Puisque $\Delta_{kl} = 0$, $k \neq l$, la variance du π-estimateur du total t_y est:

$$Var(\hat{t}_y) = \sum_{k,l \in U} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl} \quad , k = l$$

$$= \sum_{k \in U} y_k^2 \pi_k (1 - \pi_k) \times \frac{1}{\pi_k^2}$$

$$= \sum_{k \in U} \frac{y_k^2 (1 - \pi_k)}{\pi_k}$$

Et peut être estimée par $\hat{Var}(\hat{t}_y)$ qui est égale à :

$$\hat{Var}(\hat{t}_y) = \sum_{k \in U} \frac{y_k^2 (1 - \pi_k)}{\pi_k^2} \times 1_{\{k \in s\}}$$

06-11-2024

Le plan de poisson est interessant car il est simple à mettre en oeuvre ~~et il maxim~~ mais également il maximise entropie (nous introduisons maintenant une mesure du desordre).

Definition

On appelle entropie d'un plan $P(\cdot)$, la quantité $I(P) = -\sum_{s \in U} P(s) \ln p(s)$ avec la

convention que $0 \ln 0 = 0$.

L'entropie est toujours positive. De plus, comme mesure du désordre, plus $I(p)$ sera grand, plus le plan $p(\cdot)$ sera aléatoir. Pour des probabilités d'inclusion fixées, on cherchera donc le plan le plus aleatoire ou desordonné, c'est-à-dire celui maximisant l'entropie.

Lemme : $$\sum_{S \subset U} \prod_{k \in U} x_k = \prod_{k \in U} (1 + x_k)$$

Théorème : Etant donné les probabilités d'inclusion fixées π_k, $k \in U$, le plan de Poisson est le plan d'entropie maximale sur

$\Psi = \{S \subset U\}$

En définitive, le plan de Poisson est un plan de sondage respectant les probabilités d'inclusion d'ordre 1 fixées à priori et étant le plus aleatoire possible (au sens de l'entropie).

III - Sondage systématique à probabilités inégales

Ce plan de sondage a été introduit vers 1950 et est toujours largement utilisé

puisqu'il a le mérite d'être simple et exacte. Contrairement au plan de Poisson, il est également de taille fixe.

On désire tirer des échantillons dont les probabilités d'inclusion d'ordre 1 sont fixés à priori et telsque que $0 < \pi_i < 1$, $k \in U$ et $\sum_{k \in U} \pi_k = n$.

Définissons les probabilités d'inclusion cumulées C_k avec $C_k = \sum_{l=1}^{k} \pi_l$, $k \in U$, $C_0 = 0$. L'approche consiste à générer $\mu \to U(0,1)$ et de sélectionner les unités à partir de cette unique réalisation. La première unité sélectionnée notée k_1 sera telle que $C_{k_1 - 1} \leq \mu < C_{k_1}$. La deuxième unité sélectionnée notée k_2 sera cette fois-ci telle que $C_{k_2 - 1} \leq 1 + \mu < C_{k_2}$ et ainsi de suite ... De manière générale, la j-ième unité sélectionnée notée k_j sera alors $C_{k_j - 1} \leq j - 1 + \mu < C_{k_j}$

Exemple: Prennons la situation où $N = 6$, $n = 3$, $\pi_1 = 0,2$, $\pi_2 = 0,7$, $\pi_3 = 0,8$, $\pi_4 = 0,5$, $\pi_5 = \pi_6 = 0,4$

1) Déterminer l'échantillon sélectionné sachant que $\mu = 0,3658$

solution

l'on a : $C_1 = 0,2$ $C_2 = 0,9$; $C_3 = 1,7$

$C_4 = 2,2$ $C_5 = 2,6$ $C_6 = 3$; $C_0 = 0$

1ᵉ : $0 \leq 0,3658 < 0,2$ ✗

2ᵉ : $0 \leq 0,3658 < 0,9$ ✓

3ᵉ : $C_2 \leq 1 + 0,3658 < C_3$

$0,9 < 1,3658 < 1,7$ ✓

4ᵉ : $C_3 \leq 1 + 1,3658 < C_4$

$1,7 \leq 2,3658 < 2,2$ ✗

5ᵉ : $C_4 \leq 2,3658 < C_5$

$2,2 \leq 2,3658 < 2,6$ ✓

6ᵉ : $C_5 \leq 1 + 2,3658 < C_6$

$2,6 \leq 3,3658 < 3$

Chapitre 4: Le Sondage stratifié

I- Principes et justification

Dans un sondage aléatoire simple, tous les échantillons de taille aléatoir n sont possibles avec la même probabilité. On imagine de certains d'entre-eux puissent s'avérer a priori indésirable. Prenons par exemple le cas où nous disposons de cinq jetons: -1, 2, 4, 10 et 20 dont nous souhaitons évaluer la moyenne ($\mu = 7$), à l'aide d'un échantillon de taille 2. Parmi les échantillons à deux unités, on trouve les cas extrêmes: $\{-1; 2\}$ et $\{10, 20\}$ qui sont particulièrement mauvais.

Plus concrètement dans l'étude du lancement d'un nouveau produit financier, on peut supposer des différences de comportements entre les petits et gros clients de la banque. Il serait malencontreux que les hasards de l'échantillon conduisent à n'interroger que les clients appartenant à une seule de ces catégories, ou simplement que l'échantillon soit trop déséquilibré en faveur de l'une d'elle. S'il existe dans la base de sondage une information auxiliaire permettant de distinguer, a priori, les catégories de petits et gros clients, on aura tout à gagner à utiliser cette information pour

répartir l'échantillon dans chaque sous-population. C'est le principe de la stratification : de couper la population en sous-ensembles appelés strates et réaliser un sondage dans chacune d'elle. L'intérêt de cette méthode en comparaison des plans simples, est qu'elle permet d'améliorer la précision des estimateurs. Elle nécessite l'utilisation d'une information auxiliaire connue pour l'en ensemble de la population.

Exemple : Nous souhaitons estimer l'âge moyen de toutes les personnes évoluant sur le site de l'UY I. La base de sondage est composée de l'ensemble des personnes de l'UY I. Supposons que nous disposons de la répartition des éléments de la base suivant les catégories :

- Etudiants ;
- enseignants ;
- personnels administratif.

Autrement dit, nous connaissons la repartition des personnes de l'UY I suivant ces 03 catégories. Il y a fort à parier que la variable âge ne se comporte pas de la même manière dans ces 03 classes (en moyenne, on peut en effet penser que la population enseignant ou personnel administratif est plus âgé que la population Etudiant). Il paraît dès lors pertinent d'

essayer de prendre en compte cette information dans le plan de sondage.

La répartition des personnes de l'UY1 fournit une information auxiliaire à notre problématique. L'objectif principale consiste donc à mettre à profit cette information pour obtenir des résultats précis. L'information auxiliaire peut être utilisées à deux moments:

- A l'étape de la conception du plan de sondage
- à l'étape de l'estimation des paramètres

Dans ce chapitre, nous utiliserons cette information pour bâtir le plan de sondage.

II- Population et Strate

Supposons que la population U soit partitionnée en H sous-ensembles : $U_1, \dots U_H$ appelés strates et tels que $\bigcup_{i=1}^{H} U_i = U$, $U_i \cap U_k = \phi$, $\forall i \neq k$. Chaque strate U admet une taille N_k et l'on a: $\sum_{k=1}^{H} N_k = N$ où N est la taille de la population U.

Remarque: Les tailles des strates sont ici supposées connues et constituent l'information auxiliaire.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (respectivement la moyenne) s'écrit à l'aide des strates $t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} t_{y,h}$

où t_y est la valeur prise par le caractère y sur la strate U_h c'à-d $t_{y,h} = \sum_{k \in U_h} y_k$. De même la moyenne sur la population s'écrit

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_{y,h}$$

où $\mu_{y,h}$ est la moyenne des valeurs prises par le caractère y sur la strate U_h, c'est-à-dire

$$\sigma^2_{y,h} = \frac{1}{N_h} \sum_{k \in U_h} (y_k - \mu_{y,h})^2 \qquad \mu_{y,h} = \frac{1}{N_h} \sum_{k \in U_h} y_k$$

$$S^2_{y,h} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{y,h})^2$$

Remarque: la variance sur la population (totale) σ^2_y s'écrit :

$$\sigma^2_y = \frac{1}{N} \sum_{k \in U} (y_k - \mu_y)^2$$

$$= \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} \left[(y_k - \mu_{y,h}) + (\mu_{y,h} - \mu_y) \right]^2$$

$$\sigma_y^2 = \frac{1}{N} \sum_{h=1}^{H} \left\{ \sum_{k \in U_h} (y_k - \mu_{y,h})^2 + 2(\mu_{y,h} - \mu_y) \underbrace{\sum_{k \in U_h} (y_k - \mu_{y,h})}_{0} + N_h (\mu_{y,h} - \mu_y)^2 \right\}$$

$$= \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_{y,h}^2 + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2$$

$$= \sigma_{y,\text{intra}}^2 + \sigma_{y,\text{inter}}^2$$

où $\sigma_{y,\text{intra}}^2$ est la variance intra strate (moyenne des variances des strates) et

et $\sigma_{y,\text{inter}}^2$ la variance inter strate (dû aux differences entre strate).