Introduction à la théorie des sondages - Cours 2

Gaspar Massiot
gaspar.massiot@ined.fr

2024-2025

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 :

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 $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Partie 4

L'estimateur d'Horvitz-Thompson

Définition

Définition

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

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

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

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

Estimation sans biais

Theorem

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

La condition signifie que toutes les unités de la population ont une chance non nulle d'être dans l'échantillon.

Estimation sans biais

Démonstration.

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

Rappel : Variance / Précision

$$\mathrm{Var}(\hat{\Phi}) = \sum_{s} p(s) \cdot \left[\mathbb{E}(\hat{\Phi}) - \hat{\Phi}(s)\right]^2$$

C'est une mesure de la dispersion des valeurs $\hat{\Phi}(s)$ autour de leur moyenne.

Variance de l'estimateur de Horvitz-Thompson

Propriété

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

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

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

Partie 5

Autres estimateurs

Estimateur de Hájek, ...

L'estimateur de Horvitz-Thompson de la moyenne nécessite la **connaissance de** N, la taille de la population. Si on ne la connaît pas, on peut utiliser dans ce cas l'**estimateur de Hájek :**

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

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

... et d'autres

L'estimateur de Horvitz-Thompson constitue le fondement de l'estimation par sondage mais ce n'est pas le seul estimateur sans biais.
C'est l'estimateur le plus consensuel mais la théorie des sondages en a développé de nombreux autres pour répondre aux problématiques spécifiques de chaque concept étudié (population, taille d'échantillon, etc.)

Chapitre 3

Notion de base de sondage et d'erreur de sondage

Partie 1

Base de sondage

Propriétés de la base parfaite

Une base de sondage parfaite :

1. permet d'identifier les individus de façon non ambiguë
2. est exhaustive (on parle sinon de défaut de couverture)
3. est sans double-compte
4. contient de l'information auxiliaire

Défauts potentiels d'une base de sondages

Défauts potentiels d'une base de sondage :

- Sous-couverture
- Sur-couverture
- Répétition
- Classification erronée

Exemples

On veut mesurer la taille moyenne des français.es. Les bases suivantes sont-elles idéales ?

- L'annuaire
- Les listes électorales

Exemples

On veut mesurer la taille moyenne des français.es. Les bases suivantes sont-elles idéales ?

- L'annuaire $\rightarrow$ Pas ou peu d'information auxiliaire
- Les listes électorales $\rightarrow$ Pas exhaustif sur l'ensemble des français, nécessite une inscription

Partie 2

Erreur de sondage

Erreur d'échantillonnage

On étudie seulement une partie de la population : différence entre la vraie valeur dans la population et la valeur estimée à l'aide de l'échantillon.

Facteurs :

- Taille de l'échantillon
- Variabilité du paramètre d'intérêt
- Plan d'échantillonnage
- Estimateur utilisé

Erreur de mesure / d'observation

La valeur recueillie est différente de la vraie valeur attachée à l'individu k.

- Erreur de l'enquêté.e (mémoire)
- Formulation de la question
- Influence de l'enquêteur.rice et/ou du mode de passation
- Erreur de codification ou de saisie

Erreur due à la non-réponse

Non-réponse totale : Refus total de réponse ou absence

Non-réponse partielle : Refus / absence de réponse à certaines questions

Autres

Erreur de la base de sondage. En cas de défaut de couverture, biais de l'estimateur non mesurable.

Le sondage aléatoire simple (SAS)

Chapitre 1

Introduction

Partie 1

Définition

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 :

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

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 dans un SAS ? 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 d'un total dans un SAS

Retour sur l'estimateur d'Horvitz-Thompson

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

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

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

Poids de sondage dans le SAS

Dans le cas du SAS, 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 !)

Calcul de la précision dans le SAS

Theorem

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

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

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

Calcul de la précision dans le SAS

Démonstration.

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

Estimation de la précision dans le SAS

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

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

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

Chapitre 3

Quelle taille pour l'échantillon ?

Partie 1

Cas 1 : forte contrainte de coût

Cas 1 : forte contrainte de coût

$$n = \frac{C}{c}$$

Où :

- C : Budget total disponible pour l'enquête
- c : Coût unitaire du questionnaire

$\Rightarrow$ Attention, le coût unitaire du questionnaire peut comprendre beaucoup de choses : salaire des enquêteurs, affranchissement des lettres-avis, programmation du questionnaire, etc.

Partie 2

Cas 2 : faible contrainte de coût

Cas 2 : faible contrainte de coût - Estimation d'une proportion

Dans ce cas, avant de lancer notre enquête, on va chercher à évaluer la proportion d'individus portant la caractéristique d'intérêt :
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.

Cas 2 : faible contrainte de coût - Variance de p

Sa *vraie* variance vaut :

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

On l'estime par :

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

Cas 2 : faible contrainte de coût - Estimation de la précision L

On appelle "précision absolue", la demi-longueur de l'intervalle de confiance :

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

Le coefficient de variation est une autre mesure de précision que l'on appelle "précision relative", il est estimé par :

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

Cas 2 : faible contrainte de coût - 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.

Cas 2 : faible contrainte de coût - Taille pour une précision relative donnée

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

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

Dans le cas où on ne peut pas mesurer p, on se basera sur d'autres enquêtes, une enquête pilote ou une variable dites proxy qui permette de l'approcher.

Cas 2 : faible contrainte de coût - Taille pour une précision relative donnée

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

$\delta \backslash p$	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

Application

Un patron de chaîne veut connaître le nombre de personnes qui regardent l'émission de télévision qu'il diffuse en *access prime time*. Il commande ainsi une étude à un institut de sondages.

Celui-ci choisit d'échantillonner par sondage aléatoire simple n individus. Si la véritable audience (inconnue) de l'émission est de 1%, combien faut-il tirer de personnes pour obtenir un coefficient de variation (CV) de 5% ?

Application

Correction :
On connait la précision relative et la proportion visée et dans le cas où $f \approx 0$:

$$n \approx \frac{1-p}{p(CV(p))^2} = \frac{1-0.01}{0.01 * 0.05^2} = 39600$$

Question subsidiaire : Pourquoi est-il vraisemblable de supposer que le rapport entre la taille d'échantillon n et la taille de la population N ; $f = n/N \approx 0$?

Chapitre 4

Panel, domaine, ratio. . .

Partie 1

Évolution et Panel

Estimation de l'évolution d'une quantité dans un SAS

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 - Tirage de deux échantillons

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

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

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

Méthode 2 - Construction d'un 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 : $\widehat{\Delta Y} = \bar{y}_2 - \bar{y}_1$ un estimateur sans biais de ΔY, de variance :

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

Dans les bons cas, on a : $\mathrm{Cov}(\bar{y}_1, \bar{y}_2) > 0$, d'où :

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

Exemple : enquête emploi à l'INSEE

L'enquête emploi de l'INSEE est un exemple complexe d'utilisation de panels combiné à une rotation des échantillons
On pourra par exemple utiliser les individus du sous-échantillon 6 pour étudier l'évolution des résultats entre le 1er trimestre 2016 et le 2ème trimestre 2017.

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

Partie 2

Domaine

Estimation sur un domaine dans un SAS

- Un domaine d est un sous-ensemble de la population $\mathcal{U}$. Par exemple :
  - Les hommes ou les femmes ;
  - Les entreprises d'un secteur particulier ;
  - Les auditeurs d'une radio...
- On peut connaître ou non la taille de ce domaine
- On cherche à estimer cette taille, ou à estimer le total ou la moyenne d'une variable sur le domaine :
  - Le salaire moyen des femmes ;
  - Les productions des entreprises dans l'automobile ;
  - Le temps d'écoute d'une radio...

Estimation sur un domaine dans un SAS

Lorsque l'on veut estimer le total de Y sur le domaine d :

- La taille totale de l'échantillon n ne joue pas.
- C'est le nombre d'unités de l'échantillon qui sont dans le domaine d.
- Ce nombre est aléatoire : on ne sait pas combien de femmes seront interrogées.
- Ce nombre peut être petit : si on veut des résultats sur une seule ville.

Partie 3

Ratio

Estimation d'un ratio dans un SAS

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

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

Estimation d'un ratio

On peut utiliser l'estimateur :

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

mais il est biaisé !

$\Rightarrow$ Attention ! L'estimateur d'Horvitz-Thompson est sans biais quand on estime un total ou une moyenne, mais ce n'est pas le cas pour l'estimation d'un ratio...

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