Échantillonnage

Chapitre 9 : Le sondage par grappes

Jaouad Madkour

30 octobre 2026

Plan du chapitre

Tirer des paquets de ménages **2**
Quand la liste des ménages manque . . . 2
Les notations . . . 2
Les cinquante districts du cours . . . 3

Estimer à partir de grappes entières **3**
Tirer des districts entiers . . . 3
L'estimateur du total . . . 4
La moyenne par ménage : deux estimateurs . . . 5
Un tirage de cinq districts . . . 6
Estimer la variance sur l'échantillon . . . 6

L'effet de grappe **7**
La corrélation intra-grappe . . . 7
L'effet de plan des grappes . . . 7
Sur la population du cours . . . 8
L'effet de plan selon la taille des grappes . . . 9

Des grappes de tailles inégales **9**
Pourquoi Horvitz-Thompson souffre . . . 9
Horvitz-Thompson contre ratio . . . 9
Tous les échantillons de cinq districts . . . 10
Faire mieux . . . 10

Coût et précision **11**
Le prix d'un ménage . . . 11
À budget égal, deux terrains . . . 12
Quand tirer par grappes . . . 12

S'entraîner **12**
Exercice 1 : trois immeubles . . . 12
Exercice 1 : solution . . . 13
Exercice 2 : des douars de 25 ménages . . . 13
Exercice 2 : solution . . . 13

Exercice 3 : grappes ou aléatoire simple ? 14
Exercice 3 : solution 14

Ce qu'il faut retenir **14**
Le résumé du chapitre 14

Tirer des paquets de ménages

Quand la liste des ménages manque

Le sondage par grappes

La population est partagée en N_I **grappes** : des groupes d'unités voisines (district, douar, immeuble, classe). On tire n_I grappes, puis l'on interroge **toutes** les unités des grappes tirées.

Deux raisons de tirer par grappes

1. **Pas de liste de ménages** : une liste des districts suffit. On ne dénombre les ménages que dans les districts tirés.
2. **Le coût des déplacements** : 200 ménages tirés au hasard sont dispersés dans toute la région ; 5 districts de 40 ménages se visitent en quelques jours, par une équipe qui reste sur place.

Les notations

Une grappe, deux nombres

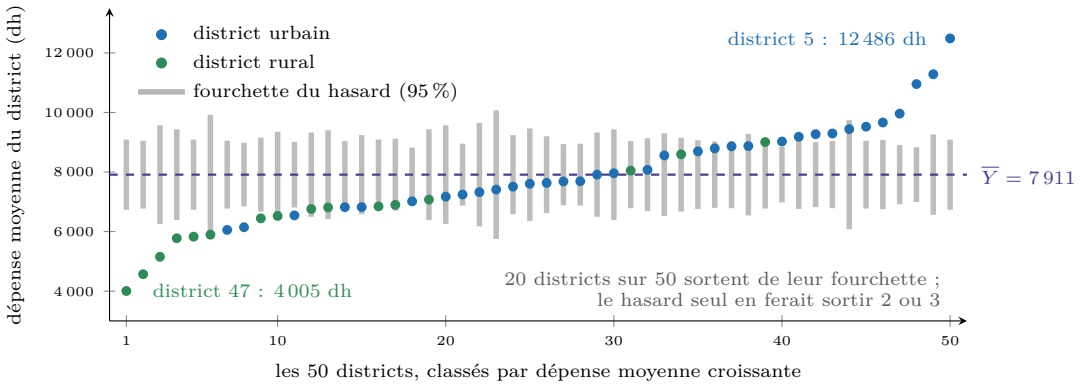
La grappe i compte M_i unités ; son total est $Y_i = \sum_{k \in i} y_k$ et sa moyenne $\overline{Y}_i = Y_i / M_i$. Alors

$$N = \sum_{i=1}^{N_I} M_i \qquad Y = \sum_{i=1}^{N_I} Y_i \qquad \overline{M} = \frac{N}{N_I}$$

Les districts du cours

Grandeur	Valeur
Nombre de districts N_I	50
Taille moyenne $\overline{M} = 2\,000/50$	40 ménages
Plus grand : district 1 (Tanger-Assilah)	71 ménages
Plus petit : district 46 (Al Hoceïma)	13 ménages
Total du district 1, Y_1	498 040 dh

Les cinquante districts du cours



Les voisins se ressemblent

Si les ménages étaient répartis au hasard entre les districts, les moyennes resteraient presque toutes dans les fourchettes grises. Elles vont de 4 005 à 12 486 dh, et 20 **districts sur** 50 en sortent : les ménages d'un même district partagent un quartier, des prix, un accès aux services.

Estimer à partir de grappes entières

Tirer des districts entiers

Le plan

On tire n_I districts parmi N_I par sondage aléatoire simple. Un ménage est tiré si et seulement si son district l'est :

$$\pi_k = \frac{n_I}{N_I} \qquad \pi_{kl} = \begin{cases} n_I/N_I & \text{même district} \\ \dfrac{n_I\,(n_I-1)}{N_I\,(N_I-1)} & \text{districts différents} \end{cases}$$

Démonstration : Les probabilités d'inclusion du tirage de grappes

Les grappes. Le tirage est un sondage aléatoire simple de n_I grappes parmi N_I : il y a $\binom{N_I}{n_I}$ échantillons de grappes, équiprobables. Une grappe donnée figure dans $\binom{N_I-1}{n_I-1}$ d'entre eux, et deux grappes données dans $\binom{N_I-2}{n_I-2}$. Donc

$$P(i \in S_I) = \frac{\binom{N_I-1}{n_I-1}}{\binom{N_I}{n_I}} = \frac{n_I}{N_I}, \qquad P(i \in S_I \text{ et } i' \in S_I) = \frac{\binom{N_I-2}{n_I-2}}{\binom{N_I}{n_I}} = \frac{n_I\,(n_I-1)}{N_I\,(N_I-1)}.$$

Les ménages. Un ménage k du district i est tiré si et seulement si son district l'est : $\pi_k = P(i \in S_I) = n_I/N_I$. Deux ménages du même district sont tirés ensemble si et seulement si leur district l'est : $\pi_{kl} = n_I/N_I$. Deux ménages de districts différents i et i' sont tirés ensemble si et seulement si les deux districts le sont : $\pi_{kl} = n_I(n_I-1)/(N_I(N_I-1))$.

La taille de l'échantillon est aléatoire. Elle vaut $n = \sum_{i\in S_I} M_i$, et $E(n) = \sum_{k\in U} \pi_k =$

$N \dfrac{n_I}{N_I} = n_I \overline{M}$.

Ce que cela dit. Pour $n_I = 5$ districts sur 50 : $\pi_k = 0{,}1$ pour tous les ménages, $\pi_{kl} = 0{,}1$ pour deux voisins et $5 \times 4/(50 \times 49) = 0{,}0082$ pour deux ménages de districts différents. La taille attendue est $5 \times 40 = 200$ ménages ; selon les districts tirés, elle varie de 88 (les cinq plus petits) à 325 (les cinq plus grands).

■

Cinq districts parmi cinquante

$\pi_k = 5/50 = 0{,}1$ pour tous les ménages, comme dans un sondage aléatoire simple de 200 ménages. Mais deux voisins sont tirés ensemble avec la probabilité 0,1, deux ménages de districts différents avec 0,0082 seulement. Et n n'est pas fixé : de 88 ménages (les cinq plus petits districts) à 325 (les cinq plus grands), 200 en moyenne.

L'estimateur du total

Horvitz-Thompson appliqué aux grappes

Chaque ménage d'un district tiré a le poids $d_k = N_I/n_I$:

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{N_I}{n_I} \sum_{i \in S_I} Y_i = N_I \, \overline{y}_I$$

où S_I est l'ensemble des districts tirés et $\overline{y}_I$ la moyenne de leurs **totaux**.

La grappe devient l'unité

C'est un sondage aléatoire simple de n_I unités parmi N_I, sur la variable Y_i. Le chapitre 3 donne donc, sans nouveau calcul, l'absence de biais et

$$V(\widehat{Y}_{HT}) = N_I^2 \, (1 - f_I) \, \frac{S_T^2}{n_I} \qquad \text{avec } f_I = \frac{n_I}{N_I}$$

où S_T^2 est la variance des N_I totaux Y_i.

Démonstration : Horvitz-Thompson avec des grappes, sans biais et sa variance

Le poids. Tous les ménages d'un district tiré sont interrogés et ont la même probabilité d'inclusion $\pi_k = n_I/N_I$, donc le même poids N_I/n_I :

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{N_I}{n_I} \sum_{i \in S_I} \sum_{k \in i} y_k = \frac{N_I}{n_I} \sum_{i \in S_I} Y_i = N_I \, \overline{y}_I,$$

où $\overline{y}_I$ est la moyenne des totaux Y_i des n_I districts tirés.

La grappe devient l'unité. Les n_I districts sont tirés par sondage aléatoire simple parmi N_I, et la variable observée sur le district i est Y_i. Par le chapitre 3, $\overline{y}_I$ est sans biais pour la

moyenne Y/N_I des Y_i, et de variance $(1-f_I)\,S_T^2/n_I$. En multipliant par N_I :

$$E(\widehat{Y}_{HT}) = Y, \qquad V(\widehat{Y}_{HT}) = N_I^2\,(1-f_I)\,\frac{S_T^2}{n_I}.$$

Ce que cela dit. Rien de nouveau à démontrer : une grappe tirée est une observation. La précision dépend du nombre de **grappes**, n_I, pas du nombre de ménages. Sur les trois immeubles de l'exercice 1 (totaux 10, 16, 14 ; $N_I = 3$, $n_I = 2$), les trois échantillons donnent $\widehat{Y}_{HT} = \frac{3}{2} \times (26, 24, 30) = 39, 36, 45$: leur moyenne est $40 = Y$ et leur variance $14 = 3^2 \times \frac{1}{3} \times \frac{28/3}{2}$.

■

La moyenne par ménage : deux estimateurs

Diviser par le nombre attendu ou effectif de ménages

$$\frac{\widehat{Y}_{HT}}{N} = \frac{\sum_{S_I} Y_i}{n_I\,\overline{M}} \qquad\qquad \overline{y}_R = \frac{\sum_{S_I} Y_i}{\sum_{S_I} M_i}$$

Le second, l'**estimateur par le ratio**, est la moyenne des ménages effectivement enquêtés. C'est aussi $\widehat{Y}_{HT}/\widehat{N}_{HT}$, avec $\widehat{N}_{HT} = \frac{N_I}{n_I}\sum_{S_I} M_i$: il n'exige pas de connaître N.

Les deux variances

$$V\Big(\frac{\widehat{Y}_{HT}}{N}\Big) = \frac{1-f_I}{n_I\,\overline{M}^2}\,S_T^2 \qquad\qquad V(\overline{y}_R) \approx \frac{1-f_I}{n_I\,\overline{M}^2}\,S_e^2$$

S_e^2 est la variance des $e_i = Y_i - \overline{Y}M_i = M_i\,(\overline{Y}_i - \overline{Y})$. Le ratio a un léger biais, d'ordre $1/n_I$ (chapitre 12).

Démonstration : Le ratio, sa variance et le rôle des résidus

L'écart du ratio, exactement. Posons $e_i = Y_i - \overline{Y}\,M_i = M_i\,(\overline{Y}_i - \overline{Y})$. Alors

$$\overline{y}_R - \overline{Y} = \frac{\sum_{i\in S_I} Y_i}{\sum_{i\in S_I} M_i} - \overline{Y} = \frac{\sum_{i\in S_I}(Y_i - \overline{Y}\,M_i)}{\sum_{i\in S_I} M_i} = \frac{\sum_{i\in S_I} e_i}{\sum_{i\in S_I} M_i}.$$

Sur toute la population, $\sum_i e_i = Y - \overline{Y}\,N = 0$: les e_i ont pour moyenne 0.

La linéarisation. Le dénominateur $\sum_{S_I} M_i$ est aléatoire, d'espérance $n_I\,\overline{M}$. Remplaçons-le par cette espérance :

$$\overline{y}_R - \overline{Y} \approx \frac{1}{\overline{M}} \cdot \frac{1}{n_I} \sum_{i\in S_I} e_i.$$

Le second facteur est la moyenne de n_I valeurs e_i tirées par sondage aléatoire simple dans une population de N_I valeurs de moyenne 0 : par le chapitre 3, elle est centrée et de variance $(1-f_I)\,S_e^2/n_I$. D'où

$$V(\overline{y}_R) \approx \frac{1-f_I}{n_I\,\overline{M}^2}\,S_e^2.$$

La comparaison avec $\widehat{Y}_{HT}/N$. Comme $\widehat{Y}_{HT}/N = \sum_{S_I} Y_i/(n_I\,\overline{M})$, sa variance est la même avec S_T^2 à la place de S_e^2. Or $Y_i = \overline{Y}\,M_i + e_i$: le total d'un district contient sa **taille**, dont le ratio retire l'effet.

L'estimation. Sur l'échantillon, les e_i sont inconnus ; on les remplace par $\hat{e}_i = Y_i - \overline{y}_R\,M_i$, dont la somme sur les districts tirés est nulle par construction de $\overline{y}_R$.

Ce que cela dit. Le ratio corrige la taille de l'échantillon obtenu, et sa variance ne retient que les écarts de moyennes de districts. Sur les 2 118 760 échantillons de cinq districts, l'écart-type exact du ratio est 768 dh (la formule approchée donne 774) contre 1 516,6 pour $\widehat{Y}_{HT}/N$; le biais du ratio est de −25 dh. Sur les trois immeubles, les ratios 6,5 ; 8 ; 10 ont pour moyenne 8,17 ≠ 8 : un biais de 0,17, sensible parce que $n_I = 2$.

■

Un tirage de cinq districts

Graine 2026 : les districts 2, 9, 19, 32 et 40

District	2	9	19	32	40	Total
Ménages M_i	46	31	68	41	29	215
Y_i (milliers)	438	265	614	283	189	1 789

- Horvitz-Thompson : $10 \times 1\,789\,000/2\,000 = 8\,945$ dh, soit 1 034 dh de trop.
- Ratio : $1\,789\,000/215 = 8\,321$ dh, soit 409 dh de trop ($\overline{Y} = 7\,911$).

Pourquoi les deux estimations diffèrent

Le tirage a ramené 215 ménages au lieu des 200 attendus. Horvitz-Thompson divise leur dépense totale par 200, le ratio par 215 : il **corrige** la taille de l'échantillon obtenu.

Estimer la variance sur l'échantillon

Remplacer S par s, calculé sur les grappes tirées

$$\widehat{V}\Big(\frac{\widehat{Y}_{HT}}{N}\Big) = \frac{1-f_I}{n_I\,\overline{M}^2}\,s_T^2 \qquad \widehat{V}(\overline{y}_R) = \frac{1-f_I}{n_I\,\overline{M}^2}\,s_e^2$$

s_T^2 : variance des totaux tirés ; s_e^2 : variance des résidus $\hat{e}_i = Y_i - \overline{y}_R\,M_i$.

Sur les cinq districts

Résidus : 55 287 ; 7 321 ; 47 847 ; −58 378 ; −52 077. D'où $\widehat{V}(\overline{y}_R) = 323\,989$, un écart-type de 569 dh (contre 1 795 dh pour Horvitz-Thompson), et l'intervalle à 95 % : $8\,321 \pm 2{,}776 \times 569$, soit [6 741 ; 9 901].

Cinq observations, pas deux cents

On a interrogé 215 ménages, mais la variance se calcule sur **cinq** grappes : d'où la loi de Student à $n_I - 1 = 4$ degrés de liberté (chapitre 4 de Statistique inférentielle), $t = 2{,}776$ au lieu de 1,96.

L'effet de grappe

La corrélation intra-grappe

Se ressembler entre voisins

ρ est la **corrélation entre les valeurs de deux unités distinctes d'une même grappe**. $\rho = 0$: les ménages sont répartis au hasard entre les grappes ; $\rho = 1$: ils sont identiques dans chaque grappe. Sa valeur minimale, $-1/(M-1)$, correspond à des grappes qui sont chacune un modèle réduit de la population.

Sur les districts du cours

Par l'analyse de la variance (carrés moyens en millions de dh², $M = 39{,}9$ en moyenne ajustée) :

$$\rho = \frac{CM_{\text{inter}} - CM_{\text{intra}}}{CM_{\text{inter}} + (M-1)\,CM_{\text{intra}}} = \frac{116{,}11 - 13{,}36}{116{,}11 + 38{,}9 \times 13{,}36} = 0{,}162$$

Les écarts entre districts font 17,9 % de la dispersion totale ; le hasard seul en ferait 2,5 %.

L'effet de plan des grappes

La formule de Kish

n_I grappes de même taille M, tirées par sondage aléatoire simple ; $n = n_I M$:

$$V(\overline{y}) \approx \left[1 + (M-1)\,\rho\right] \times (1-f)\,\frac{S^2}{n} \qquad \textit{deff} = 1 + (M-1)\,\rho$$

La démonstration courte

$\overline{y}$ est la moyenne des n_I moyennes de grappes tirées : $V(\overline{y}) = (1-f)\,S_b^2/n_I$, où S_b^2 est la variance des moyennes de grappes. Une moyenne de M valeurs de variance S^2, deux à deux corrélées à ρ, a pour variance

$$S_b^2 \approx \frac{1}{M^2}\left[MS^2 + M(M-1)\,\rho\,S^2\right] = \frac{S^2}{M}\left[1 + (M-1)\,\rho\right]$$

Il reste à remplacer, avec $n = n_I M$.

Démonstration : L'effet de grappe, calcul exact

Les notations. Les N_I grappes ont la même taille M, $N = N_I M$; y_{ij} est la valeur de la j-ième unité de la grappe i, et $\sigma^2 = \frac{1}{N}\sum(y_{ij} - \overline{Y})^2 = \frac{N-1}{N}S^2$. La corrélation intra-grappe ρ est la moyenne, sur les paires ordonnées de deux unités distinctes d'une même grappe, du produit $(y_{ij} - \overline{Y})(y_{ik} - \overline{Y})$, divisée par σ^2.

Un calcul exact. La somme des écarts d'une grappe est $\sum_j (y_{ij} - \overline{Y}) = M\,(\overline{Y}_i - \overline{Y})$. En élevant au carré,

$$M^2\,(\overline{Y}_i - \overline{Y})^2 = \sum_j (y_{ij} - \overline{Y})^2 + \sum_{j \neq k} (y_{ij} - \overline{Y})(y_{ik} - \overline{Y}).$$

Sommons sur les N_I grappes. Le premier terme donne $N\sigma^2$, le second $N_I M(M-1)\,\rho\,\sigma^2 = N\,(M-1)\,\rho\,\sigma^2$:

$$M^2 \sum_i (\overline{Y}_i - \overline{Y})^2 = N\sigma^2\,[1 + (M-1)\,\rho], \quad \text{soit} \quad M^2\,(N_I - 1)\,S_b^2 = (N-1)\,S^2\,[1 + (M-1)\,\rho],$$

où S_b^2 est la variance des moyennes de grappes.

La variance de la moyenne. $\overline{y}$ est la moyenne des n_I moyennes de grappes tirées : $V(\overline{y}) = (1 - f_I)\,S_b^2/n_I$. Le sondage aléatoire simple de même taille $n = n_I M$ a $V_{SAS} = (1-f)\,S^2/n$, avec $f = f_I$. Le rapport est

$$deff = \frac{S_b^2/n_I}{S^2/(n_I M)} = \frac{M\,S_b^2}{S^2} = [1 + (M-1)\,\rho] \cdot \frac{N-1}{M\,(N_I - 1)}.$$

Le dernier facteur vaut $\dfrac{N_I M - 1}{N_I M - M}$, voisin de 1 (il vaut 1,02 pour $N_I = 50$ et $M = 40$) : c'est l'approximation $deff \approx 1 + (M-1)\,\rho$ du cours.

La valeur minimale de ρ. Comme $S_b^2 \geq 0$, on a $1 + (M-1)\,\rho \geq 0$, soit $\rho \geq -\dfrac{1}{M-1}$, avec égalité si et seulement si toutes les moyennes de grappes valent $\overline{Y}$: chaque grappe est un modèle réduit de la population.

Ce que cela dit. Le coût des grappes se lit dans $(M-1)\,\rho$: même un petit ρ est multiplié par $M - 1$. Pour les districts du cours, $\rho = 0{,}162$ et $deff \approx 1 + 39 \times 0{,}162 = 7{,}32$: 200 ménages en cinq districts valent $200/7{,}32 = 27$ ménages tirés au hasard.

■

Sur la population du cours

Quarante ménages par district

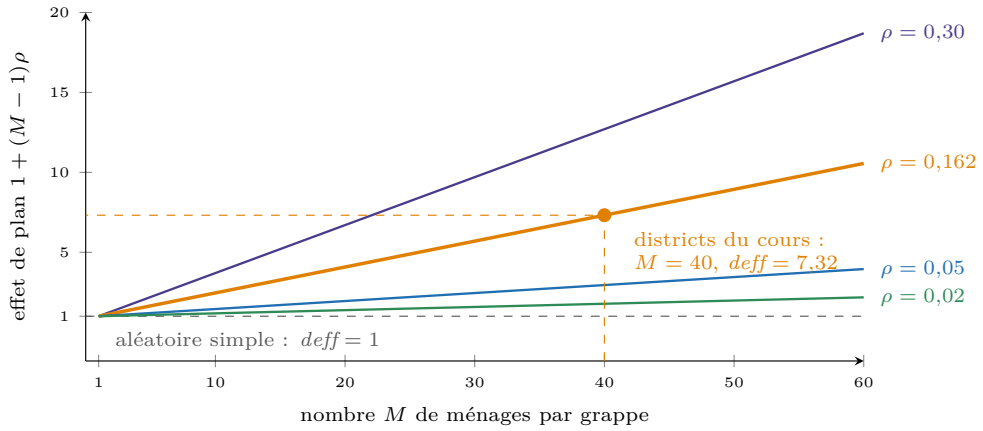
$$deff \approx 1 + 39 \times 0{,}162 = 7{,}32$$

200 ménages pris dans 5 districts donnent la précision de $200/7{,}32 = 27$ ménages tirés au hasard : la **taille effective** du chapitre 7. Chaque district de 40 ménages vaut 5,5 ménages indépendants.

Un petit rho, un grand effet

$\rho = 0{,}162$ paraît faible, mais il est multiplié par $M - 1 = 39$. À précision égale, il faut 7,32 **fois plus** de ménages : dans une grande population ($1 - f \approx 1$), $200 \times 7{,}32 = 1\,464$ ménages en districts de 40 pour égaler 200 ménages tirés au hasard.

L'effet de plan selon la taille des grappes



Deux leçons pour choisir les grappes

- **Petites** : à ρ égal, des grappes de 10 ménages donnent $deff = 1 + 9 \times 0{,}162 = 2{,}46$.
- **Hétérogènes** : à $M = 40$, $deff$ vaut 1,78 si $\rho = 0{,}02$, mais 12,7 si $\rho = 0{,}30$. Un équipement partagé par tout un douar (eau, route) donne un ρ élevé.

Des grappes de tailles inégales

Pourquoi Horvitz-Thompson souffre

Les cinq plus grands et les cinq plus petits districts

Districts tirés	Ménages	HT/N	Ratio
1, 6, 7, 14 et 19 (les plus grands)	325	14 532	8 943
15, 33, 38, 39 et 46 (les plus petits)	88	3 295	7 489

Vraie moyenne : $\overline{Y} = 7\,911$ dh.

Le total d'un district dépend surtout de sa taille

La corrélation entre Y_i et M_i vaut 0,87. Horvitz-Thompson subit toute la variation des tailles : $S_T = 142\,991$ dh. Le ratio n'en garde que les écarts de moyennes : $S_e = 72\,958$ dh, et S_e^2 ne fait que 26 % de S_T^2.

Horvitz-Thompson contre ratio

Cinq districts, deux cents ménages en moyenne

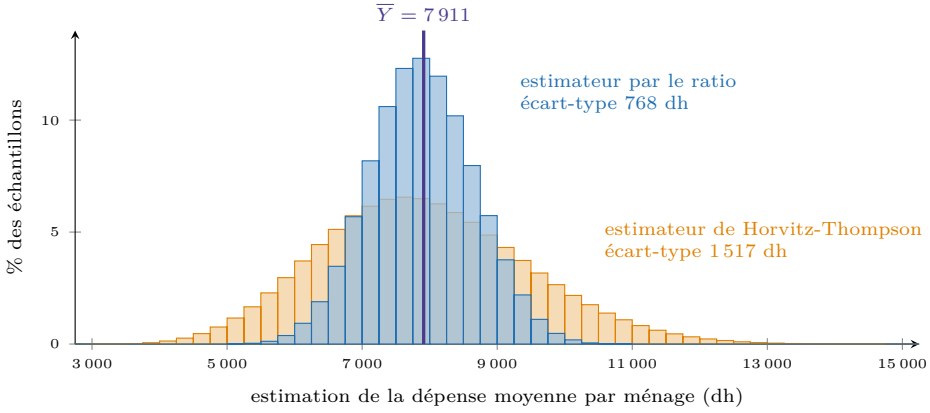
Plan et estimateur	Écart-type	*deff*
Aléatoire simple de 200 ménages	267,3	1
5 districts, Horvitz-Thompson	1 516,6	32,2

5 districts, ratio	773,8	8,4

Le ratio divise l'effet de plan par quatre

Mais 8,4 reste lourd : 200 ménages en cinq districts valent 24 ménages tirés au hasard. C'est plus que les 7,3 de la formule de Kish : dans S_e^2, chaque district pèse M_i^2, et les grands districts comptent davantage que dans une population de districts égaux.

Tous les échantillons de cinq districts



Les 2 118 760 échantillons possibles

Horvitz-Thompson est exactement centré, mais se trompe de plus de 2 000 dh une fois sur cinq (18,8 %). Le ratio a un biais de −25 dh, un écart-type de 768 dh (la formule approchée donnait 774), et ne se trompe de plus de 2 000 dh que dans 0,7 % des cas.

Faire mieux

Trois remèdes

1. **Plus de districts** : avec le ratio, l'écart-type passe de 774 dh (5 districts) à 516 dh (10) et 316 dh (20), mais chaque district coûte un déplacement.
2. **Tirer les grands districts plus souvent**, avec une probabilité proportionnelle à M_i : 754 dh avec 5 districts, sans le ratio (chapitre 11).
3. **Ne pas interroger tout le district** : tirer des ménages dans les districts tirés (chapitre 10).

La règle de l'estimation

Avec des grappes de tailles inégales, on estime une moyenne par le **ratio**, jamais par $\widehat{Y}_{HT}/N$. Le léger biais du ratio est le prix, très faible, d'une forte baisse de variance.

Coût et précision

Le prix d'un ménage

Une fonction de coût

$$C = c_1\, n_I + c_2\, n$$

c_1 : coût d'un lieu visité (déplacement, repérage, dénombrement) ; c_2 : coût d'un entretien. Un ménage tiré au hasard est isolé : il coûte $c_1 + c_2$. Dans une grappe de M ménages, il coûte $c_1/M + c_2$.

La règle de décision

À budget égal, dans une grande population ($1 - f \approx 1$), on compare les **tailles effectives** n/deff. Les grappes l'emportent si et seulement si

$$\frac{c_1 + c_2}{c_1/M + c_2} > \mathit{deff}$$

un ménage en grappe doit coûter *deff* fois moins cher qu'un ménage isolé.

Démonstration : La règle du rapport des coûts

Les deux budgets. Avec un budget C : un ménage tiré au hasard coûte $c_1 + c_2$ (le lieu, puis l'entretien), d'où $n_{SAS} = C/(c_1 + c_2)$ ménages. Un district de M ménages coûte $c_1 + M\,c_2$, d'où $n_I = C/(c_1 + M\,c_2)$ districts, soit

$$n = M\, n_I = \frac{C}{c_1/M + c_2}$$

ménages, de taille effective n/deff.

La comparaison. Les grappes l'emportent si et seulement si

$$\frac{n}{\mathit{deff}} > n_{SAS} \iff \frac{C}{(c_1/M + c_2)\,\mathit{deff}} > \frac{C}{c_1 + c_2} \iff \frac{c_1 + c_2}{c_1/M + c_2} > \mathit{deff}.$$

Le seuil. L'égalité $c_1 + c_2 = \mathit{deff}(c_1/M + c_2)$ s'écrit $c_1\,(1 - \mathit{deff}/M) = c_2\,(\mathit{deff} - 1)$; comme $\mathit{deff} = 1 + (M-1)\,\rho < M$ quand $\rho < 1$,

$$c_1^* = c_2\,\frac{\mathit{deff} - 1}{1 - \mathit{deff}/M}.$$

Les grappes l'emportent quand $c_1 > c_1^*$: un déplacement cher les favorise.

Ce que cela dit. Avec $c_2 = 60$ dh, $M = 40$ et $\mathit{deff} = 8{,}38$: $c_1^* = 60 \times 7{,}38/(1 - 0{,}2095) = 560$ dh, le seuil du cours. À l'exercice 3 ($c_2 = 80$, $M = 30$, $\mathit{deff} = 3{,}9$), $c_1^* = 80 \times 2{,}9/(1 - 0{,}13) = 266{,}7$ dh.

■

À budget égal, deux terrains

Budget de 100 000 dh, entretien à 60 dh, districts de 40 ménages

	Rural dispersé	Ville
Coût d'un déplacement c_1	800 dh	200 dh
Aléatoire simple : ménages enquêtés	116	384
Grappes : districts (ménages)	31 (1 240)	38 (1 520)
Grappes : taille effective, $n/8{,}38$	148	181

Une grande région, aux districts semblables à ceux du cours : $deff = 8{,}38$ avec le ratio.

Le terrain décide

En zone rurale, un ménage isolé coûte 860 dh, un ménage en grappe 80 dh : rapport $10{,}75 > 8{,}38$, les grappes l'emportent. En ville, 260 dh contre 65 dh : rapport 4, l'aléatoire simple l'emporte. Le seuil se situe vers $c_1 = 560$ dh.

Quand tirer par grappes

Les conditions favorables

1. Il n'existe pas de liste des unités, mais une liste des grappes.
2. Les déplacements coûtent cher par rapport aux entretiens.
3. La variable étudiée est peu homogène dans les grappes : ρ faible.

Interroger tout le district est souvent trop

Le quarantième ménage d'un district apporte peu : il ressemble aux trente-neuf autres. Mieux vaut visiter **plus** de districts et y interroger **moins** de ménages. C'est le sondage à deux degrés.

S'entraîner

Exercice 1 : trois immeubles

Cinq ménages en trois grappes

La mini-population A, B, C, D, E (dépenses 4, 6, 7, 9, 14 milliers de dh) habite trois immeubles : A et B dans le premier, C et D dans le deuxième, E seul dans le troisième. On tire deux immeubles par sondage aléatoire simple, et l'on interroge tous leurs ménages.

1. Énumérer les échantillons possibles, les ménages interrogés et $\widehat{Y}_{HT}$.
2. Vérifier que $E(\widehat{Y}_{HT}) = Y = 40$. Calculer $V(\widehat{Y}_{HT})$ par l'énumération, puis par $N_I^2(1 - f_I)S_T^2/n_I$.
3. Calculer les π_k, puis π_{AB} et π_{AC}.
4. Calculer l'estimateur par le ratio de la moyenne pour chaque échantillon. Est-il sans biais ?

Exercice 1 : solution

Les trois échantillons

1. Totaux des immeubles : 10, 16, 14 ; poids $N_I/n_I = 3/2$.

Immeubles	Ménages	$\sum Y_i$	$\widehat{Y}_{HT}$	Ratio
1 et 2	A, B, C, D	26	39	6,5
1 et 3	A, B, E	24	36	8
2 et 3	C, D, E	30	45	10

2. $E(\widehat{Y}_{HT}) = (39 + 36 + 45)/3 = 40$. $V = (1 + 16 + 25)/3 = 14$. Formule : $S_T^2 = 9{,}33$ et $9 \times \frac{1}{3} \times 9{,}33/2 = 14$.

Probabilités et ratio

3. Chaque immeuble figure dans deux échantillons sur trois : $\pi_k = 2/3$ pour tous. $\pi_{AB} = 2/3$: A et B sont toujours tirés ensemble ; $\pi_{AC} = 1/3$ (immeubles 1 et 2).
4. $(6{,}5 + 8 + 10)/3 = 8{,}17 \neq 8$: un biais de 0,17, sensible ici parce que $n_I = 2$.

Exercice 2 : des douars de 25 ménages

L'effet de grappe dans une enquête rurale

Une enquête tire des douars entiers de $M = 25$ ménages. Pour la dépense, la corrélation intra-douar vaut $\rho = 0{,}08$. La population est grande : on néglige $1 - f$.

1. Calculer l'effet de plan.
2. On enquête 16 douars. Quelle est la taille effective de l'échantillon ?
3. Combien de douars faut-il pour obtenir la précision d'un sondage aléatoire simple de 300 ménages ?
4. Pour l'accès à l'eau courante, $\rho = 0{,}5$. Que devient l'effet de plan ? Pourquoi ρ est-il si élevé ?

Exercice 2 : solution

Effet de plan et taille effective

1. $deff = 1 + 24 \times 0{,}08 = 2{,}92$.
2. $16 \times 25 = 400$ ménages, qui valent $400/2{,}92 = 137$ ménages tirés au hasard.
3. $300 \times 2{,}92 = 876$ ménages, soit $876/25 = 35{,}04$: il faut 36 douars.

Une variable très homogène

4. $deff = 1 + 24 \times 0{,}5 = 13$: les 400 ménages ne valent plus que 31 ménages indépendants. L'eau courante dépend du réseau du douar : presque tous les ménages d'un même douar ont la même réponse, et chaque ménage de plus n'apprend presque rien.

Exercice 3 : grappes ou aléatoire simple ?

Un budget de 90 000 dh

Une enquête régionale peut tirer des districts entiers de $M = 30$ ménages ($\rho = 0{,}1$), ou des ménages au hasard. Un déplacement coûte $c_1 = 1\,000$ dh, un entretien $c_2 = 80$ dh. On néglige $1 - f$.

1. Calculer l'effet de plan des districts entiers.
2. Combien de ménages l'aléatoire simple permet-il d'enquêter ? Et combien de districts, de ménages, et quelle taille effective pour les grappes ?
3. Quel plan choisir ? Vérifier avec la règle du rapport des coûts.
4. En dessous de quel coût de déplacement l'aléatoire simple l'emporterait-il ?

Exercice 3 : solution

Les deux plans

1. $deff = 1 + 29 \times 0{,}1 = 3{,}9$.
2. Aléatoire simple : 1 080 dh par ménage, $90\,000/1\,080 = 83$ ménages. Grappes : $1\,000 + 30 \times 80 = 3\,400$ dh par district, 26 districts, 780 ménages, de taille effective $780/3{,}9 = 200$.

La règle et le seuil

3. Les grappes : $200 > 83$. Un ménage en grappe coûte $3\,400/30 = 113{,}3$ dh, et $1\,080/113{,}3 = 9{,}53 > 3{,}9$.
4. On résout $(c_1 + 80)/(c_1/30 + 80) = 3{,}9$: $c_1 + 80 = 0{,}13\,c_1 + 312$, d'où $c_1 = 266{,}7$ dh. En dessous, l'aléatoire simple l'emporte.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. On tire des **grappes entières** : une liste des grappes suffit, et les déplacements diminuent.
2. Chaque unité a $\pi_k = n_I/N_I$; $\widehat{Y}_{HT} = (N_I/n_I)\sum_{S_I} Y_i$ est sans biais.
3. La grappe est l'unité statistique : variances sur les totaux Y_i, à $n_I - 1$ degrés de liberté.
4. Moyenne par ménage : le ratio $\sum Y_i / \sum M_i$, presque sans biais et bien plus précis.
5. Grappes égales : $deff = 1 + (M-1)\rho$, soit 7,3 pour les districts du cours.
6. Un petit ρ suffit à un grand effet : préférer des grappes petites et hétérogènes.
7. Les grappes l'emportent si un ménage groupé coûte $deff$ fois moins qu'un ménage isolé.

La suite

Le chapitre 10 tire des ménages **dans** les districts tirés : le sondage à deux degrés garde l'économie des déplacements sans payer tout l'effet de grappe.