

Introduction à la théorie des sondages - Cours 1

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

INSEE, département des méthodes statistiques

20 mars 2018



Organisation

- 5 séances (cours et TD)
- Notation : devoir maison à rendre
- 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>

Sommaire I

- 1 Pourquoi le sondage ?
 - Concept
 - Utilisations
 - Un échantillon "représentatif" ?
 - Pondération
- 2 Notion de base de sondage et d'erreur de sondage
 - Base de sondage
 - Erreur de sondage
 - Plan de sondage
- 3 Notion d'estimateur
 - Définitions
 - Pondération et probabilités d'inclusion

Sommaire II

- L'estimateur d'Horvitz-Thompson
- L'estimateur de Hájek
- Recherche d'un estimateur optimal

Chapitre 1

Pourquoi le sondage ?

Pourquoi le sondage ?

Notion de base de sondage et d'erreur de sondage

Notion d'estimateur

Concept

Utilisations

Pourquoi faire une enquête ?

Un échantillon "représentatif" ?

Pondération

Partie 1

Concept

Concept

Qu'est-ce que l'échantillonnage / l'estimation par sondage ?

- Une population de grande taille
- Compter ou interroger est coûteux
- On sélectionne quelques individus qui répondent " pour tout le monde"

Idée cruciale : sélectionner **aléatoirement** ces individus.

Historique

Historiquement et conceptuellement, rien d'évident !

- Laplace (1785) : recensement par une sous-partie de la population
- Kiaer (1895) : échantillon "représentatif"
... puis 1925 : acceptation de l'échantillonnage aléatoire
- Gallup (1936) : élections américaines

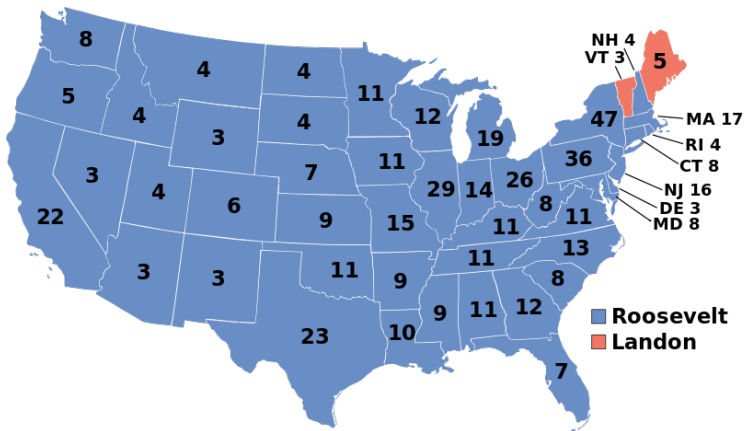
Élections américaines de 1936

- Duel entre Alfred Landon (Républicain) et Franklin Roosevelt (Démocrate)
- Un magazine interroge ses 2 millions de lecteurs : victoire de Landon
- Gallup fait un sondage sur 50 000 personnes : il prédit la victoire de Roosevelt

Pourquoi le sondage ?
Notion de base de sondage et d'erreur de sondage
Notion d'estimateur

Concept
Utilisations
Pourquoi faire une enquête ?
Un échantillon "représentatif" ?
Pondération

Élections américaines de 1936



Jusqu'en 2016 ?

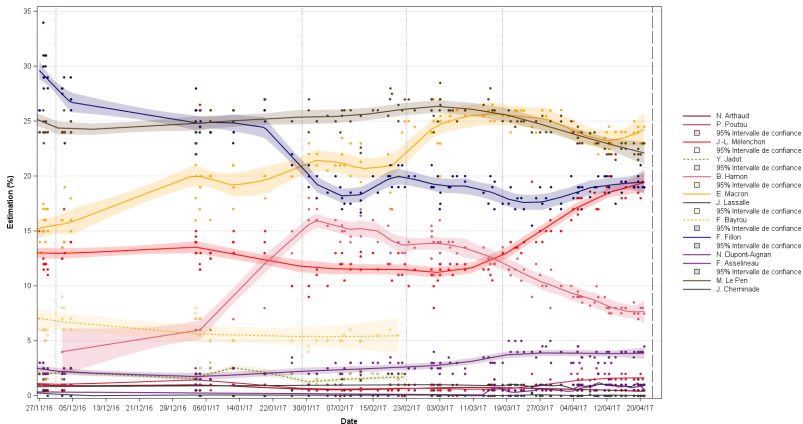
Est-ce la fin des sondages en 2016 ?

- Brexit
- Élection de Donald Trump
- Primaires de la droite en France

Ces "échecs" s'expliquent par des choix de méthode : ils ne remettent pas en cause la notion de sondages.

Un rebond en 2017

Les sondages avaient parfaitement prévu le score du premier tour des élections présidentielles de 2017 :



Pourquoi le sondage ?

Notion de base de sondage et d'erreur de sondage

Notion d'estimateur

Concept

Utilisations

Pourquoi faire une enquête ?

Un échantillon "représentatif" ?

Pondération

Partie 2

Utilisations

Statistique publique

- Enquêtes auprès des ménages : le moral des ménages, le taux de chômage
- Enquêtes auprès des entreprises - ESA (Enquête Sectorielle Annuelle) : Chiffre d'affaire par secteur, chiffres d'investissement, ...

Statistique publique

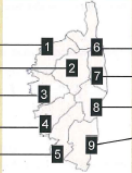
Et d'autres sujets. . .

17 Comment se répartissent vos nuitées dans ces différents modes d'hébergement (sur 2 caractères) ?

Mode d'hébergement	Nombre de nuits	Mode d'hébergement	Nombre de nuits
Chez la famille	<input type="text"/> <input type="text"/>	Chez des amis	<input type="text"/> <input type="text"/>
Dans votre résidence secondaire	<input type="text"/> <input type="text"/>	Location appartement, maison, ...	<input type="text"/> <input type="text"/>
Hôtel	<input type="text"/> <input type="text"/>	Camping	<input type="text"/> <input type="text"/>
Résidence de tourisme	<input type="text"/> <input type="text"/>	Village de vacances	<input type="text"/> <input type="text"/>
Chambre d'hôte	<input type="text"/> <input type="text"/>	Gîte, meublé de tourisme ...	<input type="text"/> <input type="text"/>
Refuge	<input type="text"/> <input type="text"/>	Bateau de plaisance	<input type="text"/> <input type="text"/>

18 Combien de nuitées avez-vous passées dans ces régions ?

Région	Nb. de nuit	Région	Nb. de nuit
1 Balagne, Calvi, Île-Rousse, Galeria...	<input type="text"/> <input type="text"/>	6 Bastia, Saint-Florent, Cap-Corse, Patrimoine...	<input type="text"/> <input type="text"/>
2 Corte, Restonica, Noli...	<input type="text"/> <input type="text"/>	7 Castagniccia, Casinca, Orezza, Corsica Verde...	<input type="text"/> <input type="text"/>
3 Cargèse, Sagone, Porto, Piana...	<input type="text"/> <input type="text"/>	8 Côte Orientale, Ghisonaccia, Aleria...	<input type="text"/> <input type="text"/>
4 Ajaccio, Porticcio, Pays ajaccien...	<input type="text"/> <input type="text"/>	9 Porto-Vecchio, Bonifacio, Alta-Rocca...	<input type="text"/> <input type="text"/>
5 Sartène, Propriano, Taravo...	<input type="text"/> <input type="text"/>		



Autres exemples

- Biologie : dénombrement d'espèces
- Politique
- Marketing



Pourquoi le sondage ?

Notion de base de sondage et d'erreur de sondage

Notion d'estimateur

Concept

Utilisations

Pourquoi faire une enquête ?

Un échantillon "représentatif" ?

Pondération

Partie 3

Pourquoi faire une enquête ?

Conception

Une enquête peut être coûteuse (en budget - 2 millions pour une enquête INSEE, mais aussi en temps des enquêtés). Il faut donc s'assurer que le sujet est :

- Pertinent (contraintes européennes, demandes d'études, sujet actuel)
- Non couvert (autres enquêtes, autres données)
- Réalisable (pas trop complexe, légalité, anonymisation)

Données administratives

Pourquoi ne pas utiliser les données des impôts pour estimer les revenus ?

- Différences de concept
- Revenus non déclarés
- Peu d'information complémentaire

Autre exemple : mesures d'audiences et Box.

Questionnaire

Une fois les objectifs identifiés, il faut réaliser un questionnaire :

- Qui colle aux concepts
- Mais compréhensible par l'enquêté : ni équivoque, ni flou
- Qui permette de la comparabilité avec d'autres sources

Questionnaire

Ce n'est pas une science exacte !

- Questions ouvertes ou fermées ?
- Quelles modalités de réponse ?
- Quel est l'ordre des questions ?

⇒ D. Verger, "Rédiger un bon questionnaire, une variante de la quadrature du cercle ?"

(<https://www.epsilon.insee.fr/jspui/handle/1/8488>)

Partie 4

Un échantillon "représentatif" ?

Un concept erroné

Un "échantillon représentatif" :

- On entend souvent cette formule
- Quel est son sens ? "Village" de 100 habitants
- Est-ce pertinent ? Si on veut connaître la production automobile en France, quelle est la bonne stratégie ?

"Sondage" devrait toujours aller de pair avec "**objectif**" (même si les objectifs pour un même échantillon peuvent être nombreux).

L'estimation naïve

Pour l'estimation du total et de la moyenne d'une variable Y , l'estimateur « naïf » est :

- Pour le total, la somme des valeurs Y des individus de l'échantillon.
- Pour la moyenne, la moyenne des valeurs Y des individus de l'échantillon.

En général, l'estimation naïve est fautive (*biaisée*), surtout quand l'échantillon est choisi de façon complexe.

Exemple d'estimation naïve

Un exemple : étude du temps quotidien passé sur Internet :

Père			15 minutes
Mère			30 minutes
Enfant 1			215 minutes
Enfant 2			240 minutes

Vraie moyenne : 125 minutes.

Exemple d'estimation naïve

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon		15 minutes
Mère	Dans l'échantillon		30 minutes
Enfant 1	Dans l'échantillon		215 minutes
Enfant 2	/	/	? minutes

Estimateur naïf = ...

Exemple d'estimation naïve

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon		15 minutes
Mère	Dans l'échantillon		30 minutes
Enfant 1	Dans l'échantillon		215 minutes
Enfant 2	/	/	? minutes

Estimateur naïf : $(15 + 30 + 215) / 3 \approx 87$ minutes

Exemple d'estimation naïve

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon		15 minutes
Mère	Dans l'échantillon		30 minutes
Enfant 1	/	/	? minutes
Enfant 2	Dans l'échantillon		240 minutes

Estimateur naïf = ...

Exemple d'estimation naïve

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon		15 minutes
Mère	Dans l'échantillon		30 minutes
Enfant 1	/	/	? minutes
Enfant 2	Dans l'échantillon		240 minutes

Estimateur naïf : $(15 + 30 + 240) / 3 = 95$ minutes

Partie 5

Pondération

Pondérer ?

Pour éviter d'utiliser l'estimateur naïf, on utilise généralement ce qu'on appelle des poids, qu'on note w (pour *weight* en anglais).

Le poids d'un individu correspond au nombre d'individus que l'individu de l'échantillon représente dans la population. Si l'on interroge 1 individu sur 100, le poids est alors de 100.

L'estimateur pondéré du total est alors la somme des $w_i y_i$ sur l'échantillon.

Retour sur l'exemple

Retour sur l'exemple du temps quotidien passé sur Internet :

Père			15 minutes
Mère			30 minutes
Enfant 1			215 minutes
Enfant 2			240 minutes

Vraie moyenne : 125 minutes.

Retour sur l'exemple

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon	Poids = 1	15 minutes
Mère	Dans l'échantillon	Poids = 1	30 minutes
Enfant 1	Dans l'échantillon	Poids = 2	215 minutes
Enfant 2	/	/	? minutes

Estimateur naïf : $(15 + 30 + 215) / 3 \approx 87$ minutes

Estimateur pondéré : ...

Retour sur l'exemple

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon	Poids = 1	15 minutes
Mère	Dans l'échantillon	Poids = 1	30 minutes
Enfant 1	Dans l'échantillon	Poids = 2	215 minutes
Enfant 2	/	/	? minutes

Estimateur naïf : $(15 + 30 + 215) / 3 \approx 87$ minutes

Estimateur pondéré : $(15 + 30 + 2*215) / 4 = 118,75$ minutes

Retour sur l'exemple

On interroge les deux parents, et un des enfants au hasard.

Père	Dans l'échantillon	Poids = 1	15 minutes
Mère	Dans l'échantillon	Poids = 1	30 minutes
Enfant 1	/	/	? minutes
Enfant 2	Dans l'échantillon	Poids = 2	240 minutes

Estimateur naïf : $(15 + 30 + 240) / 3 = 95$ minutes

Estimateur pondéré : $(15 + 30 + 2*240) / 4 = 131,25$ minutes

À retenir

- On construit notre sondage et donc notre échantillon dans un but précis.
- On utilise les résultats obtenus en se rappelant de notre méthode de sondage.

Pourquoi le sondage ?

Notion de base de sondage et d'erreur de sondage

Notion d'estimateur

Concept

Utilisations

Pourquoi faire une enquête ?

Un échantillon "représentatif" ?

Pondération

Exercice 1 - TD 1

Chapitre 2

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 (voir cours suivants)

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. Les bases suivantes sont-elles idéales ?

- L'annuaire
- Les listes électorales

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 (mémoire)
- Formulation de la question
- Influence de l'enquêteur
- 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.

Partie 3

Plan de sondage

Notations - Définitions

- Population $\mathcal{U} = \{u_1, \dots, u_k, \dots, u_N\}$
- L'individu $u_k \in \mathcal{U}$ est repéré sans ambiguïté par son identifiant k .
- Variable d'intérêt Y , qui prend la valeur y_k pour l'individu k
- Objectif du sondage : Mesurer $\Phi(Y)$, une fonction dépendant de Y .

Notations - Définitions

Y peut être

- quantitative (exemple : revenu). Dans ce cas Φ peut être le total, la moyenne, etc.
- qualitative, c'est-à-dire prendre un nombre fini de valeurs (exemple : sexe). Dans ce cas, Φ peut être la répartition dans la population.

Notations - Définitions

- Échantillon $s \subset \mathcal{U}$
- Si $s = \mathcal{U}$, recensement
- Chaque individu $u_k, k \in s$ est interrogé, et on relève y_k
- Les $y_k, k \in s$ seront utilisés pour construire un **estimateur** $\hat{\Phi}$ de Φ (voir partie 3)
- Les **unités d'échantillonnage** peuvent ne pas être les individus de la population eux-mêmes (proxy)

Notations - Définitions

La **base de sondage** donne les moyens d'identifier et de joindre les unités d'échantillonnage.

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} telle que :

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

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

Plan de sondage sans remise - 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 :

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

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

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

Plan de sondage avec remise - définition

On note $\tilde{\mathcal{S}}$ l'ensemble des échantillons avec remise ordonnés de \mathcal{U} .
 $\tilde{\mathcal{S}}$ est de cardinal **infini**.

Plan de sondage avec remise - définition

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

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

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

Plan de sondage avec remise - exemple

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

$$\tilde{p}(\{2\}) = 0 \quad \tilde{p}(\{1, 3\}) = \frac{1}{6} \quad \tilde{p}(\{2, 2\}) = \frac{1}{12}$$

$$\tilde{p}(\{3\}) = 0 \quad \tilde{p}(\{2, 3\}) = \frac{1}{12} \quad \tilde{p}(\{3, 3\}) = \frac{1}{6}$$

Plans avec remise

Dans ce cours, on s'intéresse principalement aux plans de sondages sans remise.

Chapitre 3

Notion d'estimateur

Partie 1

Définitions

Paramètre d'intérêt

Retour sur la slide 50. Y est la **variable d'intérêt** et $\Phi(Y)$ est le **paramètre d'intérêt**.

Attention, Y n'est **pas aléatoire** !

Estimateur

Une fois l'échantillon s tiré, on **estime** $\Phi(Y)$ à l'aide d'une fonction, notée $\hat{\Phi}(s)$, qui dépend de l'échantillon.

$\hat{\Phi}(s)$ est appelé un **estimateur** de $\Phi(Y)$.

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**.

Biais

$$B(\hat{\phi}) = \mathbb{E}(\hat{\phi}) - \phi$$

Si $B(\hat{\phi}) = 0$, alors on parle **d'estimateur sans biais**.

Variance / Précision

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

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

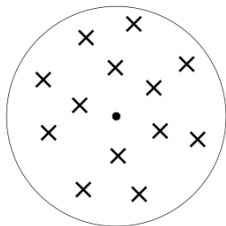
Variance / Précision

Quantités liées :

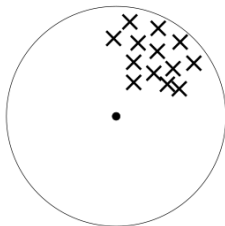
$$\sigma(\hat{\Phi}) = \sqrt{\text{Var}(\hat{\Phi})}, \text{écart-type}$$

$$CV(\hat{\Phi}) = \frac{\sigma(\hat{\Phi})}{\mathbb{E}(\hat{\Phi})}, \text{coefficient de variation}$$

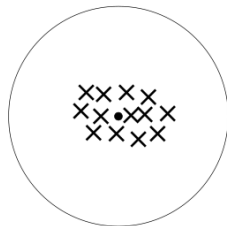
Schéma



Cas 1



Cas 2



Cas 3

Erreur quadratique moyenne

$$\begin{aligned}EQM(\hat{\Phi}) &= \sum_s p(s) \cdot [\Phi - \hat{\Phi}(s)]^2 \\ &= \text{Var}(\hat{\Phi}) + B(\hat{\Phi})^2\end{aligned}$$

Entre deux estimateurs sans biais, celui qui a la plus petite variance est de meilleure qualité.

Construction d'un intervalle de confiance

La **vraie variance** $\text{Var}(\hat{\Phi})$ n'est pas connue (il faudrait pour cela pouvoir tirer tous les échantillons).

Il faudra donc estimer la variance à partir des données de l'échantillon. L'estimateur sera noté $\hat{V}(\hat{\Phi})$ ou $\hat{\text{Var}}(\hat{\Phi})$.

Construction d'un intervalle de confiance

Estimateurs des quantités liées à la variance :

$$\hat{\sigma}(\hat{\Phi}) = \sqrt{\hat{\text{Var}}(\hat{\Phi})}, \text{ écart-type}$$

$$\hat{C}V(\hat{\Phi}) = \frac{\hat{\sigma}(\hat{\Phi})}{\hat{\Phi}}, \text{ coefficient de variation}$$

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 :

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

Partie 2

Pondération et probabilités d'inclusion

L'estimateur naïf

Rappel : pour l'estimation du total et de la moyenne d'une variable Y , l'estimateur « naïf » s'écrit :

$$\hat{T}(Y)_{naif} = \sum_{k \in S} y_k$$
$$\hat{y}_{naif} = \frac{1}{n} \sum_{k \in S} y_k$$

L'estimateur naïf

En général, l'estimation naïve est biaisée :

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

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

Probabilités d'inclusion π_k et π_{kl}

Pour résoudre le problème de biais, on doit utiliser une pondération adaptée à l'échantillon. L'outil à mobiliser : les **probabilités d'inclusion** de premier et de second degré : pour $k \in \mathcal{U}$,

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

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

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

Probabilités d'inclusion π_k et π_{kl} - Propriétés

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

$$\mathbb{E}(\delta_k) = \pi_k$$

$$\mathbb{E}(\delta_k \delta_l) = \pi_{kl}$$

$$\text{Var}(\delta_k) = \pi_k(1 - \pi_k) \quad \text{Cov}(\delta_k \delta_l) = \Delta_{kl}$$

Probabilités d'inclusion π_k et π_{kl} - Propriétés

Pour un plan à **taille fixe** n , on a :

$$\begin{aligned}\sum_{k \in \mathcal{U}} \pi_k &= n \\ \sum_{\substack{k, l \in \mathcal{U} \\ k \neq l}} \pi_{kl} &= n(n-1) \\ \sum_{\substack{l \in \mathcal{U} \\ l \neq k}} \pi_{kl} &= \pi_k(n-1)\end{aligned}$$

Exercice 2 - TD 1

Partie 3

L'estimateur d'Horvitz-Thompson

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

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

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

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 :

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

(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. □

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 :

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

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.

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 :

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

Construction d'un intervalle de confiance

Pour l'estimateur de variance de l'estimateur d'Horvitz-Thompson pour un plan à taille fixe, cela revient à calculer :

La borne inférieure de l'intervalle de confiance

$$\sum_{k \in s} \frac{y_k}{\pi_k} - 2 \sqrt{\frac{1}{2} \sum_{k \in s} \sum_{l \in s, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right)}$$

et la borne supérieure de l'intervalle de confiance

$$\sum_{k \in s} \frac{y_k}{\pi_k} + 2 \sqrt{\frac{1}{2} \sum_{k \in s} \sum_{l \in s, l \neq k} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right)}$$

Exercice 3 - TD 1

Partie 4

L'estimateur de Hájek

Définition

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 :

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

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 5

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.

Difficile à répondre car cela dépend de la population, de la taille d'échantillon, des concepts mesurés. . .

Exercice 4 - TD 1