

Introduction à la théorie des sondages - Cours 2

Antoine Rebecq
antoine.rebecq@insee.fr

INSEE, direction de la méthodologie

17 mars 2015



Sommaire I

- 1 Correction du QCM
- 2 Stratégie d'estimation
 - L'estimateur de Horvitz-Thompson pour un total ou une moyenne
 - L'estimateur de Hájek
 - Recherche d'un estimateur optimal
- 3 Sondage aléatoire simple
 - Définitions
 - Estimation
 - Estimation d'une proportion
 - Exercice
 - Autres estimations
 - Estimation sur un domaine
 - Estimation d'un ratio

Sommaire II

- Échantillonnage dans le temps
- Conclusion sur le SAS

Chapitre 1

Correction du QCM

Rappel sur les π_k

$$\pi_k = \mathbb{P}(k \in \mathcal{S}) = \mathbb{P}(\delta_k = 1) = \sum_{s \ni k} p(s)$$

$$\pi_{kl} = \mathbb{P}(k, l \in \mathcal{S}) = \mathbb{P}(\delta_k \delta_l = 1) = \sum_{s \ni k, l} p(s)$$

(où δ_k est l'indicatrice d'appartenance de k à \mathcal{S} , appelée aussi variable de Cornfield)

Rappel sur les estimateurs pondérés

Estimateur pondéré :

$$\hat{\phi} = \sum_{k \in s} w_k y_k$$

Estimateur pondéré de la moyenne :

$$\hat{\phi} = \frac{1}{N} \sum_{k \in s} w_k y_k$$

Chapitre 2

Stratégie d'estimation

Plan de sondage sans remise - définition

On note \mathcal{S} l'ensemble des parties de \mathcal{U} .

Le plan de sondage p est une loi de probabilité sur \mathcal{S} tel que :

$$\forall s \in \mathcal{S}, p(s) \geq 0$$

$$\sum_{s \in \mathcal{S}} p(s) = 1$$

Exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :

$$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

On peut définir un plan de sondage p par :

$$p(\{1\}) = 0 \quad p(\{1, 2\}) = \frac{1}{2} \quad p(\{1, 2, 3\}) = 0$$

$$p(\{2\}) = 0 \quad p(\{1, 3\}) = \frac{1}{3}$$

$$p(\{3\}) = 0 \quad p(\{2, 3\}) = \frac{1}{6}$$

Exemple d'estimation

Avec ce plan de sondage, on tente d'estimer le revenu moyen à l'aide de la moyenne dans l'échantillon (estimation plugin). Le revenu dans la population est :

$$Y_1 = 1000$$

$$Y_2 = 2000$$

$$Y_3 = 3000$$

On a donc :

$$\bar{Y} = 2000$$

Exemple d'estimation

Loi de l'estimateur $\hat{\Phi}$:

$$\begin{aligned}\hat{\Phi}_1 &= \frac{1000 + 2000}{2} = 1500, \text{ probabilité } \frac{1}{2} \\ \hat{\Phi}_2 &= \frac{1000 + 3000}{2} = 2000, \text{ probabilité } \frac{1}{3} \\ \hat{\Phi}_3 &= \frac{2000 + 3000}{2} = 2500, \text{ probabilité } \frac{1}{6}\end{aligned}$$

Exemple d'estimation

L'espérance de l'estimateur est donc :

$$\begin{aligned}\mathbb{E}[\hat{\Phi}] &= \frac{1}{2} \cdot 1500 + \frac{1}{3} \cdot 2000 + \frac{1}{6} \cdot 2500 \\ &\approx 1833 \\ &\neq \bar{Y} = 2000\end{aligned}$$

$\hat{\Phi}$ est donc biaisé (ce n'est pas une surprise !)

Espérance

$$\mathbb{E}(\hat{\Phi}) = \sum_s p(s) \cdot \hat{\Phi}(s)$$

C'est la valeur moyenne de $\hat{\Phi}$ obtenue avec le plan de sondage considéré **sur tous les échantillons possibles**.

Partie 1

L'estimateur de Horvitz-Thompson pour un total ou une moyenne

Définition

Définition

L'estimateur d'Horvitz-Thompson (ou π -estimateur) est défini :

$$\text{pour un total : } \hat{T}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

$$\text{pour une moyenne : } \hat{y}_{\pi} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$$

*C'est donc un **estimateur pondéré** utilisant les poids $w_k = \frac{1}{\pi_k}$*

Estimation sans biais

Théorème

*Si $\forall k \in \mathcal{U}, \pi_k > 0$, alors l'estimateur d'Horvitz-Thompson est **sans biais** pour le total et la moyenne.*

Estimation sans biais

Démonstration.

$$\begin{aligned}\mathbb{E}[\hat{T}_{y\pi}] &= \mathbb{E}\left[\sum_{k \in s} \frac{y_k}{\pi_k}\right] \\ &= \mathbb{E}\left[\sum_{k \in \mathcal{U}} \frac{y_k \delta_k}{\pi_k}\right] \\ &= \sum_{k \in \mathcal{U}} \frac{y_k \mathbb{E}[\delta_k]}{\pi_k} \\ &= \sum_{k \in \mathcal{U}} y_k \\ &= T(y)\end{aligned}$$

Exemple

On reprend le plan de la slide 9. Le plan est de taille fixe 2 et :

$$\pi_1 = \frac{5}{6}$$

$$\pi_2 = \frac{2}{3}$$

$$\pi_3 = \frac{1}{2}$$

Exemple

La loi de l'estimateur d'Horvitz-Thompson Φ_{HT} est donc :

$$\hat{\Phi}_{HT1} = \frac{6}{5} \cdot 1000 + \frac{3}{2} \cdot 2000 = 4200, \text{ probabilité } \frac{1}{2}$$

$$\hat{\Phi}_{HT2} = \frac{6}{5} \cdot 1000 + 2 \cdot 3000 = 7200, \text{ probabilité } \frac{1}{3}$$

$$\hat{\Phi}_{HT3} = \frac{3}{2} \cdot 2000 + 2 \cdot 3000 = 9000, \text{ probabilité } \frac{1}{6}$$

Exemple d'estimation

L'espérance de l'estimateur est donc :

$$\begin{aligned}\mathbb{E}[\hat{\Phi}_{HT}] &= \frac{1}{2} \cdot 4200 + \frac{1}{3} \cdot 7200 + \frac{1}{6} \cdot 9000 \\ &= 6000 = Y\end{aligned}$$

Et :

$$\begin{aligned}\mathbb{E}[\hat{Y}_{HT}] &= \frac{1}{3} \cdot \hat{\Phi}_{HT} \\ &= 2000 = \bar{Y}\end{aligned}$$

... ce qui permet de vérifier que $\hat{\Phi}_{HT}$ est sans biais.

Variance de l'estimateur de Horvitz-Thompson

Propriété

La variance de l'estimateur de Horvitz-Thompson s'écrit :

$$\text{Var}[\hat{T}_{y\pi}] = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl}$$

(où : $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$)

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

$$\begin{aligned}\text{Var}(\hat{t}_{y\pi}) &= \text{Var}\left(\sum_{k \in U} \frac{y_k}{\pi_k} \delta_k\right) \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \text{Var}(\delta_k) + \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \text{Cov}(\delta_k, \delta_l) \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \pi_k \cdot (1 - \pi_k) + \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l) \\ &= \sum_{k, l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}\end{aligned}$$



Variance pour un plan de taille fixe

Propriété

Si le plan de sondage est de taille fixe (formule de Yates-Grundy) :

$$\text{Var}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl}$$

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

Découle de la formule de Horvitz-Thompson quand le plan de sondage est de taille fixe. Pour démontrer la formule, il vaut mieux procéder à rebours :

$$\begin{aligned}
 & -\frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl} \\
 &= \frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 (\pi_k \pi_l - \pi_{kl}) \\
 &= \frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} \left(\frac{y_k^2}{\pi_k^2} + \frac{y_l^2}{\pi_l^2} - 2 \frac{y_k y_l}{\pi_k \pi_l} \right) (\pi_k \pi_l - \pi_{kl}) \\
 &= \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k^2}{\pi_k^2} (\pi_k \pi_l - \pi_{kl}) - \sum_{k \in U} \sum_{l \in U, l \neq k} y_k y_l \left(1 - \frac{\pi_{kl}}{\pi_k \pi_l} \right) \\
 &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \left(\sum_{l \in U, l \neq k} \pi_l - \frac{1}{\pi_k} \frac{y_k^2}{\pi_k} \pi_{kl} \right) - \sum_{k \in U} \sum_{l \in U, l \neq k} y_k y_l \left(1 - \frac{\pi_{kl}}{\pi_k \pi_l} \right)
 \end{aligned}$$

Variance de l'estimateur de Horvitz-Thompson

Démonstration.

...

Or, d'après le cours 1, on a dans le cas taille fixe : $\sum_{k \in U} \pi_k = n$ et $\sum_{l \in U, l \neq k} \pi_{kl} = \pi_k(n-1)$, cela donne :

$$\begin{aligned} & -\frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl} \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \left(n - \pi_k - \frac{\pi_k(n-1)}{\pi_k} \right) - \sum_{k \in U} \sum_{l \in U, l \neq k} y_k y_l \left(1 - \frac{\pi_{kl}}{\pi_k \pi_l} \right) \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \pi_k (1 - \pi_k) - \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k y_l}{\pi_k \pi_l} (\pi_k \pi_l - \pi_{kl}) \\ &= \sum_{k, l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl} \end{aligned}$$

Et on retombe bien sur la formule d'Horvitz-Thompson.

Exercice

Exercice : calculer la variance de l'estimateur d'Horvitz-Thompson dans le cas de la slide 9, et vérifier les formules de Yates-Grundy et de Horvitz-Thompson dans ce cas.

Estimation de variance

Les quantités précédentes sont les **vraies variances**. On peut utiliser les estimateurs suivants, qui sont sans biais dès lors que $\forall k, l, \pi_{kl} > 0$:

$$\hat{\text{Var}}(\hat{t}_{y\pi}) = \sum_{k \in s} \frac{y_k^2}{\pi_k^2} (1 - \pi_k) - \sum_{k \in s} \sum_{l \in s, l \neq k} \frac{y_k y_l}{\pi_k \pi_l \pi_{kl}} (\pi_k \pi_l - \pi_{kl})$$

Pour un plan de taille fixe :

$$\hat{\text{Var}}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}}$$

Construction d'un intervalle de confiance

On fait l'**hypothèse** : $\hat{\Phi}(s) \sim \mathcal{N}(\Phi, \text{Var}(\Phi))$

L'intervalle de confiance à 95% est défini par :

$$IC_{95\%} = \left[\hat{\Phi} - 2\sigma(\hat{\Phi}); \hat{\Phi} + 2\sigma(\hat{\Phi}) \right]$$

L'intervalle de confiance **estimé** est défini par :

$$\hat{IC}_{95\%} = \left[\hat{\Phi} - 2\hat{\sigma}(\hat{\Phi}); \hat{\Phi} + 2\hat{\sigma}(\hat{\Phi}) \right]$$

Remarque

Remarque : si le plan de sondage ne vérifie pas :

$$\forall k \neq l \in \mathcal{U}, \pi_{kl} - \pi_k \pi_l \geq 0$$

(condition de Sen-Yates-Grundy), ces estimateurs de variance peuvent prendre des valeurs négatives.

Partie 2

L'estimateur de Hájek

Définition

L'estimateur de Horvitz-Thompson de la moyenne nécessite la connaissance de N , la taille de la population. On peut utiliser dans ce cas l'estimateur :

$$\hat{y}_H = \frac{\sum_{k \in S} \frac{y_k}{\pi_k}}{\sum_{k \in S} \frac{1}{\pi_k}}$$

Propriété

L'estimateur de Hájek est biaisé, mais en général, le biais est négligeable.

Estimateur de Hájek du total

L'estimateur de Hájek peut être utilisé pour estimer un total :

$$T(\hat{Y})_H = N \cdot \frac{\sum_{k \in s} \frac{y_k}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}}$$

... mais cela impose de connaître N .

Partie 3

Recherche d'un estimateur optimal

Estimateur de Horvitz-Thompson

L'estimateur de Horvitz-Thompson constitue le fondement de l'estimation par sondage (même si d'autres estimateurs peuvent être utilisés, la logique de construction découle souvent de celle de Horvitz-Thompson, voir cours suivants)

Estimateur de Horvitz-Thompson

L'estimateur de Horvitz-Thompson n'est pas le seul estimateur sans biais.

Recherche d'optimalité

Existe-t-il un estimateur optimal en sondages ?

Question centrale pour les théoriciens des sondages dans les années 1950 à 1970 : Godambe, Hanurav, Basu, etc.

Recherche d'optimalité

Théorème (Godambe, 1955)

Dans la classe des estimateurs sans biais, pour un plan sans remise avec $n < N$ tel que $\forall k \in \mathcal{U}, \pi_k > 0$, il n'existe pas d'estimateur optimal de \bar{y}

Recherche d'optimalité

Démonstration.

Si les $\forall k \in \mathcal{U}, \pi_k > 0$, alors il existe toujours au moins un estimateur sans biais : l'estimateur d'Horvitz-Thompson. Mais on peut également définir :

$$\hat{y}_2 = \hat{y}_\pi + \bar{x} - \hat{x}_\pi$$

, où x est un total supposé connu sur la population. \hat{y}_2 est la somme d'estimateurs sans biais, il est donc également sans biais. De plus, si $\forall k, x_k = y_k$, alors :

$$EQM(\hat{y}_2) = EQM(\bar{x}) = 0$$

Ainsi, un estimateur optimal doit avoir une variance inférieure ou égale à 0, ce qui n'est possible que par recensement.



Admissibilité

Définition (Admissibilité)

*Pour un plan donné p , un estimateur $\hat{\Phi}$ est dit **admissible** si et seulement s'il n'existe pas d'estimateur meilleur que $\hat{\Phi}$ pour toute valeur de y*

Admissibilité

Découle sur des résultats assez pauvres : beaucoup d'estimateurs sont admissibles (Horvitz-Thompson, différence, etc.)

Admissibilité

Définition (Hyperadmissibilité)

*Un estimateur est dit **hyperadmissible** s'il est admissible pour tout domaine non-vide de \mathcal{U}*

Admissibilité

L'estimateur d'Horvitz-Thompson est le seul estimateur sans biais hyperadmissible.

Mais ce résultat n'est pas si intéressant en soi ! (voir éléphants de Basu)

Estimation sans biais

L'estimation par un estimateur sans biais est un choix, qui peut parfois ne pas être judicieux.

Lien intéressant à ce sujet :

http://www.johndcook.com/blog/bias_consistency/

Chapitre 3

Sondage aléatoire simple

Partie 1

Définitions

Définition

Sondage aléatoire simple sans remise (SAS) de taille n : plan de sondage sans remise de taille fixe n tel que tous les échantillons de taille n ont la même probabilité d'être tirés. Cette probabilité vaut :

$$p(s) = \frac{1}{\binom{n}{N}} \quad \text{si } |s| = n$$
$$= 0 \quad \text{sinon.}$$

On note le taux de sondage : $f = \frac{n}{N}$

Probabilités d'inclusion

$$\forall k \in \mathcal{U}, \pi_k = \mathbb{P}(k \in s) = \frac{n}{N} = f$$

$$\forall k \neq l \in \mathcal{U}, \pi_{k,l} = \mathbb{P}(k \wedge l \in s) = \frac{n(n-1)}{N(N-1)}$$

Notations

On note, **dans la population** :

$$\text{Total : } T(Y) = \sum_{k \in \mathcal{U}} Y_k$$

$$\text{Moyenne : } \bar{Y} = \frac{1}{N} \sum_{k \in \mathcal{U}} Y_k$$

$$\text{Variance : } S^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (Y_k - \bar{Y})^2$$

Notations

On note, **dans l'échantillon s** :

$$\text{Total : } n\bar{y} = \sum_{k \in s} y_k$$

$$\text{Moyenne : } \bar{y} = \frac{1}{n} \sum_{k \in s} y_k$$

$$\text{Variance : } s^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$$

Partie 2

Estimation

Estimateur d'Horvitz-Thompson

L'estimateur d'Horvitz-Thompson pour le total et la moyenne s'écrit :

$$T(\hat{Y}) = \sum_{k \in s} \frac{1}{\pi_k} y_k = \frac{N}{n} \sum_{k \in s} y_k = N\bar{y}$$
$$\hat{Y} = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} y_k = \bar{y}$$

Poids de sondage

Les poids pour l'estimation par Horvitz-Thompson sont :

$$w_k = \frac{1}{\pi_k} = \frac{N}{n}$$

On peut dire que l'individu k "représente" $w_k = \frac{N}{n}$ individus de la population \mathcal{U} .

Attention, w_k n'est pas un effectif (en particulier, w_k n'est pas forcément entier !)

Précision

Théorème

En utilisant la formule de Yates-Grundy, la **vraie** variance des estimateurs d'Horvitz-Thompson s'écrit :

$$\text{Var}(\bar{y}) = (1 - f) \frac{S^2}{n}$$
$$\text{Var}(T(\hat{Y})) = N^2(1 - f) \frac{S^2}{n}$$

Précision

Démonstration.

$$\begin{aligned}\text{Var}[\hat{Y}] &= \frac{1}{N^2} \text{Var}[T(\hat{Y})] \\ &= \frac{-1}{2N^2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl} \\ &= \frac{1}{2N^2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} \left(\frac{y_k N}{n} - \frac{y_l N}{n} \right)^2 \frac{n(N-n)}{N^2(N-1)} \\ &= \frac{N-n}{nN} \frac{1}{2N(N-1)} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} (y_k - y_l)^2 \\ &= \frac{N-n}{nN} S^2 \\ &= (1-f) \frac{S^2}{n}\end{aligned}$$

Estimation de la précision

Théorème

La variance empirique (ou dispersion) dans l'échantillon

$s^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$ est un estimateur sans biais de

$$S^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (Y_k - \bar{Y})^2$$

Estimation de la précision

Démonstration.

$$\begin{aligned}\mathbb{E}[s^2] &= \mathbb{E} \left[\frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2 \right] \\ &= \mathbb{E} \left[\frac{1}{2n(n-1)} \sum_{k \in s} \sum_{l \in s, l \neq k} (y_k - y_l)^2 \right] \\ &= \frac{1}{2n(n-1)} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} (y_k - y_l)^2 \mathbb{E}(\delta_k \delta_l) \\ &= \frac{1}{2n(n-1)} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} (y_k - y_l)^2 \frac{n(n-1)}{N(N-1)} \\ &= \frac{1}{2N(N-1)} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}, l \neq k} (y_k - y_l)^2 \\ &= S^2\end{aligned}$$

Estimation de la précision

On peut estimer sans biais la variance de l'estimateur d'Horvitz-Thompson par :

$$\hat{\text{Var}}(\bar{y}) = (1 - f) \frac{s^2}{n}$$
$$\hat{\text{Var}}(T(\hat{Y})) = N^2(1 - f) \frac{s^2}{n}$$

Partie 3

Estimation d'une proportion

Estimation d'une proportion

On cherche à estimer P la proportion d'individus portant une caractéristique dans la population \mathcal{U} .

p , la proportion dans s d'individus portant la caractéristique, est un estimateur sans biais de P .

Variance

$$\text{Var}(p) = (1 - f) \frac{N}{N - 1} \frac{P(1 - P)}{n}$$
$$\hat{\text{Var}}(p) = (1 - f) \frac{p(1 - p)}{n - 1}$$

Précision

Demi-longueur de l'intervalle de confiance :

$$L = 2\sqrt{\frac{p(1-p)}{n-1}}$$

Coefficient de variation estimé :

$$\begin{aligned}\hat{C}V(p) &= \frac{\sqrt{\hat{\text{Var}}(p)}}{p} \\ &= \sqrt{(1-f)\frac{1}{n-1}\frac{1-p}{p}}\end{aligned}$$

Taille pour une précision absolue donnée

On fixe L ("précision absolue"). Si $f \approx 0$, on a :

$$n \approx \frac{4p(1-p)}{L^2}$$

Cas général (f pas forcément petit, et niveau de confiance z , quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$) :

$$n = \frac{1 + n_0}{1 + \frac{n_0}{N}}$$

$$\text{avec : } n_0 = \frac{z^2 p(1-p)}{L^2}$$

Taille pour une précision relative donnée

On se fixe une précision relative δ , définie par le rapport de la demi-longueur de l'intervalle de confiance à l'estimation :

$$\delta = \frac{2\hat{\sigma}}{p}$$

Taille pour une précision relative donnée

De manière équivalente, on peut fixer le coefficient de variation :

$$\hat{C}V(p) = \frac{\delta}{2} = \sqrt{(1-f) \frac{1-p}{p(n-1)}}$$

Si $f \approx 0$:

$$n \approx \frac{1-p}{p(\hat{C}V(p))^2}$$

Taille pour une précision relative donnée

Taille de l'échantillon pour une précision relative de $\pm\delta\%$ selon la valeur de la proportion recherchée :

	0,05	0,10	0,20	0,30	0,40	0,50
1 %	760000	360000	160000	93333	60000	40000
2 %	190000	90000	40000	23333	15000	10000
3 %	84444	40000	17778	10370	6667	4444
4 %	47500	22500	10000	5833	3750	2500
5 %	30400	14400	6400	3733	2400	1600
10 %	7600	3600	1600	933	600	400

Exercice : surface cultivée

Exemple d'application : la législation sur la méthode des quotas, en France.

- <http://www.commission-des-sondages.fr/oblig/instituts.htm>
- <http://www.ipsos.fr/faq>
- <http://www.20minutes.fr/politique/1567767-20150320-elections-departementales-ump-udi-30->

Partie 4

Exercice

Exercice : surface cultivée

On veut estimer la surface moyenne cultivée dans les fermes d'un canton rural. Sur $N = 2010$ fermes que comprend ce canton, on en tire 100 par sondage aléatoire simple. On mesure y_k la surface cultivée dans la ferme k (en hectares), et on trouve :

$$\sum_{k \in s} y_k = 2907 \text{ ha}$$

$$\sum_{k \in s} y_k^2 = 154593 \text{ ha}^2$$

- 1 Donner un estimateur sans biais pour la moyenne
- 2 Donner un intervalle de confiance à 95% pour cet estimateur

Partie 5

Autres estimations

Paragraphe 1

Estimation sur un domaine

Notations

$\mathcal{U}_d \subset \mathcal{U} =$ sous-population d'intérêt

$N_d =$ taille de \mathcal{U}_d (connue ou inconnue)

$P_d = \frac{N_d}{N} =$ taille relative de \mathcal{U}_d

$Q_d = 1 - P_d$

$s_d = s \cap \mathcal{U}_d$

$n_d =$ taille de s_d

$p_d = \frac{n_d}{n} =$ taille relative de s_d

$q_d = 1 - p_d$

Estimation de la taille d'un domaine

On définit sur \mathcal{U} la variable Z indicatrice d'appartenance au domaine :

$$Z_k = 1 \text{ si } k \in \mathcal{U}_d$$

$$Z_k = 0 \text{ sinon}$$

Estimation de la taille d'un domaine

Alors :

$$T(Z) = \sum_{k \in \mathcal{U}} Z_k = N_d$$

$$\bar{Z} = \frac{N_d}{N} = P_d$$

$$\bar{z} = \frac{1}{n} \sum_{k \in s} z_k = p_d$$

$$S^2 = \frac{N}{N-1} P_d Q_d$$

$$s^2 = \frac{n}{n-1} p_d q_d$$

Estimation de la taille d'un domaine

Théorème

$$\hat{N}_d = N \cdot p_d = N \cdot \frac{n_d}{n} \text{ est un estimateur sans biais de } N_d$$
$$\hat{P}_d = p_d \text{ est un estimateur sans biais de } P_d$$

Estimation de la taille d'un domaine

Démonstration.

Toutes ces quantités s'écrivent sous la forme d'un total (via Z) et correspondent à l'estimateur d'Horvitz-Thompson, qui est sans biais. □

Estimation de la taille d'un domaine

On a aussi :

$$\text{Var}(\hat{N}_d) = N^2(1-f) \frac{N}{N-1} \frac{P_d Q_d}{n}$$

$$\text{Var}(\hat{P}_d) = \text{Var}(p_d) = (1-f) \frac{N}{N-1} \frac{P_d Q_d}{n}$$

$$\hat{\text{Var}}(\hat{N}_d) = N^2(1-f) \frac{p_d q_d}{n-1}$$

$$\hat{\text{Var}}(\hat{P}_d) = \hat{\text{Var}}(p_d) = (1-f) \frac{p_d q_d}{n-1}$$

Estimation d'un total sur un domaine

On veut estimer le total $T_{\mathcal{U}_d}(Y)$ d'une variable Y sur le domaine \mathcal{U}_d . On définit sur \mathcal{U} la variable Y^d par :

$$Y_k^d = Y_k \text{ si } k \in \mathcal{U}_d$$

$$Y_k^d = 0 \text{ sinon}$$

Alors le total à estimer s'écrit :

$$T(Y^d) = \sum_{k \in \mathcal{U}} Y_k^d = \sum_{k \in \mathcal{U}_d} Y_k = T_{\mathcal{U}_d}(Y)$$

Estimation d'un total sur un domaine

Un estimateur sans biais de $T_{U_d}(Y)$ est :

$$\hat{T}_{U_d}(Y) = \frac{n_d}{n} N \bar{y}_d$$

où : $\bar{y}_d = \frac{1}{n_d} \sum_{k \in S_d} y_k$

Estimation d'un total sur un domaine

Et pour ce qui est de la précision :

$$\text{Var}(\hat{T}_{\mathcal{U}^d}(Y)) = N^2(1-f) \frac{S_{Y^d}^2}{n} \text{ avec } S_{Y^d}^2 = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (Y_k^d - \bar{Y}^d)^2$$

$$\hat{\text{Var}}(\hat{T}_{\mathcal{U}^d}(Y)) = N^2(1-f) \frac{s_{Y^d}^2}{n} \text{ avec } s_{Y^d}^2 = \frac{1}{n-1} \sum_{k \in s} (y_k^d - \bar{y}^d)^2$$

avec :

\bar{Y}^d = moyenne de Y^d sur \mathcal{U}

\bar{y}^d = moyenne de Y^d sur s

Estimation d'un total sur un domaine

Remarque sur la précision : Si on pose :

$$\bar{Y}_d = \frac{1}{N_d} \sum_{k \in \mathcal{U}_d} Y_k = \text{moyenne de } Y \text{ sur } \mathcal{U}_d$$

$$\bar{S}_d^2 = \frac{1}{N_d - 1} \sum_{k \in \mathcal{U}_d} (Y_k - \bar{Y}_d)^2 = \text{dispersion de } Y \text{ sur } \mathcal{U}_d$$

alors on a :

$$\text{Var}(\hat{T}_{\mathcal{U}_d}(Y)) \sim N_d^2 \left(\frac{1}{\mathbb{E}(n_d)} - \frac{1}{N_d} \right) \left[\frac{1 - \frac{1}{N_d}}{1 - \frac{1}{N}} S_d^2 + \frac{N - N_d}{N - 1} \bar{Y}_d^2 \right]$$

C'est donc la taille (attendue) de l'échantillon dans le domaine qui est déterminante et non n .

Estimateur alternatif pour le total

Si on connaît la taille du domaine N_d , un autre estimateur "naturel" de $T_{\mathcal{U}_d}(Y)$ est :

$$\hat{T}_{\mathcal{U}_d}^{alt}(Y) = N_d \bar{y}_d$$

C'est-à-dire que l'on remplace un estimateur sans biais de N_d :

$\hat{N}_d = \frac{n_d}{n} N$ par N_d . En général, $\hat{T}_{\mathcal{U}_d}^{alt}(Y)$ est préférable à $\hat{T}_{\mathcal{U}_d}(Y)$.

Estimation de la moyenne sur un domaine

On veut estimer : $\bar{Y}_d = \frac{T_{U_d}(Y)}{N_d}$. On peut utiliser :

$$\hat{Y}_d = \frac{\hat{T}_{U_d}(Y)}{N_d} \text{ si on connaît } N_d$$

$$\hat{Y}_d^{alt} = \frac{\hat{T}_{U_d}^{alt}(Y)}{N_d} = \bar{y}_d \text{ que l'on connaisse } N_d \text{ ou non !}$$

Ce dernier estimateur est assez intuitif (plugin !), et est en général meilleur que le premier.

Paragraphe 2

Estimation d'un ratio

Estimation d'un ratio

On cherche à estimer le rapport des totaux (ou des moyennes) de deux variables X et Y :

$$R = \frac{T(X)}{T(Y)} = \frac{\bar{X}}{\bar{Y}}$$

Attention ! L'estimateur d'Horvitz-Thompson est sans biais quand on estime un total ou une moyenne.

Estimation d'un ratio

On peut utiliser l'estimateur :

$$\hat{R} = \frac{T(\hat{X})}{T(\hat{Y})} = \frac{\hat{X}}{\hat{Y}} = \frac{\bar{x}}{\bar{y}}$$

Son biais s'écrit :

$$B(\hat{R}) \approx -\frac{1}{\bar{X}^2}(1-f)\frac{S_{XY} - RS_X^2}{n}$$

où : $S_{XY} = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (y_k - \bar{y})(x_k - \bar{x})$

Précision de l'estimateur du ratio

Son écart quadratique moyen et l'EQM estimé s'écrivent :

$$EQM(\hat{R}) = \frac{1-f}{n\bar{X}^2} (S_Y^2 + R^2 S_X^2 - 2RS_{XY})$$

$$E\hat{Q}M(\hat{R}) = \frac{1-f}{n\bar{X}^2} (s_Y^2 + \hat{R}^2 s_X^2 - 2\hat{R}s_{XY})$$

Paragraphe 3

Échantillonnage dans le temps

Problème

On veut estimer l'évolution de la moyenne d'une variable Y entre deux dates 1 et 2 : $\Delta Y = \bar{Y}_1 - \bar{Y}_2$

Méthode 1

Méthode 1 : On tire deux échantillons indépendants aux dates 1 et 2, selon un sondage aléatoire simple.

On a alors : $\Delta\hat{Y} = \bar{y}_2 - \bar{y}_1$ un estimateur sans biais de ΔY , de variance :

$$\text{Var}(\Delta\hat{Y}) = \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2)$$

Méthode 2 : panel

Méthode 2 : On utilise un panel, c'est-à-dire que l'on tire un échantillon en date 1, et on le réinterroge à la date 2. On a alors : $\hat{\Delta Y} = \bar{y}_2 - \bar{y}_1$ un estimateur sans biais de ΔY , de variance :

$$\text{Var}(\hat{\Delta Y}) = \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2) - 2\text{Cov}(\bar{y}_1, \bar{y}_2)$$

$$\text{où : } \text{Cov}(\bar{y}_1, \bar{y}_2) = (1 - f) \frac{S_{12}}{n}$$

$$\text{et : } S_{12} = \frac{1}{N-1} \sum_{k \in \mathcal{U}} (Y_{1k} - \bar{Y}_1)(Y_{2k} - \bar{Y}_2)$$

Méthode 2 : panel

Dans les bons cas, on a : $S_{12} > 0$, d'où :

$$\text{Var}(\hat{\Delta Y}) < \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2)$$

Exemple : enquête emploi à l'INSEE

Année	Trimestre	Sous-échantillons					
n	T1	6	5	4	3	2	1 →
	T2	→ 7	6	5	4	3	2
	T3	8	7	6	5	4	3
	T4	9	8	7	6	5	4
n+1	T1	10	9	8	7	6	5
	T2	11	10	9	8	7	6
	T3	12	11	10	9	8	7
	T4	13	12	11	10	9	8

Paragraphe 4

Conclusion sur le SAS

Conclusion sur le SAS

- Les estimateurs ont une forme simple
- Ne nécessite aucune information sur les individus de la base de sondage
- Est essentiel pour comprendre les plans de sondage plus complexes
- Peut permettre d'approximer les plans de sondage plus complexes