Introduction à la théorie des sondages
Cours 3 : Stratification

Martin Chevalier
martin.chevalier@insee.fr

INSEE, Département des méthodes statistiques

28 mars 2017

Sommaire

Chapitre 1

Rappel des épisodes précédents

Principe du sondage

Objectif Construire un estimateur $\Phi(Y)$ d'une variable Y à partir d'un échantillon s de taille n tiré dans une population $\mathcal{U}$ de taille N.

Plan de sondage On définit un plan de sondage p comme une loi de probabilité sur l'ensemble des échantillons possibles $\mathcal{S}$.

Exemple : $\mathcal{U} = \{1, 2, 3\}$. On définit le plan de sondage p_1 par :

$$p_1(\{1\}) = p_1(\{2\}) = p_1(\{3\}) = 0$$
$$p_1(\{1,2\}) = 0,5 \quad p_1(\{1,3\}) = 0,2 \quad p_1(\{2,3\}) = 0,2$$
$$p_1(\{1,2,3\}) = 0,1$$

Remarque p_1 n'est pas un plan de sondage de taille fixe.

Probabilités d'inclusion

Le plan de sondage permet de déterminer des probabilités d'inclusion pour chaque unité de la population.

Probabilité d'inclusion simple $\pi_k = \sum_{s \in \mathcal{S}} \delta_k p(s)$

Probabilité d'inclusion double $\pi_{k,l} = \sum_{s \in \mathcal{S}} \delta_k \delta_l p(s)$

avec $\delta_k(s) = \mathbf{1}(k \in s)$

Exemple : Avec le plan de sondage p_1

$$\pi_1 = 0,8 \quad \pi_2 = 0,8 \quad \pi_3 = 0,5$$
$$\pi_{1,2} = 0,6 \quad \pi_{1,3} = 0,3 \quad \pi_{2,3} = 0,3$$

Estimateur de Horvitz-Thompson

Pour un plan de sondage p, un échantillon s et une variable d'intérêt Y, on définit les estimateurs de Hotvitz-Thompson du total et de la moyenne de Y par :

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

Si $\forall k \in \mathcal{U} \quad \pi_k > 0$ alors l'estimateur de Horvitz-Thompson est **sans biais**.

On dispose d'**estimateurs de la variance de ces estimateurs** :

$$\hat{V}(\hat{T}_{HT}(Y)) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s, l \neq k} \frac{\pi_k \pi_l - \pi_{k,l}}{\pi_{k,l}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Sondage aléatoire simple de taille fixe

Un sondage aléatoire simple (SAS) de taille fixe est un plan de sondage particulier défini par : $\forall s \in \mathcal{S} \quad p(s) = \frac{1}{\binom{N}{n}}$. Ainsi :

$$f = \frac{n}{N} \qquad \forall k \quad \pi_k = \frac{n}{N} \quad et \quad \forall k, l \quad \pi_{k,l} = \frac{n(n-1)}{N(N-1)}$$

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

$$\hat{\bar{Y}}_{SAS} = \bar{y} \quad \text{et} \quad \hat{T}_{SAS}(Y) = N\bar{y}$$

Leur variance est estimée par :

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

Cas particulier d'une proportion

Une proportion P est un **cas particulier de moyenne** (où la variable Y est une indicatrice), aussi :

$$\hat{P}_{SAS} = p$$

où p est la proportion calculée dans l'échantillon s.

Cas particulier d'une proportion

Une proportion P est un **cas particulier de moyenne** (où la variable Y est une indicatrice), aussi :

$$\hat{P}_{SAS} = p$$

où p est la proportion calculée dans l'échantillon s.

Par ailleurs, on peut montrer que la **variance empirique** dans l'échantillon associée à une proportion p est :

$$s_p^2 = \frac{n}{n-1} p(1-p) \quad \text{et ainsi} \quad \hat{V}(\hat{P}_{SAS}) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}$$

Cas particulier d'une proportion

Une proportion P est un **cas particulier de moyenne** (où la variable Y est une indicatrice), aussi :

$$\hat{P}_{SAS} = p$$

où p est la proportion calculée dans l'échantillon s.

Par ailleurs, on peut montrer que la **variance empirique** dans l'échantillon associée à une proportion p est :

$$s_p^2 = \frac{n}{n-1} p(1-p) \quad \text{et ainsi} \quad \hat{V}(\hat{P}_{SAS}) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}$$

Sous l'hypothèse que le taux de sondage $f = \frac{n}{N}$ est négligeable, la taille d'échantillon n^* **pour obtenir le coefficient de variation** CV_0 est alors :

$$n^* \approx \frac{1 - p_0}{p_0 \times CV_0^2}$$

Chapitre 2

Principe de la stratification

Principe de la stratification

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

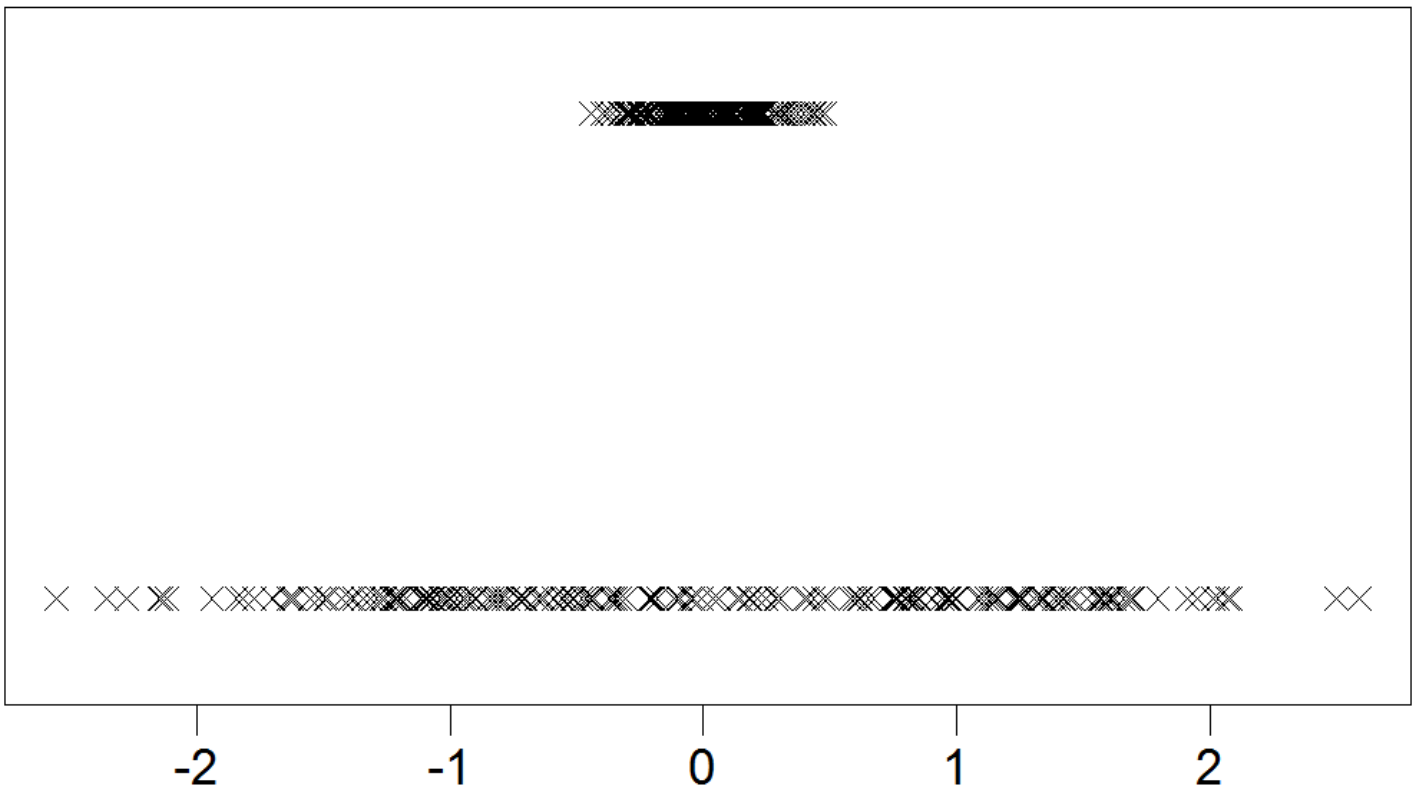
La variance des estimateurs de Horvitz-Thompson dépend directement de la dispersion de la variable d'intérêt Y.

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.

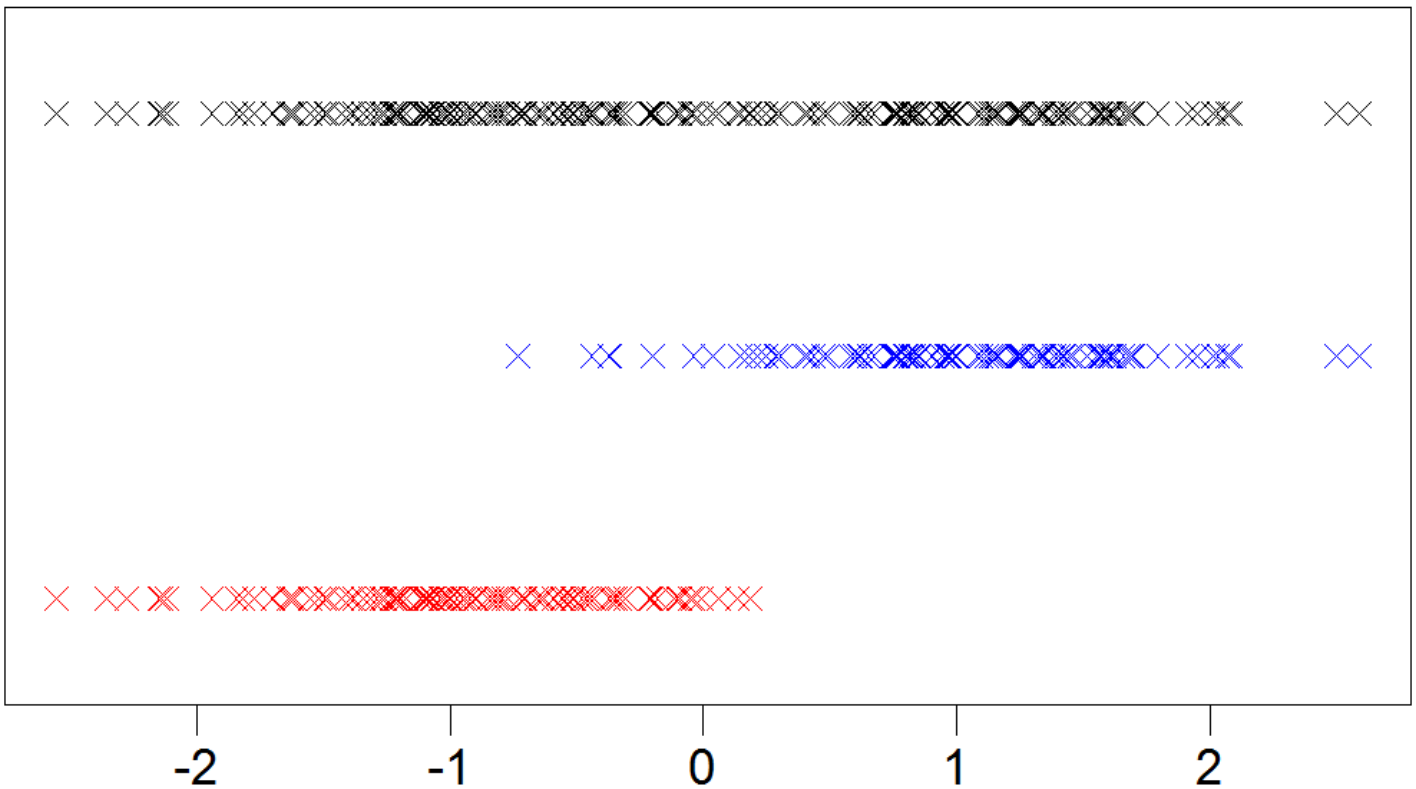
Principe de la stratification

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.

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 3

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 ;
- Nombre de nuits passées pour un séjour touristique, selon l'origine du vol ;
- Temps d'audience de certaines radios, selon l'âge.

Partie 1

Estimateur de Horvitz-Thompson

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.

Partie 2

Sondage aléatoire simple stratifié

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)

Plan de sondage stratifié

Exemple : Tirage de 2 individus par strate, estimateur de variance

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
Variance	8	8	8	32	32	32	8	8	8
Strate II	8 10	8 12	10 12	8 10	8 12	10 12	8 10	8 12	10 12
Variance	2	8	2	2	8	2	2	8	2
Estimateur	0,4	0,7	0,4	1,4	1,7	1,4	0,4	0,7	0,4

Moyenne de l'estimateur de variance : 0,83 (non biaisé)

Variance de l'estimateur de variance : 0,236 (SAS non-stratifié : 0,251)

Chapitre 4

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.

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, estimation de variance

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
Variance	32	32	32	50	50	50	2	2	2
Strate II	6 8	6 10	8 10	6 8	6 10	8 10	6 8	6 10	8 10
Variance	2	8	2	2	8	2	2	8	2
Estimateur	1,4	1,7	1,4	2,2	2,4	2,2	0,2	0,4	0,2

Moyenne de l'estimateur de variance : 1,33 (non biaisé)

Variance de l'estimateur de variance : 0,944 (SAS non-stratifié : 0,251)

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

Constitution des strates

Exemple : Bornes optimales pour les strates de nombre de salariés

La variable « nombre de salariés » est en général disponible dans la base de sondage.

Afin de l'utiliser comme variable de stratification, il faut déterminer des bornes de définition pour les strates.

Les limites usuelles à l'INSEE sont : 10-19, 20-49, 50-99, 100-249, 250-499, 500 et plus.

Ce choix de stratification a été optimisé par une procédure adaptée.

Constitution des strates

Exemple : Bornes optimales pour les strates de nombre de salariés

Il existe un certain nombre de méthodes pour déterminer les limites $b_0, b_1, \ldots b_H$ optimales pour une variable y.

Une des plus simples est la **méthode géométrique**. L'idée consiste à remarquer qu'à l'optimum, les coefficients de variation devraient être égaux au sein de chaque strate.

$$\forall h \in \{1, \ldots H\}, \quad \frac{s_h}{\bar{y}_h} = constant$$

Comme les CV ne peuvent pas toujours être calculés, on suppose que les y suivent une loi uniforme au sein de chaque strate h. Alors :

$$\bar{y}_h \approx \frac{b_h + b_{h+1}}{2} \quad \text{and} \quad s_h \approx \frac{b_h - b_{h-1}}{\sqrt{12}}$$

Constitution des strates

Exemple : Bornes optimales pour les strates de nombre de salariés

$\forall h < H$:

$$\frac{s_h}{\bar{y}_h} = \frac{s_{h+1}}{\bar{y}_{h+1}} \quad \Rightarrow \quad \frac{b_h - b_{h-1}}{b_h + b_{h-1}} = \frac{b_{h+1} - b_h}{b_{h+1} + b_h}$$

$$\Rightarrow \quad b_h^2 = b_{h+1} b_{h-1}$$

Avec $b_0 > 0$, ceci implique :

$$\forall h \in \{1, \dots H\}, \quad b_h = b_0 \left(\frac{b_H}{b_0} \right)^{\frac{h}{H}}$$

où b_0 et b_H sont respectivement les valeurs min et max de y.

Constitution des strates

Exemple : Bornes optimales pour les strates de nombre de salariés

Application à des données INSEE

Les limites obtenues par cette méthode (exemple précédent) sont : 10-24, 25-59, 60-143, 144-348, 349-846, 847 et plus.

À précision donnée du nombre de salariés, il est possible de *comparer* le nombre d'individus requis pour un SAS, un plan stratifié avec des limites usuelles et un plan stratifié avec limites déterminées par la méthode géométrique.

CV	**SAS**	**Stratification usuelle**	**Méthode géométrique**
1 %	57 922	666	611
5 %	3 276	156	151
10 %	925	138	136

Constitution des strates

Exemple : Bornes optimales pour les strates de nombre de salariés

Dans cette situation théorique, la variable à estimer est connue sur toute la population (via la base de sondage), voici pourquoi les gains associés à la stratification sont importants.

En général, la varible d'intérêt est corrélée avec la variable de stratification. Le choix des limites peut influencer l'efficacité de la stratification.

Le package **R** `stratification` implémente différentes méthodes de calculer les regroupements/limites de stratification.

Chapitre 5

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.

Partie 1

Allocation proportionnelle

Choix des allocations

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**.

Choix des allocations

Allocation proportionnelle : Sondage aléatoire simple au sein de chaque strate

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 :

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

Choix des allocations

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 et allocation optimale

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

Optimalité de l'allocation de Neyman.

Le problème est le suivant : $\begin{cases} \min_{n_h} \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ \text{sous la contrainte} \quad n = \sum_h n_h \end{cases}$

En ne gardant que les termes en n_h, on écrit le Lagrangien :

$$L(n_1, n_2, \ldots, n_H, \lambda) = \sum_h \frac{N_h^2 S_h^2}{n_h} - \lambda \left(\sum_h n_h - n \right)$$

Choix des allocations

Allocation de Neyman

Optimalité de l'allocation de Neyman.

Les conditions du premier ordre s'écrivent :

$$\frac{\delta L}{\delta n_h} = 0 \Rightarrow \frac{N_h^2 S_h^2}{n_h^2} = \lambda \Rightarrow n_h = \frac{N_h S_h}{\sqrt{\lambda}}$$
$$\frac{\delta L}{\delta \lambda} = 0 \Rightarrow n = \sum_h n_h = \sum_h \frac{N_h S_h}{\sqrt{\lambda}} \Rightarrow \frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_h N_h S_h}$$

D'où le résultat : $n_h = n \times \frac{N_h S_h}{\sum_{h'=1}^{H} N_{h'} S_{h'}}$ □

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 optimum.

Choix des allocations

Allocation de Neyman et règle de Dalenius

Quand le mécanisme d'allocation au sein des strates est l'allocation de Neyman, alors il est utile de constituer les strates en suivant la **règle de Dalenius**.

Règle de Dalenius : Constituer les strates de telle sorte que $N_h S_h$ soit le même pour chaque strate.

De cette façon, les allocations sont les mêmes pour chaque strate :

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

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

Allocation de Neyman et strate exhaustive

Quand la variance de Y diffère beaucoup d'une strate à une autre, **l'allocation de Neyman de certaines strates peut excéder la taille de leur population**.

Toutes les unités de ces strates doivent être échantillonnées : elles sont regroupées dans une **strate exhaustive** et leur **probabilité d'inclusion est 1**.

Si l'on ne procédait à aucun ajustement, **l'échantillon final aurait alors une taille inférieure à la taille souhaitée**, car trop peu d'individus seraient échantillonnés dans la strate exhaustive.

Choix des allocations

Allocation de Neyman et strate exhaustive

Afin d'obtenir la taille d'échantillon souhaitée, ces strates peuvent être traitées par un algorithme itératif :

1. Calculer les allocations sur l'ensemble des unités.
2. Tant que les allocations donnent un échantillon de taille inférieure à n :
   1. Saturer les strates exhaustives ;
   2. Calculer une nouvelle allocation conduisant à la taille d'échantillon souhaitée à partir des strates restantes.
3. Interroger toutes les unités des strates exhaustives et échantillonner les strates non-exhaustives en utilisant l'allocation calculée.

Choix des allocations

Exemple : Choix d'allocation dans une enquête auprès des entreprises

On cherche à mener une enquête sur l'investissement des entreprises (nature, destination, etc.) d'un secteur donné par un SAS stratifié par tranches d'effectif de 300 unités parmi1 060.

Grâce aux données fiscales, on dispose d'informations sur la moyenne ($\bar{y}_h$) et la dispersion (S_h^2) du montant total d'investissement pour chaque tranche d'effectif.

On souhaite évaluer l'impact du mode d'allocation de l'échantillon sur la précision des estimateurs issus de l'enquête.

Choix des allocations

Allocation optimale : Meilleure précision possible à coût total donné

L'allocation optimale est plus générale que l'allocation de Neyman, dans la mesure où elle prend en compte des **coûts de collecte variables d'une strate à l'autre**.

Exemple : interrogation de tous les membres d'une famille, de tous les sites de production...

On définit ainsi le coût total de l'enquête $C = c_0 + \sum_{h=1}^{H} n_h c_h$ et on peut montrer que la précision maximale pour un coût donné est atteinte avec l'allocation :

$$n_h = \frac{C}{\sqrt{c_h}} \times \frac{N_h S_h}{\sum_{h'=1}^{H} \sqrt{c_{h'}} N_{h'} S_{h'}}$$

Choix des allocations

Allocation optimale

Démonstration.

Le problème est le suivant : $\begin{cases} \min_{n_h} \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ \text{sous la contrainte} \quad C = \sum_h n_h c_h \end{cases}$

En ne gardant que les termes en n_h, on écrit le Lagrangien :

$$L(n_1, n_2, \ldots, n_H, \lambda) = \sum_h \frac{N_h^2 S_h^2}{n_h} - \lambda \left(\sum_h n_h c_h - C \right)$$

Choix des allocations

Allocation optimale

Démonstration.

Les conditions du premier ordre s'écrivent :

$$\begin{cases} \dfrac{\delta L}{\delta n_h} = 0 \Rightarrow \dfrac{N_h^2 S_h^2}{n_h^2} = \lambda c_h \Rightarrow n_h = \dfrac{N_h S_h}{\sqrt{\lambda c_h}} \\ \dfrac{\delta L}{\delta \lambda} = 0 \Rightarrow C = \sum_h n_h c_h = \sum_h \dfrac{N_h S_h \sqrt{c_h}}{\sqrt{\lambda}} \Rightarrow \dfrac{1}{\sqrt{\lambda}} = \dfrac{C}{\sum_h N_h S_h \sqrt{c_h}} \end{cases}$$

D'où le résultat : $n_h = \dfrac{C}{\sqrt{c_h}} \times \dfrac{N_h S_h}{\sum_{h'=1}^{H} \sqrt{c_{h'}} N_{h'} S_{h'}}$ □

Choix des allocations

Allocation optimale : interprétation

Le taux de sondage par strate ainsi obtenu est affecté par la variance de Y et le coût de collecte au sein de chaque strate :

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

En d'autre termes :

1. comme avec l'allocation de Neyman, on doit sur-représenter les strates où la dispersion de Y est la plus forte ;
2. de surcroît, on doit sur-représenter les strates où le coût de collecte c_h est le plus faible.

Partie 3

Exemples d'autres allocations

Choix des allocations

Allocation optimale pour plusieurs variables

L'allocation optimale pour une variable Y peut détériorer la précision des estimateurs d'autres variables.

Quand l'enquête vise à mesurer plusieurs variables faiblement corrélées, l'allocation peut chercher à les prendre en compte simultanément en minimisant la quantité :

$$V = \sum_{j=1}^{J} \alpha_j V(\hat{T}_{str}(Y^j))$$

Le vecteur des α_j pondère l'importance de chacune des J variables intervenant dans le calcul de l'allocation.

Choix des allocations

Allocation optimale pour plusieurs variables

L'allocation qui minimise la quantité V ainsi obtenue conduit à des taux de sondage par strate tels que :

$$\frac{n_h}{N_h} \propto \frac{\sqrt{\sum_{j=1}^{J} \alpha_j S^2_{Y_h^j}}}{\sqrt{c_h}}$$

Problème : Comment choisir les α_j ?

Choix des allocations

Atteindre la même précision au sein de chaque strate

Quand un SAS est mené au sein de chaque strate, la variance de $\bar{Y}$ dans chaque strate est fonction de S_h^2 et n_h (f est supposé négligeable) :

$$V(\bar{Y}) \approx \frac{S_h^2}{n_h}$$

Pour obtenir la même précision dans chaque strate, l'allocation doit ainsi être proportionnelle à la variance de Y dans chaque strate.

$$n_h = n \times \frac{S_h^2}{\sum_{h'=1}^{H} S_{h'}^2}$$

Partie 4

Conclusion

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 6

Tirage systématique et stratification implicite

Tirage systématique et stratification implicite

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 :

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

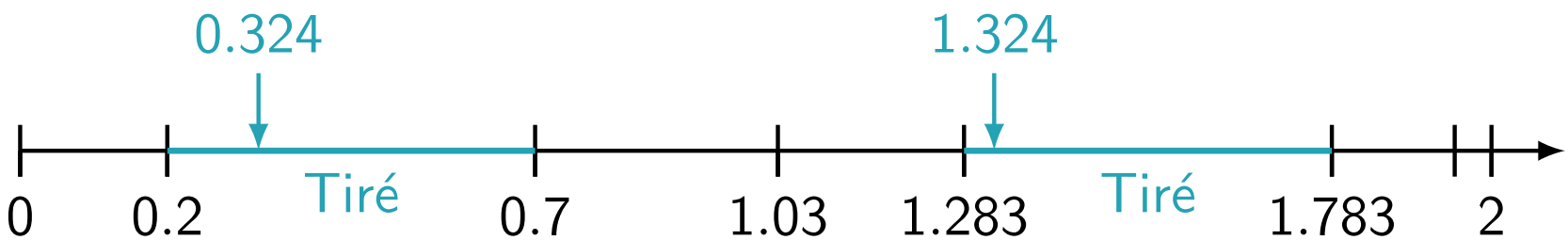
où j parcourt $1, \ldots, n$.

Tirage systématique et stratification implicite

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

Tirage systématique et stratification implicite

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.

Tirage systématique et stratification implicite

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.

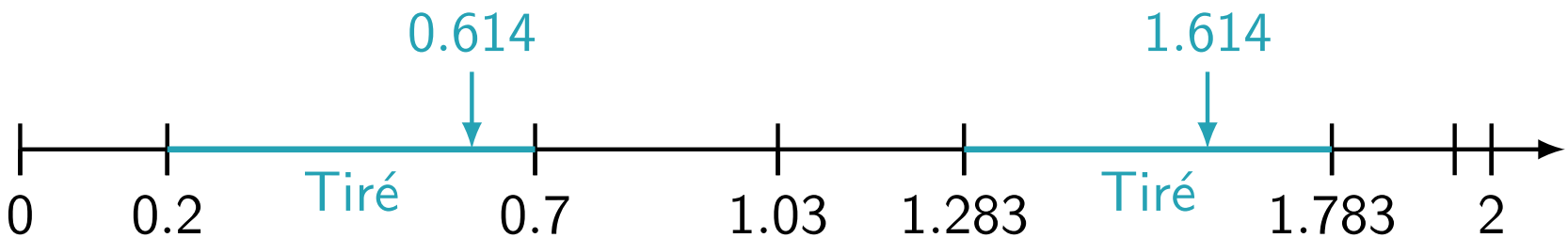
Un tirage systématique sur fichier trié ne peut donc qu'améliorer la précision de tous les estimateurs de l'enquête.

Tirage systématique et stratification implicite

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.

Tirage systématique et stratification implicite

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.