Échantillonnage

Chapitre 8 : Le sondage systématique

Jaouad Madkour

31 octobre 2026

Plan du chapitre

Un ménage sur k **2**
- Le principe . . . 2
- La liste rangée en lignes de k . . . 2
- Pourquoi le pratiquer . . . 3
- Quand N n'est pas un multiple de n . . . 3

Probabilités d'inclusion et estimateur **4**
- Une chance sur k pour chaque unité . . . 4
- Sur la mini-population . . . 5
- La moyenne de l'échantillon est sans biais . . . 5
- Pas d'estimateur sans biais de la variance . . . 6

La variance exacte **7**
- Une moyenne parmi k . . . 7
- Les dix échantillons de la population du cours . . . 7
- Les dix moyennes face au sondage aléatoire simple . . . 8
- Hétérogènes à l'intérieur, semblables entre eux . . . 8

L'ordre de la liste **9**
- L'ordre du fichier : une stratification implicite . . . 9
- Une précision qui dépend de l'ordre exact . . . 10
- Trier selon une variable liée à la dépense . . . 10

Le danger de la périodicité **11**
- Un cycle dans la liste . . . 11
- Le pas tombe sur la période . . . 11
- Se prémunir . . . 12

Estimer la variance en pratique **12**
- Faire comme si c'était un sondage aléatoire simple . . . 12
- Prudent sur une liste rangée, trompeur sur un cycle . . . 13
- Les différences successives . . . 13
- Deux estimateurs face à trois ordres . . . 14
- Le tirage systématique répété . . . 14

S'entraîner **15**
Exercice 1 : douze boutiques rangées 15
Exercice 1 : solution 16
Exercice 2 : un immeuble de dix étages 16
Exercice 2 : solution 16
Exercice 3 : les ménages connectés 17
Exercice 3 : solution 17

Ce qu'il faut retenir **17**
Le résumé du chapitre 17

Un ménage sur k

Le principe

Le tirage systématique

Les N unités sont rangées dans l'ordre d'une **liste**. On fixe le **pas** $k = N/n$, supposé entier pour l'instant.

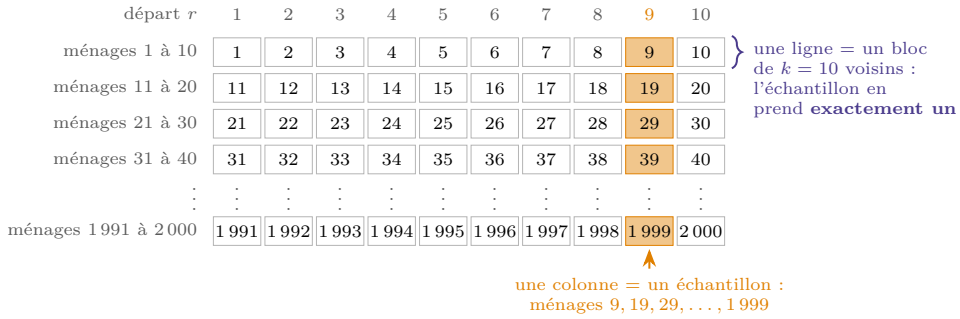
1. Tirer un **départ** r au hasard parmi $1, 2, \ldots, k$.
2. Prendre les unités de rangs $r,\ r+k,\ r+2k,\ \ldots,\ r+(n-1)\,k$.

Un seul nombre aléatoire suffit pour tout l'échantillon.

Sur la population du cours

$N = 2\,000$ ménages et $n = 200$, donc $k = 10$. Le départ tiré (graine 2026) est $r = 9$: l'échantillon comprend les ménages $9, 19, 29, \ldots, 1\,999$.

La liste rangée en lignes de k



Dix échantillons possibles, pas plus

Rangée en lignes de $k = 10$, la liste devient un tableau : un échantillon systématique est **une colonne**. Il n'y a que $k = 10$ échantillons possibles, contre environ 10^{281} pour un sondage aléatoire simple de 200 ménages (chapitre 1).

Pourquoi le pratiquer

Simple sur le terrain

Un enquêteur qui parcourt une rue, une liste papier ou la sortie d'un magasin n'a qu'à **compter** : un logement sur k, un client sur 25. Inutile de numéroter toute la population à l'avance, ni même de connaître N exactement. Le travail se contrôle facilement.

Un échantillon étalé sur toute la liste

Chaque ligne du tableau est un **bloc** de k voisins, et l'échantillon en prend exactement un. Aucune partie de la liste n'est oubliée, aucune n'est surreprésentée. Cet étalement fait la **force** du tirage systématique quand la liste est bien rangée, et son **danger** quand elle cache un cycle.

Quand N n'est pas un multiple de n

Deux solutions

- **Pas arrondi** : prendre pour k l'entier le plus proche de N/n. Selon le départ, l'échantillon a l'une de deux tailles voisines : la taille devient aléatoire (chapitre 1).
- **Pas fractionnaire** : tirer u uniformément dans $]0\,;k]$ et prendre les rangs $\lceil u\rceil, \lceil u + k\rceil, \dots, \lceil u + (n-1)\,k\rceil$, où $\lceil\cdot\rceil$ arrondit à l'entier supérieur. La taille vaut exactement n.

Deux ménages parmi cinq, pas de deux et demi

Sur la mini-population A, B, C, D, E, $k = 5/2 = 2{,}5$. Avec $u = 1{,}3$, les rangs sont $\lceil 1{,}3\rceil = 2$ et $\lceil 3{,}8\rceil = 4$: B et D. Les cinq intervalles de longueur 0,5 qui composent $]0\,;2{,}5]$ donnent cinq échantillons équiprobables : AC, AD, BD, BE, CE.

Démonstration : Le pas fractionnaire donne exactement n unités

Les rangs sont distincts. Le pas vérifie $k = N/n \geq 1$. Les nombres $u,\ u+k,\ \dots,\ u+(n-1)\,k$ sont espacés d'au moins 1 et le dernier vaut au plus $k + (n-1)\,k = N$: leurs plafonds $\lceil\cdot\rceil$ sont n entiers distincts de 1 à N. L'échantillon compte donc **exactement** n unités, quel que soit u.

Chaque unité a la probabilité n/N. L'unité j est tirée si et seulement si l'un des nombres $u + i\,k$ tombe dans $]j-1\,;j]$. Tout réel $x \in]0\,;N]$ s'écrit d'une seule façon $x = u + i\,k$ avec $u \in]0\,;k]$ et $i \in \{0,\dots,n-1\}$. Les valeurs de u qui font tirer j sont donc obtenues en translatant des morceaux de $]j-1\,;j]$ (de $-i\,k$, avec i qui dépend du morceau). Deux points de cet intervalle ne peuvent pas donner le même u, car leur écart est inférieur à $1 \leq k$: les morceaux ne se chevauchent pas, et l'ensemble des u qui font tirer j a pour longueur exactement 1. Comme u est uniforme sur un intervalle de longueur k,

$$\pi_j = \frac{1}{k} = \frac{n}{N}.$$

Sur la mini-population. Pour $N = 5$, $n = 2$ et $k = 2{,}5$, les plafonds $\lceil u\rceil$ et $\lceil u + 2{,}5\rceil$ ne changent qu'en $u = 0{,}5\,;\ 1\,;\ 1{,}5\,;\ 2\,;\ 2{,}5$: cinq intervalles de longueur 0,5, de probabilité $0{,}5/2{,}5 = 1/5$ chacun.

$u \in$	$]0\,;0{,}5]$	$]0{,}5\,;1]$	$]1\,;1{,}5]$	$]1{,}5\,;2]$	$]2\,;2{,}5]$
rangs	1 et 3	1 et 4	2 et 4	2 et 5	3 et 5
échantillon	AC	AD	BD	BE	CE

Ce que cela dit. Chaque ménage figure dans deux des cinq échantillons : $\pi_j = 2/5 = n/N$, comme pour un pas entier. Mais les paires AB, BC, CD, DE et AE ne sont jamais tirées ensemble : le tirage systématique sépare les voisins.

■

Probabilités d'inclusion et estimateur

Une chance sur k pour chaque unité

Les probabilités d'inclusion

Dans ce chapitre, k désigne le pas : les unités sont notées j et l. Chaque unité appartient à **une seule** des k colonnes :

$$\pi_j = \frac{1}{k} = \frac{n}{N}, \qquad \pi_{jl} = \begin{cases} 1/k & \text{si } j \text{ et } l \text{ sont dans la même colonne,} \\ 0 & \text{sinon.} \end{cases}$$

Démonstration : Les probabilités d'inclusion du tirage systématique

Une unité, une seule colonne. Avec $N = k\,n$, le départ $r \in \{1, \dots, k\}$ détermine l'échantillon $s_r = \{r,\ r+k,\ \dots,\ r+(n-1)\,k\}$, dont tous les rangs sont congrus à r modulo k. Deux départs différents donnent deux échantillons disjoints, et tout rang j appartient à l'échantillon du départ congru à j modulo k : les k échantillons forment une **partition** de U. Chacun a la probabilité $1/k$, donc chaque unité est tirée avec $\pi_j = 1/k = n/N$.

Les paires. Deux unités $j \neq l$ sont tirées ensemble si et seulement si elles sont dans le même échantillon, c'est-à-dire congrues modulo k : alors $\pi_{jl} = 1/k$. Sinon, aucun échantillon ne les contient toutes les deux et $\pi_{jl} = 0$.

Combien de paires peuvent être tirées ensemble ? Chaque échantillon contient $\binom{n}{2}$ paires, et ces paires sont distinctes d'un échantillon à l'autre : $k\binom{n}{2}$ paires sur $\binom{N}{2}$. Pour $N = 2\,000$, $n = 200$ et $k = 10$:

$$10 \times \binom{200}{2} = 199\,000 \quad \text{sur} \quad \binom{2\,000}{2} = 1\,999\,000,$$

soit à peine un dixième ; environ 90 % des paires ont $\pi_{jl} = 0$.

Ce que cela dit. Les probabilités simples sont celles du sondage aléatoire simple, pas les probabilités doubles. Or la variance de l'estimateur de Horvitz-Thompson et son estimateur dépendent des π_{jl} (chapitre 2) : le tirage systématique est un autre plan, que les formules du sondage aléatoire simple ne décrivent pas.

■

Sur la population du cours

$\pi_j = 1/10$, comme pour un sondage aléatoire simple de 200 ménages. Mais sur les 1 999 000 paires de ménages, $10 \times \binom{200}{2} = 199\,000$ seulement peuvent être tirées ensemble : 90 % **des paires** ont $\pi_{jl} = 0$.

Sur la mini-population

Cinq échantillons de probabilité un cinquième

Échantillon	AC	AD	BD	BE	CE
Moyenne	5,5	6,5	7,5	10	10,5

Chaque ménage figure dans deux échantillons : $\pi_j = 2/5 = 0{,}4 = n/N$. Espérance : $(5{,}5 + 6{,}5 + 7{,}5 + 10 + 10{,}5)/5 = 8 = \overline{Y}$.

Mêmes probabilités simples, autres probabilités doubles

Les π_j sont ceux du sondage aléatoire simple (chapitre 1), mais 5 paires sur 10 ne sont **jamais** tirées ensemble : AB, BC, CD et DE, voisines dans la liste, et AE. Le tirage systématique sépare les voisins.

La moyenne de l'échantillon est sans biais

L'estimateur

Puisque $\pi_j = n/N$, l'estimateur de Horvitz-Thompson du total vaut $\widehat{Y}_{HT} = \sum_{j\in S} y_j/\pi_j = N\,\overline{y}$: on estime $\overline{Y}$ par la moyenne $\overline{y}$ des unités tirées, sans biais (chapitre 2).

La démonstration directe

Notons $\overline{y}_r$ la moyenne de l'échantillon de départ r. Les k échantillons forment une **partition** de U, chacun de taille n, et $nk = N$:

$$E(\overline{y}) = \frac{1}{k}\sum_{r=1}^{k}\overline{y}_r = \frac{1}{k}\cdot\frac{1}{n}\sum_{j\in U} y_j = \frac{Y}{N} = \overline{Y}$$

Sur la population du cours, les dix moyennes systématiques ont pour moyenne 7 911,44 dh : exactement $\overline{Y}$.

Démonstration : La moyenne systématique est sans biais

Une partition. Les k échantillons $s_1, \ldots, s_k$ partitionnent U et chacun compte n unités, avec $k\,n = N$. La somme des totaux des échantillons est donc le total de la population :

$$\sum_{r=1}^{k}\sum_{j\in s_r} y_j = \sum_{j\in U} y_j = N\,\overline{Y}.$$

L'espérance. Chaque départ a la probabilité $1/k$:

$$E(\overline{y}) = \frac{1}{k}\sum_{r=1}^{k}\overline{y}_r = \frac{1}{k}\cdot\frac{1}{n}\sum_{r=1}^{k}\sum_{j\in s_r} y_j = \frac{N\,\overline{Y}}{k\,n} = \overline{Y}.$$

Le lien avec Horvitz-Thompson. Avec $\pi_j = n/N$ pour tout j, $\widehat{Y}_{HT} = \sum_{j\in S} y_j/\pi_j = \frac{N}{n}\sum_{j\in S} y_j = N\,\overline{y}$: estimer $\overline{Y}$ par $\overline{y}$ revient à diviser l'estimateur de Horvitz-Thompson par N. Pour un pas fractionnaire, les échantillons ne forment plus une partition, mais $\pi_j = n/N$ suffit : $E(\widehat{Y}_{HT}) = \sum_{j\in U}\pi_j\, y_j/\pi_j = Y$.

Ce que cela dit. Le tirage systématique estime la moyenne sans biais, comme le sondage aléatoire simple. Sur la mini-population, les cinq moyennes 5,5 ; 6,5 ; 7,5 ; 10 ; 10,5 ont pour moyenne $8 = \overline{Y}$; sur la population du cours, les dix moyennes ont pour moyenne 7 911,44.

■

Pas d'estimateur sans biais de la variance

Une seule valeur observée

Le chapitre 2 estime la variance sans biais à condition que $\pi_{jl} > 0$ pour **toutes** les paires. Ici, la plupart sont nulles. Plus simplement : le plan choisit **une** colonne parmi k. On observe une seule des k valeurs $\overline{y}_r$, et une seule valeur ne dit rien de leur dispersion.

Démonstration : Aucun estimateur ne peut être sans biais pour la variance

Ce que l'on observe. Si le départ est r, on observe les valeurs c_r de la colonne r, et rien d'autre. Un estimateur de variance est donc une fonction $v_r(c_r)$ de ces valeurs (qui peut dépendre du départ r, par les rangs). Il est sans biais pour **toutes** les populations si et seulement si, pour toutes les colonnes $c_1, \dots, c_k$,

$$\frac{1}{k}\sum_{r=1}^{k} v_r(c_r) = V(\overline{y}) = \frac{1}{k}\sum_{r=1}^{k}(\overline{y}_r - \overline{Y})^2.$$

Comme $\overline{Y} = \frac{1}{k}\sum_r \overline{y}_r$, cela s'écrit

$$\sum_{r=1}^{k} v_r(c_r) = \sum_{r=1}^{k}\overline{y}_r^{\,2} - \frac{1}{k}\Big(\sum_{r=1}^{k}\overline{y}_r\Big)^2.$$

Un obstacle. Le membre de gauche est une somme de termes dont **chacun ne dépend que d'une colonne**. Le développement du carré à droite contient le terme $-\frac{2}{k}\,\overline{y}_1\,\overline{y}_2$, un produit des moyennes de deux colonnes différentes.

Le test. Remplaçons la colonne 1 par une autre colonne c_1' (moyenne $\overline{y}_1'$) et la colonne 2 par c_2' (moyenne $\overline{y}_2'$), les autres colonnes restant fixes. Formons la combinaison des quatre égalités obtenues, avec les signes + pour (c_1, c_2) et (c_1', c_2'), − pour (c_1', c_2) et (c_1, c_2'). À gauche, chaque terme $v_r(c_r)$ apparaît une fois avec le signe + et une fois avec le signe − : tout s'annule. À droite, les carrés $\overline{y}_r^{\,2}$ s'annulent de même, et il reste

$$-\frac{2}{k}\,(\overline{y}_1 - \overline{y}_1')(\overline{y}_2 - \overline{y}_2'),$$

non nul dès que les moyennes diffèrent. L'égalité est impossible.

Ce que cela dit. Ce n'est pas une faiblesse des formules, mais une limite du plan : on n'observe qu'**une** colonne sur k, alors que la variance compare les colonnes entre elles. D'où les trois solutions du chapitre : appliquer la formule du sondage aléatoire simple (approchée), les différences successives (approchées), ou tirer plusieurs départs (sans biais).

■

Une grappe tirée parmi k

Chaque colonne est un paquet de n unités dispersées dans la liste : le tirage systématique est un sondage **par grappes** où l'on ne tire qu'une grappe (chapitre 9). Sa précision existe et se calcule sur la population ; c'est son **estimation**, à partir de l'échantillon, qui pose problème.

La variance exacte

Une moyenne parmi k

La variance du tirage systématique

$\overline{y}$ prend les k valeurs $\overline{y}_1, \dots, \overline{y}_k$, chacune avec la probabilité $1/k$:

$$V(\overline{y}) = \frac{1}{k} \sum_{r=1}^{k} \left(\overline{y}_r - \overline{Y}\right)^2$$

Sur la mini-population

Écarts des cinq moyennes à $\overline{Y} = 8$: $-2{,}5$; $-1{,}5$; $-0{,}5$; 2 ; $2{,}5$. D'où

$$V(\overline{y}) = \frac{6{,}25 + 2{,}25 + 0{,}25 + 4 + 6{,}25}{5} = \frac{19}{5} = 3{,}8$$

contre $(1-f)\,S^2/n = 4{,}35$ pour le sondage aléatoire simple (chapitre 3).

Les dix échantillons de la population du cours

Les dix moyennes, arrondies au dirham

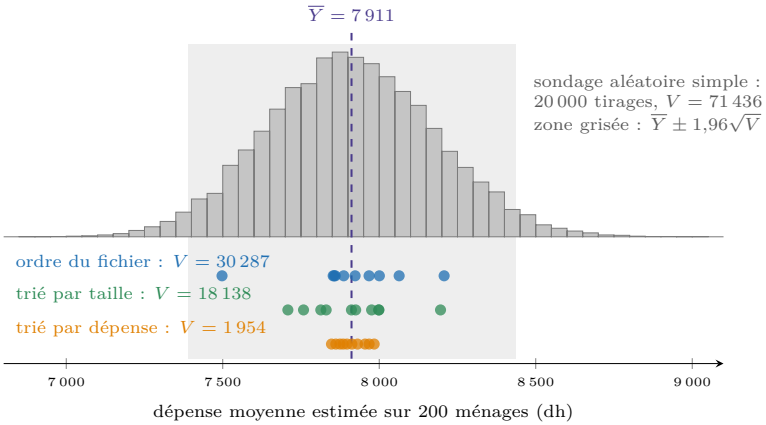
Départ r	1	2	3	4	5
Moyenne	8 000	7 853	7 860	7 856	8 064
Départ r	6	7	8	9	10
Moyenne	7 887	8 207	7 498	7 923	7 968

$$V(\overline{y}) = 30\,287 \qquad \text{contre} \qquad V_{SAS} = (1-f)\,\frac{S^2}{n} = 71\,435$$

Une taille effective de 472 ménages

Rapport des variances : 30 287/71 435 = 0,424. Erreur-type de 174 dh au lieu de 267 ; demi-largeur de l'intervalle à 95 % : 341 dh au lieu de 524. Taille effective (chapitre 7) : 200/0,424 ≈ 472 ménages.

Les dix moyennes face au sondage aléatoire simple



Lecture

En gris, $\overline{y}$ sur 20 000 sondages aléatoires simples. Les dix moyennes de l'ordre du fichier (bleu) sont plus resserrées, celles d'une liste triée par taille (vert) plus encore ; triée par dépense (orange), elles se confondent presque avec $\overline{Y}$.

Hétérogènes à l'intérieur, semblables entre eux

Décomposer la dispersion totale

Les k échantillons partitionnent U : la somme des carrés totale se décompose en

$$\sum_{j\in U}(y_j-\overline{Y})^2 = \underbrace{\sum_{r=1}^{k}\sum_{j\in s_r}(y_j-\overline{y}_r)^2}_{\text{intra}} + \underbrace{n\sum_{r=1}^{k}(\overline{y}_r-\overline{Y})^2}_{\text{inter} = N\,V(\overline{y})}$$

C'est la décomposition de la variance de la statistique descriptive, appliquée aux k échantillons. Le total étant fixé, $V(\overline{y})$ est faible quand **chaque échantillon est aussi dispersé que la population**.

Démonstration : La variance du tirage systématique et sa décomposition

On découpe la somme par échantillons. Les k échantillons partitionnent U : $\sum_{j\in U}(y_j-\overline{Y})^2 = \sum_{r=1}^{k}\sum_{j\in s_r}(y_j-\overline{Y})^2$. Dans l'échantillon s_r, on écrit $y_j-\overline{Y} = (y_j-\overline{y}_r)+(\overline{y}_r-\overline{Y})$; en élevant

au carré, le double produit disparaît car $\sum_{j\in s_r}(y_j-\overline{y}_r)=0$, et

$$\sum_{j\in s_r}(y_j-\overline{Y})^2=\sum_{j\in s_r}(y_j-\overline{y}_r)^2+n\,(\overline{y}_r-\overline{Y})^2.$$

En additionnant sur r, on obtient la somme des carrés intra plus la somme des carrés inter.

L'inter vaut $N\,V(\overline{y})$. Comme $E(\overline{y})=\overline{Y}$, $V(\overline{y})=\frac{1}{k}\sum_r(\overline{y}_r-\overline{Y})^2$, donc

$$n\sum_{r=1}^{k}(\overline{y}_r-\overline{Y})^2=n\,k\,V(\overline{y})=N\,V(\overline{y})\qquad\text{d'où}\qquad V(\overline{y})=\frac{1}{N}\Big[(N-1)\,S^2-\text{intra}\Big].$$

Une liste rangée au hasard. Si l'ordre de la liste est aléatoire, chaque colonne est un échantillon aléatoire simple de n unités : $E(\overline{y}_r-\overline{Y})^2=(1-f)\,S^2/n$ pour chaque r. Donc $E\big[V(\overline{y})\big]=(1-f)\,S^2/n$, la variance du sondage aléatoire simple, et la part inter de la dispersion totale vaut en moyenne

$$\frac{N\,(1-f)\,S^2/n}{(N-1)\,S^2}=\frac{N\,(1-f)}{n\,(N-1)}.$$

Ce que cela dit. Le total étant fixé, $V(\overline{y})$ est faible quand l'intra est presque égal au total : quand chaque échantillon est aussi dispersé que la population. Sur la population du cours, l'inter vaut $2\,000\times30\,287=60{,}6$ millions, soit 0,19 % de $(N-1)S^2=31{,}7$ milliards ; pour une liste au hasard, la formule donne $0{,}9\times2\,000/(200\times1\,999)=0{,}45\,\%$. Le gain du fichier vient de son ordre, qui rend les échantillons plus semblables entre eux.

■

Sur la population du cours

Inter : $2\,000\times30\,287=60{,}6$ millions, soit 0,19 % de la somme totale (31,7 milliards). Pour une liste rangée au hasard, ce serait en moyenne 0,45 % : les échantillons du fichier sont plus dispersés à l'intérieur, donc plus semblables entre eux.

L'ordre de la liste

L'ordre du fichier : une stratification implicite

Ce que contient chaque échantillon

Le fichier range les ménages par province, milieu et district. Chaque échantillon compte donc **exactement** 82, 56, 38 et 24 ménages des quatre provinces (l'allocation proportionnelle du chapitre 6), et de 1 à 8 ménages de chaque district.

Stratification implicite

Trier la liste selon une variable puis tirer systématiquement revient à **stratifier** selon elle : chaque bloc de k voisins est une petite strate où l'on tire une unité.

Plan, $n = 200$	Variance	Rapport au SAS
Sondage aléatoire simple	71 435	1
Stratifié proportionnel par province	68 140	0,954
Stratifié proportionnel par district	60 057	0,841
Systématique, ordre du fichier	30 287	0,424

Une précision qui dépend de l'ordre exact

Les mêmes ménages, rangés autrement

Ordre de la liste, $k = 10$	Rapport au SAS
Entièrement aléatoire, moyenne de 2 000 ordres	0,98
Aléatoire dans chaque district, moyenne	0,84
Ordre du fichier	0,424

Le même fichier avec un autre pas : 1,34 pour $k = 4$, 0,51 pour $k = 5$, 0,78 pour $k = 20$.

Ne pas compter sur la chance

Mélanger les ménages **à l'intérieur** de chaque district donne en moyenne le gain de la stratification par district : c'est ce que garantit le rangement. Le 0,424 tient en plus à l'ordre exact des ménages, favorable ici (seuls 12 % des ordres font aussi bien) ; avec $k = 4$, le même fichier fait **pire** que le hasard. Pour prévoir une taille d'échantillon, on ne retient que le gain garanti.

Trier selon une variable liée à la dépense

Trier par la taille, puis par la dépense

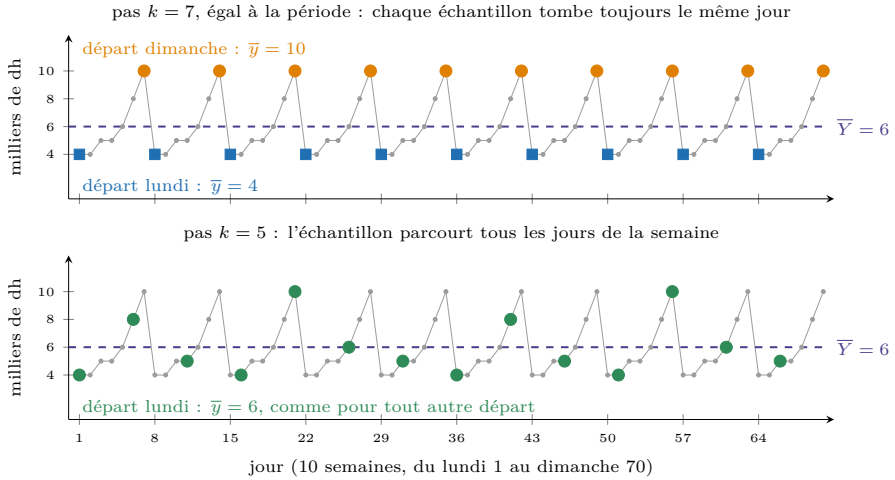
- **Par la taille du ménage**, connue du recensement et corrélée à 0,622 avec la dépense : $V = 18\,138$, rapport 0,254. À taille égale, les ménages restent dans l'ordre du fichier : la liste est stratifiée par taille, puis par province et district.
- **Par la dépense elle-même** : $V = 1\,954$, rapport 0,027. Les dix moyennes vont de 7 847 à 7 984 dh.

Irréaliste, mais instructif

Personne ne connaît la dépense avant l'enquête : sinon, on ne la ferait pas. Mais la leçon est générale : plus l'ordre de la liste suit y, plus chaque bloc de k voisins est homogène, et plus le tirage systématique est précis. En pratique, on trie la base selon une **variable auxiliaire** connue et liée à y.

Le danger de la périodicité

Un cycle dans la liste



La recette d'un commerce sur dix semaines

Recette journalière, en milliers de dh : 4 le lundi et le mardi, 5 le mercredi et le jeudi, 6 le vendredi, 8 le samedi, 10 le dimanche, jour de souk. $\overline{Y} = 6$. On tire un jour sur k.

Le pas tombe sur la période

Deux pas, deux résultats opposés

Pas	n	Moyennes possibles	$V(\overline{y})$	SAS
$k = 7$	10	une par jour, de 4 à 10	4,286	0,373
$k = 5$	14	6 pour les cinq départs	0	0,248

Avec $k = 7$, la variance est 11,5 fois celle du sondage aléatoire simple. Avec $k = 5$, chaque échantillon parcourt deux fois toute la semaine.

Démonstration : Le pas tombe sur la période, et comment l'éviter

La liste. Supposons la liste périodique, de période p (ici $p = 7$) : $y_{j+p} = y_j$, avec p qui divise N. La moyenne $\overline{Y}$ est celle d'une période.

Le pas est un multiple de la période. Si p divise k, les rangs r, $r + k$, $r + 2k$, ... sont tous congrus à r modulo p : ils portent la **même valeur** y_r. Donc $\overline{y}_r = y_r$: l'échantillon est constant, et $\overline{y}$ prend la valeur d'un seul jour de la période. Sa variance est celle du cycle, et ne diminue plus quand n augmente. Pour la recette de la semaine ($k = 7$) :

$$V(\overline{y}) = \frac{1}{7}\left[(4-6)^2 + (4-6)^2 + (5-6)^2 + (5-6)^2 + 0 + (8-6)^2 + (10-6)^2\right] = \frac{30}{7} = 4{,}286.$$

Le pas est premier avec la période. Si k et p n'ont aucun diviseur commun et que p divise n, les rangs $r + i\,k$ pour $i = 0, \ldots, p-1$ ont des restes tous différents modulo p. En effet, si

$r + i\,k \equiv r + i'\,k$, alors p divise $(i - i')\,k$, donc $(i - i')$, ce qui impose $i = i'$. Les $n = m\,p$ rangs parcourent donc m fois chaque reste, soit m fois chaque jour de la période : $\overline{y}_r = \overline{Y}$ pour tous les départs, et $V(\overline{y}) = 0$. Pour $k = 5$ et $n = 14$, les cinq moyennes valent 6.

La comparaison. La variance des 70 jours est $S^2 = 300/69$; le sondage aléatoire simple donne $(1 - 10/70)\,S^2/10 = 0{,}373$ pour $n = 10$ et $(1 - 14/70)\,S^2/14 = 0{,}248$ pour $n = 14$. Pour $k = 7$, le rapport est $4{,}286/0{,}373 = 11{,}5$.

L'échantillon ne voit rien. Avec le départ dimanche, les dix valeurs tirées valent toutes 10 : $\overline{y} = 10$, soit $+67\,\%$, et $s^2 = 0$, donc la variance estimée est nulle.

Ce que cela dit. Le danger vient du rapport entre le pas et la période, pas de la taille de l'échantillon. Un pas multiple de la période est le pire cas ; un pas premier avec elle (et p divisant n) est le meilleur, d'où la parade du cours. Dans l'immeuble de l'exercice 2 (période 4), $k = 4$ donne $V = 0{,}75$ contre 0,058 pour le sondage aléatoire simple, et $k = 5$ avec $n = 8$ donne $\overline{y} = 3{,}5$ pour tous les départs.

■

Une erreur que l'échantillon ne voit pas

Le départ dimanche donne $\overline{y} = 10$: une surestimation de $67\,\%$. Pire, les dix recettes tirées sont égales, donc $s^2 = 0$ et la variance estimée est **nulle** : l'échantillon se déclare parfaitement précis. Rien, dans les données recueillies, ne signale le problème.

Se prémunir

Trois parades

1. **Connaître la liste** : chercher les cycles avant de tirer.
2. **Choisir un pas premier avec la période** : $k = 5$ ou 10 plutôt que 7 ou 14.
3. **Tirer plusieurs départs**, ou mélanger la liste à l'intérieur de blocs.

D'où viennent les cycles

De la façon dont la liste a été **fabriquée** : relevés quotidiens (sept jours), relevés mensuels (douze mois), immeubles dont chaque étage a le même plan, rues numérotées avec les pairs d'un côté et les impairs de l'autre. Une périodicité approchée suffit à dégrader la précision.

Estimer la variance en pratique

Faire comme si c'était un sondage aléatoire simple

La formule du chapitre 3

Faute d'estimateur sans biais, on applique celui du sondage aléatoire simple :

$$\widehat{V}(\overline{y}) = (1 - f)\,\frac{s^2}{n}$$

L'échantillon de départ 9

$\overline{y} = 7\,922{,}95$ dh et $s = 3\,995{,}46$ dh, d'où $\widehat{V}(\overline{y}) = 0{,}9 \times s^2/200 = 71\,836{,}8$ et une erreur-type de 268,0 dh. Intervalle à 95 % : [7 397,6 ; 8 448,3].

Prudent sur une liste rangée, trompeur sur un cycle

Une variance surestimée

Selon le départ, $\widehat{V}$ va de 54 571 à 92 488 ; il vaut 71 622 en moyenne sur les dix départs, alors que la vraie variance est 30 287 : 2,4 fois trop. L'intervalle annoncé est trop large : l'erreur est **du bon côté**, et l'on perd seulement le bénéfice de la stratification implicite dans l'affichage de la précision.

Une variance sous-estimée

Sur une liste périodique, c'est l'inverse : le départ dimanche donne $\widehat{V} = 0$ alors que la vraie variance est 11,5 fois celle du sondage aléatoire simple. La formule du sondage aléatoire simple n'est prudente qu'**après** avoir vérifié que la liste ne cache pas de cycle.

Les différences successives

Comparer des voisins

Soit $z_1, z_2, \ldots, z_n$ les valeurs tirées, **dans l'ordre de la liste** :

$$s_d^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (z_{i+1} - z_i)^2, \qquad \widehat{V}_d(\overline{y}) = (1-f)\,\frac{s_d^2}{n}$$

Pourquoi diviser par deux

Deux voisins indépendants, de même moyenne et de variance σ^2, vérifient $E\big[(z_{i+1} - z_i)^2\big] = V(z_{i+1}) + V(z_i) = 2\sigma^2$. Voisins dans la liste, ils sont dans la même strate implicite : s_d^2 mesure la dispersion **à l'intérieur** des strates, sans la tendance de la liste.

Démonstration : Pourquoi les différences successives suivent l'ordre de la liste

Deux voisins indépendants. Si z_i et z_{i+1} sont indépendants, de même moyenne et de même variance σ^2, alors $E\big[(z_{i+1} - z_i)^2\big] = V(z_{i+1} - z_i) = 2\sigma^2$. Donc

$$E(s_d^2) = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} 2\sigma^2 = \sigma^2,$$

et s_d^2 est sans biais pour σ^2.

Avec une tendance. Supposons que la liste monte régulièrement : $z_i = a + b\,i + \varepsilon_i$, les ε_i étant indépendants, centrés, de variance σ^2.

- Pour les différences, $z_{i+1} - z_i = b + (\varepsilon_{i+1} - \varepsilon_i)$, donc $E[(z_{i+1} - z_i)^2] = b^2 + 2\sigma^2$ et

$$E(s_d^2) = \sigma^2 + \frac{b^2}{2}.$$

- Pour la variance ordinaire, $z_i - \overline{z} = b\,(i - \overline{\imath}) + (\varepsilon_i - \overline{\varepsilon})$. Le double produit est d'espérance nulle et $E\sum_i (\varepsilon_i - \overline{\varepsilon})^2 = (n-1)\,\sigma^2$; avec $\sum_{i=1}^{n} (i - \overline{\imath})^2 = n(n^2-1)/12$,

$$E(s^2) = \sigma^2 + b^2\,\frac{n\,(n+1)}{12}.$$

Ce que cela dit. La tendance de la liste gonfle s^2 de $b^2\,n(n+1)/12$ (plus de $3\,000\,b^2$ pour $n = 200$) mais s_d^2 de $b^2/2$ seulement : les différences comparent des voisins, qui se trouvent dans la même strate implicite, et ignorent la pente de la liste. C'est ce qui permet à $\widehat{V}_d$ de suivre le gain de la stratification implicite. Mais si la liste est triée sur y lui-même, la pente est tout ce qui reste : $\widehat{V}_d$ tombe à 427 contre 1 954 pour la vraie variance.

■

Deux estimateurs face à trois ordres

Moyennes sur les dix départs

Départ 9 dans l'ordre du fichier : $s_d = 3\,739{,}0$ dh, $\widehat{V}_d = 62\,910$, erreur-type 250,8 dh.

Ordre de la liste	V exacte	$\widehat{V}$	$\widehat{V}_d$
Ordre du fichier	30 287	71 622	62 452
Trié par taille	18 138	71 677	37 998
Trié par dépense	1 954	71 750	427

Ce que voit chaque estimateur

$\widehat{V}$ ignore l'ordre de la liste ; $\widehat{V}_d$ le suit. Dans l'ordre du fichier, $\widehat{V}_d$ retrouve à peu près la variance du stratifié par district (60 057) : le gain garanti, pas la chance. Mais si la liste est triée sur y lui-même, il **sous-estime** : 427 contre 1 954.

Le tirage systématique répété

Plusieurs départs au lieu d'un

Avec un pas K plus grand, on tire m **départs** par sondage aléatoire simple parmi K. Les m moyennes forment un sondage aléatoire simple parmi les K possibles, d'où un estimateur de variance **sans biais** :

$$\overline{y} = \frac{1}{m}\sum_{i=1}^{m} \overline{y}_i, \qquad \widehat{V}(\overline{y}) = \Big(1 - \frac{m}{K}\Big)\frac{s_c^2}{m}, \quad s_c^2 : \text{variance des } m \text{ moyennes}$$

Démonstration : Plusieurs départs, un sondage aléatoire simple de colonnes

Des colonnes. Avec le pas K, la liste se partage en K colonnes disjointes de $n' = N/K$ unités, chacune étant un échantillon systématique. Tirer m départs parmi K par sondage aléatoire simple revient à tirer m colonnes parmi K. Notons $\overline{y}_r$ la moyenne de la colonne r ; l'estimateur est la moyenne $\overline{y}$ des m moyennes tirées. Les colonnes partitionnent U et ont la même taille : la moyenne des K moyennes de colonnes est $\overline{Y}$.

Le chapitre 3 sur les moyennes de colonnes. Les K moyennes de colonnes forment une population de K nombres, de moyenne $\overline{Y}$ et de variance $S_c^2 = \frac{1}{K-1}\sum_r (\overline{y}_r - \overline{Y})^2$; on en tire m par sondage aléatoire simple. Les résultats du chapitre 3 donnent :

- $E(\overline{y}) = \overline{Y}$: l'estimateur est sans biais ;
- $V(\overline{y}) = \left(1 - \dfrac{m}{K}\right)\dfrac{S_c^2}{m}$;
- la variance s_c^2 des m moyennes tirées est sans biais pour S_c^2, donc $\widehat{V}(\overline{y}) = \left(1 - \dfrac{m}{K}\right)\dfrac{s_c^2}{m}$ est un estimateur **sans biais** de $V(\overline{y})$.

Ce que cela dit. On obtient ce que le départ unique refuse : un estimateur sans biais de la variance, parce que l'on observe plusieurs colonnes et que l'on peut comparer leurs moyennes. Le prix est une précision moindre, car des colonnes de n' unités équivalent à des blocs de K voisins, une stratification moins fine : pour $K = 50$ et $m = 5$, la vraie variance est 62 168 (rapport 0,870 au sondage aléatoire simple), contre 30 287 avec un seul départ de pas 10.

■

Cinq départs de quarante ménages

$K = 50$, $m = 5$, départs 2, 9, 19, 32, 40 (graine 2026) ; moyennes 8 043 ; 7 558 ; 8 635 ; 8 441 ; 8 595. Estimation 8 254,5 dh, $s_c = 454{,}1$, $\widehat{V} = 0{,}9 \times s_c^2/5 = 37\,125$. Le prix : la vraie variance du plan est 62 168 (rapport 0,870) contre 30 287 avec un seul départ, car des blocs de 50 voisins stratifient moins finement.

S'entraîner

Exercice 1 : douze boutiques rangées

Une liste triée par chiffre d'affaires

Douze boutiques, rangées par chiffre d'affaires croissant, réalisent 2, 3, 3, 4, 5, 6, 6, 7, 8, 8, 9 et 11 milliers de dh par mois. On en tire 3 par tirage systématique.

1. Quel est le pas ? Énumérer les échantillons possibles et leurs moyennes.
2. Vérifier que $E(\overline{y}) = \overline{Y}$.
3. Calculer $V(\overline{y})$ et la comparer à celle d'un sondage aléatoire simple de même taille, sachant que $S^2 = 82/11$.
4. D'où vient la différence ?

Exercice 1 : solution

Les quatre échantillons

1. $k = 12/3 = 4$. Rangs $1, 5, 9$: $2, 5, 8$, moyenne 5. Rangs $2, 6, 10$: $3, 6, 8$, moyenne $17/3 = 5{,}67$. Rangs $3, 7, 11$: $3, 6, 9$, moyenne 6. Rangs $4, 8, 12$: $4, 7, 11$, moyenne $22/3 = 7{,}33$.
2. $(5 + 17/3 + 6 + 22/3)/4 = 24/4 = 6$, et $\overline{Y} = 72/12 = 6$.

Variance et explication

3. Écarts à 6 : -1 ; $-1/3$; 0 ; $4/3$. $V(\overline{y}) = (1+1/9+0+16/9)/4 = 26/36 = 0{,}722$. Sondage aléatoire simple : $(1 - 3/12) \times (82/11)/3 = 1{,}864$. Rapport : 0,39.
4. La liste est triée selon y : chaque échantillon prend une boutique dans chacun des trois blocs de quatre (petites, moyennes, grandes). C'est une stratification implicite en trois strates.

Exercice 2 : un immeuble de dix étages

Le nombre de pièces des appartements

Un immeuble de 10 étages compte 4 appartements par étage, listés étage par étage. Le premier de chaque étage, à l'angle, a 5 pièces ; les trois autres en ont 3.

1. Calculer $\overline{Y}$ et S^2.
2. Tirage systématique de pas $k = 4$: donner les moyennes possibles, calculer $V(\overline{y})$ et la comparer à celle d'un sondage aléatoire simple de même taille.
3. Même question avec $k = 5$.
4. Quelle leçon pratique en tirer ?

Exercice 2 : solution

Un cycle de période quatre

1. $\overline{Y} = (10 \times 5 + 30 \times 3)/40 = 3{,}5$; $S^2 = (10 \times 1{,}5^2 + 30 \times 0{,}5^2)/39 = 30/39 = 0{,}769$.
2. $n = 10$. Le départ 1 ne prend que des appartements d'angle ($\overline{y} = 5$), les trois autres départs n'en prennent aucun ($\overline{y} = 3$). $V(\overline{y}) = (1{,}5^2 + 3 \times 0{,}5^2)/4 = 0{,}75$, contre $(1 - 10/40) \times 0{,}769/10 = 0{,}058$: 13 fois plus.

Changer de pas

3. $n = 8$. Avec $k = 5$, les rangs tirés parcourent deux fois les quatre positions de l'étage : chaque échantillon contient 2 appartements d'angle et 6 autres, et $\overline{y} = 28/8 = 3{,}5$ pour tous les départs. $V(\overline{y}) = 0$.
4. Repérer la période de la liste et choisir un pas premier avec elle.

Exercice 3 : les ménages connectés

Une proportion dans l'ordre du fichier

On tire 200 ménages de la population du cours par tirage systématique dans l'ordre du fichier ($k = 10$). Les dix proportions de ménages connectés possibles sont : 0,710 ; 0,685 ; 0,665 ; 0,635 ; 0,675 ; 0,685 ; 0,665 ; 0,630 ; 0,695 ; 0,695.

1. Vérifier que leur moyenne est $P = 0{,}674$.
2. Calculer $V(\widehat{P})$ et la comparer à celle du sondage aléatoire simple, 0,000989 (chapitre 4).
3. En déduire la marge d'erreur à 95 % et la taille effective.

Exercice 3 : solution

Moyenne et variance

1. La somme des dix proportions vaut 6,74, d'où la moyenne $0{,}674 = P$.
2. Écarts à 0,674, en millièmes : 36 ; 11 ; −9 ; −39 ; 1 ; 11 ; −9 ; −44 ; 21 ; 21. Somme des carrés : 6 040 millionièmes, d'où $V(\widehat{P}) = 0{,}000604$ et une erreur-type de 0,0246, contre 0,0315. Rapport : 0,611.

Marge et taille effective

3. Marge : $1{,}96 \times 0{,}0246 = 0{,}048$, soit ±4,8 points au lieu de ±6,2. Taille effective : $200/0{,}611 \approx 328$ ménages. L'accès à internet dépend fortement du milieu (0,768 en ville, 0,434 à la campagne), et le fichier est rangé par province et par milieu : stratification implicite.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Un départ r au hasard parmi $k = N/n$, puis une unité tous les k rangs.
2. k échantillons possibles : $\pi_j = 1/k = n/N$, mais $\pi_{jl} = 0$ pour la plupart des paires.
3. $\overline{y}$ est sans biais ; sa variance $\frac{1}{k}\sum_r (\overline{y}_r - \overline{Y})^2$ n'a pas d'estimateur sans biais.
4. Ranger la liste selon une variable liée à y crée une **stratification implicite** : rapport 0,424 (fichier), 0,254 (tri par taille).
5. Le gain exact dépend de l'ordre précis : pour prévoir, on ne compte que sur le gain garanti.
6. Une liste **périodique** peut ruiner l'estimation sans que l'échantillon le révèle.
7. Variance estimée : formule du SAS, différences successives ou plusieurs départs.

La suite

Le tirage systématique tire une seule grappe d'unités dispersées. Le chapitre 9 tire au contraire plusieurs **grappes** d'unités voisines, les districts : moins cher sur le terrain, mais moins précis.