Introduction à la théorie des sondages - Cours 3

Gaspar Massiot
gaspar.massiot@ined.fr

2024-2025

Sources d'erreur en sondage

- Erreur d'échantillonnage
  - Taille de l'échantillon
  - Plan d'échantillonnage
  - Estimateur choisi
  - Variabilité du paramètre
- Erreur de mesure
  - Enquêté.e
  - Questionnaire
  - Saisie
- Non-réponse
  - Totale : entrée vide
  - Partielle : seule une partie du questionnaire est remplie.

Sondage aléatoire simple de taille fixe

Un sondage aléatoire simple (SAS) de **taille fixe** est un plan de sondage particulier défini par $p(s) = \frac{1}{\binom{N}{n}}$ dès que s est de taille n.

Exemple : $\mathcal{U} = \{1, 2, 3\}$. On définit le plan de sondage p_{SAS} par :

$$p_{SAS}(\{1\}) = p_{SAS}(\{2\}) = p_{SAS}(\{3\}) = 0$$
$$p_{SAS}(\{1, 2\}) = p_{SAS}(\{1, 3\}) = p_{SAS}(\{2, 3\}) = 1/3$$
$$p_{SAS}(\{1, 2, 3\}) = 0.$$

Ainsi :

$$\forall k \in \mathcal{U}, \pi_k = \mathbb{P}(k \in s) = \frac{n}{N} = f$$
$$\forall k \neq l \in \mathcal{U}, \pi_{k,l} = \mathbb{P}(k \wedge l \in s) = \frac{n(n-1)}{N(N-1)}$$

Sondage aléatoire simple de taille fixe

Dans ce contexte, les estimateurs de Horvitz-Thompson se réécrivent :

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

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} y_k = \bar{y}$$

La vraie variance est données par la formule de Yates-Grundy :

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

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

Sondage aléatoire simple de taille fixe

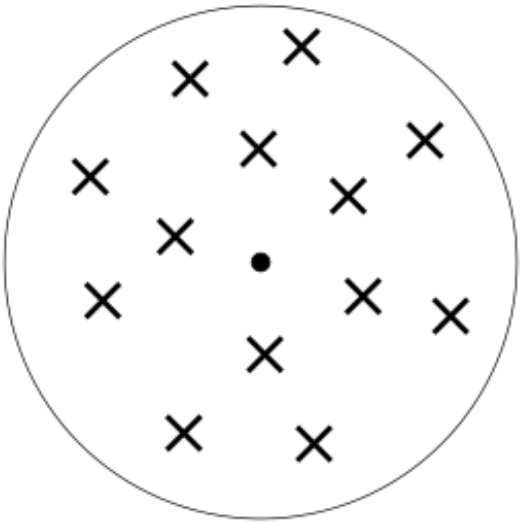
Pour choisir la taille de l'échantillon, on peut construire un estimateur p (par proxy ou en s'appuyant sur des études précédentes) de la proportion d'individus portant la caractéristique d'intérêt dans la population étudiée.
On peut montrer que, si $f \approx 0$,

$$n \approx \frac{4p(1-p)}{L^2},$$

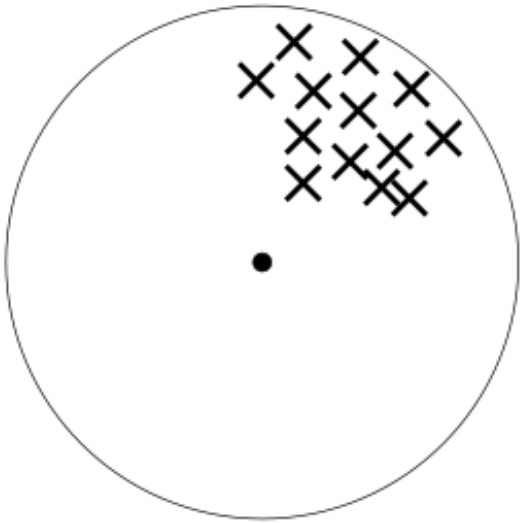
est une bonne approximation de la taille d'échantillon à tirer pour atteindre une "précision absolue" L donnée.
On peut également fixer la valeur du coefficient de variation CV_0 recherchée. Dans ce cas :

$$n \approx \frac{1-p}{p(CV_0)^2}$$

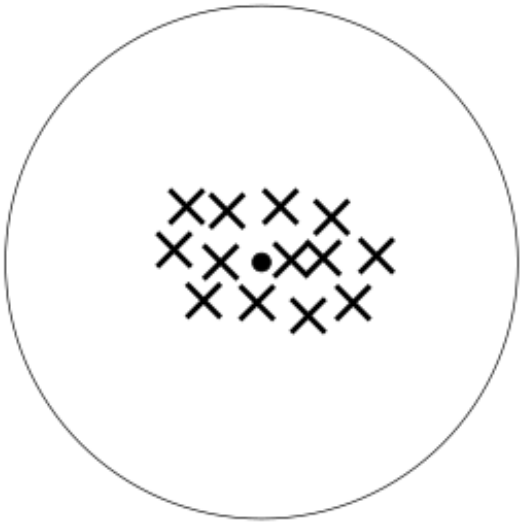
Schéma



Cas 1



Cas 2



Cas 3

Formule de Yates-Grundy :

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

Stratification

Chapitre 1

Principe de la stratification

Principe de la 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.

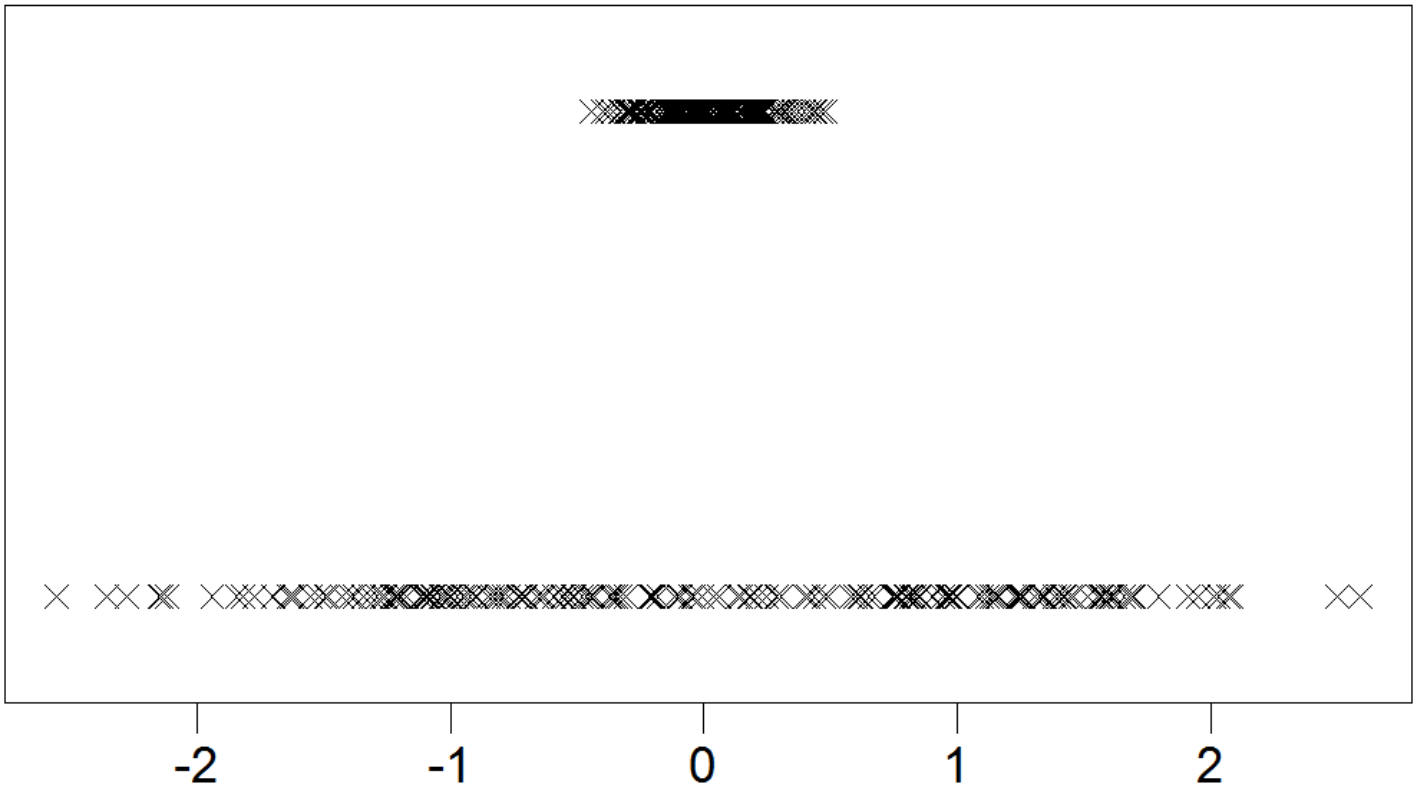
$$\mathrm{Var}(\hat{\bar{Y}}) = (1 - f)\frac{S^2}{n}$$

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.

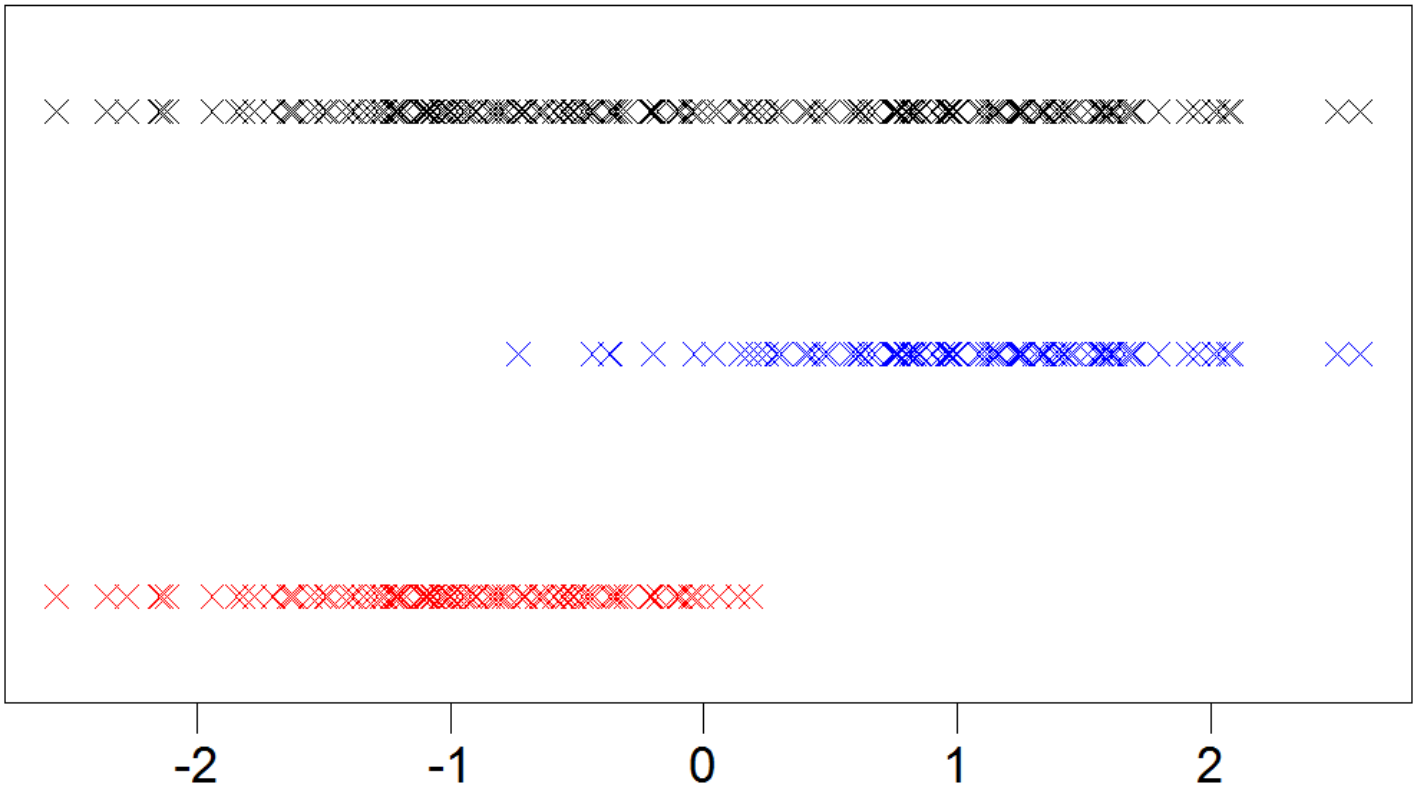
Principe de la stratification

Dispersion de la variable et précision de ses estimateurs



Principe de la stratification

Dispersion de la variable et précision de ses estimateurs



Principe de la 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**.

Principe de la 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.

Principe de la stratification

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

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.

Principe de la 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.

Principe de la 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.

Principe de la 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.

Principe de la 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 !

Chapitre 2

Plan de sondage stratifié

Plan de sondage stratifié

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 :

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

Plan de sondage stratifié

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.

Plan de sondage stratifié

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).

Plan de sondage stratifié

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 :

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

Plan de sondage stratifié

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**.

Plan de sondage stratifié

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.

Plan de sondage stratifié

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 :

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

Plan de sondage stratifié

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.

Plan de sondage stratifié

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.

Plan de sondage stratifié

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)

Chapitre 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**, ou autrement dit à **minimiser l'hétérogénéité** des strates.

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 ; i.e. les plus **homogènes** possibles.

Constitution des strates

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)

Constitution des strates

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)

Constitution des strates

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.

Constitution des strates

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

Chapitre 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**.

Partie 1

Allocation proportionnelle

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.

Partie 2

Allocation de Neyman

Choix des allocations

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**.

Choix des allocations

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.

Choix des allocations

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)

Choix des allocations

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)

Choix des allocations

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 l'optimum.

Choix des allocations

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).

Choix des allocations

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.

Choix des allocations

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 ?

Chapitre 5

Pour aller plus loin

Partie 1

Tirage systématique et stratification implicite

Pour aller plus loin

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 :

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

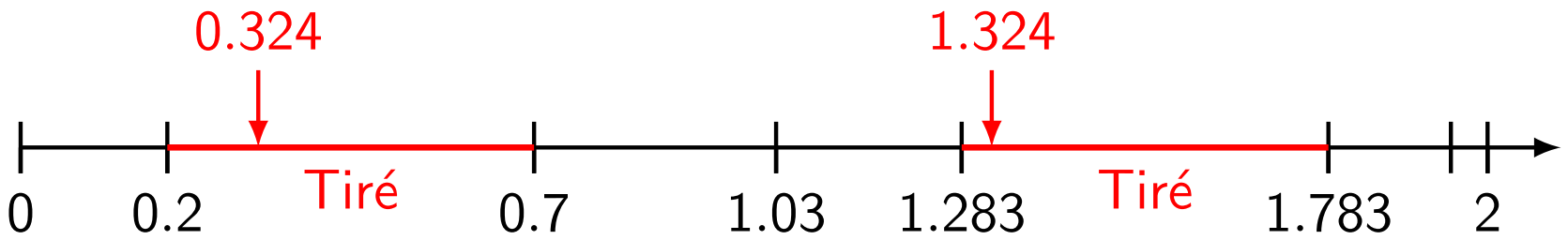
où j parcourt $1, \ldots, n$.

Pour aller plus loin

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 $s = \{2, 5\}$.

Pour aller plus loin

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.

Pour aller plus loin

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.

Pour aller plus loin

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.

Pour aller plus loin

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 2

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*.