Introduction à la théorie des sondages - Cours 4

Antoine Rebecq
antoine.rebecq@insee.fr

INSEE, direction de la méthodologie

21 avril 2015

Sommaire I

Sommaire II

Chapitre 1

Exemple introductif

Partie 1

Rappel sur le sondage à probabilités inégales

Partie 2

Quelques simulations

Quelques simulations

Imaginons une enquête visant à mesurer plusieurs variables sur une population :

- Y_1 : salaire mensuel
- Y_2 : tranche de surface du logement (variable catégorielle, sur 5 positions)
- Y_3 : couleur des cheveux (variable catégorielle, sur 3 positions)
- Y_4 : montant mensuel d'allocations sociales reçues

On dispose également d'une variable auxiliaire X, qui est le revenu fiscal de l'année précédente

Quelques simulations

On cherche à comparer le plan à probabilités inégales décrit au cours précédent au SAS (sondage aléatoire simple)

On va procéder par simulations (méthode de Monte-Carlo).

Lois des variables

On tire les variables dans les lois suivantes :

- Y_1 : $Pareto(2, 1000)$
- Y_2 : recodification par quantiles de Y_1 + aléa
- Y_3 : multinomiale $(0.1, 0.6, 0.3)$
- Y_4 : $\frac{2}{X} \cdot Weibull(10, 5)$
- X : $Pareto(2, 1000) + |\mathcal{N}(0, 2000)|$

Corrélations entre variables

	Y1	Y2	Y3	Y4	X
Y1	1.00	0.34	0.00	-0.45	0.26
Y2	0.34	1.00	0.00	-0.95	0.09
Y3	0.00	0.00	1.00	-0.00	0.00
Y4	-0.45	-0.95	-0.00	1.00	-0.12
X	0.26	0.09	0.00	-0.12	1.00

Mesure

On mesure l'évolution de l'écart quadratique moyen relatif :

$$EQM(\hat{\Phi}) = B(\hat{\Phi})^2 + \mathrm{Var}(\hat{\Phi})$$

$$EQM_{relatif}(\hat{\Phi}) = \frac{\sqrt{EQM(\hat{\Phi})}}{\Phi}$$

$$\Delta(EQM_{relatif}) = \frac{EQM_{relatif}(INEG) - EQM_{relatif}(SAS)}{EQM_{relatif}(SAS)}$$

Résultats des simulations : Y_1

Y1
-100.00

Résultats des simulations : Y_2

Y2_1	Y2_2	Y2_3	Y2_4	Y2_5
47.51	39.81	18.41	8.68	-32.56

Résultats des simulations : Y_3

Y3_1	Y3_2	Y3_3
18.19	36.02	22.15

Résultats des simulations : Y_4

Y_4
150.09

Partie 3

Conclusion

Conclusion

Conclusion : À moins de n'utiliser une enquête que pour mesurer une variable quantitative Y (si l'on possède une variable auxiliaire X qui lui est bien corrélée), le plan à probabilités inégales est **risqué**.

Conclusion

Dans les chapitres suivants, on présentera différents plans à probabilités inégales parmi les plus utilisés en pratique, ainsi que leurs avantages et inconvénients.

Chapitre 2

Plan stratifié

Partie 1

Principe et notations

Plan stratifié

Soit Y une variable quantitative définie sur $\mathcal{U}$.

La formule de variance d'un SAS dépend de la dispersion S^2 : la précision de l'estimation diminue quand la dispersion de la variable d'intérêt augmente.

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.

Plan stratifié

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

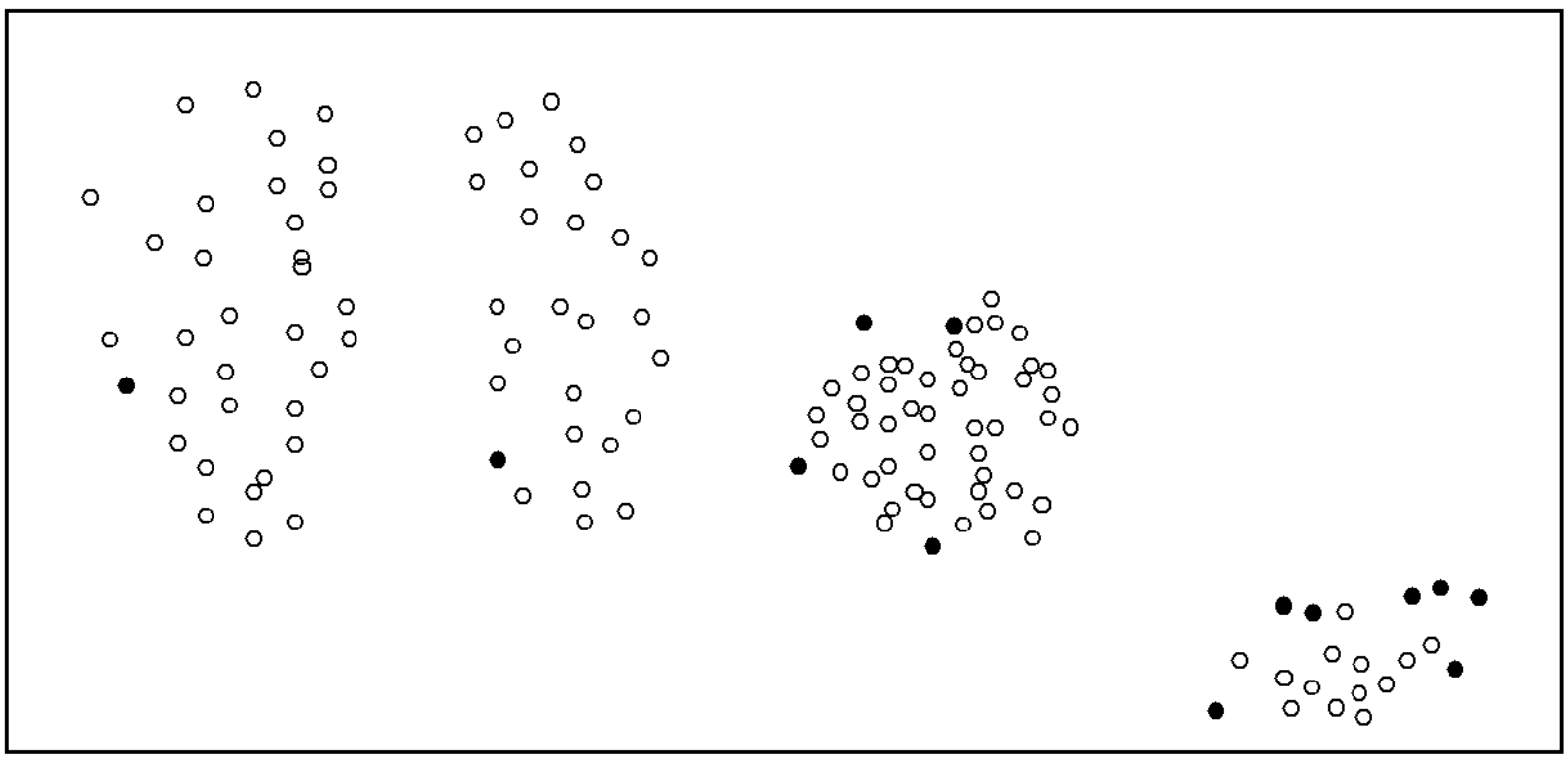
Remarque : Contrairement au SAS, le tirage stratifié requiert de l'**information auxiliaire** dans la base de sondage.

On suppose que les tailles N_h des strates sont connues (typiquement, grâce à la base de sondage).

Plan stratifié

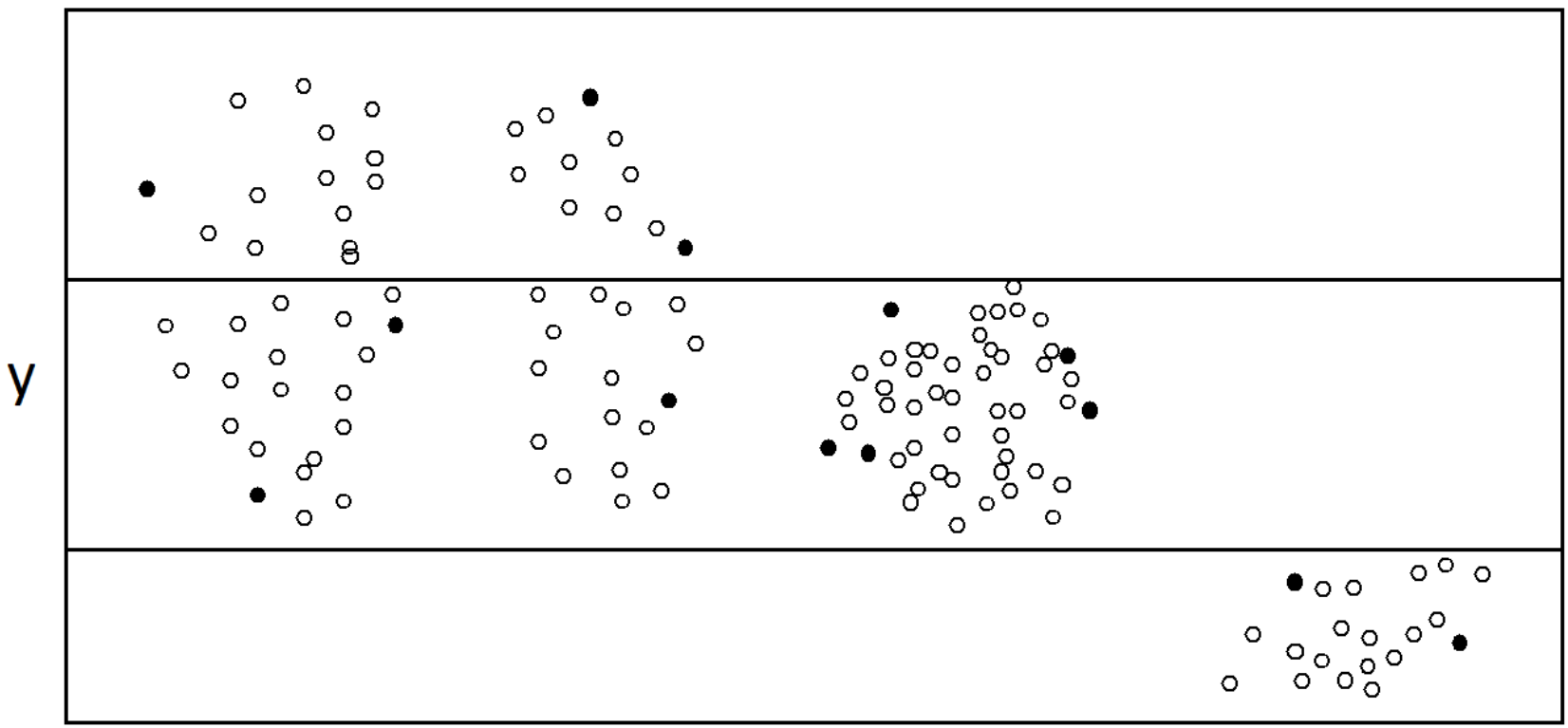
Représentation graphique

SAS de $n = 13$ individus dans une population de taille $N = 130$.



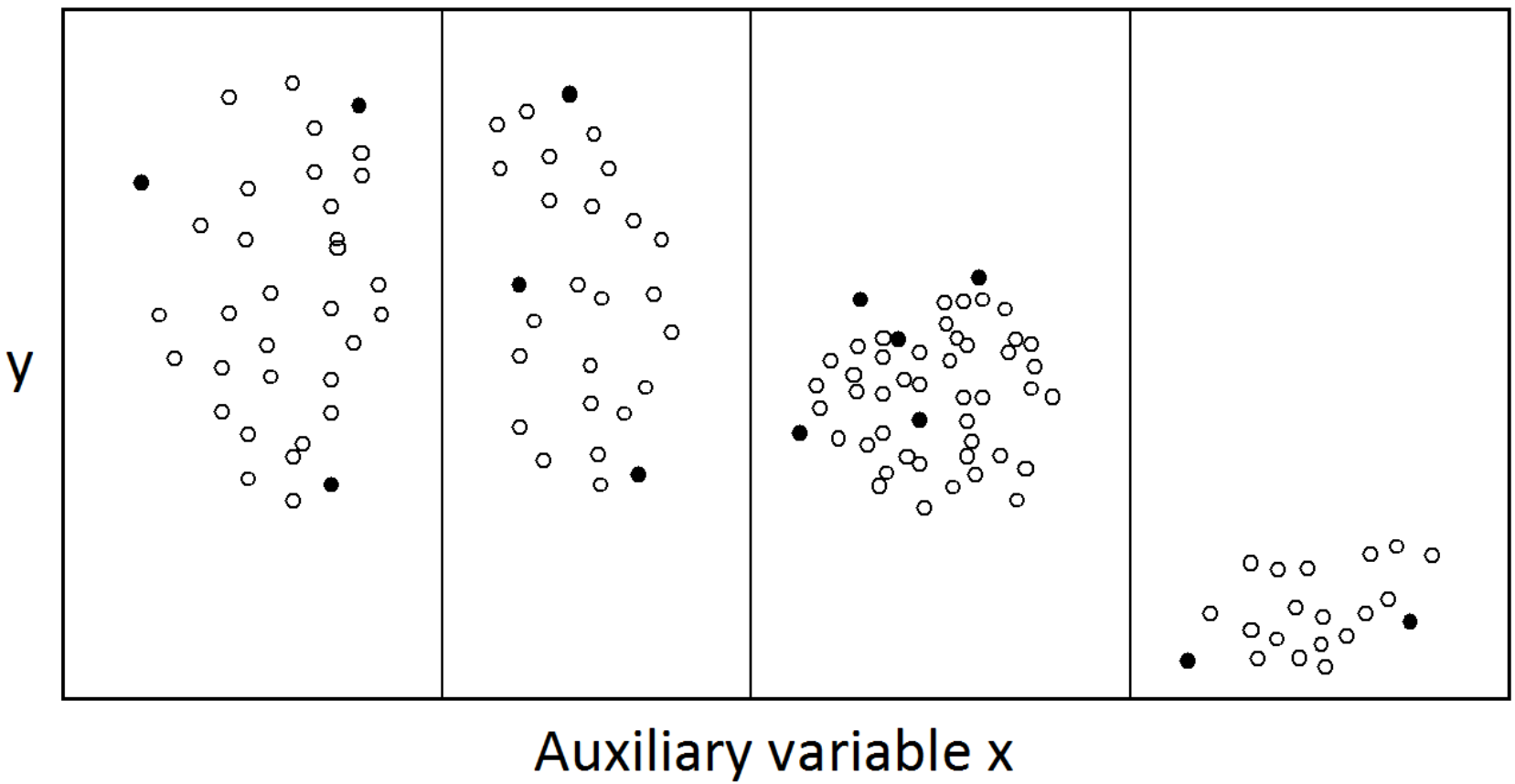
Plan stratifié

Représentation graphique : constitution des strates



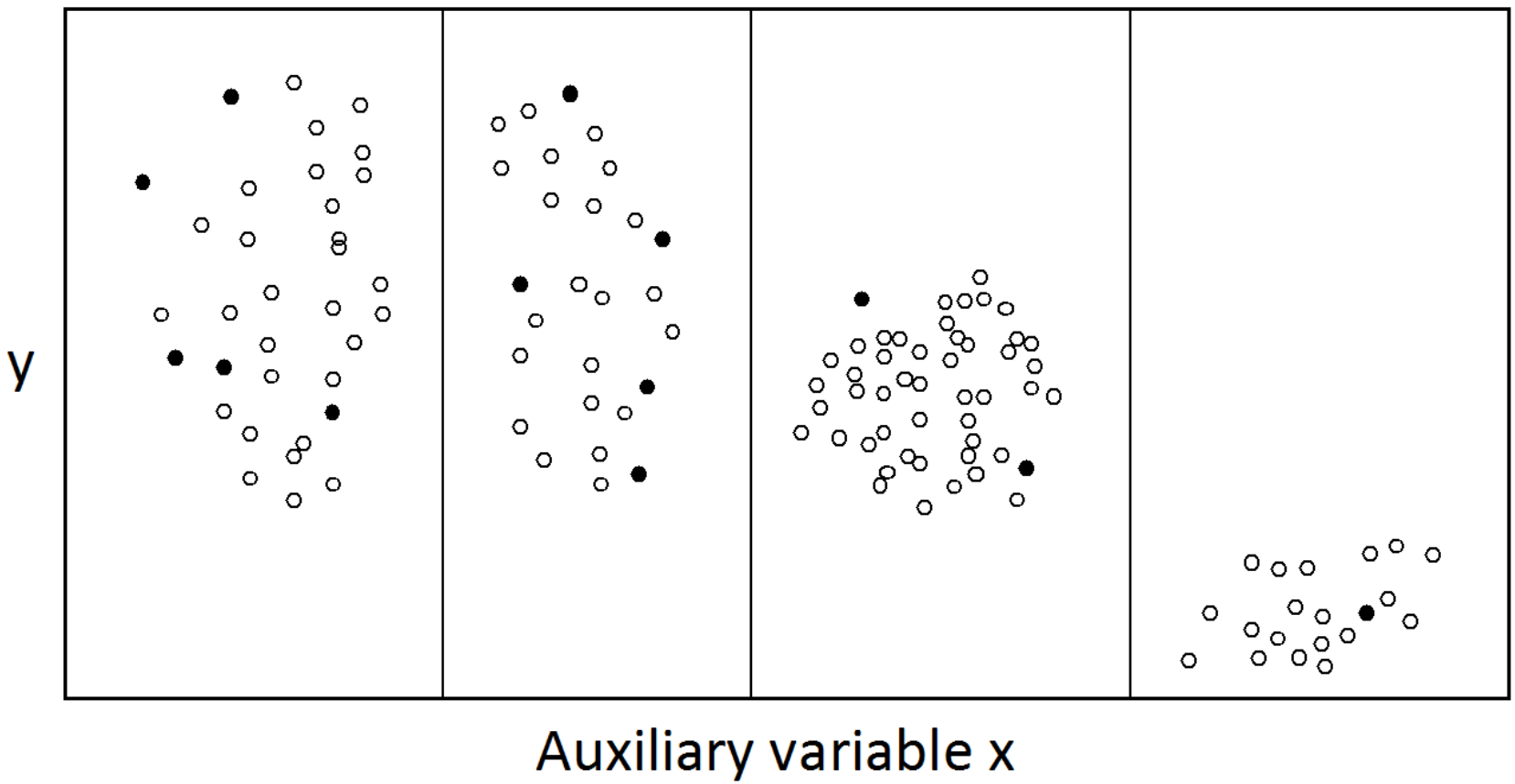
Plan stratifié

Représentation graphique : constitution des strates



Plan stratifié

Représentation graphique : allocation par strate



Plan stratifié

Algorithme pour créer un plan de sondage stratifié de taille fixe n

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'échantillondans 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 stratifié

	Population U	Échantillon s
Taille de la strate h	N_h	n_h
Nombre d'observations	$N = \sum_h N_h$	$n = \sum_h n_h$
Total de Y dans la strate h	$T_h(Y) = \sum_{i \in \mathcal{U}_h} Y_i$	$t_h(Y) = \sum_{i \in s_h} Y_i$
Total si Y dans $\mathcal{U}$	$T(Y) = \sum_h T_h(Y)$	$t(Y) = \sum_h t_h(Y)$
Moyenne de Y dans la strate h	$\bar{Y}_h = \frac{T_h(Y)}{N_h}$	$\bar{y}_h = \frac{t_h(Y)}{n_h}$
Moyenne de Y dans $\mathcal{U}$	$\bar{Y} = \sum_h \frac{N_h}{N} \bar{Y}_h$	$\bar{y} = \sum_h \frac{n_h}{n} \bar{y}_h$

Partie 2

Estimation et précision

Plan stratifié

Estimateur Le total de Y est estimé sans biais par :

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

où $\hat{T}_h(Y)$ est l'estimateur d'Horvitz-Thompson de $T_h(Y)$:

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

Plan stratifié

Précision : Les $\hat{T}_h(Y)$ sont indépendants, d'où :

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

avec $V(\hat{T}_h(Y)) = \sum_{i \in U_h} \sum_{j \in U_h} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$

et $\hat{V}(\hat{T}_{str}(Y))$ un estimateur sans biais de la variance (Horvitz-Thompson or Yates-Grundy).

$V(\hat{T}_h(Y))$ peut être calculé à partir de l'estimateur d'Horvitz-Thomspon de la variance, en remarquant que :

$$\pi_{ij} - \pi_i \pi_j = 0 \quad \text{si} \quad i \in U_h \quad \text{and} \quad j \in U_{h'}, \quad h \neq h'$$

Paragraphe 1

Plan stratifié avec SAS dans chaque strate

Plan stratifié

Estimation et précision

On suppose que le plan de sondage utilisé au sein de chaque strate est un SAS de taux de sondage

$$f_h = \frac{n_h}{N_h}$$

Estimateurs : Le total $T(Y)$ et la moyenne $\bar{Y}$ sont estimés sans biais par :

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

Plan stratifié

Estimation et précision

Remarques

1. $\hat{\bar{Y}}_{str} \neq \bar{y}$ L'estimateur en plan de sondage stratifié diffère de la moyenne empirique.

2. $\hat{T}_{str}(Y) = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} N_h \left(\frac{1}{n_h} \sum_{i \in s_h} y_i \right) =$
$\sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} y_i$
Pour chaque observation de h, le poids est $\frac{N_h}{n_h}$.

Plan stratifié

Estimation et précision

Précision : La variance de l'estimateur stratifié du total est :

$$V(\hat{T}_{str}(Y)) = \sum_{h=1}^{H} N_h^2 V(\bar{y}_h) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

Remarque : La précision dépend seulement de la dispersion de Y **au sein de chaque strate** : plus la dispersion intra est faible, plus stratification est efficace.

La variance estimée est :

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

Remarque : Pour pouvoir être calculé, cet estimateur nécessite au moins deux observations par strate.

Plan stratifié

Estimation et précision

La variance de l'estimateur de la moyenne est :

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

Cette variance est estimée sans biais par :

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

Plan stratifié

Exemple : 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 (1.07 pour un SAS)

Plan stratifié

Exemple : 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 (0.251 pour un SAS)

Partie 3

Constitution des strates

Plan stratifié

Ces résultats donnent l'intuition des règles à suivre pour constituer les strates ainsi que les allocations.

La variance de l'estimation de Y étant directement reliée à la dispersion intra de Y, une bonne stratification se doit de minimiser cette dispersion intra.

Afin d'obtenir la stratification la plus efficace, *les valeurs de Y doivent être les plus proches possibles à l'intérieur de chaque strate*.

Plan stratifié

Exemple : 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 (1.07 pour un SAS)

Plan stratifié

Exemple : 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 (0.251 pour un SAS)

Plan stratifié

Exemple : 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 0.44 (1.07 pour un SAS)

Plan stratifié

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.

Comme pour le sondage à probabilités inégales, il s'agit donc d'utiliser de **l'information auxiliaire** provenant de la base de sondage, si possible *bien corrélée à* Y

En fonction des variables auxiliaires disponibles dans la base de sondage, la stratification pourra être constituée à l'aide d'une ou plusieurs variables, et ce afin de :

- Maximiser l'homogénéité intra
- Maximiser l'hétérogénéité inter

Remarque : Un choix de stratification peut être efficient pour une variable Y, mais inefficace pour une autres.

Plan stratifié

Combien de strates ?

En théorie, plus le nombre de strates est élevé, meilleure est la stratification (puisqu'ainsi on diminue la dispersion intra).

En pratique, il existe un "seuil critique" :

- La collecte sur le terrain serait plus délicate, annulant les gains de précision
- Pour effectuer les estimations de précision, au moins deux unités sont nécessaires par strate (incluant la non-réponse).

Plan stratifié

Critères usuels pour le choix de stratification

Enquêtes ménages

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

Enquêtes entreprises

- Secteur d'activité
- Taille de l'entreprise
- Région

Paragraphe 1

Exemple : strates pour le nombre de salariés

Plan stratifié

Constitution des strates

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, 500-999, 1,000-4,999, 5,000 et plus.

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

Plan stratifié

Constitution des strates

Il existe un certain nombre de méthodes pour déterminer les limites $b_0, b_1, \dots b_H$ 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.

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

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

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

Plan stratifié

Constitution des strates

$\forall h < H$:

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

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

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.

Plan stratifié

Constitution des strates

Application à des données INSEE

Les limites obtenues par cette méthode (exemple précédent) sont : 10-24, 25-59, 60-143, 144-348, 349-846, 847-2,055, 2,056-4,999,5,000 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

Plan stratifié

Constitution des strates

Dans cette situation, 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.

Partie 4

Allocation par strate

Plan stratifié

Une fois les strates définies (et en supposant que l'échantillon est de taille fixe n), y a-t-il un moyen optimal de déterminer les allocations de chaque strate dans l'échantillon ?

La réponse à cette question dépend du but de l'enquête :

- Obtenir la meilleure précision possible pour une variable
- Obtenir la meilleure précision possible pour plusieurs variables simultanément
- Obtenir une bonne précision dans chaque strate afin de pouvoir comparer les estimateurs par strate

Paragraphe 1

Allocation optimale

Plan stratifié

Allocation par strate

Supposons que le coût de l'enquête s'écrive :

$$C = \sum_{h=1}^{H} n_h c_h \quad (+c_0)$$

où : c_h est le coût de réaliser un questionnaire dans la strate h.

Problèmes

- Déterminer n_h qui minimise $\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y}))$ à coût donné C
- Déterminer n_h qui minimise le coût C à précision $\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y}))$ donnée

Plan stratifié

Allocation par strate

Précision optimale à coût donné

Les n_h qui minimise la variance $\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y}))$ à coût C donné sont :

$$n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{k=1}^{H} \sqrt{c_k} N_k S_k}$$

et la variance minimale obtenue est :

$$\mathrm{Var}_{\mathrm{opt}}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y})) = \frac{1}{\mathrm{C}} \left(\sum_{\mathrm{h}=1}^{\mathrm{H}} \sqrt{\mathrm{c_h}} \mathrm{N_h} \mathrm{S_h} \right)^2 - \sum_{\mathrm{h}=1}^{\mathrm{H}} \mathrm{N_h} \mathrm{S_h^2}$$

Plan stratifié

Allocation par strate

Preuve

$$\begin{cases} \min_{n_h} \sum_h N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \\ \text{with constraint} \quad C = \sum_h n_h c_h \end{cases}$$

En ne gardant que les termes en n_h, on écrit le Lagrangien du problème de minimisation :

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

Les conditions du premier ordre s'écrivent :

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

D'où : $n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{k=1}^{H} \sqrt{c_k} N_k S_k}$

Plan stratifié

Allocation par strate

Coût optimal à précision donnée

Les n_h qui minimisent le coût C à précision $\mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y}))$ donnée sont :

$$n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{\sum_{k=1}^{H} \sqrt{c_k} N_k S_k}{V(\hat{T}_{str}(Y)) + \sum_{k=1}^{H} N_k S_k^2}$$

et le coût minimal est :

$$C_{opt} = \frac{\left(\sum_{h=1}^{H} \sqrt{c_h} N_h S_h\right)^2}{V(\hat{T}_{str}(Y)) + \sum_{h=1}^{H} N_h S_h^2}$$

Plan stratifié

Allocation par strate

Interprétation

Dans les deux cas :

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

- On doit sur-représenter la strate où la dispersion de Y est la plus forte : en d'autres termes, il s'agit d'aller chercher l'information là où elle se trouve !
- On doit sur-représenter la strate où le coût unitaire c_h est le plus faible

Plan stratifié

Allocation par strate

Allocation de Neyman : Si l'on suppose que le coût unitaire (d'un questionnaire) c_h est uniforme au sein de chaque strate, l'allocation optimale se nomme **allocation de Neyman** :

$$n_h = n \times \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}$$

Règle de Dalenius : Si l'on utilise l'allocation de Neyman, il est utile de définir les strates de telle sorte que $N_h S_h$ est le même pour chaque strate. Cela donne des allocations égales dans chaque strate :

$$n_h = \frac{n}{H}$$

Plan stratifié

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 : 0.67 (1.07 pour un SAS)

Plan stratifié

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 : $8/3 = 2.67$ (1.07 pour un SAS)

Plan stratifié

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 :

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

Il s'ensuit :

$$\begin{cases} n_I = 4 \times \dfrac{3 \times 4}{3 \times 4 + 3 \times 2} = \dfrac{48}{18} = 2.7 \\ n_{II} = 4 \times \dfrac{3 \times 2}{3 \times 4 + 3 \times 2} = \dfrac{24}{18} = 1.3 \end{cases}$$

Ce qui explique le résultat précédent.

Plan stratifié

Allocation par strate

Estimation du terme S_h

La variance intra-strate de Y est inconnue. Afin de pouvoir utiliser l'allocation optimale, elle peut être estimée de multiples façons :

- Dire d'expert
- Information auxiliaire de la base de sondage
- Enquêtes précédentes
- Réalisation d'une petite enquête préliminaire (si le coût n'est pas trop élevé en regard des objectifs)

Paragraphe 2

Allocation proportionnelle

Plan stratifié

Allocation par strate

Définition : L'allocation proportionnelle est identique à la répartition des individus dans la population :

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

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

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

Il s'agit donc d'un sondage à probabilités égales.

Plan stratifié

Allocation par strate

Estimateurs : identiques au SAS ...

$$\hat{\bar{Y}}_{prop} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y}$$

Variance : ... mais les variances diffèrent !

$$V(\hat{\bar{Y}}_{prop}) = \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2 \simeq (1-f) \frac{S_{within}^2}{n}$$

Plan stratifié

Allocation par strate

Comparaison avec le SAS

On a : $\mathrm{Var}(\hat{\bar{\mathrm{Y}}}_{\mathrm{SAS}}) = (1-\mathrm{f})\dfrac{\mathrm{S}^2}{\mathrm{n}}$ et $S^2_{within} \leq S^2$ (formule de décomposition de la variance) :

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

Le plan stratifié avec allocation proportionnelle *possède toujours une précision meilleure que le SAS*.

Plus grande est la dispersion inter strates, plus grand est le gain associé à la stratification.

Plan stratifié

Allocation par strate

Comparaison avec l'allocation de Neyman

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 variable** Y : elle peut éventuellement être néfaste pour l'estimation d'une autre variable d'intérêt.

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

Paragraphe 3

Autres allcations

Plan stratifié

Allocation par strate

Même précision au sein de chaque strate

La variance de $\bar{Y}$ dans chaque strate est fonction de S_h^2 et n_h (on considère que $f \approx 0$) :

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

Si l'on cherche à obtenir la même précision dans chaque strate, l'allocation doit être proportionnelle à la variance intra de Y dans chaque strate.

$$n_h = n \times \frac{S_h^2}{\sum_{k=1}^{H} S_k^2}$$

Plan stratifié

Allocation par strate

Allocation optimale pour plusieurs variables

L'allocation optimale pour une variable Y peut déteriorer la précision pour l'estimation d'autres variables, au point que le plan stratifié possède une variance supérieure à celle d'un SAS.

Une façon de procéder afin d'éviter cela est de pondérer les variances des J variables d'intérêt de l'enquête :

$$\mathrm{Var} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \alpha_{\mathrm{j}} \mathrm{Var}(\hat{\mathrm{T}}_{\mathrm{str}}(\mathrm{Y}^{\mathrm{j}}))$$

Plan stratifié

Allocation par strate

Allocation optimale pour plusieurs variables
Ain de minimiser Var à coût C donné :

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

Problème : Comment choisir les α_j ?...

Paragraphe 4

Strate exhaustive

Plan stratifié

Allocation par strate

Il peut arriver (à moins d'utiliser l'allocation proportionnelle), que l'allocation choisie soit de taille supérieure à celle de la population dans certaines strates (voir cours probabilités d'inclusion inégales).

Tous les individus appartenant à cette strate doivent être échantillonnés : elle est alors appelée **strate exhaustive**.

Cela peut engendrer une taille d'échantillon finale inférieure au n prévu, car alors trop peu d'individus seraient échantillonnés dans la strate exhaustive.

Plan stratifié

Allocation par strate

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

1. Calculer les allocations utilisant toutes les strates
2. Tant que les allocations donnent un échantillon de taille inférieure à n :
   1. Saturer les strates exhaustives
   2. Calculer une nouvelle allocation pour toutes les strates restantes, en enlevant les individus appartenant à une strate exhaustive.
3. Échantillonner les strates non exhaustives en utilisant l'allocation calculée.

Paragraphe 5

Exemple : plan de sondage pour une enquête entreprises

Plan stratifié

Allocation par strate

But : Échantillonner $n = 300$ entreprises parmi une population $\mathcal{U}$ de taille $N = 1060$ (par exemple un secteur particulier).

Variables auxiliaires disponibles : Le nombre de salariés est connu par tranches. Pour chaque entreprise, on connaît la moyenne $(\bar{y})$ et la variance (s_h^2) du nombre de changements d'effectifs.

Plan stratifié

Allocation par strate

Taille de l'entreprise	$\mathbf{N_h}$	$\bar{\mathbf{y}}_h$	$\mathbf{S_h^2}$	*Prop.*	*Opti.*
0-9	500	10	2		
10-19	300	50	15		
20-49	150	200	50		
50-499	100	500	100		
500 et plus	10	1,000	2,500		

Todo : Déterminer les allocations optimale et proportionnelle (la variation d'effectifs étant la variable auxiliaire) Pour chaque cas, calculer la variance dans l'estimation de la variation d'effectifs.

Partie 5

Tirage systématique et stratification

Plan stratifié

On se donne un plan de sondage taille fixe n d'une population N, défini par ses probabilités d'inclusion (simples) π_i. On définit : $a_i = \sum_{j=1}^{i} \pi_j$. *Principe :*

1. On tire un réel η dans $\mathcal{U}_{[0;1]}$.
2. On sélectionne toutes les unités i vérifiant :

$$a_{i-1} \leq X + j - 1 < a_i$$

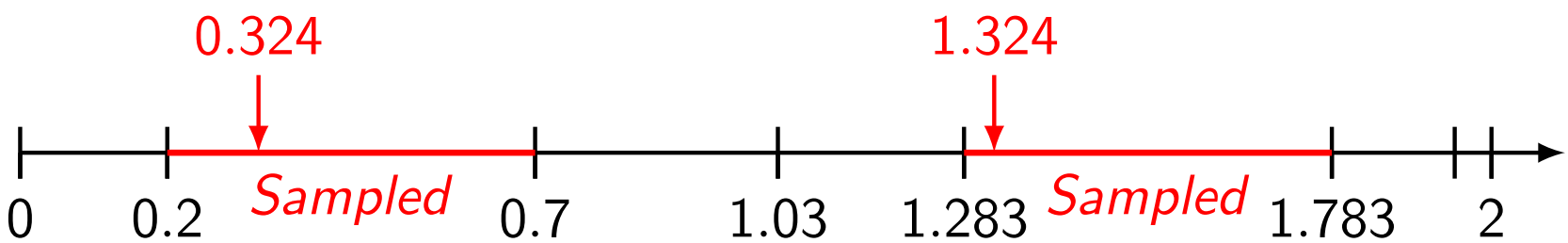
où j parcourt $1 \cdots n$.

Plan stratifié

Exemple

$N = 7 \quad n = 2 \quad \sum_{i=1}^{7} \pi_i = 2 \quad \eta = 0.324$

i	1	2	3	4	5	6	7
π_i	0.2	0.5	0.33	0.25	0.5	0.166	0.05
a_i	0.2	0.7	1.03	1.283	1.783	1.950	2.00



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

Plan stratifié

Propriétés

- On obtient la taille désirée n, et on respecte les π_k
- Efficace : un seul parcours de la base de sondage nécessaire
- Suivant l'ordre du fichier, quelques probabilités d'inclusion doubles π_{kl} peuvent être nulles : les estimateurs de variance de l'estimateur d'Horvitz-Thompson pourront être biaisés.
- Le tirage systématique est souvent utilisé pour réduire la variance des estimations (voir cours 4)

Plan stratifié

Tirage systématique et plan stratifié

Quand la base de sondages est triée selon la variable de stratification, le tirage systématique à probabilités égales est approximativement égale **en termes de précision** à un sondage stratifié :

- avec allocation proportionnelle
- et un tirage par SAS dans chaque strate

MAIS : les probabilités d'inclusion doubles π_{kl} diffèrent, en particulier car une bonne partie d'entre elles est nulle.

Plan stratifié

Justifications

- Le tirage systématique sur fichier trié est équivalent à une stratification implicite, qui ne peut qu'être meilleure que le SAS en termes de précision.
- On peut ainsi réaliser une stratification à un niveau très fin (avec simplement quelques individus par strate), tandis qu'une stratification explicite aurait donné des strates de tailles parfois trop faibles.

Exemples à l'INSEE

- Dans les enquêtes entreprises, la région est souvent introduite comme variable de stratification implicite dans un tirage systématique, en triant le fichier par région.
- Dans les enquêtes ménages, la stratification liée au sujet de l'enquête est introduite via tirage systématique.

Plan stratifié

Trade-off entre la variance de l'estimateur d'Horvitz-Thompson estimator et le biais de l'estimateur de la variance

Les propriétés du tirage systématiques s'apparentent à un trade-off :

- D'un côté, le tirage systématique sur fichier trié **améliore la variance de l'estimateur d'Horvitz-Thompson**
- De l'autre, il donne un nombre élevé de probabilités d'inclusion doubles nulles, ce qui donne un **estimateur de la variance biaisé**.

En pratique, on préfère souvent une variance plus faible, quitte à ne pas pouvoir l'estimer sans biais.

Chapitre 3

Exercice

Tirage systématique et stratification

Un directeur de cirque possède 100 éléphants, et veut estimer le poids total de son troupeau car il veut traverser un fleuve en bateau. Le directeur avait déjà fait peser tous les éléphants de son troupeau et avait obtenu les résultats suivants :

	Effectifs N_h	Moyennes $\bar{y}_h$	Variances S^2_{yh}
Mâles	60	6	4
Femelles	40	4	2.25

Tirage systématique et stratification

1. Calculer la variance dans la population du caractère “poids de l'éléphant” pour l'année précédente.
2. Si, l'année précédente, le directeur avait procédé à un SAS de 10 éléphants, quelle aurait été la variance de l'estimateur du poids total du troupeau ?
3. Si le directeur avait procédé à un tirage stratifié à allocation proportionnelle de 10 éléphants, quelle aurait été la variance de l'estimateur du poids total du troupeau ?
4. Si le directeur avait procédé à un tirage stratifié à allocation optimale de 10 éléphants, quels auraient été les effectifs de strates, et quelle aurait été la variance de l'estimateur du poids total du troupeau ?