différentes. Et pourtant, ce plan est utilisé beaucoup car parfois, la situation pratique en impose son usage (enquête auprès des médecins) ou puisque il coûte moins cher d'interroger des individus de la même grappe.

Améliorer son efficacité :
• en faisant un sondage πps des grappes avec $\pi_{Ii} \propto u_i$ pour u_i une variable proportionnelle à t_i et connue pour chaque grappe.
• en stratifiant les grappes selon la variable u_i : à l'intérieur de chaque strate la variation de u_i est petite.

Exemple

Chapitre 3.2 : Sondages à deux degrés

1. U est composée de N_I **unités primaires UPs** $U_1, \ldots, U_i, \ldots, U_{N_I}$ de tailles $|U_i| = N_i$ for $i = 1, \ldots, N_I$ avec N_i souvent inconnu :

$$U_I = \{1, \ldots, i, \ldots, N_I\}.$$

2. **Premier degré** : Un échantillon $s_I, s_I \subset \mathcal{U}_I$ de **UP** est obtenu selon un plan $p_I(\cdot)$. Le nombre de **UP** est n_{s_I}.
3. **Second degré** : Pour chaque $i \in s_I$, un échantillon $s_i \subset U_i$ **d'individus** est obtenu selon un plan $p(\cdot|s_I)$.

L'échantillon final est $s = \cup_{i \in s_I} s_i$ de taille $n_s = \sum_{i \in s_I} n_{s_i}$.

Les probabilités d'inclusion

- **Pour les UP ou pour le premier degré** : $i, j \in U_I$

$$\pi_{Ii} = P(i \in s_I) \quad \text{et} \quad \pi_{Iij} = P(i\&j \in s_I)$$

$$\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}.$$

- **Pour les individus ou pour le deuxième degré** : On suppose *l'indépendance* : $P(\cup s_i|s_I) = \Pi_{i\in s_I} P(s_i|s_I)$, la sélection des individus dans une **UP** se réalise de façon indépendante de la sélection dans les autres **UP**-s.
l'invariance : $p_i(\cdot|s_I) = p_i(\cdot)$ pour $i \in s_I$,
Les probabilités correspondantes au plan $p_i(\cdot)$ sont pour $k, l \in U_i$:

$$\pi_{k|i} = P(k \in s_i|i \in s_I) \quad \text{et} \quad \pi_{kl|i} = P(k\&l \in s_i|i \in s_I)$$

$$\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$$

Les probabilités d'inclusion relatives au plan à deux degrés :

$$\pi_k = \pi_{I i}\pi_{k|i} \quad \text{si} \quad k \in U_i$$

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{kli} & \text{si} \quad k = l \in U_i \\ \pi_{Ii}\pi_{kl|i} & \text{si} \quad k \neq l \in U_i \\ \pi_{Iij}\pi_{k|i}\pi_{l|j} & \text{si} \quad k \in U_i, l \in U_j, i \neq j \end{cases}$$

On veut construire le π- estimateur $\hat{t}_\pi$ pour le total

$$t_y = \sum_U y_k = \sum_{U_I} t_i \quad \text{avec} \quad t_i = \sum_{U_i} y_k$$

Dans le plan à deux degrés, on conditionne par rapport au plan utilisé lors du premier degré et on utilise :

$$E(\hat{t}_\pi) = E_I E_{II}(\hat{t}_\pi)$$

$$V(\hat{t}_\pi) = E_I(V_{II}(\hat{t}_\pi)) + V_I(E_{II}(\hat{t}_\pi))$$

Conditionnellement au premier degré et pour $i \in s_I$, le total t_i est estimé par

$$\hat{t}_{i\pi} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$$

avec la variance

$$V_i = V(\hat{t}_{i\pi}) = \sum_{U_i} \sum_{U_i} \Delta_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

$$\hat{V}_i = \sum_{s_i} \sum_{s_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

Le π estimateur pour le total dans le cas d'un plan à deux degrés

Résultat

Pour un plan à deux degrés, le π estimateur pour le total $t_y = \sum_U y_k$ est

$$\hat{t}_\pi = \sum_{s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}}$$

avec $\hat{t}_{i\pi}$ le π estimateur pour le total t_i par rapport au deuxième degré. La variance est égale à $V(\hat{t}_\pi) = V_{UP} + V_{US}$ avec

$$V_{UP} = \sum_{U_I} \sum_{U_I} \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}} \quad et \quad V_{US} = \sum_{U_I} \frac{V_i}{\pi_{Ii}}.$$

Un estimateur de la variance est donné par

$$\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$$

Remarques

1. les conditions dans lesquelles les deux termes de la variance sont nuls :
   - si $s_I = U_I$ avec une probabilité égale à 1, alors $\pi_{Ii} = \pi_{Iij} = 1$ pour tous i, j. On a $V_{UP} = 0$ et $V_{US} = \sum_{U_I} V_i$.
     le sondage stratifié avec des strates = UP
   - si $s_i = U_i$ avec probabilité 1 pour tous i, alors $V_{US} = 0$;
     le sondage par grappes.
2. le deuxième terme de l'estimateur de la variance $\sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$ est (très) long à calculer, il est éliminé de l'expression de l'estimateur de la variance et on prend comme estimateur :

$$\hat{V}^*(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}}$$

Le plan à deux degrés : SAS, SAS

On considère le plan particulier qui consiste à sélectionner un échantillon aléatoire simple à chaque degré.
Premier degré avec un plan SAS : on prend un échantillon s_I dans U_I de taille n_I parmi N_I UP :

$$\pi_{Ii} = \frac{n_I}{N_I}, \quad \pi_{Iij} = \frac{n_I(n_I-1)}{N_I(N_I-1)}$$

Deuxième degré avec un plan SAS : pour chaque $i \in s_I$, on prend un échantillon s_i dans U_i de taille n_i parmi N_i,

$$\pi_{k|i} = \frac{n_i}{N_i}, \quad \pi_{kl|i} = \frac{n_i(n_i-1)}{N_i(N_i-1)}$$

Le π estimateur : $\hat{t}_\pi = \frac{N_I}{n_I} \sum_{s_I} N_i \overline{y}_{s_i} = \frac{N_I}{n_I} \sum_{s_I} \hat{t}_{i\pi}$
La variance est :

$$V(\hat{t}_\pi) = N_I^2 \frac{1 - f_I}{n_I} S_{tU_I}^2 + \frac{N_I}{n_I} \sum_{U_I} N_i^2 \frac{1 - f_i}{n_i} S_{yU_i}^2$$

avec $S_{tU_I}^2 = \frac{1}{N_I - 1} \sum_{U_I} (t_i - \overline{t}_{U_I})^2$, $S_{yU_i}^2 = \frac{1}{N_I - 1} \sum_{U_i} (y_k - \overline{y}_{U_i})^2$
Exemple : Premier degré avec un plan SAS : $n_I = 5$ parmi $N_I = 50$ UPs ;
Deuxième degré avec un plan SAS : pour chaque $i \in s_I$, on prend un échantillon s_i dans U_i de taille $n_i = 3$ parmi N_i;

i	N_i	y_k
19	5	41,49,49
45	8	49,49,45
47	5	31, 31, 35
50	9	39, 41, 61
31	7	49, 51, 33