Introduction à la théorie des sondages - Cours 2

Thomas Merly-Alpa
thomas.merly-alpa@insee.fr

INSEE, département des méthodes statistiques

22 janvier 2018

Organisation

- 8 cours, 4 TD en demi-groupes
- 1/3 de la note : devoir maison à rendre **le 5 mars**
- 2/3 de la note : examen final le 19 mars
- 2 intervenants :
  - Thomas Merly-Alpa - `thomas.merly-alpa@insee.fr`
  - Martin Chevalier - `martin.chevalier@insee.fr`
- Les slides et TD du cours sont à l'adresse `http://nc233.com/teaching`

Principe du sondage

Objectif Construire un estimateur $\Phi(Y)$ d'une variable Y à partir d'un échantillon s de taille n tiré dans une population $\mathcal{U}$ de taille N.

Plan de sondage On définit un plan de sondage p comme une loi de probabilité sur l'ensemble des échantillons possibles $\mathcal{S}$.

Exemple : $\mathcal{U} = \{1, 2, 3\}$. On définit le plan de sondage p_1 par :

$$p_1(\{1\}) = p_1(\{2\}) = p_1(\{3\}) = 0$$
$$p_1(\{1,2\}) = 0,5 \quad p_1(\{1,3\}) = 0,2 \quad p_1(\{2,3\}) = 0,2$$
$$p_1(\{1,2,3\}) = 0,1$$

Remarque p_1 n'est pas un plan de sondage de taille fixe.

Probabilités d'inclusion

Le plan de sondage permet de déterminer des probabilités d'inclusion pour chaque unité de la population.

Probabilité d'inclusion simple $\pi_k = \sum_{s \in \mathcal{S}} \delta_k p(s)$

Probabilité d'inclusion double $\pi_{k,l} = \sum_{s \in \mathcal{S}} \delta_k \delta_l p(s)$

avec $\delta_k(s) = \mathbf{1}(k \in s)$

Enfin, on note $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Probabilités d'inclusion

L'estimateur d'Horvitz-Thompson est défini :

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

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

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

Sommaire

Chapitre 1

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 :

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

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

Un petit rappel

Combien vaut $\binom{N}{n}$? On rappelle que cette notation, n parmi N, signifie "le nombre de façons de choisir n éléments parmi N", noté aussi C_N^n. On a ainsi :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

où $n! = 1 \times 2 \times 3 \times \ldots \times n$.

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** :

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

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

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

Notations

On note, **dans l'échantillon s** :

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

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

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

Partie 2

Réaliser un tirage

Tirage aléatoire simple

Comment procéder en pratique pour tirer un échantillon ? Il y a plusieurs possibilités.

- En R, par exemple, on peut utiliser la fonction `sample` qui réalise un sondage aléatoire simple.
- Sinon, le moyen le plus simple consiste à trier la population complètement au hasard, et choisir les n premiers individus.

Tirage aléatoire simple

Comment trier aléatoirement une population ?

A		
B		
C		
D		
E		
F		
G		
H		
I		
J		
K		

Tirage aléatoire simple

On génère pour chaque individu une variable aléatoire uniforme, entre 0 et 1.

A	0.123	
B	0.245	
C	0.654	
D	0.987	
E	0.015	
F	0.975	
G	0.126	
H	0.745	
I	0.811	
J	0.626	
K	0.413	

Tirage aléatoire simple

On trie la population sur cette variable, et on prend les $n = 4$ premiers (par exemple)

E	0.015	Sélection
A	0.123	Sélection
G	0.126	Sélection
B	0.245	Sélection
K	0.413	
J	0.626	
C	0.654	
H	0.745	
I	0.811	
F	0.975	
D	0.987	

Chapitre 2

Estimation dans un SAS

Partie 1

Estimation d'un total

Estimateur d'Horvitz-Thompson

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

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

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

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 :*

$$\mathrm{Var}(\bar{\mathrm{y}}) = (1-f)\frac{S^2}{n}$$

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

Précision

Démonstration.

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

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 :

$$\hat{\mathrm{Var}}(\bar{y}) = (1-f)\frac{s^2}{n}$$

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

Partie 2

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

Sa *vraie* variance vaut :

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

On l'estime par :

$$\hat{\mathrm{Var}}(p) = (1 - f)\frac{p(1-p)}{n-1}$$

Précision

Demi-longueur de l'intervalle de confiance :

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

Coefficient de variation estimé :

$$\hat{CV}(p) = \frac{\sqrt{\hat{\mathrm{Var}}(p)}}{p}$$
$$= \sqrt{(1-f)\frac{1}{n-1}\frac{1-p}{p}}$$

Taille pour une précision absolue donnée

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

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

C'est souvent le cas lorsque qu'on s'intéresse à une grande population.

Taille pour une précision absolue donnée

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

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

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

Souvent $z = 2$, quantile à 95% de la normale centrée réduite.

Taille pour une précision relative donnée

La précision relative δ est définie par le rapport de la demi-longueur de l'intervalle de confiance à l'estimation :

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

On peut se ramener au coefficient de variation simplement :

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

Taille pour une précision relative donnée

De manière équivalente à la précision absolue L, on peut fixer le coefficient de variation $\hat{CV}(p)$. Dans ce cas, et si $f \approx 0$:

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

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

Exemple

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`

Partie 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 : $\hat{\Delta Y} = \bar{y_2} - \bar{y_1}$ un estimateur sans biais de ΔY, de variance :

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

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 :

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

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

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

Méthode 2 : panel

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

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

Exemple : enquête emploi à l'INSEE

Année	Trimestre	Sous-échantillons					
2016	T1	6	5	4	3	2	1 $\rightarrow$
	T2	$\rightarrow$ 7	6	5	4	3	2
	T3	8	7	6	5	4	3
	T4	9	8	7	6	5	4
2017	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

Partie 4

Estimation sur un domaine

Notations

$$\mathcal{U}_d \subset \mathcal{U} = \text{ sous-population d'intérêt}$$
$$N_d = \text{taille de } \mathcal{U}_d \text{ (connue ou inconnue)}$$
$$P_d = \frac{N_d}{N} = \text{ taille relative de } \mathcal{U}_d$$
$$Q_d = 1 - P_d$$
$$s_d = s \cap \mathcal{U}_d$$
$$n_d = \text{ taille de } s_d$$
$$p_d = \frac{n_d}{n} = \text{ taille relative de } s_d$$
$$q_d = 1 - p_d$$

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 :

$$\mathrm{Var}(\hat{\mathrm{N}}_\mathrm{d}) = N^2(1-f)\frac{N}{N-1}\frac{P_d Q_d}{n}$$

$$\mathrm{Var}(\hat{\mathrm{P}}_\mathrm{d}) = \mathrm{Var}(\mathrm{p_d}) = (1-\mathrm{f})\frac{\mathrm{N}}{\mathrm{N}-1}\frac{\mathrm{P_d Q_d}}{\mathrm{n}}$$

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

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

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 $T_{\mathcal{U}_d}(Y)$ est :

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

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

Estimation d'un total sur un domaine

Et pour ce qui est de la précision :

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

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

avec :

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

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

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 :

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

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 :

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

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, $\hat{T}^{alt}_{\mathcal{U}_d}(Y)$ est préférable à $\hat{T}_{\mathcal{U}_d}(Y)$.

Estimation de la moyenne sur un domaine

On veut estimer : $\bar{Y}_d = \dfrac{T_{\mathcal{U}_d}(Y)}{N_d}$. On peut utiliser :

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

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

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

Partie 5

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 :

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

Son biais s'écrit :

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

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

Précision de l'estimateur du ratio

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

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

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

Conclusion sur le SAS

- Les estimateurs ont une forme simple
- Ne nécesssite 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