Échantillonnage

Chapitre 8 : Le sondage systématique

Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 31 octobre 2026

Plan du chapitre

Un ménage sur k

Probabilités d'inclusion et estimateur

La variance exacte

L'ordre de la liste

Le danger de la périodicité

Estimer la variance en pratique

S'entraîner

Ce qu'il faut retenir

Un ménage sur k

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 $1, 2, \dots, k$.
2. Prendre les unités de rangs $r,\ r+k,\ r+2k,\ \dots,\ r+(n-1)\,k$.

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 $9, 19, 29, \dots, 1\,999$.

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).

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.

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 $\lceil u \rceil, \lceil u+k \rceil, \ldots, \lceil u+(n-1)\,k \rceil$, 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.

Probabilités d’inclusion et estimateur

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 :

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

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 $\pi_{jl} = 0$.

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 $\overline{Y}$.

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 $\overline{y}_{r\cdot}$, et une seule valeur ne dit rien de leur dispersion.

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

La variance du tirage systématique

$\overline{y}$ prend les k valeurs $\overline{y}_1, \ldots, \overline{y}_{k'}$ 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 moyennes, arrondies au dirham

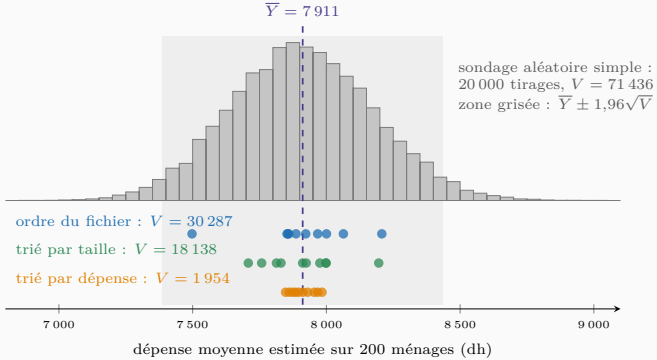
Départ r	1	2	3	4	5
Moyenne	8 000	7 853	7 860	7 856	8 064
Départ r	6	7	8	9	10
Moyenne	7 887	8 207	7 498	7 923	7 968

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

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) : $200/0{,}424 \approx 472$ 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 $\overline{Y}$.

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**.

Sur la population du cours

Inter : $2\,000 \times 30\,287 = 60{,}6$ 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

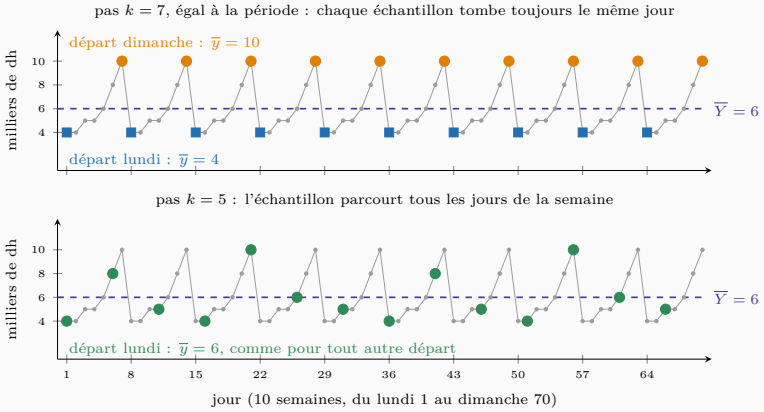
- **Par la taille du ménage**, connue du recensement et corrélée à $0{,}622$ avec la dépense : $V = 18\,138$, rapport $0{,}254$. À taille égale, les ménages restent dans l'ordre du fichier : la liste est stratifiée par taille, puis par province et district.
- **Par la dépense elle-même** : $V = 1\,954$,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

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.

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.

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

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 $95\,\%$: $[7\,397{,}6\,;\,8\,448{,}3]$.

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.

Comparer des voisins

Soit $z_1, z_2, \dots, z_n$ 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 $E\left[(z_{i+1} - z_i)^2\right] = V(z_{i+1}) + V(z_i) = 2\sigma^2$. 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.

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 $1\,954$.

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** :

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

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

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.

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.

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.

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 $\pm 4{,}8$ points au lieu de $\pm 6{,}2$. 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

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.