Plans à probabilités inégales

Enquêtes et Sondages - CNAM - UE STA 108

Sylvie Rousseau

28/10/05

Sommaire

1. Pourquoi échantillonner à probabilités inégales ?
   - Généralités
   - Exemple
2. Comment choisir les probabilités d'inclusion ?
   - Le cas des plans proportionnels à la taille
3. Comment estimer un total ?
   - Définition de l'estimateur de Horvitz-Thompson
   - Propriétés de l'estimateur de Horvitz-Thompson
     - Espérance
     - Variance
     - Estimateur de la variance
4. Comment estimer une moyenne ?
   - L'estimateur de Horvitz-Thompson
   - L'estimateur de Hájek
   - L'exemple de Basu
5. Comment construire des intervalles de confiance pour le total ou la moyenne ?

1.1. Pourquoi échantillonner à probabilités inégales ?

- Pour retenir de préférence les unités les plus porteuses de l'information

➔ Gain de précision par rapport à un SAS où toutes les unités ont même importance

1.1. Quand échantillonner à probabilités inégales ?

- Lorsque les unités de la population étudiée contribuent inégalement au total d'intérêt.

Exemple :

Pour estimer la production d'un secteur que l'on sait assurée par 2 géants du secteur et des centaines de PME, il est légitime de sélectionner d'office les 2 grandes entreprises et échantillonner de manière aléatoire quelques PME.

1.1 Principe des plans à probabilités inégales

Aller chercher l'information là où elle se trouve.

➔ Ce qui suppose de disposer
avant l'échantillonnage
d'information auxiliaire
connue sur toute la population
et liée avec le caractère d'intérêt.

1.2. Exemple

- Population de 4 entreprises A, B, C et D de 500, 100, 30 et 20 salariés
- Admettons que l'on veuille estimer le nombre total de salariés (certes connu : 650) à partir d'un échantillon de taille 2
- Comparons les 2 tirages suivants :
  - un sondage aléatoire simple sans remise
  - un échantillonnage à probabilités inégales

1.2. Sondage aléatoire simple

- Il y a $C_4^2 = 6$ échantillons possibles

Echantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés des 4 entreprises
{A, B}	1/6	1 200
{A, C}		1 060
{A, D}		1 040
{B, C}		260
{B, D}		240
{C, D}		100

- En moyenne, on estime parfaitement bien le vrai effectif total de 650 salariés.
- Mais l'estimateur est très dispersé : sa variance vaut 207 567 (CV ≅ 70%).

1.2. Plan à probabilités inégales n°1

- Avec les probabilités d'inclusion suivantes :

Entreprise k	Effectif salarié X_k	Probabilité d'inclusion π_k
A	500	1
B	100	0,5
C	30	0,25
D	20	0,25

- 3 échantillons sont possibles :

Echantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés des 4 entreprises
{A, B}	0,5	700
{A, C}	0,25	620
{A, D}	0,25	580

- En moyenne, l'estimateur est aussi sans biais.
- Et sa variance est beaucoup plus faible :
  elle vaut 2 700 (CV ≅ 8%).

1.2. Plan à probabilités inégales n°2

- Avec les probabilités d'inclusion suivantes :

Entreprise k	Effectif salarié X_k	Probabilité d'inclusion π_k
A	500	0.25
B	100	0,25
C	30	0,5
D	20	1

- 3 échantillons sont possibles :

Echantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés des 4 entreprises
{A, D}	0,25	2 020
{B, D}	0,25	420
{C, D}	0,5	80

- En moyenne, l'estimateur est aussi sans biais.
- Mais il s'avère extrêmement dispersé : sa variance vaut 644 900 (CV ≅ 124%).

1.2. Conclusion

Il peut donc s'avèrer très intéressant de sélectionner les unités avec des probabilités d'inclusion proportionnelles aux valeurs prises pour un caractère *x*,

- lié positivement avec le caractère d'intérêt
- et connu pour tous les individus de la base de sondage.

2. Comment choisir les probabilités d'inclusion ? (1)

- Cas des plans proportionnels à la taille (ppt)
- Exemple d'une enquête qui s'intéresse au chiffre d'affaires d'entreprises d'un secteur donné.
  - Si l'on dispose du nombre de salariés de toutes les entreprises de ce secteur,
  - Et si on pressent que le chiffre d'affaires est plus ou moins proportionnel au nombre de salariés,

  ➔ il est légitime de calculer les probabilités d'inclusion de toutes les entreprises de manière proportionnelle à leur effectif salarié

2. Comment choisir les probabilités d'inclusion ? (2)

- Pour un échantillon de taille fixe *n*, on calcule la probabilité de sélectionner la *k*^{ème} entreprise :

$$\forall k \in U,\ \pi_k = P(k \in S) = n \frac{X_k}{\sum_{k \in U} X_k}$$

où X_k désigne la variable de taille, ici le nombre de salariés de la *k*^{ème} entreprise de la population *U*.

- On vérifie que : $\sum_{k \in U} \pi_k = n$ (plan de taille fixe).

2. Comment choisir les probabilités d'inclusion ? (3)

- Cependant certaines probabilités d'inclusion π_k peuvent dépasser 1.

Dans ce cas,

- On sélectionne d'office les unités en question : $\pi_k = 1$ (strate dite exhaustive)
- On recalcule les probabilités d'inclusion des autres individus proportionnellement à :
  - la taille de l'échantillon restant à sélectionner
  - et à leur contribution dans le nouveau total

Quitte à réitérer la démarche jusqu'à ce que $\pi_k \leq 1, \quad \forall k \in U$

Notations dans la population

- Population finie *U* de *N* objets identifiables (ou individus, unités statistiques) :

$$U = \{1,2,...,k,...,N\}$$

- Variable d'intérêt Y de caractéristique individuelle Y_k

- Total : $T_y = \sum_{k \in U} Y_k$

- Moyenne : $\mu_y = \frac{T_y}{N} = \frac{1}{N} \sum_{k \in U} Y_k$

Notations dans l'échantillon

- Echantillon *s* : sous-ensemble de *U* de taille *n*
- Ensemble des échantillons possibles : $\mathcal{S}$
- Plan de sondage probabiliste : loi de probabilité sur $\mathcal{S}$

$$p(s) \geq 0, \forall s \in S, \text{ et } \sum_{s \in S} p(s) = 1.$$

- Probabilité d'inclusion d'ordre un de *k* :

$$\pi_k = P(k \in s) = \sum_{k \ni s} p(s) = E(I_k) \quad \text{où} \quad I_k = \begin{cases} 1 & k \in S \\ 0 & k \notin S \end{cases}$$

- Probabilité d'inclusion ou double de *k* et *l* ($k \neq l$) :

$$\pi_{kl} = p(k \in s, l \in s) = \sum_{k,l \ni s} p(s) = E(I_k I_l)$$

- Covariance entre I_k et I_l pour *k* et *l* :

$$\Delta_{kl} = E(I_{kl}) - E(I_k)E(I_l) = \pi_{kl} - \pi_k \pi_l$$

3.1. Comment estimer un total ?

- En 1952, Horvitz et Thompson ont proposé l'estimateur suivant du total T_y de la variable *Y :*

$$\hat{t}_{y\pi} = \sum_{k \in s} \frac{Y_k}{\pi_k}$$

- On l'appelle aussi estimateur par les valeurs dilatées ou π-estimateur.

3.1. Remarques sur le π-estimateur

- C'est un estimateur linéaire.
- Les poids de sondage ne dépendent pas de l'échantillon.
- Il permet d'estimer la taille *N* de la population, qu'elle soit connue ou non :

$$\hat{N}_\pi = \sum_{k \in s} \frac{1}{\pi_k}$$

- Il est valable quel que soit le plan de sondage.
- Il généralise les résultats du sondage aléatoire simple sans remise de taille fixe *n*, où $\pi_k = \frac{n}{N}$ pour tout *k* de *U*.

3.2. Espérance du π-estimateur de T_y

- Si $\pi_k > 0$ pour tout individu k de la population U, alors l'estimateur d'Horvitz et Thompson du total est sans biais.
- Si certaines probabilités d'inclusion sont nulles, alors l'estimateur est biaisé. Ce biais ne dépend que des unités qui n'ont aucune chance d'être échantillonnées : on parle de problème de couverture.

3.3. Variance du π-estimateur de T_y *(1)*

- *Dans le cas général*

  Si $\pi_k > 0$ pour tout *k* de *U*, alors la variance de l'estimateur d'Horvitz et Thompson du total vaut :

$$\mathrm{V}ar\left(\hat{t}_{y\pi}\right) = \sum_{k\in U}\sum_{l\in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$$

3.3. Variance du π-estimateur de T_y *(2)*

- *Dans le cas d'un plan de taille fixe*

  Si $\pi_k > 0$ pour tout *k* de *U* et si le plan est de taille fixe, alors Sen, Yates et Grundy ont montré que la variance de l'estimateur d'Horvitz et Thompson du total peut aussi s'écrire :

$$\mathrm{V}ar\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2 \Delta_{kl}$$

3.4. Comment estimer la variance du π-estimateur de T_y ?

- Cas général : Si $\pi_{kl} > 0$ pour tous *k* et *l* de *U*, alors la variance de l'estimateur d'Horvitz et Thompson du total peut être estimée sans biais par :

$$\hat{\mathrm{V}}ar_1\left(\hat{t}_{y\pi}\right) = \sum_{k\in s}\sum_{l\in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$$

- Cas d'un plan de taille fixe : Si $\pi_{kl} > 0$ pour tous *k* et *l* de *U* et si le plan est de taille fixe, alors la variance de l'estimateur d'Horvitz et Thompson du total peut aussi être estimée sans biais par :

$$\hat{\mathrm{V}}ar_2\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$$

3.4. Comparaison de ces 2 estimateurs

- Le 1er estimateur est valable dans le cas général.
- Le 2ème estimateur n'est valable que si le plan est de taille fixe.
- Dans le cas où est le plan est de taille fixe, on dispose généralement de 2 estimateurs concurrents et différents.
- Ils sont tous les 2 sans biais dès que tous les π_{kl} sont strictement positifs quels que soient les individus *k* et *l* de la population.
- Les 2 estimateurs peuvent prendre des valeurs négatives, mais il existe une condition suffisante pour que le second estimateur soit positif. Cette condition, dite condition de Sen-Yates-Grundy, est

$$\forall k \neq l \in U \quad \Delta_{kl} \leq 0 \quad \text{soit :} \quad \pi_{kl} - \pi_k \pi_l \leq 0$$

4.1. Comment estimer une moyenne ?

- Lorsque la taille de la population est connue, on peut estimer la moyenne de *Y* avec l'estimateur de Horvitz et Thompson :

$$\hat{\mu}_{y\pi} = \frac{1}{N}\sum_{k\in s}\frac{Y_k}{\pi_k} = \frac{\hat{t}_{y\pi}}{N}$$

- Dans le cas particulier où la variable d'intérêt *Y* est dichotomique et vaut 1 dans *p%* des cas, le π-estimateur de la proportion *p* vaut :

$$\hat{p}_{\pi} = \frac{1}{N}\sum_{k\in s}\frac{Y_k}{\pi_k}$$

4.1. Propriétés du π-estimateur de μ_y

- Les propriétés vues pour l'estimateur de Horvitz et Thompson d'un total s'appliquent à une moyenne, en adaptant les formules pour tenir compte du coefficient *N*.
- Estimateur sans biais : $\mathrm{E}\left(\hat{\mu}_{y\pi}\right) = \mu_y$
- De variance :

$$\mathrm{V}ar\left(\hat{\mu}_{y\pi}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl} \quad \text{ou} \quad \mathrm{V}ar\left(\hat{\mu}_{y\pi}\right) = -\frac{1}{2N^2}\sum_{k\in U}\sum_{l\in U}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$$

- Estimée par :

$$\hat{\mathrm{V}}ar_1\left(\hat{\mu}_{y\pi}\right) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \quad \text{ou} \quad \hat{\mathrm{V}}ar_2\left(\hat{\mu}_{y\pi}\right) = -\frac{1}{2N^2}\sum_{k\in s}\sum_{l\in s}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$$

4.2. L'estimateur de Hájek de μ_y

- Lorsque la taille de la population est inconnue, on peut estimer la moyenne de *Y* avec l'estimateur de Hájek (1971) :

$$\hat{\mu}_{yH} = \frac{1}{\hat{N}_\pi} \sum_{k \in s} \frac{Y_k}{\pi_k} = \frac{\sum_{k \in s} \frac{Y_k}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}} = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}$$

4.2. Remarques sur l'estimateur de Hájek

- L'estimateur de Hájek est un ratio de 2 π- estimateurs.
- Il est biaisé : on montre que le biais est asymptotiquement nul et on le considéra négligeable lorsque la taille de l'échantillon est grande.
- Sa précision peut s'avèrer supérieure à celle de l'estimateur d'Horvitz-Thompson.
- Si $\forall k \in U, \quad Y_k = a$ (a constante réelle quelconque), alors : $\begin{cases} \hat{\mu}_{yH} = a = \mu_y \\ \hat{\mu}_{y\pi} = a \dfrac{\hat{N}_\pi}{N} \end{cases}$

4.3. Exemple de Basu (1)

- Basu (1971) relate l'exemple suivant : le propriétaire d'un cirque souhaite estimer le poids total de son troupeau de 50 éléphants.
- Il y a 3 années de cela Sambo représentait le poids moyen du troupeau.
- En présumant que c'est toujours le cas, il propose de peser à nouveau Sambo et d'en déduire l'estimation suivante du poids du troupeau :

$$50 \times (\text{Poids de Sambo})$$

4.3. Exemple de Basu (2)

- Le statisticien du cirque propose quant à lui d'estimer sans biais le poids total en échantillonnant un éléphant à probabilités inégales avec :

$$\begin{cases} \pi_{\text{Sambo}} = {}^{99}\!/_{100} = 0{,}99 \\ \pi_{\text{Autres}} = {}^{1}\!/_{100} \times {}^{1}\!/_{49} = {}^{1}\!/_{4900} \end{cases}$$

- et en calculant :
  - $(\text{Poids de Sambo}) / 0{,}99$ si Sambo est échantillonné
  - $(\text{Poids de l'éléphant tiré}) \times 4900$ sinon

5. Comment construire des intervalles de confiance ?

- En général, on considère que le π-estimateur suit approximativement une loi normale lorsque la taille de l'échantillon est grande.
- L'intervalle de confiance au niveau de confiance *1-*α pour le total T_y est donc donné par :

$$IC_{1-\alpha}(t_y) = \left[\hat{t}_{y\pi} - u_{1-\frac{\alpha}{2}} \sqrt{\hat{V}ar(\hat{t}_{y\pi})}; \hat{t}_{y\pi} + u_{1-\frac{\alpha}{2}} \sqrt{\hat{V}ar(\hat{t}_{y\pi})} \right]$$

où $u_{1-\frac{\alpha}{2}}$ désigne le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$

- De même, l'intervalle de confiance pour la moyenne est :

$$IC_{1-\alpha}(\mu_y) = \left[\hat{\mu}_{y\pi} - u_{1-\frac{\alpha}{2}} \sqrt{\hat{V}ar(\hat{\mu}_{y\pi})}; \hat{\mu}_{y\pi} + u_{1-\frac{\alpha}{2}} \sqrt{\hat{V}ar(\hat{\mu}_{y\pi})} \right]$$