Méthodologie d'enquête et sondages - Partie 1

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

INSEE, département des méthodes statistiques

8 et 9 janvier 2018

Sommaire I

Sommaire II

Sommaire III

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

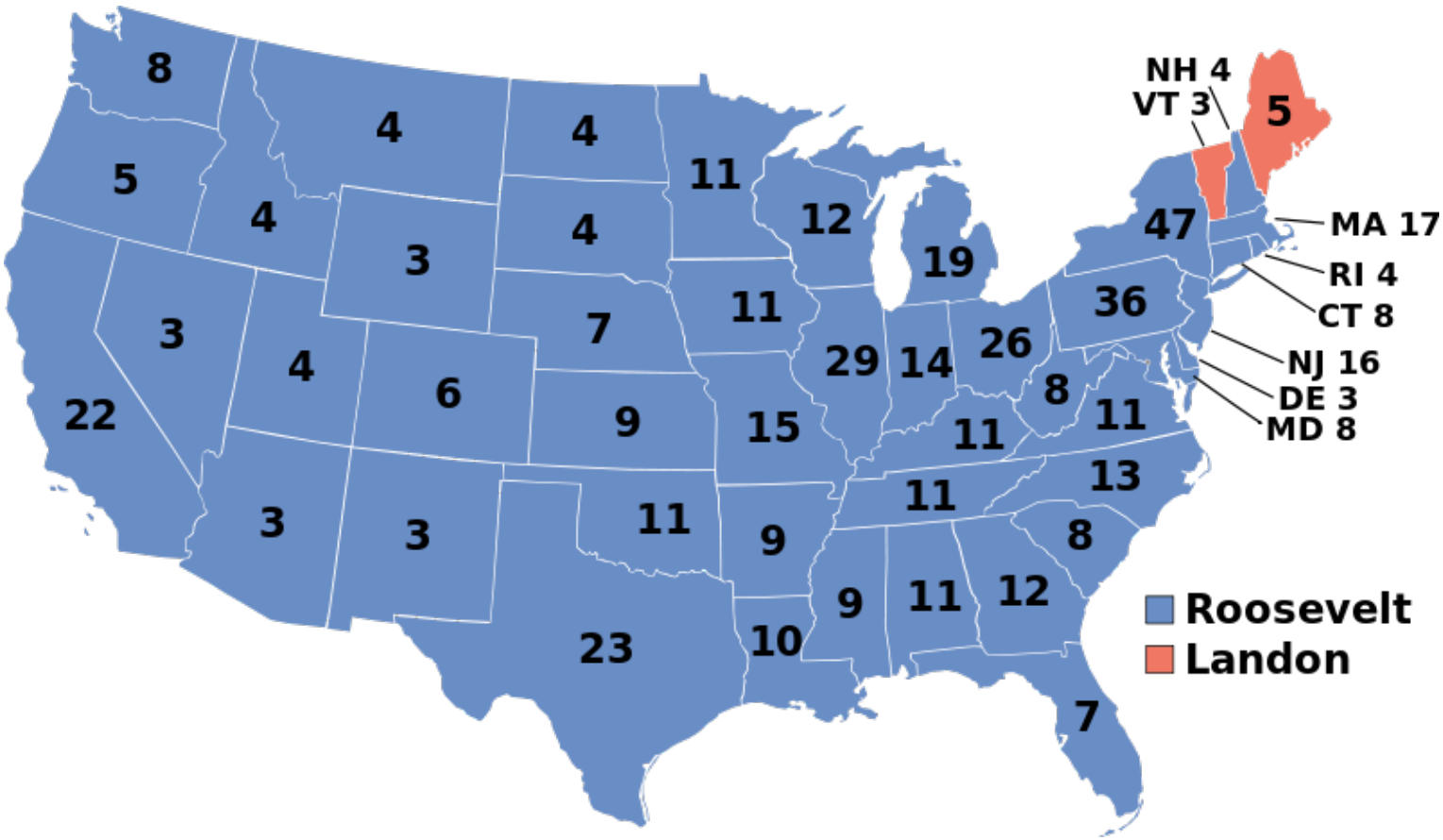
Historiquement et conceptuellement, rien d'évident !

- 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 ?

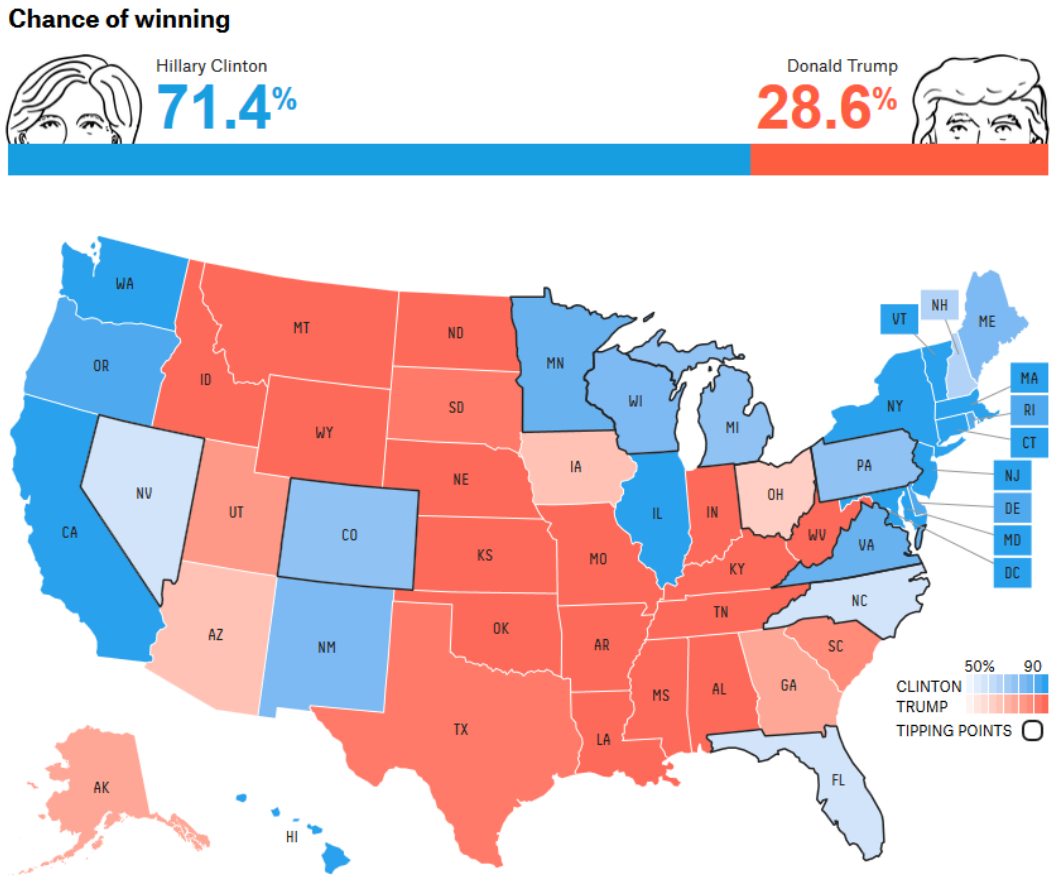
Est-ce la fin des sondages en 2016 ?

- 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.

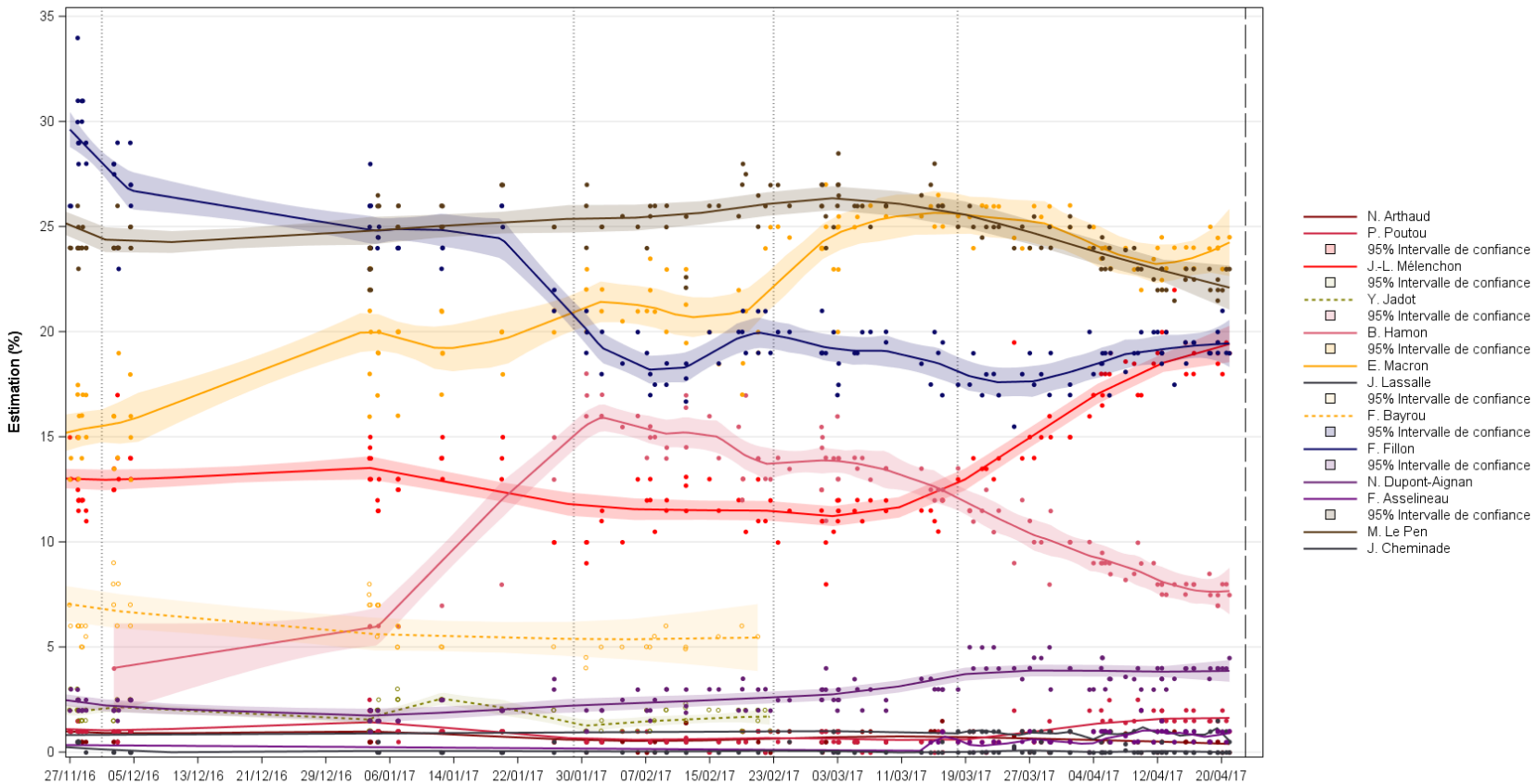
Élections américaines de 2016

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



Un rebond en 2017

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



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 :

- EHIS : Enquête sur le Santé des populations (INSEE-DREES-IRDES) ;
- Panel ELIPSS : Panel de sciences sociales (Sciences Po) ;
- EGT : Enquête Globale Transport (Mobilité IDF-CEREMA) ;
- CT/RPS : Conditions de Travail (INSEE-DARES-DREES-DGAFP)...

Autres exemples

- Biologie : dénombrement d'espèces
- Politique
- Marketing

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 ?

$\Rightarrow$ 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 fausse (*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.

Chapitre 2

Vocabulaire : plan, probabilités

Partie 1

Notations

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 $y_k, k \in s$ sont utilisés pour construire un **estimateur** $\hat{\Phi}$ de Φ
- 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 (et la plupart du temps), on s'intéresse principalement aux plans de sondages sans remise.

Partie 3

Les probabilités d'inclusion

Probabilité d'inclusion π_k

En pratique, p est peu utile. On utilise plutôt les probabilités d'inclusion. La probabilité d'inclusion simple d'un individu k est la probabilité que cet individu soit dans l'échantillon. Ainsi, pour $k \in \mathcal{U}$,

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

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

Probabilité d'inclusion π_{kl}

La probabilité d'inclusion double de deux individus k et l est la probabilité que ces deux individu soient ensemble dans l'échantillon. Ainsi, pour $k, l \in \mathcal{U}$,

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

Attention : on n'a pas $\pi_{kl} = \pi_k \pi_l$ en général ! On note par ailleurs $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

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

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

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

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 générales

Paramètre d'intérêt

Retour sur la slide 40. 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

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

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

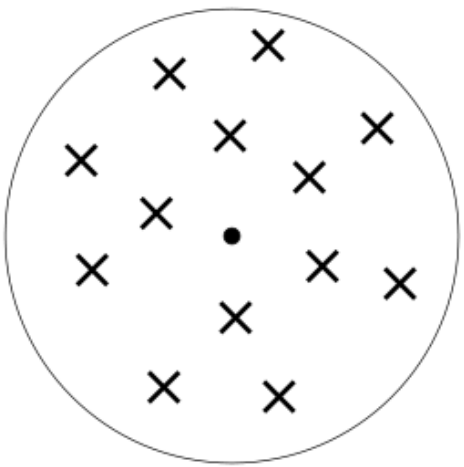
Variance / Précision

Quantités liées :

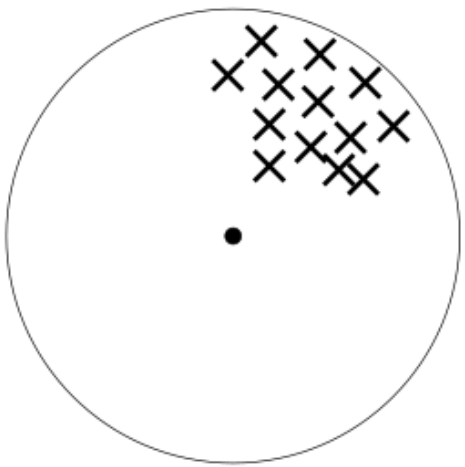
$$\sigma(\hat{\Phi}) = \sqrt{\mathrm{Var}(\hat{\Phi})}, \text{écart-type}$$

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

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 :

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

Partie 2

Retour sur l'estimateur naïf

L'estimateur naïf

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

$$\hat{T}(Y)_{naif} = \sum_{k \in s} y_k$$

$$\hat{\bar{y}}_{naif} = \frac{1}{n} \sum_{k \in s} y_k$$

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**.

Partie 3

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* $w_k = \dfrac{1}{\pi_k}$

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

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

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

Variance de l'estimateur de Horvitz-Thompson

Propriété

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

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

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

Partie 4

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.

Estimateur optimal ?

L'estimateur de Horvitz-Thompson constitue le fondement de l'estimation par sondage mais ce 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...

Éléphants de Basu

Exemple classique en Sondages :

- Le responsable d'un cirque veut connaître le poids de ses 50 éléphants.
- Il sait que Sambo est à peu près l'éléphant moyen.
- Stratégie du responsable : $Est_1 = 50 Poids_{Sambo}$
- Un statisticien propose d'utiliser Horvitz-Thompson.
  - Sambo a 99% de chances d'être sélectionné.
  - Chaque autre éléphant a $\frac{1}{100} \cdot \frac{1}{49}$ chances d'être sélectionné.

Éléphants de Basu

Si Sambo est sélectionné (ce qui arrive 99% du temps !) :

$$Est_2 = \frac{100}{99} Poids_{Sambo}$$

ce qui est étonnant. Et si c'est un autre éléphant :

$$Est_2 = 4900 Poids_{Autre}$$

ce qui donne un résultat déraisonnable ! Pourtant, l'estimateur est sans biais. . . ce n'est clairement pas l'estimateur optimal.

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

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.

Application

Un sondeur réalise une enquête ayant pour but de mesurer le patrimoine moyen des ménages d'Île-de-France. Les individus sont tirés parmi la liste des titulaires d'une carte Navigo (carte d'abonnement aux transports en commun d'Île-de-France) pour l'année 2016.
Le plan de sondage donne une probabilité de sélection plus forte aux individus habitant des communes aux revenus médians les plus élevés : Paris 16, Neuilly-sur-Seine, Paris 7, Versailles. L'estimateur utilisé est celui d'Horvitz-Thompson.
70 % des individus échantillonnés répondent au questionnaire, mais l'estimateur utilisé ne prend pas en compte la non-réponse.

Application

Que penser des affirmations suivantes ?

1. Donner une probabilité de sélection plus forte pour certains individus conduit à un échantillon non représentatif de la population ;
2. Le poids de sondage d'un individu sélectionné qui vit à Paris 13 est supérieur à celui d'un individu sélectionné habitant à Paris 16 ;
3. L'estimateur d'Horvitz-Thompson est biaisé, il vaudrait mieux utiliser l'estimateur naïf ;
4. La base de sondage présente un défaut de couverture, l'estimation est potentiellement biaisée ;
5. Ne pas prendre en compte la non-réponse n'a pas d'impact sur l'estimation.

Partie 3

Bases imparfaites et partage de poids

Base imparfaite

Retour sur la base de sondage parfaite :

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

Pas toujours le cas. Exemple : si l'on souhaite faire une enquête auprès des sans domicile ?

Enquête Sans Domicile

Enquête menée par l'INSEE en 2001 et en 2012. Comment atteindre la population ?

- on liste les services d'aide aux SDF (hébergement, repas) ;
- on en échantillonne certains ;
- on se rend dans le centre tiré ;
- on interroge une personne sur p.

Quel est le poids de sondage d'un individu ?

Partage des poids

Supposons qu'un individu A est allé dans un centre d'hébergement une seule nuit. Un individu B lui par contre y dort tous les soirs. A t-on autant de chance de sélectionner A ou B ?

Idée : prendre en compte le nombre de nuits passées dans des centres ; c'est le nombre de **liens**. On utilise ensuite ce qu'on appelle le **partage des poids** : on divise le poids par le nombre de liens.

Chapitre 5

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 :

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

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 ? 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	

Partie 3

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

Theorem

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

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

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

Précision

Démonstration.

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

Estimation de la précision

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 4

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 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`

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 :

$$n \approx \frac{1-p}{p(CV(p))^2} = \frac{0.99}{0.01 * 0.05^2} = 39600$$

Et on vérifie que $f \approx 0$.

Partie 5

Panel, domaine, ratio...

Évolution

On veut estimer l'évolution de la moyenne d'une variable Y entre deux dates 1 et 2 : $\Delta Y = \bar{Y}_1 - \bar{Y}_2$

Évolution

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 :

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

Panels

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 :

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

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

Estimation sur un domaine

- Un domaine d est un sous-ensemble de la population 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

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.

Petits domaines

On souhaite parfois obtenir des résultats sur des domaines d qui sont de très petite taille. Par exemple :

- Les habitants d'un département ou d'une ville spécifique ;
- Les spectateurs d'une chaîne télé spécialiste ;
- Les salons de coiffure.

Quand la taille de l'intersection entre l'échantillon et le domaine n_d est très petite, il est difficile d'estimer avec précision. La solution optimale est souvent d'augmenter cette taille en amont du sondage, mais ce n'est pas toujours possible.

Estimation sur petits domaines

Il existe néanmoins des méthodes pour obtenir des estimations avec une précision correcte sur ces petits domaines. L'idée est la suivante :

1. On détermine un lien entre des variables X et les Y observés sur la population entière, par exemple par régression linéaire ;
2. On calcule via le modèle sur les X (connus) du domaine d une valeur de Y_d.

Rien ne garantit que le modèle calculé à l'étape 2 soit valide pour le domaine sur lequel nous travaillons ; il convient donc de traiter les résultats obtenus avec précaution.

Estimation d'un ratio

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

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

Estimation d'un ratio

On peut utiliser l'estimateur :

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

mais il est biaisé !

$\Rightarrow$ Attention ! L'estimateur d'Horvitz-Thompson est sans biais quand on estime un total ou une moyenne, mais pas toujours...

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

Chapitre 6

Stratification

Partie 1

Principe de la stratification

Stratification

Dispersion de la variable d'intérêt et précision de ses estimateurs

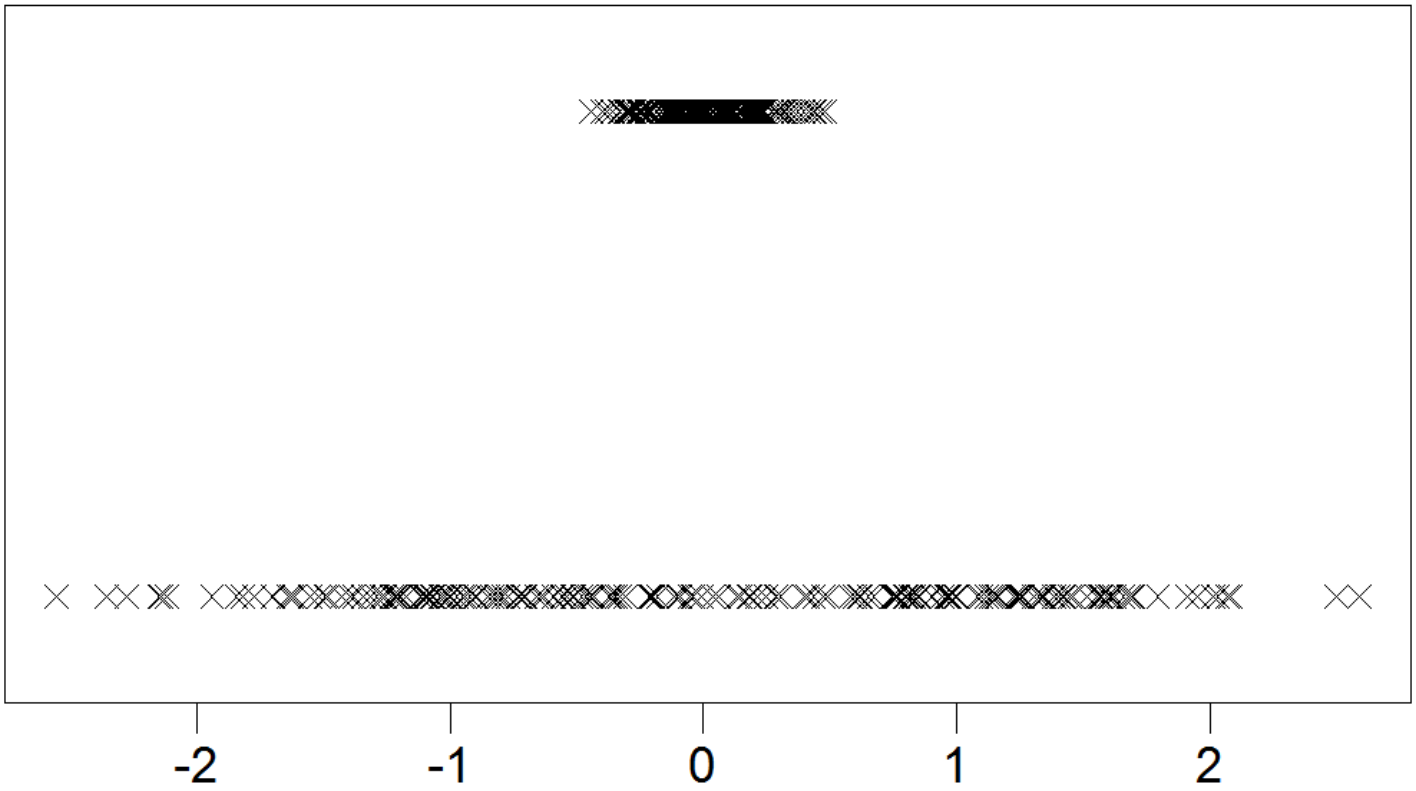
La variance des estimateurs de Horvitz-Thompson dépend directement de la dispersion de la variable d'intérêt Y.

Plus Y est dispersée, plus ses estimateurs sont imprécis (à plan de sondage et taille d'échantillon identiques).

Dans certains cas cependant, des variables de la base de sondage permettent de ventiler l'échantillon en groupes au sein desquels la variance de la variable d'intérêt est plus faible.

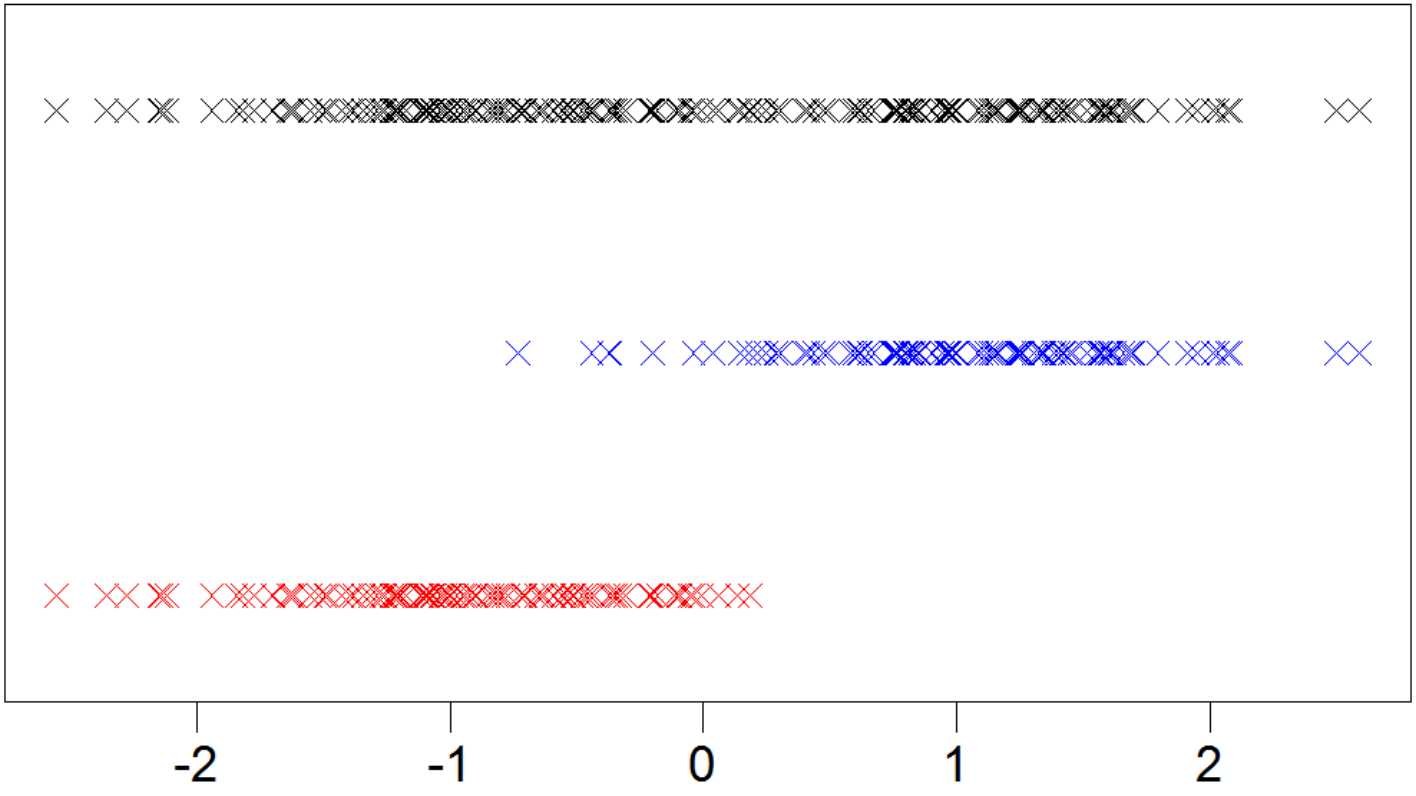
Stratification

Dispersion de la variable et précision de ses estimateurs



Stratification

Dispersion de la variable et précision de ses estimateurs



Stratification

Décomposition de la variance

En toute généralité, la variance de la variable Y peut en effet être décomposée selon H groupes, par exemple des modalités d'une variable X catégorielle :

$$S^2 = \underbrace{\sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_h^2}_{S^2_{intra} = \text{ Variance intra}} + \underbrace{\sum_{h=1}^{H} \frac{N_h}{N - 1} (\bar{Y}_h - \bar{Y})^2}_{S^2_{inter} = \text{ Variance inter}}$$

Il s'agit de la **formule de décomposition de la variance**.

Stratification

Exploiter les liens entre une variable de la base de sondage et la variable d'intérêt

La stratification consiste à :

- partitionner $\mathcal{U}$ en H groupes (les **strates**), notés $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_h, \ldots, \mathcal{U}_H$ telles que, à l'intérieur de chaque strate h, la dispersion S_h^2 de Y est faible ;
- à l'intérieur de chaque strate h, tirer des échantillons indépendants selon un plan p_h.

Justification Grâce à la faible dispersion dans chaque strate, les estimateurs devraient être plus précis, ce qui donnera une variance globale plus faible.

But secondaire Le plan stratifié va permettre de poser *a priori* une exigence de précision minimale par strate, en choisissant judicieusement les tailles d'échantillons dans chaque strate.

Stratification

Exemple : Enquête sur les loyers

Dans le cadre d'une enquête sur les loyers, on cherche à déterminer le meilleur moyen de tirer 40 logements parmi 1 000.

On dispose dans la base de sondage d'une information auxiliaire : on sait si chaque logement appartient au secteur libre (privé) ou au secteur social (HLM).

Il y a en tout 250 logements sociaux dans la base de sondage.

Stratification

Exemple : Enquête sur les loyers

4 plans de sondages sont mis en œuvre indépendamment :

1. sondage aléatoire simple (SAS) de 40 logements ;
2. SAS de 20 logements du secteur libre d'une part et SAS de 20 logements du secteur social d'autre part ;
3. SAS de 30 logements du secteur libre d'une part et SAS de 10 logements du secteur social d'autre part ;
4. SAS de 36 logements du secteur libre d'une part et SAS de 4 logements du secteur social d'autre part.

Stratification

Exemple : Enquête sur les loyers

On obtient les résultats suivants :

Plan	Secteur	n	$1/\pi_k$	Estimation	Variance estimée
1	L	31	25	12,71	0,39
	S	9	25		
2	L	20	37,5	12,69	0,28
	S	20	12,5		
3	L	30	25	12,51	0,22
	S	10	25		
4	L	36	20,8	12,78	0,18
	S	4	62,5		

Note : Les formules d'estimation et de variance utilisées pour les plans 2, 3 et 4 sont présentées plus loin dans ce cours.

Stratification

Exemple : Enquête sur les loyers

La stratification permet de réaliser des gains en termes de variance : les estimations semblent en général plus précises.

Deux éléments conditionnent l'efficacité de la stratification :

1. le lien entre variable d'intérêt et information auxiliaire : c'est parce que le loyer d'un logement est statistiquement lié à son secteur que l'on observe des gains de variance ;
2. l'allocation entre les strates : le gain est plus important si la plus grande part de l'échantillon est tirée dans le secteur libre, où les loyers sont plus variables.

Attention : Certaines allocations peuvent conduire à augmenter la variance par rapport au SAS !

Partie 2

Plan de sondage stratifié

Stratification

Méthode pour tirer un échantillon stratifié de taille fixe

1. Partitionner la population $\mathcal{U}$ en H strates. Chaque individu de la base de sondage doit être affecté à une (unique) strate.
2. Déterminer les allocations de l'échantillon dans chaque strate, sous la contrainte :
$$\sum_{h=1}^{H} n_h = n$$
n est supposé connu (les sondages de taille fixe permettent de fixer le budget nécessaire à l'enquête).
3. Dans chaque strate $\mathcal{U}_h$, tirer un échantillon s_h de taille n_h avec un plan p_h.

L'échantillon final s est l'union de tous les s_h :

$$s = s_1 \cup s_2 \cup \ldots \cup s_H$$

Stratification

Exemples

- Exemple précédent : loyers dans le secteur privé ou HLM ;
- Chiffre d'affaire des entreprises selon leur secteur d'activité ;
- Durée du trajet domicile-travail, selon la zone de résidence ;
- Temps d'audience de certaines radios, selon l'âge.

Stratification

Estimateur de Horvitz-Thompson

Les plans de sondage p_1, p_2,...,p_H menés au sein des H strates conduisent pour chaque unité échantillonnée k à une probabilité d'inclusion π_k.

On reste donc dans le cadre de Horvitz-Thompson :

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

sont des estimateurs sans biais respectivement du total et de la moyenne de la variable Y.

Leur variance peut être estimée à l'aide des formules de Horvitz-Thompson ou de Yates-Grundy (plan de sondage à taille fixe).

Stratification

Estimateur de Horvitz-Thompson

Il est néanmoins intéressant pour la suite de réécrire ces estimateurs pour faire apparaître la stratification.

On peut ainsi réécrire l'estimateur du total de Y :

$$\hat{T}_{str}(Y) = \sum_{h=1}^{H} \hat{T}_h(Y)$$

où $\hat{T}_h(Y)$ est l'estimateur du total de Y au sein de la strate h :

$$\hat{T}_h(Y) = \sum_{i \in s_h} \frac{y_i}{\pi_i}$$

Stratification

Estimateur de Horvitz-Thompson

De même, la variance de $\hat{T}_{str}(Y)$ peut être réécrite :

$$\begin{aligned} V(\hat{T}_{str}(Y)) &= V\left(\sum_{h=1}^{H} \hat{T}_h(Y)\right) \\ &= \sum_{h=1}^{H} V(\hat{T}_h(Y)) + 2 \sum_{\substack{h,h'=1 \\ h' \neq h}}^{H} Cov(\hat{T}_h(Y), \hat{T}_{h'}(Y)) \\ &= \sum_{h=1}^{H} V(\hat{T}_h(Y)) \end{aligned}$$

car les tirages réalisés au sein de chaque strate sont indépendants.

Stratification

Sondage aléatoire simple stratifié

Un sondage aléatoire simple stratifié est un plan de sondage stratifié avec au sein de chaque strate un sondage aléatoire simple.

Au sein de chaque strate de taille N_h connue, un échantillon de n_h unités est donc tiré par sondage aléatoire simple. On définit $f_h = \frac{n_h}{N_h}$ le taux de sondage de la strate h.

Particulièrement facile à mettre en œuvre, ce plan de sondage est très utilisé en pratique : c'est le cas par exemple de la quasi-totalité des enquêtes auprès des entreprises réalisées par l'Insee.

Stratification

Sondage aléatoire simple stratifié

Au sein de chaque strate h, le total et la moyenne de la variable Y sont estimés sans biais par :

$$\hat{T}_h(Y) = N_h \bar{y}_h \quad \text{et} \quad \hat{\bar{Y}}_h = \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{n_h} \sum_{h \in s_h} y_k$$

On estime la variance respective de ces deux estimateurs par :

$$\hat{V}(\hat{T}_h(Y)) = N_h^2 (1 - f_h) \frac{s_h^2}{n_h} \quad \text{et} \quad \hat{V}(\hat{\bar{Y}}_h) = (1 - f_h) \frac{s_h^2}{n_h}$$

Stratification

Sondage aléatoire simple stratifié

On estime sans biais le total et la moyenne de Y sur l'ensemble de l'échantillon par :

$$\hat{T}_{SAS-str}(Y) = \sum_{h=1}^{H} N_h \bar{y}_h \quad \text{et} \quad \hat{\bar{Y}}_{SAS-str} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h$$

Remarques :

1. Pour chaque observation de h, le poids est $\frac{N_h}{n_h}$.
2. Si $\frac{n_h}{n} \neq \frac{N_h}{N}$ alors $\hat{\bar{Y}}_{SAS-str} \neq \bar{y}$: l'estimateur en plan de sondage stratifié n'est pas toujours la moyenne empirique.

Stratification

Sondage aléatoire simple stratifié

La variance de ces estimateurs est estimée sans biais par :

$$\hat{V}(\hat{T}_{SAS-str}(Y)) = \sum_{h=1}^{H} N_h^2 (1-f_h) \frac{s_h^2}{n_h} \text{ et } \hat{V}(\hat{\bar{Y}}_{SAS-str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h) \frac{s_h^2}{n_h}$$

Remarques :

1. Pour pouvoir être calculé, cet estimateur nécessite au moins deux observations par strate.
2. La précision dépend seulement de la dispersion de Y **au sein de chaque strate** : plus les strates sont homogènes pour la variable Y, plus la stratification est efficace.

Stratification

Exemple : Tirage de 2 individus par strate

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Échantillon	1	2	3	4	5	6	7	8	9
Strate I	2 6	2 6	2 6	2 10	2 10	2 10	6 10	6 10	6 10
Moyenne	4	4	4	6	6	6	8	8	8
Strate II	8 10	8 12	10 12	8 10	8 12	10 12	8 10	8 12	10 12
Moyenne	9	10	11	9	10	11	9	10	11
Estimateur	6,5	7	7,5	7,5	8	8,5	8,5	9	9,5

Variance d'échantillonnage : 0,83 (SAS non-stratifié : 1,07)

Partie 3

Constitution des strates

Constitution des strates

Ces résultats donnent l'intuition des règles à suivre pour constituer les strates afin de maximiser l'efficacité de la stratification.

La variance de l'estimation de Y étant directement reliée à l'homogénéité de Y au sein des strates, une bonne stratification doit chercher à maximiser cette homogénéité.

Autrement dit, la stratification doit être choisie de telle sorte que les valeurs de Y soient les plus proches possibles les unes des autres à l'intérieur de chaque strate.

Stratification

Exemple : Tirage de 2 individus par strate

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification B	I	II	II	I	II	I

Échantillon	1	2	3	4	5	6	7	8	9
Strate I	2 10	2 10	2 10	2 12	2 12	2 12	10 12	10 12	10 12
Moyenne	6	6	6	7	7	7	11	11	11
Strate II	6 8	6 10	8 10	6 8	6 10	8 10	6 8	6 10	8 10
Moyenne	7	8	9	7	8	9	7	8	9
Estimateur	6,5	7	7,5	7	7,5	8	9	9,5	10

Variance d'échantillonnage : 1,33 (SAS non-stratifié : 1,07)

Stratification

Exemple : Tirage de 2 individus par strate

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification C	I	I	I	II	II	II

Échantillon	1	2	3	4	5	6	7	8	9
Strate I	2 6	2 6	2 6	2 8	2 8	2 8	6 8	6 8	6 8
Mean	4	4	4	5	5	5	7	7	7
Strate II	10 10	10 12	10 12	10 10	10 12	10 12	10 10	10 12	10 12
Mean	10	11	11	10	11	11	10	11	11
Estimateur	7	7,5	7,5	7,5	8	8	8,5	9	9

Variance d'échantillonnage : 0,44 (SAS non-stratifié : 1,07)

Stratification

Comment connaître S_h^2 ?

Y étant la variable que l'on veut estimer à l'aide de l'enquête, on ne connaît pas S_h^2.

Il s'agit donc d'utiliser l'information auxiliaire provenant de la base de sondage, sous l'hypothèse qu'elle est statistiquement liée à Y.

L'objectif est de constituer une partition de la population à partir des variables de la base de sondage de façon à ce que Y soit le moins dispersée possible dans les strates de tirage.

Remarque : Un choix de stratification peut être judicieux pour une variable Y mais pas pour d'autres.

Stratification

Quelques critères usuels pour le choix de stratification

Enquêtes ménages

- Région
- Type d'aire urbaine : urbaine, péri-urbaine, rurale
- Diplôme

Enquêtes entreprises

- Secteur d'activité
- Nombre de salariés
- Région

Partie 4

Choix des allocations

Choix des allocations

Une fois les strates définies, existe-t-il une façon optimale de répartir l'échantillon entre les strates ?

La réponse à cette question dépend de l'objectif que l'on donne à la stratification :

- **améliorer la précision par rapport à un SAS non-stratifié** pour l'ensemble des variables de l'enquête ;
- **atteindre la meilleure précision possible pour une variable**, quitte à perdre en précision sur d'autres.

D'autres objectifs sont également possibles : gagner en précision sur un ensemble de variables, intégrer des contraintes de précision sur certains domaines de diffusion, etc.

Allocation proportionnelle

L'allocation proportionnelle consiste à répartir l'échantillon entre les strates à proportion de leur taille dans la population :

$$\forall h \in \{1, \ldots, H\} \quad n_h = n \times \frac{N_h}{N}$$

Le **taux de sondage est identique** au sein de chaque strate :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

Autrement dit, toutes les unités ont le même poids $\frac{N}{n}$: il s'agit d'un **sondage à probabilités égales**.

Allocation proportionnelle

Quand un SAS est mené au sein de chaque strate avec allocation proportionnelle, l'estimateur de Horvitz-Thompson **coïncide avec celui du SAS non-stratifié** :

$$\hat{\bar{Y}}^{prop}_{SAS-str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \frac{1}{n_h} \sum_{h \in s_h} y_k = \frac{1}{n} \sum_{h \in s} y_k = \bar{y}$$

Mais **sa variance diffère** du fait de la stratification :

$$V(\hat{\bar{Y}}^{prop}_{SAS-str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (1-f_h) \frac{S_h^2}{n_h} = \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2 \simeq (1-f) \frac{S^2_{intra}}{n}$$

Allocation proportionnelle : Comparaison avec le SAS

On sait que : $V(\hat{\bar{Y}}_{SAS}) = (1-f)\frac{S^2}{n}$.

D'autre part $V(\hat{\bar{Y}}^{prop}_{SAS-str}) \simeq (1-f)\frac{S^2_{intra}}{n}$

Or par définition $S^2_{intra} \leq S^2$ donc :

$$V(\hat{\bar{Y}}^{prop}_{SAS-str}) \leq V(\hat{\bar{Y}}_{SAS})$$

Un SAS stratifié avec allocation proportionnelle conduit toujours à des estimateurs au moins aussi précis qu'un SAS non-stratifié de même taille.

Stratification

Allocation de Neyman : Meilleure précision possible à taille d'échantillon donnée

L'objectif de l'allocation de Neyman est de minimiser la variance de l'estimateur d'une variable Y à taille d'échantillon donnée.

On suppose dans un premier temps que **la variance de Y au sein de chaque strate h (notée S_h) est connue**.

L'**allocation de Neyman** est alors :

$$n_h = n \times \frac{N_h S_h}{\sum_{h'=1}^{H} N_{h'} S_{h'}}$$

On peut montrer que **cette allocation minimise la variance de l'estimateur du total de Y**.

Stratification

Allocation de Neyman : interprétation

Avec l'allocation de Neyman, le taux de sondage par strate est proportionnel à la variance de Y dans cette strate :

$$\frac{n_h}{N_h} \propto S_h$$

En d'autres termes, ce mécanisme d'allocation conduit à aller chercher l'information là où elle se trouve :

- Les strates homogènes (S_h petit) sont peu enquêtées ;
- Les strates dans lesquelles les unités ont des comportements variés (S_h grand) sont beaucoup enquêtées.

Stratification

Exemple : 3 individus dans la strate I, 1 individu dans la strate II

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Échantillon	1	2	3
Strate I	2 6 10	2 6 10	2 6 10
Moyenne	6	6	6
Strate II	8	10	12
Moyenne	8	10	12
Estimateur	7	8	9

Variance d'échantillonnage : 0,67 (SAS non-stratifié : 1,07)

Stratification

Exemple : 1 individu dans la strate I, 3 individus dans la strate II

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Échantillon	1	2	3
Strate I	2	6	10
Moyenne	2	6	10
Strate II	8 10 12	8 10 12	8 10 12
Moyenne	10	10	10
Estimateur	6	8	10

Variance d'échantillonnage : 2,67 (SAS non-stratifié : 1,07)

Stratification

Exemple : Allocation de Neyman

Population $\mathcal{U}$	A	B	C	D	E	F
Valeurs	2	6	8	10	10	12
Stratification A	I	I	II	I	II	II

Pour cet exemple, les données sont :

$$n = 4, \quad N_I = N_{II} = 3, \quad S_I = 4, \quad \text{and} \quad S_{II} = 2$$

Les allocations de Neyman sont donc :

$$\begin{cases} n_I = 4 \times \dfrac{3 \times 4}{3 \times 4 + 3 \times 2} = \dfrac{48}{18} = 2,7 \\ \\ n_{II} = 4 \times \dfrac{3 \times 2}{3 \times 4 + 3 \times 2} = \dfrac{24}{18} = 1,3 \end{cases}$$

La première allocation est donc très proche de optimum.

Stratification

Comment estimer les S_h ?

Dans tous ces calculs, on suppose les variances intra-strates de Y S_h connues, ce qui n'est pas le cas.

Afin de pouvoir utiliser l'allocation de Neyman, ces quantités doivent être estimées :

- dire d'expert ;
- information auxiliaire de la base de sondage ;
- enquêtes précédentes ;
- petite enquête préliminaire (si le coût n'est pas trop élevé en regard des objectifs).

Stratification

Allocation de Neyman et allocation proportionnelle

Pour une variable d'intérêt Y, l'allocation de Neyman est significativement meilleure que l'allocation proportionnelle dès lors que les S_h varient beaucoup d'une strate à l'autre.

Toutefois, l'allocation de Neyman est optimale **pour la seule variable** Y : elle peut être néfaste pour l'estimation d'une autre variable d'intérêt.

On peut également choisir un compromis entre ces deux allocations. L'optimum de l'allocation de Neyman est réputé « plat » : s'en éloigner un peu ne déteriore pas trop la précision.

Stratification

Conclusion

Comment choisir les allocations ?

- Il faut bien connaître l'objectif de l'enquête Y ;
- Il faut disposer d'information auxiliaire corrélée à Y ;
- Les strates qui sont très atypiques (par exemple, les très grandes entreprises) ont vocation à être dans l'exhaustif ;
- Les autres strates sont représentées selon leur influence sur Y :
  - Les unités de la strate sont-elles similaires ?
  - Est-ce que connaître leur valeur de Y apporte beaucoup d'information ?

Partie 5

Tirage systématique et stratification implicite

Stratification

Algorithme de tirage systématique

On cherche à effectuer un tirage de taille fixe n dans une population N. Chaque unité k de $\mathcal{U}$ dispose d'une probabilité d'inclusion simple π_k.

L'ordre des unités dans la base de sondage est fixé : on définit le cumul des probabilités d'inclusion $a_k = \sum_{k'=1}^{k} \pi_{k'}$.

L'algorithme de tirage systématique est alors le suivant :

1. On tire un réel η dans une loi uniforme sur $[0\,;1]$.
2. On sélectionne toutes les unités k vérifiant :

$$a_{k-1} \leq \eta + j - 1 < a_k$$

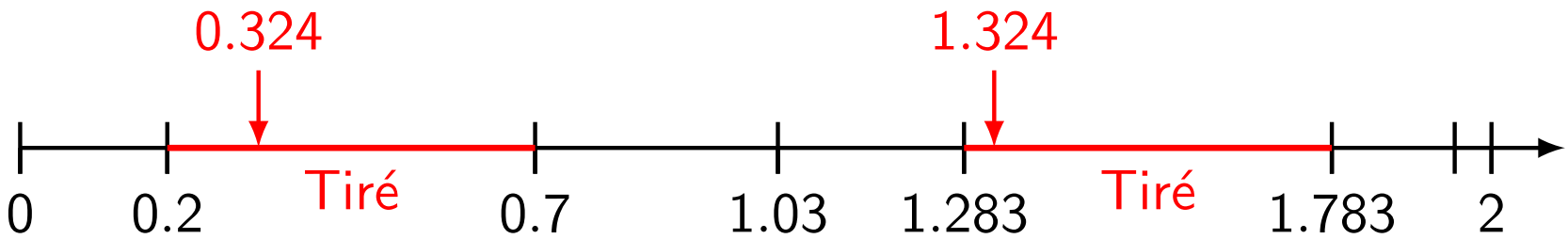
où j parcourt $1, \ldots, n$.

Stratification

Exemple de tirage systématique

$N = 7 \quad n = 2 \quad \sum_{k=1}^{7} \pi_k = 2 \quad \eta = 0,324$

k	1	2	3	4	5	6	7
π_k	0,2	0,5	0,33	0,25	0,5	0,166	0,05
a_k	0,2	0,7	1,03	1,283	1,783	1,950	2,00



L'échantillon tiré est $s = \{2, 5\}$.

Stratification

Propriétés du tirage systématique

1. Le sondage est à taille fixe et respecte les π_k.
2. C'est un algorithme efficace : un seul parcours de la base de sondage est nécessaire.
3. Selon l'ordre du fichier, des probabilités d'inclusion doubles π_{kl} peuvent être nulles : les estimateurs de variance de l'estimateur de Horvitz-Thompson sont alors biaisés.

Stratification

Stratification implicite

Quand la base de sondages est triée selon une ou plusieurs variables, mettre en œuvre un algorithme de tirage systématique sur l'ensemble de la base induit une **stratification implicite**.

En termes de précision, on obtient en effet un plan de sondage approximativement équivalent à un **sondage stratifié** :

1. dans les strates composées par les variables de tri ;
2. avec un SAS au sein de chaque strate ;
3. et une allocation proportionnelle.

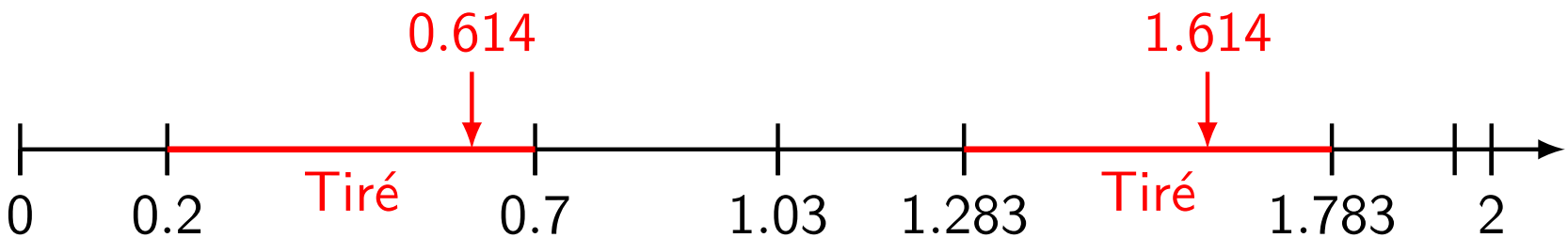
Un tirage systématique sur fichier trié ne peut donc qu'améliorer la précision de tous les estimateurs de l'enquête.

Stratification

Retour sur l'exemple de tirage systématique

$N_H = 3 \quad N_F = 4 \quad n = 2 \quad \sum_{k=1}^{7} \pi_k = 2 \quad \eta = 0,614$

k	1	2	3	4	5	6	7
Sexe	H	H	H	F	F	F	F
π_k	0,2	0,5	0,33	0,25	0,5	0,166	0,05
a_k	0,2	0,7	1,03	1,283	1,783	1,950	2,00



L'échantillon tiré est $s = \{2, 5\}$, stratifié entre hommes et femmes.

Stratification

Arbitrage entre précision des estimateurs et estimation sans biais de la précision

Intérêt : Quand la stratification devient trop fine, les estimateurs deviennent instables. On peut alors recourir à une stratification implicite par tirage systématique.

Arbitrage : Certaines probabilités d'inclusion double devenant nulle, les estimateurs de variance sont biaisés. On gagne certes en variance, mais on ne peut plus l'estimer sans biais.

En pratique, on préfère souvent une variance plus faible, même si cela signifie ne plus pouvoir l'estimer sans biais.

Partie 6

Tirage équilibré

Retour sur l'échantillon "représentatif"

Lorsque l'on réalise un sondage aléatoire simple, on ne connaît pas la structure de l'échantillon obtenu : ratio homme/femme, etc.

Pour pallier ce problème, on peut stratifier ou faire un tirage systématique.

Mais comment faire si l'on souhaite une structure précise pour :

- Sexe ;
- Âge ;
- Région...

Tirage équilibré

Cela demanderait trop de strates : à chaque fois qu'on rajoute un critère, il faut le croiser avec tous les autres, ce qui augmente très rapidement le nombre de strates.

Une autre méthode est possible : l'échantillonnage équilibré

- On choisit des variables X pour la structure : qualitatives ou quantitatives
- On sélectionne un échantillon s qui est correct sur ces variables X, c'est à dire que :

$$\hat{X}_{HT} = T(X)$$

- Si ce n'est pas possible, on cherche à être le plus proche possible.

En pratique

Comment utiliser le tirage équilibré en pratique ?

- Méthode réjective : tirer des échantillons jusqu'à obtenir un échantillon qui convienne. Problème : quelles sont les vraies probabilités de sélection ?
- Méthode du Cube, qui respecte les π_i. Méthode assez complexe, qui est implémentée :
  - en SAS, via la macro Cube : `https://www.insee.fr/fr/information/2021904`
  - en R, par exemple dans les packages *sampling* et *BalancedSampling*.