

SONDAGE A PROBABILITÉS INÉGALES

- Les plans simples équiprobables ne sont utilisés qu'en l'absence de toute autre information
- Tirage à probabilités inégales: une manière d'utiliser de l'information auxiliaire
- Infinité de plans à probabilités inégales et sans remise

SONDAGE A PROBABILITÉS INÉGALES

- Estimateur de Horvitz-Thompson ou des valeurs dilatées pour un total:

$$\hat{T} = \sum_{i \in S} a_i y_i = \sum_{i=1}^N a_i Y_i \delta_i$$

$$E(\hat{T}) = \sum_{i=1}^N a_i Y_i E(\delta_i) = \sum_{i=1}^N a_i \pi_i Y_i$$

Pour que \hat{T} soit sans biais: $E(\hat{T}) = \sum_{i=1}^N Y_i$

$$a_i \pi_i = 1$$

SONDAGE A PROBABILITÉS INÉGALES

Théorème:

$\hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i}$ est le seul estimateur linéaire sans biais

de T

Pour une moyenne \bar{Y} $\hat{Y} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$

SONDAGE A PROBABILITÉS INÉGALES

Exemple (Ardilly) : nombre d'habitants Y inconnu, nombre de logements X connu.
Estimation du nombre moyen d'habitants par tirage à probabilités proportionnelles au nombre de logements

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	-

SONDAGE A PROBABILITÉS INÉGALES

Echantillons de deux communes:

Échantillon s	$\hat{Y}(s)$	$\bar{y}(s)$ (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

SONDAGE A PROBABILITÉS INÉGALES

- Si N est inconnu:

$$N = \sum_{i=1}^N 1$$

- L'estimateur de N est donc:

$$\hat{N} = \sum_{i \in S} \frac{1}{\pi_i}$$

- D'où:

$$E\left(\sum_{i \in S} \frac{1}{\pi_i}\right) = N$$

SONDAGE A PROBABILITÉS INÉGALES

- Estimateur de Hajek:

$$\hat{Y} = \left(\sum_{i \in s} \frac{1}{\pi_i} \right)^{-1} \sum_{i \in s} \frac{y_i}{\pi_i}$$

- Poids aléatoires de somme 1.
- Estimateur légèrement biaisé

SONDAGE A PROBABILITÉS INÉGALES

- Un cas gênant:

$$Y_i = C$$

$$\hat{y} = \frac{1}{N} \sum_{i \in S} \frac{Y_i}{\pi_i} = \frac{C}{N} \sum_{i \in S} \frac{1}{\pi_i}$$

Comme $\sum_{i \in S} \frac{1}{\pi_i} \neq N$ alors $\hat{y} \neq C$

- Mais: $E(\hat{y}) = C$

SONDAGE A PROBABILITÉS INÉGALES

- 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 :

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

SONDAGE A PROBABILITÉS INÉGALES

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

Première formule:

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

Deuxième formule:

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

SONDAGE A PROBABILITÉS INÉGALES

- 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...)

SONDAGE A PROBABILITÉS INÉGALES

- 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

$$x_1=300 \quad x_2=90 \quad x_3=70 \quad x_4=50 \quad x_5=20 \quad x_6=20$$

SONDAGE A PROBABILITÉS INÉGALES

- Unités sélectionnées d'office et unités tirées au hasard.
- Infinité de plans de sondage pour des π_i fixés.



Sondage systématique à probabilités inégales

- On cumule pour tous les individus les probabilités d'inclusion:
 - $V_k = \pi_1 + \pi_2 + \dots + \pi_k$
 - On génère une seule réalisation u de la loi $U[0, 1[$
 - On sélectionne k tel que $V_{k-1} \leq u < V_k$
 - puis i tel que $V_{i-1} \leq u + 1 < V_i$
 - puis j tel que $V_{j-1} \leq u + 2 < V_j$
- etc ... on obtient in fine n individus

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