Introduction aux traitements statistiques
d'enquêtes sociologiques

Cours 7: Panorama des méthodes de sondages
Méthodes d'échantillonnage

Damien Cartron et Martin Chevalier

Année universitaire 2023-2024

Panorama des méthodes de sondage

Objectif Dresser un **panorama des méthodes de sondage** :

- connaître les principes fondamentaux des sondages ;
- être en mesure d'exploiter la documentation méthodologique d'une enquête ;
- mieux comprendre l'importance de la pondération dans une enquête.

Deux parties distinctes :

- **méthodes d'échantillonnage** : tout ce qui fait qu'on aboutit à l'échantillon à enquêter ;
- **méthodes de redressement** : tout ce qui est fait après l'enquête pour améliorer les estimations.

Format magistral, pas d'exercices mais beaucoup d'exemples → **simulations**

Objectifs de la séance

1. Connaître la notion de plan de sondage et d'estimateur de Horvitz-Thompson

2. Bien maîtriser le sondage aléatoire simple

3. Voir en détails une manière d'améliorer la précision d'un sondage : la stratification

Exemple introductif : enquête Patrimoine

Présentation et objectifs de l'enquête

L'enquête Patrimoine (renommée à partir de 2017-2018 « Histoire de vie et patrimoine ») est une **grande enquête de l'Insee**.

Elle a **plusieurs objectifs** :

- observer et décrire le patrimoine des ménages ;
- fournir des éléments explicatifs sur la formation et la transmission du patrimoine ;
- mesurer les inégalités de patrimoine entre les ménages.

Du point de vue de son échantillon, une caractéristique importante est que **les foyers fiscaux à « haut patrimoine »** (supérieur à 1,3 M d'euros) y sont surreprésentés.

Cadre de simulations

Comme précédemment, on procède par simulation pour **mieux percevoir les enjeux** associés à ce sondage :

- population : 38 M de foyers fiscaux dont environ 380 000 à « haut patrimoine » ;
- patrimoine moyen dans la population : environ 300 000 euros ;
- échantillon : 16 000 foyers fiscaux dont 3 000 à « haut patrimoine ».

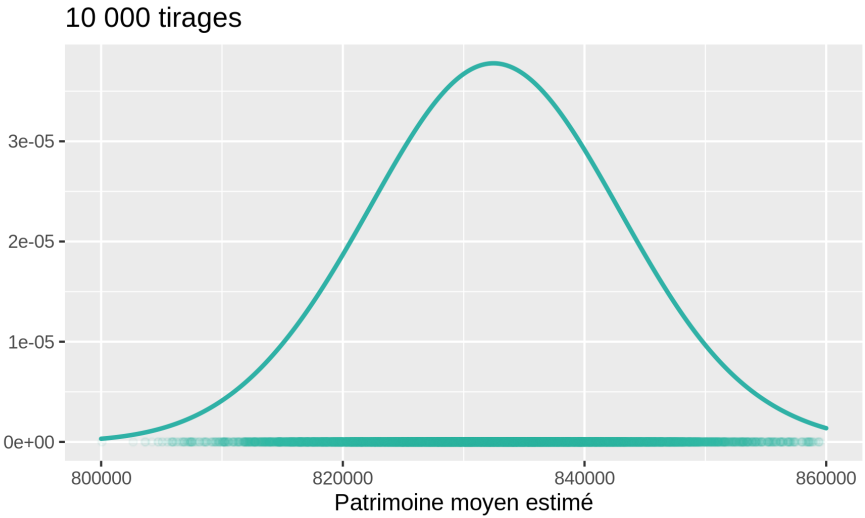
Tirage de l'échantillon :

1. tirage aléatoire de 13 000 foyers fiscaux à « bas patrimoine » ;
2. tirage aléatoire de 3 000 foyers fiscaux à « haut patrimoine » .

Objectif des simulations Analyser les propriétés de la **moyenne arithmétique simple** dans l'échantillon comme **estimateur du patrimoine moyen** dans la population.

Exemple introductif : enquête Patrimoine

Estimation du patrimoine moyen : moyenne arithmétique simple



Forte **surestimation du patrimoine moyen** : la moyenne arithmétique simple est ici un **estimateur biaisé** du patrimoine moyen.

Mais pourquoi devrait-il en être autrement ?

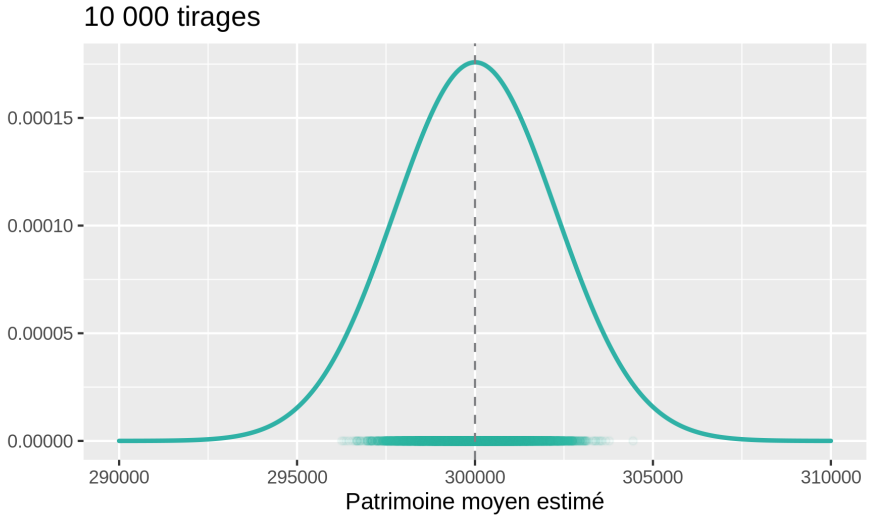
- Sur quels résultats théoriques s'appuie-t-on pour construire un estimateur de qualité dans le cadre d'une enquête par sondage ?
- Qu'est-ce qui fait qu'ici en particulier, la moyenne arithmétique simple ne présente pas les bonnes propriétés ?

Intuition

- la moyenne arithmétique simple accorde la même importance à tous les foyers fiscaux, alors que tous n'ont **pas la même probabilité d'être sélectionnés** (13 000 / 37,62 M contre 3 000 / 380 000) ;
- l'estimateur devrait en tenir compte et **pondérer les observations en fonction de cette probabilité**.

Exemple introductif : enquête Patrimoine

Estimation du patrimoine moyen : moyenne pondérée



Plan de sondage

Population et échantillon

On désigne par U la population dans laquelle est tiré l'échantillon.

Exemple $U = \{a, b, c\}$

On désigne par s un échantillon tiré dans la population et par $\mathcal{S}$ l'ensemble des échantillons possibles.

Exemple Pour $U = \{a, b, c\}$ l'ensemble des échantillons possibles est :

$$\mathcal{S} = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}, \emptyset\}$$

Plan de sondage

On appelle plan de sondage sur une population U une **distribution de probabilités sur l'ensemble des échantillons possibles $\mathcal{S}$**.

Exemple On définit le plan de sondage p_1 par :

$$\begin{array}{lll} p_1(\{a\}) = 0,1 & p_1(\{a,b\}) = 0,1 & p_1(\{a,b,c\}) = 0,2 \\ p_1(\{b\}) = 0,1 & p_1(\{a,c\}) = 0,1 & p_1(\emptyset) = 0,2 \\ p_1(\{c\}) = 0,1 & p_1(\{b,c\}) = 0,1 & \end{array}$$

Remarque Comme p_1 est une distribution de probabilités sur $\mathcal{S}$,

$$\sum_{s \in \mathcal{S}} p_1(s) = p_1(\{a\}) + p_1(\{b\}) + \ldots + p_1(\emptyset) = 1$$

Plan de sondage

Exemple On peut également définir le plan de sondage p_2 par :

$$\begin{array}{lll} p_2(\{a\}) = 0 & p_2(\{a,b\}) = 0,5 & p_2(\{a,b,c\}) = 0 \\ p_2(\{b\}) = 0 & p_2(\{a,c\}) = 0,25 & p_2(\emptyset) = 0 \\ p_2(\{c\}) = 0 & p_2(\{b,c\}) = 0,25 & \end{array}$$

Remarque Contrairement à p_1, p_2 est de **taille fixe** : tous les échantillons avec une probabilité non-nulle sont de même taille (2).

Statistique d'intérêt et estimateur

On note θ la statistique d'intérêt, qui dépend d'une ou plusieurs variables d'intérêt.

Exemple Statistiques descriptives univariées pour une variable Y :

- total : $\theta(Y) = T(Y)$
- moyenne : $\theta(Y) = \bar{Y}$

On note $\hat{\theta}$ un **estimateur** de θ sous le plan de sondage. Il s'agit d'une **variable aléatoire** dont les valeurs dépendent des échantillons tirés.

Pour un échantillon s donné, on note $\hat{\theta}_s$ l'**estimation** correspondante de θ.

Exemple Soit une variable Y dont les valeurs dans la population U sont :

	a	b	c
Y	20	10	3

La **statistique d'intérêt est la moyenne** de Y : $\theta(Y) = \bar{Y}$, qui vaut 11 dans la population.

Choix de l'estimateur On décide d'estimer la moyenne dans la population par la **moyenne arithmétique simple dans l'échantillon** : $\hat{\bar{Y}} = \bar{y}$

$$\hat{\bar{Y}}_{\{a\}} = 20 \quad \hat{\bar{Y}}_{\{a,b\}} = 15 \quad \hat{\bar{Y}}_{\emptyset} = 0$$
$$\hat{\bar{Y}}_{\{b\}} = 10 \quad \hat{\bar{Y}}_{\{a,c\}} = 11,5 \quad \hat{\bar{Y}}_{\{a,b,c\}} = 11$$
$$\hat{\bar{Y}}_{\{c\}} = 3 \quad \hat{\bar{Y}}_{\{b,c\}} = 6,5$$

Espérance d'un estimateur sous un plan de sondage

Pour un plan de sondage p et un estimateur $\theta(Y)$, on définit l'espérance de $\theta(Y)$ sous le plan de sondage p par :

$$\mathbb{E}_p(\hat{\theta}(Y)) = \sum_{s \in \mathcal{S}} p(s)\hat{\theta}_s(Y)$$

Interprétation Il s'agit de la **moyenne des valeurs prises** par l'estimateur de la statistique d'intérêt dans les différents échantillons **pondérée par la probabilité d'apparition** de chaque échantillon.

Remarque L'espérance d'sun estimateur peut également être estimée **par simulation**.

Espérance d'un estimateur sous un plan de sondage

Exemple Rappel du plan de sondage p_2

$$\begin{array}{lll} p_2(\{a\}) = 0 & p_2(\{a,b\}) = 0,5 & p_2(\{a,b,c\}) = 0 \\ p_2(\{b\}) = 0 & p_2(\{a,c\}) = 0,25 & p_2(\emptyset) = 0 \\ p_2(\{c\}) = 0 & p_2(\{b,c\}) = 0,25 & \end{array}$$

Espérance de l'estimateur par la moyenne arithmétique simple $\hat{\hat{Y}}$ sous le plan de sondage p_2 :

$$\begin{aligned} \mathbb{E}_{p_2}(\hat{\hat{Y}}) &= \sum_{s \in \mathcal{S}} p_2(s) \hat{\hat{Y}}_s \\ &= p_2(\{a\}) \hat{\hat{Y}}_{\{a\}} + \ldots + p_2(\{a,b,c\}) \hat{\hat{Y}}_{\{a,b,c\}} \\ &= 0 \times 20 + \ldots + 0,5 \times 15 + 0,25 \times 11,5 + 0,25 \times 6,5 + \ldots + 0 \times 11 \\ &= 12 \end{aligned}$$

Erreur de sondage

Le **biais d'un estimateur** sous le plan de sondage p est défini par l'écart entre l'espérance de l'estimateur et la valeur de la statistique dans la population :

$$B_p(\hat{\theta}(Y)) = \mathbb{E}(\hat{\theta}(Y)) - \theta(Y)$$

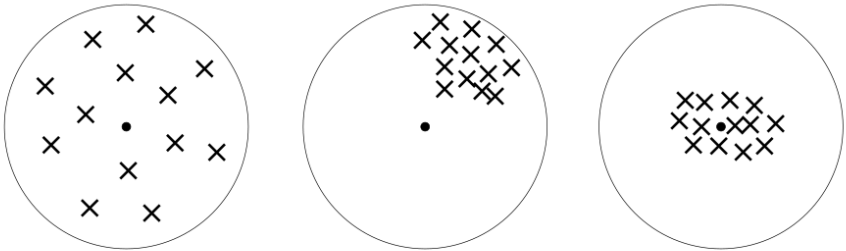
Remarque Si $B(\hat{\theta}(Y)) = 0$ alors $\hat{\theta}(Y)$ est un **estimateur sans biais**.

La **variance d'un estimateur** sous le plan de sondage p est définie comme la somme des écarts à l'espérance pour l'ensemble des échantillons possibles :

$$V_p(\hat{\theta}(Y)) = \sum_{s \in \mathcal{S}} p(s) \left[\hat{\theta}_s(Y) - \mathbb{E}(\hat{\theta}(Y))\right]^2$$

Remarque La variance d'un estimateur peut également être estimée **par simulation**.

Erreur de sondage



- Cas 1 : Pas de biais, forte variance
- Cas 2 : Biais, faible variance
- Cas 3 : Pas de biais, faible variance

Exemple La moyenne arithmétique simple n'est pas un estimateur sans biais de $\bar{Y}$ sous le plan de sondage p_2. En effet, avec $\hat{\bar{Y}} = \bar{y}$:

$$B(\hat{\bar{Y}}) = \mathbb{E}_{p_2}(\hat{\bar{Y}}) - \bar{Y} = 12 - 11 = 1 \neq 0$$

Moralité On se trouve ici exactement dans le même cas que dans l'exemple introductif : **la moyenne arithmétique simple est un estimateur biaisé de la moyenne dans la population**.

Le cadre élaboré par Horvitz et Thompson permet précisément d'**obtenir des estimateurs sans biais** sous n'importe quel plan de sondage (ou presque).

La notion de **probabilités d'inclusion** y joue un rôle fondamental.

Estimateur d'Horvitz-Thompson

Probabilités d'inclusion simple et double

Probabilités d'inclusion simple Probabilité qu'une unité i de la population se trouve dans l'échantillon :

$$\pi_i = \sum_{s \in \mathcal{S}^{(i)}} p(s)$$

où $\mathcal{S}^{(i)}$ est l'ensemble des échantillons contenant l'unité i

Probabilités d'inclusion double Probabilité que deux unités i et j de la population se trouvent conjointement dans l'échantillon :

$$\pi_{ij} = \sum_{s \in \mathcal{S}^{(ij)}} p(s)$$

où $\mathcal{S}^{(ij)}$ est l'ensemble des échantillons contenant conjointement les unités i et j

Probabilités d'inclusion simple et double

Exemples Rappel du plan de sondage p_1 :

$$\begin{array}{lll} p_1(\{a\}) = 0,1 & p_1(\{a,b\}) = 0,1 & p_1(\{a,b,c\}) = 0,2 \\ p_1(\{b\}) = 0,1 & p_1(\{a,c\}) = 0,1 & p_1(\emptyset) = 0,2 \\ p_1(\{c\}) = 0,1 & p_1(\{b,c\}) = 0,1 & \end{array}$$

Probabilités d'inclusion simple : $\pi_a = 0,50 \quad \pi_b = 0,50 \quad \pi_c = 0,50$
Probabilités d'inclusion double : $\pi_{ab} = 0,30 \quad \pi_{ac} = 0,30 \quad \pi_{bc} = 0,30$

Rappel du plan de sondage p_2 :

$$\begin{array}{lll} p_2(\{a\}) = 0 & p_2(\{a,b\}) = 0,5 & p_2(\emptyset) = 0 \\ p_2(\{b\}) = 0 & p_2(\{a,c\}) = 0,25 & p_2(\{a,b,c\}) = 0 \\ p_2(\{c\}) = 0 & p_2(\{b,c\}) = 0,25 & \end{array}$$

Probabilités d'inclusion simple : $\pi_a = 0,75 \quad \pi_b = 0,75 \quad \pi_c = 0,50$
Probabilités d'inclusion double : $\pi_{ab} = 0,50 \quad \pi_{ac} = 0,25 \quad \pi_{bc} = 0,25$

Estimateur d'Horvitz-Thompson : Définition et biais

Pour une variable d'intérêt Y dont on cherche à estimer le total $T(Y)$, on appelle **estimateur d'Horvitz-Thompson** la quantité :

$$\hat{T}_s^{HT}(Y) = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Propriété fondamentale Si le plan de sondage est tel que toutes les unités de la population ont une **probabilité non-nulle** d'apparaître dans l'échantillon, alors l'estimateur d'Horvitz-Thompson est **sans biais**.

Autrement dit : si $\forall i \in U \ \pi_i > 0$ alors $\mathbb{E}(\hat{T}^{HT}(Y)) = T(Y)$

Remarque L'estimateur Horvitz-Thompson de la moyenne est

$$\hat{\bar{Y}}_s^{HT} = \frac{1}{N} \hat{T}_s^{HT}(Y) = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$$

Estimateur d'Horvitz-Thompson : Définition et biais

Exemple Valeurs de $\hat{\hat{Y}}_s^{HT}$

$$\hat{\hat{Y}}_{\{a,b\}}^{HT} = \frac{1}{N} \sum_{i \in \{a,b\}} \frac{y_i}{\pi_i} = \frac{1}{3} \left(\frac{20}{0,75} + \frac{10}{0,75} \right) = 13,33$$

$$\hat{\hat{Y}}_{\{a,c\}}^{HT} = \frac{1}{N} \sum_{i \in \{a,c\}} \frac{y_i}{\pi_i} = \frac{1}{3} \left(\frac{20}{0,75} + \frac{3}{0,50} \right) = 10,89$$

$$\hat{\hat{Y}}_{\{b,c\}}^{HT} = \frac{1}{N} \sum_{i \in \{b,c\}} \frac{y_i}{\pi_i} = \frac{1}{3} \left(\frac{10}{0,75} + \frac{6}{0,50} \right) = 6,44$$

Estimateur d'Horvitz-Thompson : Définition et biais

Exemple Espérance de $\hat{\hat{Y}}^{HT}$ sous le plan de sondage p_2

On constate que sous le plan de sondage p_2 $\forall i \in U \quad \pi_i > 0$: on en déduit que $\hat{\hat{Y}}^{HT}$ est sans biais sous ce plan de sondage, c'est-à-dire que

$$\mathbb{E}_{p_2}(\hat{\hat{Y}}^{HT}) = \bar{Y} = 11$$

On le vérifie néanmoins :

$$\begin{aligned}
\mathbb{E}_{p_2}(\hat{\hat{Y}}^{HT}) &= \sum_{s \in \mathcal{S}} p_2(s) \hat{\hat{Y}}_s^{HT} \\
&= p_2(\{a\}) \hat{\hat{Y}}_{\{a\}}^{HT} + \ldots + p_2(\{a,b,c\}) \hat{\hat{Y}}_{\{a,b,c\}}^{HT} \\
&= 0,50 \times \hat{\hat{Y}}_{\{a,b\}}^{HT} + 0,25 \times \hat{\hat{Y}}_{\{a,c\}}^{HT} + 0,25 \times \hat{\hat{Y}}_{\{b,c\}}^{HT} \\
&= 0,50 \times 13,33 + 0,25 \times 10,89 + 0,25 \times 6,44 \\
&= 11 = \bar{Y}
\end{aligned}$$

Estimateur d'Horvitz-Thompson : Variance

Il est possible d'estimer la variance de l'estimateur d'Horvitz-Thompson par la quantité :

$$\hat{V}_s\left(\hat{T}^{HT}(Y)\right) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

Propriété fondamentale Si toutes les paires d'unités ont une **probabilité d'être tirées conjointement non-nulle**, alors l'estimateur de la variance de l'estimateur d'Horvitz-Thompson est **sans biais**.

Autrement dit, si toutes les probabilités d'inclusion double sont non-nulles, alors il est possible d'**estimer raisonnablement correctement la variance d'un estimateur d'Horvitz-Thompson à partir d'un seul échantillon**.

Estimateur d'Horvitz-Thompson : Synthèse

L'estimateur d'Horvitz-Thompson présente un côté un peu « magique » :

- il est non seulement **sans biais**... : on ne commet pas d'erreur systématique (à la hausse ou à la baisse) en l'employant ;
- ...mais en plus sa variance est **calculable à partir d'un seul échantillon**.

En d'autres termes : **en moyenne on ne se trompe pas et en plus on peut avoir une idée d'à quel point on se trompe**.

Remarque Cela suppose quand même que les probabilités d'inclusion simple et double soient toutes non-nulles, ce qui n'est pas toujours le cas.

Estimateur d'Horvitz-Thompson : Poids de sondage

Les estimateurs d'Horvitz-Thompson peut être vu comme des **statistiques pondérées par l'inverse des probabilités d'inclusion** :

$$\hat{T}_s^{HT}(Y) = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i y_i \text{ où } w_i = \frac{1}{\pi_i}$$

En vertu du cadre d'Horvitz-Thompson, ces « **poids de sondage** » permettent de **construire un estimateur sans biais de n'importe quelle variable d'intérêt**.

Il est possible d'**interpréter** le poids de chaque unité de l'échantillon comme le nombre d'unité de la population qu'elle représente...

...mais c'est bien la théorie des sondages qui permet de **démontrer** l'utilité de cette quantité pour construire des estimateurs.

Sondage aléatoire simple

Sondage aléatoire simple

Définition

Un sondage aléatoire simple est un plan de sondage de taille fixe dans lequel **tous les échantillons de la taille souhaitée** ont la **même probabilité d'être sélectionnés**.

Autrement dit, on définit le sondage aléatoire simple de taille n par :

$$\forall s \in \mathcal{S} \qquad p_{SAS}(s) = \begin{cases} \alpha \text{ si } s \in \mathcal{S}_n \\ 0 \text{ sinon} \end{cases}$$

où $\mathcal{S}_n$ désigne l'ensemble des échantillons de taille n et α est une constante.

Remarque En notant $|\mathcal{S}_n|$ le nombre d'échantillons de taille n, on peut montrer que

$$\alpha = \frac{1}{|\mathcal{S}_n|}$$

Exemple d'algorithme

Il existe une **multitude d'algorithmes** pour tirer un échantillon en respectant exactement la définition d'un sondage aléatoire simple.

L'un d'entre eux est **particulièrement intuitif** :

1. Tirer dans une loi uniforme une valeur pour chaque unité de la population.
2. Classer la population selon cette valeur.
3. Sélectionner les n premières unités.

Cet algorithme suppose cependant de tirer une valeur aléatoire pour **chaque unité de la population** puis de la **trier selon ces valeurs**, ce qui peut être très long si la population est très grande.

Probabilités d'inclusion

On définit un sondage aléatoire simple de n unités parmi une population U de N unités.

En repartant de la définition du sondage aléatoire simple et en utilisant des techniques de dénombrement, on peut montrer que :

- $\forall i \in U \quad \pi_i = \frac{n}{N} \Rightarrow w_i = \frac{N}{n}$ (poids de sondage égaux)
- $\forall i, j \in U \quad i \neq j \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}$

Démonstration Support de formation plus approfondi (en anglais), p. 25.

À retenir Une des *propriétés* du sondage aléatoire simple est que **toutes les unités ont la même probabilité d'inclusion simple**, égale au taux de sondage $f = \frac{n}{N}$.

Estimateur d'Horvitz-Thompson

Dans le cas d'un sondage aléatoire simple, l'estimateur d'Horvitz-Thompson de la moyenne **se simplifie** :

$$\hat{\bar{Y}}^{HT} = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i} = \frac{1}{N}\sum_{i\in s}\frac{y_i}{n/N} = \frac{1}{N}\frac{N}{n}\sum_{i\in s} y_i = \frac{1}{n}\sum_{i\in s} y_i = \bar{y}$$

où $\bar{y}$ est la moyenne empirique dans l'échantillon.

À retenir Dans un sondage aléatoire simple, l'estimateur d'Horvitz-Thompson de la moyenne **coïncide avec la moyenne empirique dans l'échantillon**.

Nota bene Il s'agit d'une ***propriété*** du sondage aléatoire simple :

- ce n'est pas vrai pour tous les plans de sondage ;
- cela peut être néanmoins vrai pour d'autres plans de sondage.

Erreur de l'estimateur d'Horvitz-Thompson

Dans le cas d'un sondage aléatoire simple, toutes les probabilités d'inclusion simple sont **strictement positives** (égales à n/N) : l'estimateur d'Horvitz-Thompson est donc **sans biais**.

Par ailleurs l'estimateur de la **variance de l'estimateur d'Horvitz-Thompson** de la moyenne **se simplifie** :

$$\hat{V}(\hat{\bar{Y}}^{HT}) = (1-f)\frac{s^2}{n}$$

où $f = \dfrac{n}{N}$ est le taux de sondage, s^2 la variance empirique dans l'échantillon et n la taille de l'échantillon.

Démonstration Support de formation plus approfondi (en anglais), p. 29.

Intervalle de confiance

Définition Intervalle auquel la vraie valeur d'une statistique (ici la moyenne) a une probabilité fixée d'appartenir.

Formule

$$IC_{1-\alpha\ \%}(\hat{\bar{Y}}^{HT}) = \left[\hat{\bar{Y}}^{HT} - q_{1-\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\hat{V}(\hat{\bar{Y}}^{HT})};\ \hat{\bar{Y}}^{HT} + q_{1-\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\hat{V}(\hat{\bar{Y}}^{HT})}\right]$$

$$= \hat{\bar{Y}}^{HT} \pm q_{1-\alpha/2}^{\mathcal{N}(0,1)}\sqrt{\hat{V}(\hat{\bar{Y}}^{HT})}$$

où $q_{1-\alpha/2}^{\mathcal{N}(0,1)}$ est le quantile à $1-\alpha/2$ d'une loi normale centrée réduite.

Exemple Intervalle de confiance à 95 %

- $\alpha = 0,05$
- $q_{1-\alpha/2}^{\mathcal{N}(0,1)} = q_{0,975}^{\mathcal{N}(0,1)} = 1,96$
- $IC_{95\ \%}(\hat{\bar{Y}}^{HT}) = \hat{\bar{Y}}^{HT} \pm 1,96\sqrt{\hat{V}(\hat{\bar{Y}}^{HT})}$

Cas d'une proportion

Quand la variable d'intérêt Y est dichotomique, sa moyenne $\bar{Y}$ est une proportion notée P.

Dans le cas d'un sondage aléatoire simple :

- l'estimateur d'Horvitz-Thompson de cette proportion est la proportion estimée dans l'échantillon p (cas particulier de moyenne) : $\hat{p}^{HT} = p$;
- l'estimateur de la variance de l'estimateur d'Horvitz-Thompson de cette proportion peut se réécrire :

$$\hat{V}(\hat{p}^{HT}) = (1 - f)\frac{p(1-p)}{n-1}$$

Cette formulation, dans laquelle on néglige souvent le taux de sondage, permet de **facilement faire le lien** entre **taille d'échantillon** et **précision** d'un sondage visant à estimer une proportion.

Limites du sondage aléatoire simple

Le sondage aléatoire simple est **le plus simple des plans de sondage** :

- il peut être mis en œuvre dès lors qu'on dispose d'une liste des unités de la population (rien d'autre n'est requis) ;
- les estimateurs issus de ce plan de sondage coïncident avec des statistiques classiques ;
- sa variance est facile à calculer.

Il conduit cependant souvent en pratique à des **estimateurs trop imprécis** : la taille de l'échantillon est en général **contrainte**, aussi **tout est déterminé par la variance empirique** de la variable d'intérêt.

Question Comment augmenter la précision d'un sondage sans augmenter la taille de l'échantillon ?

En règle générale, la base de sondage n'est pas qu'une liste des unités de la population : elle contient d'**autres variables potentiellement liées au sujet de l'enquête**.

Exemple L'enquête Patrimoine est tirée dans les fichiers fiscaux, qui comportent notamment des informations sommaires sur le patrimoine des personnes.

Pour augmenter la précision d'un sondage sans augmenter la taille de l'échantillon, on peut **exploiter cette information auxiliaire pour améliorer le plan de sondage**.

Plusieurs méthodes permettent d'atteindre cet objectif : la **stratification** est la plus courante d'entre elles.

Principe de la stratification

Principe de la stratification

Intuition

Plutôt que de tirer l'échantillon en une seule fois dans l'ensemble de la population :

1. **Découper la population** en sous-ensembles appelés « **strates** » ;
2. **Tirer l'échantillon indépendamment** dans chaque strate.

Exemple L'enquête Patrimoine comporte deux strates : la strate des foyers fiscaux avec un « haut patrimoine » (supérieur à 1,3 M d'euros) et celle des foyers fiscaux avec un « bas patrimoine » (inférieur à 1,3 M d'euros).

Pour pouvoir découper la population en strates pertinentes, il est nécessaire de disposer d'**information auxiliaire dans la base de sondage**.

Exemple Le plan de sondage de l'enquête Patrimoine exploite le fait que l'assujettissement des foyers fiscaux à l'Impôt de solidarité sur la fortune (ISF) est connu : c'est de là que provient le seuil de 1,3 M d'euros.

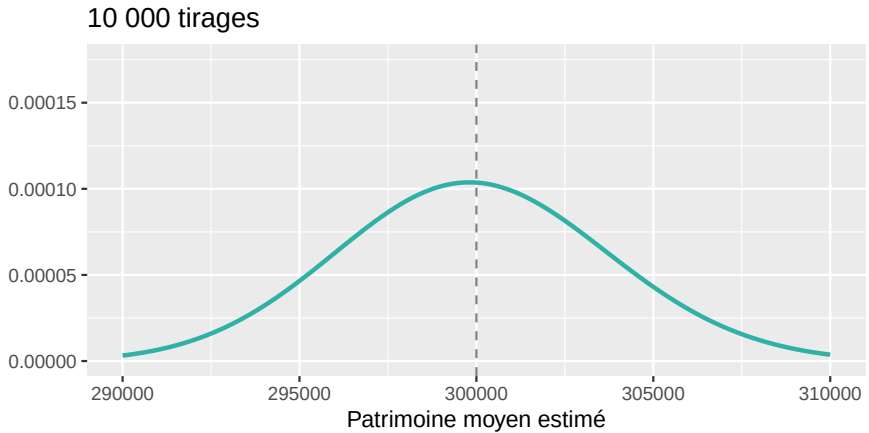
Pour mesurer l'impact de la stratification sur la précision d'un sondage, on procède **par simulation**.

On simule tout d'abord le tirage de l'enquête Patrimoine **sans stratification** : sondage aléatoire simple de 16 000 foyers fiscaux parmi 38 M.

On observe la **distribution de l'estimateur d'Horvitz-Thompson du patrimoine moyen** ainsi que le **nombre de foyers fiscaux à « haut patrimoine »** dans l'échantillon.

Principe de la stratification

Impact de la stratification sur la précision : simulations



Remarques

- distribution relativement dispersée : l'intervalle [293 500 ; 306 500] contient 95 % des valeurs ;
- le nombre de foyers fiscaux à « haut patrimoine » dans l'échantillon varie fortement : [136 ; 185] contient 95 % des valeurs.

Ces deux éléments sont liés : le fait que le nombre de foyers fiscaux à « haut patrimoine » varie fortement entraîne une plus forte variabilité de l'estimateur du patrimoine moyen.

Principe de la stratification Tirer séparément un échantillon de foyers fiscaux à « bas » et à « haut » patrimoine, de façon à maîtriser la part de foyers fiscaux à « haut patrimoine » dans l'échantillon.

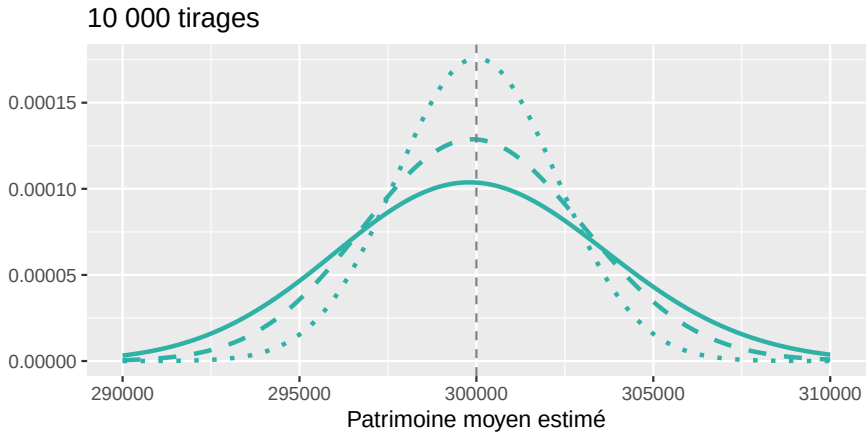
On simule alors deux scénarios de tirage stratifié :

1. Stratifié avec 160 foyers fiscaux à « haut patrimoine » : le nombre de foyers fiscaux à haut-patrimoine dans l'échantillon est déterminé **proportionnellement** à ce qu'il représente dans la population (1 %) ;

2. Stratifié avec 3 000 foyers fiscaux à « haut patrimoine » : forte **surreprésentation** des foyers fiscaux à « haut patrimoine ».

Principe de la stratification

Impact de la stratification sur la précision : simulations



Le scénario « **Stratifié (160 HP)** » garantit que les foyers fiscaux à « haut patrimoine » sont représentés dans l'échantillon à hauteur de leur part dans la population.

À ce titre, ce plan de sondage est de nature à rendre l'estimateur plus précis que le sondage aléatoire simple, en **évitant qu'il y ait trop ou trop peu** de foyers fiscaux à « haut patrimoine » dans l'échantillon.

Le scénario « **Stratifié (3 000 HP)** » surreprésente largement les foyers fiscaux à « haut patrimoine ».

Ce faisant, il est en mesure de **capter la très forte variabilité de leur patrimoine** et il améliore encore la précision de l'estimateur.

Plan de sondage stratifié

Plan de sondage stratifié

Définition

Soit U une population de taille N, partitionnée en H strates.

Le plan de sondage p correspondant aux opérations suivantes est qualifié de **stratifié** :

1. Au sein de chaque strate h, on mène **indépendamment** un sondage avec le plan de sondage p_h pour obtenir un échantillon s_h.
2. L'échantillon complet s est obtenu par union des H échantillons s_h.

Remarque En toute généralité, les plans de sondage p_h appliqués au sein de chacune des strates n'ont **pas besoin d'être identiques**, ils peuvent être de nature très différente.

Le caractère déterminant d'un plan de sondage stratifié est que les opérations de tirage sont menées **totalement indépendamment d'une strate à l'autre**.

Sondage aléatoire simple stratifié

On qualifie de sondage aléatoire simple stratifié un plan de sondage stratifié où un sondage aléatoire simple est mené au sein de chacune des strates.

L'estimateur d'Horvitz-Thompson se réécrit alors :

- du total : $\hat{T}^{HT}_{SAS-str}(Y) = \sum_{h \in H} \hat{T}^{HT}_{h}(Y) = \sum_{h \in H} N_h \bar{y}_h$
- de la moyenne : $\hat{\bar{Y}}^{HT}_{SAS-str} = \frac{1}{N} \hat{T}^{HT}(Y) = \frac{1}{N} \sum_{h \in H} N_h \bar{y}_h \underbrace{\neq}_{\text{en général}} \bar{y}$

avec N_h la taille de la strate H dans la population et $\bar{y}_h$ la moyenne empirique de Y dans la strate h dans l'échantillon.

À retenir Dans un sondage aléatoire simple stratifié, l'estimateur d'Horvitz-Thompson de la moyenne **peut ne pas coïncider avec la moyenne empirique de l'échantillon**.

Variance d'un sondage aléatoire simple stratifié

En raison de l'**indépendance des tirages entre les strates**, la variance de l'estimateur de la moyenne dans un sondage aléatoire simple stratifié se réécrit :

$$\hat{V}(\hat{T}^{HT}_{SAS-str}(Y)) = \sum_{h=1}^{H} \hat{V}(\hat{T}^{HT}_{SAS,h}(Y)) = \sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_h^2}{n_h}$$

avec n_h la taille de l'échantillon dans la strate h, $f_h = \frac{n_h}{N_h}$ le taux de sondage dans la strate h et s_h la variance empirique de la variable Y dans la strate h de l'échantillon.

Rappel Dans un sondage aléatoire simple non-stratifié :

$$\hat{V}(\hat{T}^{HT}_{SAS}(Y)) = N^2(1-f)\frac{s^2}{n}$$

Variance d'un sondage aléatoire simple stratifié

À retenir $\hat{V}(\hat{\bar{Y}}^{HT}_{SAS}) \propto \frac{s^2}{n}$ alors que $\hat{V}(\hat{\bar{Y}}^{HT}_{SAS-str}) \propto \sum_h \frac{s_h^2}{n_h}$

Interprétation

- dans un sondage aléatoire simple **non-stratifié**, la variance de l'estimateur dépend de la **variance empirique** de la variable d'intérêt ;
- dans un sondage aléatoire simple **stratifié**, la variance de l'estimateur dépend de la **variance empirique intra-strate** de la variable d'intérêt.

Conséquence Si la variable de stratification « **explique bien** » la variable d'intérêt, celle-ci **varie peu au sein des strates** et l'estimateur issu du sondage aléatoire simple stratifié a une **faible variance**.

Une « bonne » variable de stratification est donc une **variable de la base de sondage** qui ait **un lien statistique fort avec la ou les variables d'intérêt**.

Allocation de l'échantillon entre les strates

Dans un sondage stratifié, la répartition de l'échantillon entre les strates est un **paramètre du sondage**, *i.e.* quelque chose que la ou le responsable du sondage peut faire varier à sa guise.

Pour un ensemble de strates donné, une bonne allocation est susceptible d'**améliorer la précision du sondage**.

Il existe **deux mécanismes d'allocation** importants à connaître :

- **allocation proportionnelle** : allouer l'échantillon entre les strates proportionnellement à la taille des strates dans la population ;
- **allocation de Neyman** : allouer l'échantillon entre les strates en surreprésentant les strates où la variance de la variable d'intérêt est la plus grande.

Allocation proportionnelle

Définition $\forall h \quad n_h = n \times \frac{N_h}{N}$

Propriétés

- les probabilités d'inclusion se réécrivent : $\forall i \quad \pi_i = \frac{n_h}{N_h} = \frac{n}{N}$
- l'estimateur de la moyenne se réécrit :

$$\hat{\bar{Y}}^{HT}_{SAS-str-prop} = \frac{1}{N}\sum_{h\in H} N_h \bar{y}_h = \sum_{h\in H}\frac{n_h}{n}\bar{y}_h = \sum_{h\in H}\frac{n_h}{n}\frac{1}{n_h}\sum_{i\in s_h} y_i = \frac{1}{n}\sum_{i\in s} y_i \underbrace{=}_{\text{ici seulement}} \bar{y}$$

- la variance est toujours plus faible que dans un sondage aléatoire simple non-stratifié de même taille :

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

Allocation proportionnelle

À retenir Dans un SAS stratifié avec allocation proportionnelle, beaucoup de choses **se passent comme si on était dans un SAS non-stratifié** :

- les probabilités d'inclusion simples sont toutes égales à n/N et donc les poids de sondage tous égaux à N/n ;
- l'estimateur Horvitz-Thompson de la moyenne coïncide avec la moyenne empirique dans l'échantillon.

Pourtant il y a une différence essentielle avec le SAS non-stratifié : le fait que la taille de chaque strate dans l'échantillon est maîtrisée **stabilise** les estimateurs, qui sont **toujours plus précis que dans un SAS non-stratifié**.

Exemple Le scénario « Stratifié (160 HP) » correspond à une allocation proportionnelle.

Allocation de Neyman

Définition $n_h = n \times \frac{N_h S_h}{\sum_{h'=1}^{H} N_{h'} S_{h'}}$ avec S_h la variance attendue pour la variable d'intérêt Y dans la strate h

Propriétés

- Par rapport au sondage aléatoire simple, certaines unités sont **surreprésentées**, d'autre **sous-représentées** : les probabilités d'inclusion simple (et donc les poids de sondage) ne sont **pas égales**.
- L'estimateur d'Horvitz-Thompson de la moyenne **ne coïncide pas** avec la moyenne empirique dans l'échantillon.
- Ce sondage aléatoire simple fournit une **variance minimale** pour la variable Y.

Allocation de Neyman

À retenir Un SAS stratifié avec allocation de Neyman vise à **estimer au mieux une variable en particulier**, celle utilisée pour calibrer les allocations.

Exemple Le scénario « Stratifié (3 000 HP) » correspond à une allocation de Neyman.

À noter qu'il arrive que l'allocation de Neyman conduise à des **estimateurs moins précis qu'avec un sondage aléatoire simple non-stratifié**.

Cela survient quand la variable d'intérêt est **peu corrélée (voire anticorrélée)** avec la variable utilisée pour calibrer les allocations.

Pour utiliser cette allocation, il est essentiel de **bien identifier la variable d'intérêt** et de s'assurer de la **qualité de l'information auxiliaire**.

À retenir

La stratification permet d'**améliorer la précision d'un sondage** à taille donnée en exploitant l'**information auxiliaire** de la base de sondage.

La variance des estimateurs dépend de la **variabilité de la variable d'intérêt au sein des strates** : on souhaite que chaque strate soit **la plus homogène possible**.

La **manière de répartir l'échantillon entre les strates** affecte les propriétés des estimateurs :

- **allocation proportionnelle** : pondération et estimateurs identiques à ceux du SAS non-stratifié, mais précision toujours meilleure ;
- **allocation de Neyman** : pondération et estimateurs différents de ceux du SAS non-stratifié, précision meilleure pour une variable mais pas nécessairement pour toutes.

Conclusion

L'objectif des méthodes d'échantillonnage est de fournir un **cadre théorique** pour construire des estimateurs par sondage avec de bonnes propriétés.

L'estimateur d'Horvitz-Thompson permet ainsi d'obtenir des **estimateurs sans biais** pour n'importe quelle variable et dont la **variance est calculable**.

La notion de **probabilité d'inclusion** est cruciale et constitue le **fondement théorique des poids de sondage**.

Les méthodes de stratification permettent d'**améliorer la précision d'un sondage aléatoire simple** en exploitant l'**information auxiliaire** de la base de sondage.

Quelques perspectives

Méthodes pour améliorer la qualité d'une enquête après la collecte :

- correction de la non-réponse ;
- redressements.

→ **prochaine séance**

Autres méthodes pour améliorer la précision d'un plan de sondage :

- sondage à probabilités inégales ;
- sondage équilibré.

Méthodes pour réduire le coût d'une enquête (quand déplacement d'un enquêteur) :

- sondage en grappes ;
- sondage à plusieurs degrés ;
- échantillon-maître.