Introduction à la théorie des sondages - Cours 1

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

INSEE, département des méthodes statistiques

16 janvier 2017

Organisation

- 8 cours, 4 TD en demi-groupes
- 1/3 de la note : devoir maison à rendre **le 27 février**
- 2/3 de la note : examen final le 20 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`

Sommaire

Chapitre 1

Pourquoi le sondage ?

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

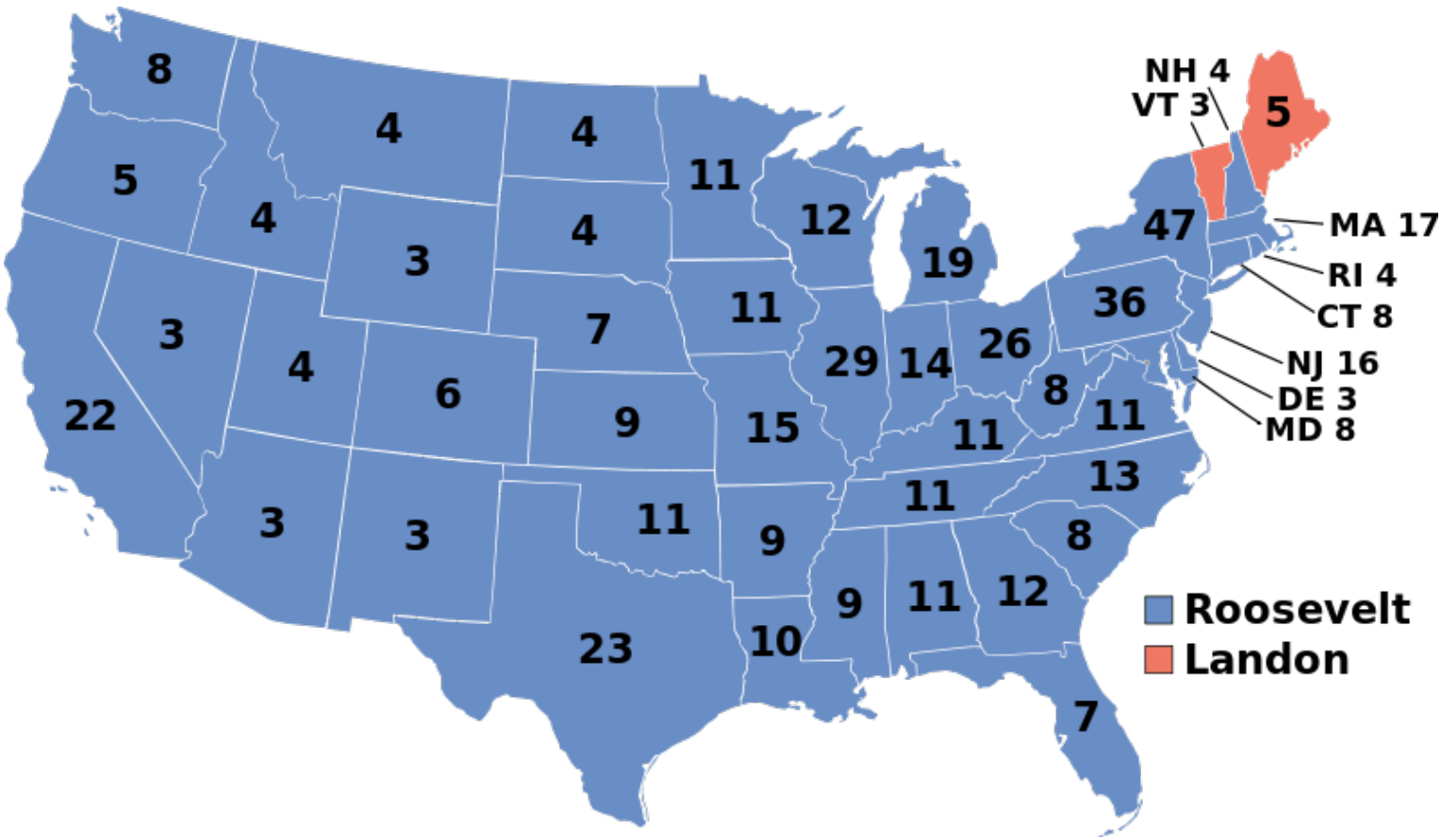
Historiquement et conceptuellement, rien d'évident !

- 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

Élections américaines de 1936



Jusqu'en 2016 ?

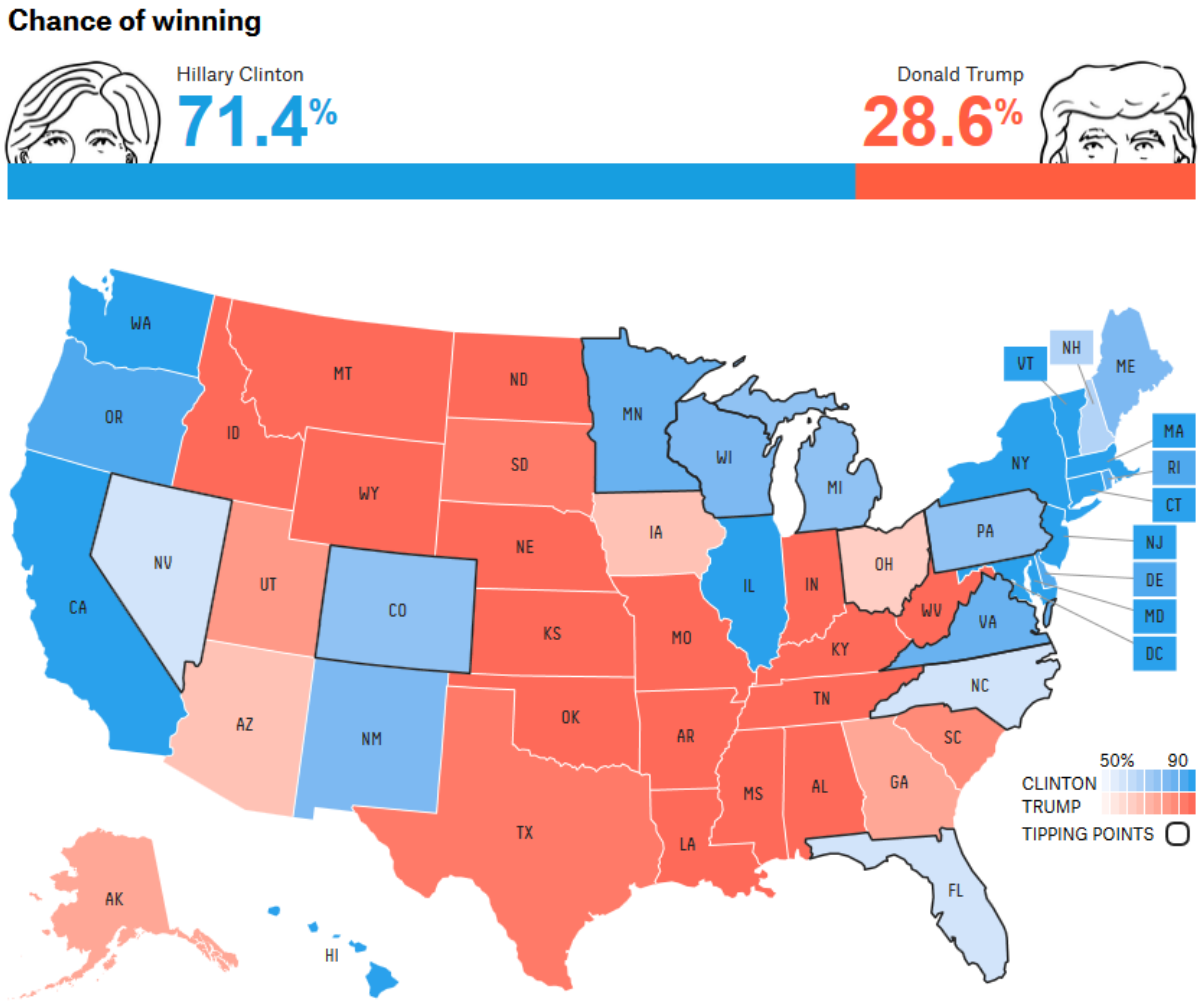
Est-ce la fin des sondages en 2016 ?

- 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.

Élections américaines de 2016

Nate Silver, `http://fivethirtyeight.com` :



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

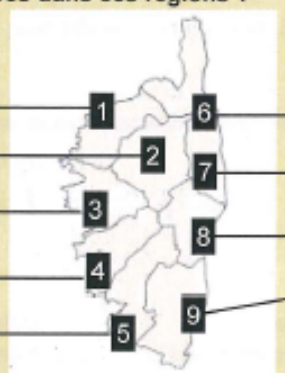
Et d'autres sujets. . .

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		Chez des amis	
Dans votre résidence secondaire		Location appartement, maison, ...	
Hôtel		Camping	
Résidence de tourisme		Village de vacances	
Chambre d'hôte		Gîte, meublé de tourisme ...	
Refuge		Bateau de plaisance	

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

	Région	Nb. de nuit
1	Balagne, Calvi, Île-Rousse, Galeria...	
2	Corte, Restonica, Niolu...	
3	Cargèse, Sagone, Porto, Piana...	
4	Ajaccio, Porticcio Pays ajaccien...	
5	Sartène, Propriano, Taravo...	



	Région	Nb. de nuit
6	Bastia, Saint-Florent, Cap-Corse, Patrimonio...	
7	Castagniccia, Casinca, Orezza, Costa Verde...	
8	Côte Orientale, Ghisonaccia, Aleria...	
9	Porto-Vecchio, Bonifacio Alta-Rocca...	

Autres exemples

- Biologie : dénombrement d'espèces
- Politique
- Marketing

Partie 3

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).

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.

Chapitre 2

Plan de sondage

Partie 1

Notations - Définitions

Notations - Définitions

- Population $\mathcal{U} = \{u_1, ..., u_k, ..., u_N\}$
- 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 $\mathcal{S}$ 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.

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 $k \in \mathcal{U}$,

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

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

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

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

$$\mathrm{Var}(\delta_\mathrm{k}) = \pi_\mathrm{k}(1 - \pi_\mathrm{k}) \qquad \mathrm{Cov}(\delta_\mathrm{k}\delta_\mathrm{l}) = \Delta_\mathrm{kl}$$

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

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

$$\sum_{k \in \mathcal{U}} \pi_k = n$$

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

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

Chapitre 3

Notion d'estimateur

Partie 1

Définitions

Paramètre d'intérêt

Retour sur la slide 21. 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

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

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

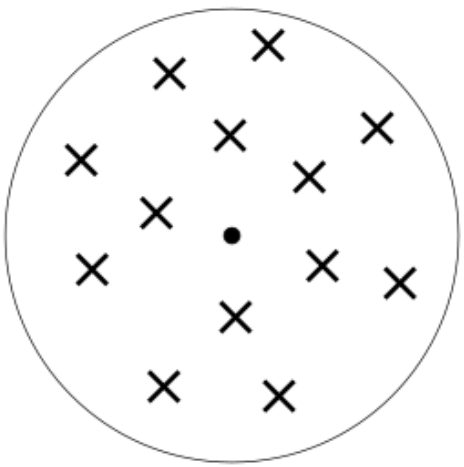
Variance / Précision

Quantités liées :

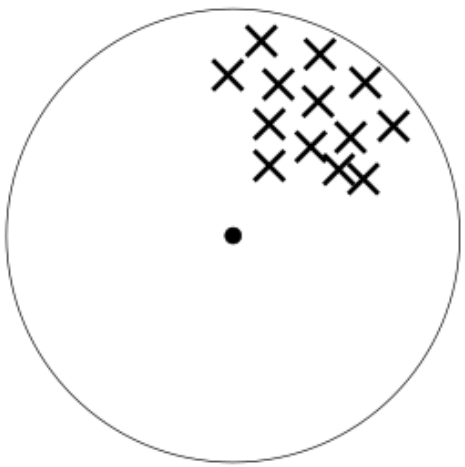
$$\sigma(\hat{\Phi}) = \sqrt{\mathrm{Var}(\hat{\Phi})}, \text{écart-type}$$

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

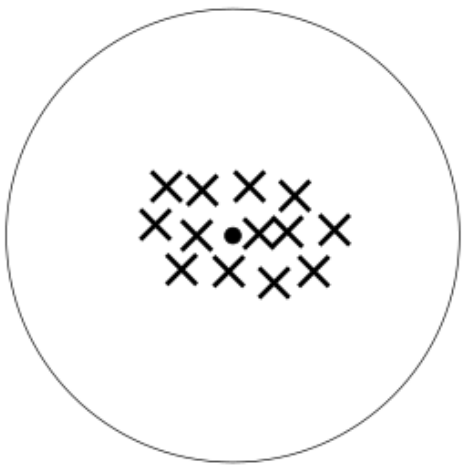
Schéma



Cas 1



Cas 2



Cas 3

Erreur quadratique moyenne

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

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** $\mathrm{Var}(\hat{\Phi})$ 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 $\hat{\mathrm{Var}}(\hat{\Phi})$.

Construction d'un intervalle de confiance

Estimateurs des quantités liées à la variance :

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

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

Construction d'un intervalle de confiance

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

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 (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.