

Introduction à la théorie des sondages - Cours 1

Antoine Rebecq

`antoine.rebecq@insee.fr`

INSEE, direction de la méthodologie

18 janvier 2016



Sommaire I

- 1 Pourquoi le sondage ?
 - Concept
 - Utilités
 - Les “horreurs” conceptuelles
- 2 Plan de sondage
 - Notations - Définitions
 - Plans avec et sans remise
 - Probabilités d'inclusion π
- 3 Notion d'estimateur
 - Définitions
 - Comment juger de la qualité d'un estimateur ?
- 4 Notion de base de sondage et d'erreur de sondage
 - Base de sondage
 - Erreur de sondage

Chapitre 1

Pourquoi le sondage ?

Pourquoi le sondage ?

Plan de sondage

Notion d'estimateur

Notion de base de sondage et d'erreur de sondage

Concept

Utilités

Les "horreurs" conceptuelles

Partie 1

Concept

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

Historiquement et conceptuellement, rien d'évident !

- Laplace
- Kiaer, 1895 ... 1925
- Gallup

Pourquoi le sondage ?

Plan de sondage

Notion d'estimateur

Notion de base de sondage et d'erreur de sondage

Concept

Utilités

Les "horreurs" conceptuelles

Élections américaines de 1936



Notations

Moralité : Méthode récente et **contre-intuitive** !

Pourquoi le sondage ?

Plan de sondage

Notion d'estimateur

Notion de base de sondage et d'erreur de sondage

Concept

Utilités

Les "horreurs" conceptuelles

Partie 2

Utilités

Statistique publique

- Exemples enquête ménages : le moral des ménages, le taux de chômage
- Exemple enquêtes entreprise : ESA (Enquête Sectorielle Annuelle) : Chiffre d'affaire par secteur, chiffres d'investissement, ...

Autres exemples

- Biologie : dénombrement d'espèces
- Sport : mesure / détection dopage

Sondage non probabiliste

- Politique
- Marketing



Partie 3

Les "horreurs" conceptuelles

Deux concepts erronés

- "Représentativité"
- Estimation "plugin"

À retenir

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

Chapitre 2

Plan de sondage

Partie 1

Notations - Définitions

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.

Partie 2

Plans avec et sans remise

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 :

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

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

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.

Partie 3

Probabilités d'inclusion π

Probabilités d'inclusion π_k et π_{kl}

En pratique, p est peu utile. On utilise plutôt les probabilités d'inclusion de premier et de second degré : pour $k \in \mathcal{U}$,

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

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

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

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 :

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

Chapitre 3

Notion d'estimateur

Partie 1

Définitions

Paramètre d'intérêt

Retour sur la slide 18. 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 $\Phi(Y)$.

Partie 2

Comment juger de la qualité d'un estimateur ?

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.

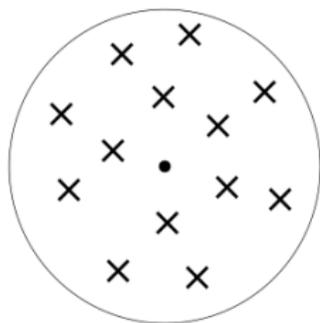
Variance / Précision

Quantités liées :

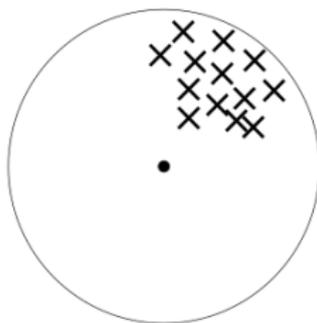
$$\sigma(\hat{\Phi}) = \sqrt{\text{Var}(\hat{\Phi})}, \text{écart-type}$$

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

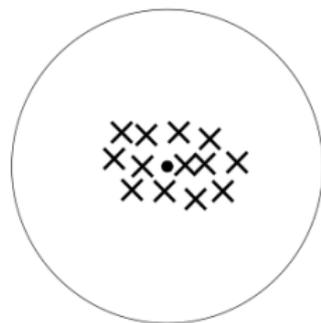
Schéma



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 :

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

Chapitre 4

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 (sinon on parle 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

Exemple

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.

Slides du cours

<http://nc233.com/teaching>