

et σ_y^2 , on tire la variance inter strate (due differences entre strate).

3-11-2024

III - L'échantillon, les probabilités d'inclusion et estimation

Un sondage est dit stratifié, si, pour chaque strate on tire un échantillon selon sondage aléatoire simple sans remise de taille fixe n_h et que les tirages au sein de chaque strate sont mutuellement indépendants.

Soit S_h l'échantillon aléatoire tiré dans la strate U_h à l'aide d'un plan de

sondage $P_h(\cdot)$. L'échantillon aléatoire S obtenu au final est donc :

$$S = \bigcup_{h=1}^H S_h$$

Le plan de sondage associé $P(S)$ n'est rien d'autre que :

$$P(S) = \prod_{h=1}^H P_h(S_h) \quad \forall S = \bigcup_{h=1}^H S_h \quad \text{et la}$$

taille de l'échantion S est : $n = \sum_{h=1}^H n_h$

Le calcul des probabilités d'inclusion pour sondage stratifié n'est pas difficile mais il faut tout de même faire attention. Pour les probabilité d'inclusion d'ordre 1 et si l'unité k appartient à la strate U_h alors :

$\pi_k = \frac{n_h}{N_h}$, puisqu'on a effectué une plan simple sans remise de taille n_h pour cette strate.

Pour les probabilités d'inclusion d'ordre 2, c'est un peu plus difficile et le résultat dépend du fait que les unités k, l appartiennent à la même strate ou non.

- Si k et l appartiennent à la même strate alors

$$\pi_{k,l} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}$$

- Si k et l appartiennent à deux strates différentes U_{k_1} et U_{k_2} alors (par indépendance entre les strates):

$$\pi_{kl} = \pi_k \pi_l = \frac{n_{k_1} n_{k_2}}{N_{k_1} N_{k_2}}$$

En conséquence, on a:

$$\Delta_{kl} = \begin{cases} \frac{n_k}{N_k} \left(1 - \frac{n_k}{N_k}\right), & k=l, k \in U_k \\ -\frac{n_k (N_k - n_k)}{N_k^2 (N_k - 1)}, & k \neq l, k, l \in U_k \\ 0, & k \in U_k, l \notin U_k \end{cases}$$

Du coup, les π -estimateurs du total t_y de la moyenne μ_y sont:

$$\begin{aligned} \hat{t}_{y, \text{strat}} &= \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k=1}^H \sum_{k \in S_k} \frac{N_k y_k}{n_k} \\ &= \sum_{k=1}^H \hat{t}_{y, k} \end{aligned}$$

Délégué Nsangou Mohamed

$$\hat{\mu}_{y, \text{strate}} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k \in S_h} y_k = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h$$

où $\hat{t}_{y,h}$ est l'estimateur du total pour la strate h
 c-à-d $\hat{t}_{y,h} = \sum_{k \in S_h} \frac{N_h y_k}{n_h} = \frac{N_h}{n_h} \sum_{k \in S_h} y_k$ et

$\bar{y}_h$ est la moyenne de l'échantillon prelevée sur la strate h c-à-d

$$\bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

Les résultats du sondage sont donnés dans le tableau suivant :

strate	1	2	1	3	1	1	2	3	2	1
Âges	20	50	25	42	23	22	35	44	38	26

1) Calculer la moyenne des individus de l'échantillon par strate.

2. Déterminer la moyenne μ

3. Calculer la variance de la moyenne

4. Calculer la variance estimée donc $\hat{S}^2$

Solution

1. Calculons la moyenne par strate :

$$\bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k$$

Pour $h=1$: $y_1 = \frac{1}{n_1} \sum_{k \in S_1} y_k$

$$= \frac{1}{5} (20 + 25 + 23 + 22 + 26)$$

$$= 23,2$$

Pour $h=2$: $y_2 = \frac{1}{n_2} \sum_{k \in S_2} y_k$

$$= \frac{1}{3} (50 + 35 + 38)$$

$$= 41$$

Pour $h=3$: $y_3 = \frac{1}{n_3} \sum_{k \in S_3} y_k$

$$= \frac{1}{2} (42 + 44)$$

$$= 43$$

2. Determinons la moyenne μ

$$\hat{\mu} = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h$$

$$\hat{\mu} = \frac{1}{10} (5 \times 23,2 + 3 \times 41 + 2 \times 43)$$

$$\boxed{\hat{\mu} = 32,5}$$

3) Calcul de la variance μ

$$\text{Var}(\hat{\mu}) = \frac{1}{N} \sum_{k \in U_k} (\bar{y}_k - \hat{\mu}_y)^2$$

$$\text{Var}(\hat{\mu}) = \frac{1}{N^2} \sum_{h=1}^H N_h \frac{N_h - n_h}{n_h} S_h^2$$

$$\text{Var}(\hat{t}) = \sum_{h=1}^H \text{Var}(\hat{t}_{y_h})$$

$$\hat{\text{Var}}(\hat{t}_0) = \sum_{h=1}^H N_h (N_h - n_h) \frac{\hat{S}_{y_h}^2}{n_h}$$

$$\hat{S}_{y_h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2, \quad h = 1, 2, \dots, H$$

IV - Plan stratifié et allocation proportionnelle

Definition

Un plan stratifié est dit à allocation proportionnelle si $\frac{n_h}{N_h} = \frac{n}{N} \quad \forall h = 1, 2, \dots, H$

c-à-d que les strates de tailles importantes auront plus d'unités dans l'échantillon que celles de tailles plus petites.

Remarque : Généralement, la taille d'échantillon pour chaque strate $n_h = \frac{n N_h}{N}$ ne sera pas entière, mais afin de simplifier les développements théoriques qui viennent nous allons tout de même le supposer.

Les π -estimateurs du total et de la moyenne sont alors :

$$\hat{t}_{y, \text{strat prop}} = \sum_{h=1}^H \hat{t}_{y, h} = \frac{N}{n} \sum_{h \in S} y_h$$

$$\hat{\mu}_{y, \text{strat prop}} = \frac{1}{n} \sum_{h \in S} y_h$$

La variance de la moyenne est alors :

$$\text{Var}(\hat{\mu}_{y, \text{strat prop}}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^H N_h S_{y, h}^2$$

$$\text{Var}(\hat{t}_{y, \text{strat prop}}) = \sum_{h=1}^H N_h (N_h - n_h) \frac{S_{y, h}^2}{n_h}$$

$$= \sum_{h=1}^H N_h \left(\frac{N}{n} - 1\right) S_{y, h}^2$$

$$= \frac{N-n}{n} \sum_{h=1}^H N_h S_{y, h}^2$$

Remarque: Lorsque les tailles des strates N_h sont suffisamment grandes alors: $S_{y,h}^2 \approx \sigma_{y,h}^2$ et donc:

$$\begin{aligned} \text{Var}(\hat{t}_{y, \text{strat prop}}) &= \frac{N-n}{n} \sum_{h=1}^H N_h \sigma_{y,h}^2 \\ &= N(N-n) \frac{\sigma_{y, \text{intra}}^2}{n} \end{aligned}$$

Alors que la variance de l'estimateur du total pour un plan simple sans remise vérifie:

$$\text{Var}(\hat{t}_{y, \pi}) \approx N(N-n) \frac{\sigma_y^2}{n}$$

Les deux expressions sont quasiment identiques mais puisque $\sigma_y^2 = \sigma_{y, \text{intra}}^2 + \sigma_{y, \text{inter}}^2$, la première expression est plus petite, c-à-d que l'on obtient de meilleurs résultats avec un plan stratifié avec allocation proportionnel qu'avec un plan simple sans remise.

Ceci sera d'autant plus vrai que la variance inter strate sera grande, ce qui est le cas lorsque le caractère d'intérêt y dépend fortement du caractère servant à la

strat stratification, ici les tailles N_h .
 estimera sans biais cette variance par :

$$\hat{\text{Var}}(\hat{t}_{y, \text{strat prop}}) = \frac{N-n}{n} \sum_{h=1}^H N_h \hat{S}_{y,h}^2$$

avec $\hat{S}_{y,h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2$, $h = 1, \dots, H$

Exemple:

On donne dans le tableau suivant pour chaque individu de l'université de Caen :
 - son âge

- Catégorie : 1 = étudiant, 2 = enseignant
 3 = personnel administratif
- sa couleur de cheveux : a = brun, b = blond
 c = châtain

Pour simplifier, on considère une population de 20 individus.

Âge	Catégorie	cheveux
24	1	c
52	2	a
42	3	b
19	1	c
38	3	a
26	1	b

45	2	c
23	1	a
39	2	a
24	1	b
22	1	c
48	2	a
24	1	a
38	3	a
26	1	b
36	3	b
46	2	b
23	1	c
39	2	a
18	1	c

$$n=10$$

- 1) On souhaite estimer la moyenne d'âge μ à l'aide d'un plan simple. Quelle est la variance de l'estimateur ?
- 2) On désire stratifier la population suivant la catégorie. Quelle est la variance de l'estimateur $\hat{\mu}$ pour un tel plan ?
- 3) On choisit maintenant de stratifier suivant la couleur de cheveux. Quelle est la variance de l'estimateur pour un tel plan ?

Délégué Nsangou Mohamed