Échantillonnage

Chapitre 10 : Le sondage à deux degrés

Jaouad Madkour

31 octobre 2026

Plan du chapitre

Deux tirages emboîtés **2**
- Pourquoi tirer en deux temps . . . 2
- Le tirage sur la population du cours . . . 2
- Les notations . . . 3

Probabilités d'inclusion et estimateur **3**
- La chance d'être tiré . . . 3
- Un estimateur sans biais du total . . . 4
- Chaque ménage tiré a un poids . . . 5
- L'estimation sur l'échantillon tiré . . . 5

La variance à deux termes **5**
- Deux sources d'aléa . . . 5
- La démonstration . . . 6
- Dix districts, vingt ménages . . . 7
- Répartir autrement les 200 ménages . . . 7
- Pourquoi le premier degré pèse tant . . . 7

Mieux estimer, estimer la variance **8**
- L'estimateur par le ratio . . . 8
- Le ratio sur l'échantillon tiré . . . 8
- Estimer la variance à partir de l'échantillon . . . 8
- Erreurs-types sur l'échantillon tiré . . . 10

Choisir le nombre de districts et de ménages **10**
- Un modèle de coût . . . 10
- Le nombre optimal de ménages par district . . . 10
- Sur la population du cours . . . 11
- Un optimum plat . . . 12
- Ce qui fait varier le nombre optimal . . . 12

Le plan autopondéré **13**
- Des poids tous égaux . . . 13
- Le piège de l'autopondération . . . 13

S'entraîner **13**
Exercice 1 : quatre douars . . . 13
Exercice 1 : solution . . . 14
Exercice 2 : le budget d'une enquête rurale . . . 14
Exercice 2 : solution . . . 14
Exercice 3 : vingt districts ou dix . . . 15
Exercice 3 : solution . . . 15

Ce qu'il faut retenir **15**
Le résumé du chapitre . . . 15

Deux tirages emboîtés

Pourquoi tirer en deux temps

Ce que le chapitre 9 a montré

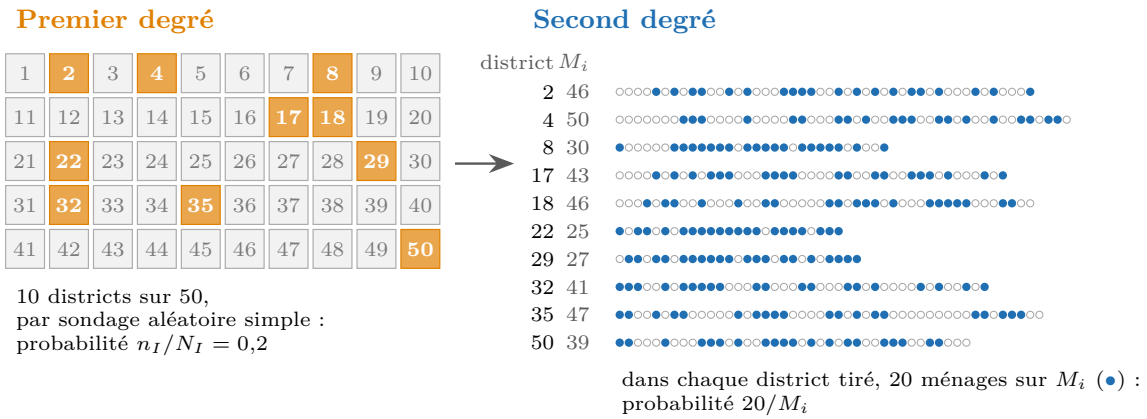
Tirer 5 districts entiers, soit 200 ménages en moyenne, limite les déplacements, mais l'effet de plan atteint 32,2 avec l'estimateur de Horvitz-Thompson et 8,4 avec le ratio. Les ménages d'un même district se ressemblent ($\rho = 0{,}162$) : le quarantième ménage interrogé apprend peu de chose que les vingt premiers n'aient déjà dit.

Le sondage à deux degrés

1. **Premier degré** : tirer n_I **unités primaires**, les districts, parmi N_I.
2. **Second degré** : dans chaque district i tiré, tirer m_i **unités secondaires**, les ménages, parmi ses M_i ménages.

Les deux tirages sont ici des sondages aléatoires simples. La liste des ménages n'est nécessaire **que dans les districts tirés**.

Le tirage sur la population du cours



Dix districts, vingt ménages dans chacun

Graine 2026 : les districts 2, 4, 8, 17, 18, 22, 29, 32, 35 et 50, puis 20 ménages dans chacun. Les 200 ménages se trouvent dans 10 districts ; un sondage aléatoire simple de 200 ménages en toucherait 48,2 en moyenne, sur 50.

Les notations

Deux niveaux

Grandeur	Population	Échantillon
Districts	$N_I = 50$	n_I, ensemble S_I
Ménages du district i	M_i	m_i, ensemble s_i
Taux de sondage		$f_1 = n_I/N_I$, $f_{2i} = m_i/M_i$
Variance du district	S_i^2	s_i^2
Total du district	Y_i	$\widehat{Y}_i = M_i\,\overline{y}_i$

Deux plans déjà vus, aux deux extrêmes

- $m_i = M_i$: on enquête **tout** le district tiré. C'est le sondage par grappes du chapitre 9.
- $n_I = N_I$: on garde **tous** les districts, et chacun devient une strate (chapitre 6).

Le sondage à deux degrés se place entre les deux : quelques districts, une partie de leurs ménages.

Probabilités d'inclusion et estimateur

La chance d'être tiré

Le produit de deux probabilités

Un ménage k du district i est tiré si son district l'est, **puis** s'il est tiré dans son district :

$$\pi_k = P(i \in S_I) \times P(k \in s_i \mid i \in S_I) = \frac{n_I}{N_I} \cdot \frac{m_i}{M_i}$$

Dix districts, vingt ménages

$f_1 = 10/50 = 0{,}2$. Un district de moins de 20 ménages (il y en a trois : 13, 15 et 18 ménages) est enquêté en entier.

- District 1, 71 ménages : $\pi_k = 0{,}2 \times 20/71 = 0{,}056$, poids 17,75.
- District 24, 40 ménages : $\pi_k = 0{,}2 \times 20/40 = 0{,}1$, poids 10.
- District 46, 13 ménages : $\pi_k = 0{,}2 \times 13/13 = 0{,}2$, poids 5.

Les ménages des grands districts ont **moins** de chances d'être tirés.

Un estimateur sans biais du total

Estimer chaque district, puis la population

Dans chaque district tiré, $\widehat{Y}_i = M_i\,\overline{y}_i$ estime le total Y_i (chapitre 4). Puis, comme au chapitre 9 :

$$\widehat{Y} = \frac{N_I}{n_I}\sum_{i\in S_I}\widehat{Y}_i = \sum_{i\in S_I}\sum_{k\in s_i}\frac{y_k}{\pi_k}$$

C'est l'estimateur de Horvitz-Thompson du chapitre 2.

La démonstration en deux temps

Les districts une fois tirés, le second degré est un sondage aléatoire simple : $E(\widehat{Y}_i \mid S_I) = Y_i$. Il reste l'estimateur du chapitre 9 ; avec les indicatrices I_i des districts, $E(I_i) = n_I/N_I$:

$$E(\widehat{Y}) = E\Big(\frac{N_I}{n_I}\sum_{i=1}^{N_I} I_i\,Y_i\Big) = \frac{N_I}{n_I}\sum_{i=1}^{N_I}\frac{n_I}{N_I}\,Y_i = Y$$

Démonstration : Probabilité d'inclusion, poids et absence de biais à deux degrés

La probabilité d'inclusion. Le ménage k du district i est tiré si son district l'est, puis s'il est tiré dans son district : $\pi_k = P(i \in S_I)\,P(k \in s_i \mid i \in S_I) = \dfrac{n_I}{N_I}\cdot\dfrac{m_i}{M_i}$. Le poids est $1/\pi_k = \dfrac{N_I}{n_I}\cdot\dfrac{M_i}{m_i}$.

L'estimateur de Horvitz-Thompson. En sommant sur les ménages tirés d'un district tiré, $\sum_{k\in s_i} M_i/m_i\; y_k = M_i\,\overline{y}_i = \widehat{Y}_i$, donc

$$\sum_{i\in S_I}\sum_{k\in s_i}\frac{y_k}{\pi_k} = \frac{N_I}{n_I}\sum_{i\in S_I} M_i\,\overline{y}_i = \frac{N_I}{n_I}\sum_{i\in S_I}\widehat{Y}_i = \widehat{Y}.$$

Sans biais, en deux temps. Conditionnellement aux districts tirés S_I, chaque $\overline{y}_i$ est la moyenne d'un sondage aléatoire simple de m_i ménages : $E(\widehat{Y}_i \mid S_I) = M_i\,\overline{Y}_i = Y_i$. Donc $E(\widehat{Y} \mid S_I) = \frac{N_I}{n_I}\sum_{i\in S_I} Y_i$, l'estimateur du chapitre 9. La loi de l'espérance totale, $E(\widehat{Y}) = E\big[E(\widehat{Y} \mid S_I)\big]$, donne alors $E(\widehat{Y}) = Y$ par le chapitre 9.

Ce que cela dit. Le poids d'un ménage est le produit des poids des deux degrés, et l'estimateur reste sans biais. Sur les quatre douars de l'exercice 1 ($n_I = 2, m = 2$) : $\pi_k = \frac{2}{4}\times\frac{2}{4} = \frac{1}{4}$ pour les 16 ménages, $\widehat{Y} = 4\times$ (somme des quatre dépenses tirées), et la moyenne des 216 échantillons possibles vaut $116 = Y$.

■

Chaque ménage tiré a un poids

Le poids de sondage

$$d_k = \frac{1}{\pi_k} = \frac{N_I}{n_I} \cdot \frac{M_i}{m_i}$$

le produit des poids des deux degrés. Alors $\widehat{Y} = \sum_{k \in S} d_k\, y_k$, et $\widehat{N} = \sum_{k \in S} d_k = \frac{N_I}{n_I} \sum_{i \in S_I} M_i$ estime le nombre de ménages.

Les poids de l'échantillon tiré

$d_k = 5 \times M_i/20$: de 6,25 (district 22, 25 ménages) à 12,5 (district 4, 50 ménages). Les dix districts tirés comptent 394 ménages, d'où $\widehat{N} = 5 \times 394 = 1\,970$, alors que $N = 2\,000$.

L'estimation sur l'échantillon tiré

Dépense moyenne des vingt ménages tirés, en dh

District	M_i	$\overline{y}_i$	District	M_i	$\overline{y}_i$
2	46	10 336,5	22	25	7 964,5
4	50	10 137,5	29	27	6 861,5
8	30	7 243,5	32	41	7 259,0
17	43	5 837,5	35	47	8 905,0
18	46	9 352,5	50	39	5 811,0

Le total et la moyenne

$\sum_{i \in S_I} M_i\, \overline{y}_i = 3\,208\,042{,}5$ dh, d'où $\widehat{Y} = 5 \times 3\,208\,042{,}5 = 16\,040\,212{,}5$ dh et $\widehat{Y}/N = 8\,020{,}11$ dh par ménage, pour $\overline{Y} = 7\,911{,}44$: une erreur de +108,7 dh.

La variance à deux termes

Deux sources d'aléa

La variance de l'estimateur du total

$$V(\widehat{Y}) = \underbrace{N_I^2\,(1 - f_1)\,\frac{S_T^2}{n_I}}_{V_1 \text{ : premier degré}} + \underbrace{\frac{N_I}{n_I} \sum_{i=1}^{N_I} M_i^2\,(1 - f_{2i})\,\frac{S_i^2}{m_i}}_{V_2 \text{ : second degré}}$$

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

Lire les deux termes

- V_1 est la variance du chapitre 9 : celle que l'on aurait si l'on connaissait **exactement** le total de chaque district tiré.
- V_2 ajoute l'erreur commise **dans** chaque district, celle d'un sondage aléatoire simple de m_i ménages (chapitres 3 et 4). Chaque district y compte avec sa probabilité f_1 d'être

tiré.

La démonstration

Conditionner par les districts tirés

Comme la variance d'une population en groupes (statistique descriptive), la variance se décompose en variance de la moyenne conditionnelle plus moyenne de la variance conditionnelle :

$$V(\widehat{Y}) = V\big[E(\widehat{Y} \mid S_I)\big] + E\big[V(\widehat{Y} \mid S_I)\big]$$

1. $E(\widehat{Y} \mid S_I) = \frac{N_I}{n_I}\sum_{i\in S_I} Y_i$: sa variance est celle du chapitre 9, V_1.
2. Les districts fixés, les $\widehat{Y}_i$ sont indépendants, de variances $V_i = M_i^2(1-f_{2i})S_i^2/m_i$: $V(\widehat{Y} \mid S_I) = \frac{N_I^2}{n_I^2}\sum_i I_i\,V_i$. Avec $E(I_i) = n_I/N_I$, son espérance est V_2.

Démonstration : La variance à deux termes

La loi de la variance totale. Pour une variable X et une information S, écrivons $X - E(X) = [X - E(X \mid S)] + [E(X \mid S) - E(X)]$. En élevant au carré et en prenant l'espérance, le double produit disparaît : le second crochet est fixé quand S l'est, et $E\big[X - E(X \mid S)\,\big|\,S\big] = 0$. Il reste

$$V(X) = E\big[V(X \mid S)\big] + V\big[E(X \mid S)\big].$$

Le premier terme. Ici $X = \widehat{Y}$ et $S = S_I$. On a $E(\widehat{Y} \mid S_I) = \frac{N_I}{n_I}\sum_{i\in S_I} Y_i$, dont la variance est celle du chapitre 9 :

$$V\big[E(\widehat{Y} \mid S_I)\big] = N_I^2\,(1 - f_1)\,\frac{S_T^2}{n_I} = V_1.$$

Le second terme. Les districts étant fixés, les tirages du second degré sont indépendants d'un district à l'autre, donc les $\widehat{Y}_i$ sont indépendants, de variances $V_i = M_i^2\,(1 - f_{2i})\,S_i^2/m_i$ (chapitres 3 et 4). Avec les indicatrices I_i des districts tirés,

$$V(\widehat{Y} \mid S_I) = \frac{N_I^2}{n_I^2}\sum_{i\in S_I} V_i = \frac{N_I^2}{n_I^2}\sum_{i=1}^{N_I} I_i\,V_i.$$

Comme $E(I_i) = n_I/N_I$, l'espérance vaut $\dfrac{N_I^2}{n_I^2}\cdot\dfrac{n_I}{N_I}\sum_i V_i = \dfrac{N_I}{n_I}\sum_{i=1}^{N_I} V_i = V_2$.

Ce que cela dit. $V(\widehat{Y}) = V_1 + V_2$: le premier terme est la variance que l'on aurait si l'on connaissait exactement les totaux des districts tirés, le second l'erreur commise dans chaque district. Sur les quatre douars de l'exercice 1, $V_1 = 314{,}67$, $V_2 = 74{,}67$ et $V = 389{,}33$, valeur que l'énumération des 216 échantillons retrouve exactement.

■

Vérification par simulation

Sur 20 000 tirages de 10 districts et 20 ménages (graine 2026), l'écart-type de $\widehat{Y}/N$ vaut 1 032,7 dh ; la formule donne 1 032 dh.

Dix districts, vingt ménages

Les deux termes sur la population du cours

Écart-type des 50 totaux de district : $S_T = 142\,991$ dh. Pour $n_I = 10$ et $m = 20$:

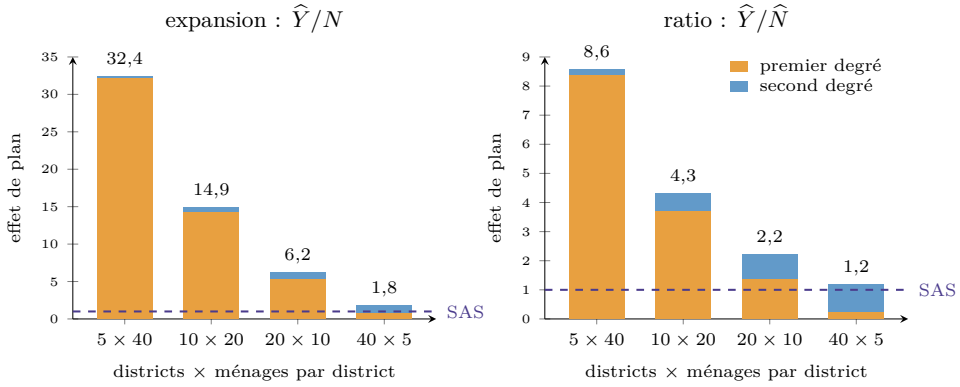
$$V_1 = 4{,}089 \cdot 10^{12}, \qquad V_2 = 1{,}70 \cdot 10^{11}, \qquad V(\widehat{Y}) = 4{,}26 \cdot 10^{12}$$

Erreur-type de la moyenne : $\sqrt{4{,}26 \cdot 10^{12}/2\,000} \;=\; 1\,032$ dh, contre 267 dh pour un sondage aléatoire simple de 200 ménages. Effet de plan : 14,9.

Le premier degré domine

96 % de la variance vient du **choix des districts**, 4 % seulement du choix des ménages. Taille effective : $200/14{,}9 \approx 13$ ménages. Mieux choisir les ménages dans les districts tirés ne changerait presque rien : c'est le premier degré qu'il faut améliorer.

Répartir autrement les 200 ménages



Lecture

De 5×40 (en fait 173 ménages en moyenne : 22 districts en comptent moins de 40) à 40×5, l'effet de plan tombe de 32,4 à 1,8 avec l'estimateur d'expansion $\widehat{Y}/N$, de 8,6 à 1,2 avec le ratio. Plus de districts réduit V_1, en $1/n_I$; V_2 croît un peu (de 0,19 à 0,95 en effet de plan), car on interroge moins de ménages par district.

Pourquoi le premier degré pèse tant

Des districts de tailles très inégales

Les totaux Y_i vont de 88 450 à 766 730 dh, surtout parce que les districts comptent de 13 à 71 ménages : la corrélation entre Y_i et M_i vaut 0,868. L'estimateur $\widehat{Y}$ ignore les tailles : un échantillon de grands districts **surestime** le total, un échantillon de petits le sous-estime.

Deux remèdes

- **Estimer autrement** : rapporter le total estimé au nombre estimé de ménages $\widehat{N}$. C'est l'estimateur par le ratio du chapitre 9.

- **Tirer autrement** : donner aux grands districts plus de chances d'être tirés, proportionnellement à leur taille (chapitre 11).

Mieux estimer, estimer la variance

L'estimateur par le ratio

Rapporter au nombre estimé de ménages

$$\widehat{\overline{Y}}_R = \frac{\widehat{Y}}{\widehat{N}} = \frac{\sum_{i \in S_I} M_i \, \overline{y}_i}{\sum_{i \in S_I} M_i}$$

Sa variance approchée garde la forme à deux termes, divisée par N^2 : dans V_1, les totaux Y_i sont remplacés par les écarts $e_i = Y_i - \overline{Y} \, M_i$, de variance S_e^2 ; V_2 ne change pas.

Dix districts, vingt ménages

$S_e = 72\,958$ dh contre $S_T = 142\,991$: V_1 est divisé par 3,8. Erreur-type 556 dh, effet de plan 4,32 au lieu de 14,9, taille effective 46 ménages. Le modèle du chapitre 9 en donne une bonne idée : $1 + (m-1)\rho = 1 + 19 \times 0{,}162 = 4{,}08$.

Le ratio sur l'échantillon tiré

Deux estimations de la dépense moyenne

- Expansion : $\widehat{Y}/N = 8\,020{,}11$ dh, erreur $+108{,}7$.
- Ratio : $3\,208\,042{,}5/394 = 8\,142{,}24$ dh, erreur $+230{,}8$.

Ici, le ratio se trompe davantage. Les dix districts enquêtés en entier lui auraient donné 7 940 dh : l'écart vient du **second degré**. Dans les districts 2 et 4, les vingt ménages tirés dépensent 814 et 867 dh de plus que la moyenne de leur district.

Un échantillon ne juge pas un estimateur

Sur 20 000 tirages 10×20 : expansion, moyenne 7 905,6 et écart-type 1 032,7 ; ratio, moyenne 7 895,1 (biais $-16{,}3$ dh) et écart-type 551,5. Le ratio a un léger biais, mais son écart-type est **presque deux fois plus petit**.

Estimer la variance à partir de l'échantillon

Un estimateur sans biais

$$\widehat{V}(\widehat{Y}) = N_I^2 \, (1 - f_1) \, \frac{s_1^2}{n_I} + \frac{N_I}{n_I} \sum_{i \in S_I} M_i^2 \, (1 - f_{2i}) \, \frac{s_i^2}{m_i}$$

où s_1^2 est la variance des n_I totaux estimés $\widehat{Y}_i$. Pour le ratio, on remplace les $\widehat{Y}_i$ par $\widehat{Y}_i - \widehat{\overline{Y}}_R \, M_i$.

Le premier terme contient presque tout

Les $\widehat{Y}_i$ varient d'un district à l'autre pour les deux raisons à la fois : l'espérance du premier terme vaut $V_1 + (1 - f_1)\,V_2$. Ne garder que lui sous-estime V de $f_1 V_2$ seulement, ici 0,8 %. C'est la pratique courante quand f_1 est petit.

Démonstration : L'estimateur de variance à deux degrés est sans biais

Le premier terme. Les districts S_I étant fixés, $\widehat{Y}_i = Y_i + \varepsilon_i$, où l'erreur ε_i du second degré est centrée, de variance V_i, indépendante d'un district à l'autre. Développons la somme des carrés des n_I valeurs $\widehat{Y}_i$ autour de leur moyenne : le double produit entre les Y_i et les ε_i est d'espérance nulle, et

$$E\sum_{i\in S_I}(\varepsilon_i-\bar{\varepsilon})^2=\sum_{i\in S_I}V_i-n_I\cdot\frac{1}{n_I^2}\sum_{i\in S_I}V_i=\frac{n_I-1}{n_I}\sum_{i\in S_I}V_i.$$

En divisant par $n_I - 1$: $E(s_1^2 \mid S_I) = s_Y^2 + \dfrac{1}{n_I}\sum_{i\in S_I} V_i$, où s_Y^2 est la variance des Y_i dans S_I. Puis, sur les tirages de districts, $E(s_Y^2) = S_T^2$ (chapitre 3) et $E\big[\frac{1}{n_I}\sum_{S_I} V_i\big] = \frac{1}{N_I}\sum_{i=1}^{N_I} V_i$:

$$E(s_1^2)=S_T^2+\frac{1}{N_I}\sum_{i=1}^{N_I}V_i.$$

Multiplions par $N_I^2\,(1-f_1)/n_I$ et rappelons $V_2 = \frac{N_I}{n_I}\sum_i V_i$:

$$E\left[N_I^2\,(1-f_1)\,\frac{s_1^2}{n_I}\right]=V_1+(1-f_1)\,\frac{N_I}{n_I}\sum_i V_i=V_1+(1-f_1)\,V_2.$$

Le second terme. Comme $E(s_i^2) = S_i^2$, on a $E(\widehat{V}_i \mid S_I) = V_i$ pour l'estimateur $\widehat{V}_i = M_i^2(1 - f_{2i})\,s_i^2/m_i$, et

$$E\left[\frac{N_I}{n_I}\sum_{i\in S_I}\widehat{V}_i\right]=\frac{N_I}{n_I}\cdot\frac{n_I}{N_I}\sum_{i=1}^{N_I}V_i=\sum_{i=1}^{N_I}V_i=f_1\,V_2.$$

La somme. $V_1 + (1 - f_1)\,V_2 + f_1\,V_2 = V_1 + V_2 = V(\widehat{Y})$: l'estimateur est sans biais.

Ce que cela dit. Le premier terme à lui seul a pour espérance $V - f_1\,V_2$: il sous-estime la variance de $f_1\,V_2$. Pour 10 districts sur 50 et 20 ménages par district, $f_1\,V_2 = 0{,}2 \times 1{,}70 \cdot 10^{11}$, soit 0,8 % de $V = 4{,}26 \cdot 10^{12}$: c'est pourquoi, quand f_1 est petit, on s'en contente.

■

Erreurs-types sur l'échantillon tiré

Deux intervalles à 95 pour cent

Écart-type des dix totaux estimés : $s_1 = 123\,935$ dh.

Estimateur	Estimation	Erreur-type	Intervalle à 95 %
Expansion	8 020	880,1	[6 029 ; 10 011]
Ratio	8 142	511,3	[6 986 ; 9 299]

Le premier terme fait 99 % de la variance estimée de l'expansion.

Dix districts, pas deux cents ménages

Les intervalles utilisent la loi de Student à $n_I - 1 = 9$ degrés de liberté, $t = 2{,}262$ (Statistique inférentielle, chapitre 4) : l'information vient de 10 districts. Traiter les 200 ménages comme un sondage aléatoire simple donnerait une erreur-type de 252 dh : un intervalle **deux fois trop étroit**.

Choisir le nombre de districts et de ménages

Un modèle de coût

Coût et variance

Un district coûte c_1 (déplacement, dénombrement des ménages), un ménage interrogé coûte c_2 :

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

Avec m ménages par district, la variance de la moyenne vaut à peu près (chapitre 9) :

$$V \approx \frac{S^2}{n_I\, m}\big[1 + (m-1)\rho\big] = \frac{S^2}{n_I}\Big[\rho + \frac{1-\rho}{m}\Big]$$

Une enquête rurale

$c_1 = 800$ dh, $c_2 = 60$ dh, budget $C = 25\,000$ dh. Avec $m = 20$, un district coûte 2 000 dh : 12,5 districts et 250 ménages. Avec $m = 4$, il coûte 1 040 dh : 24 districts et 96 ménages. Lequel est le plus précis ?

Le nombre optimal de ménages par district

La règle

$$m^* = \sqrt{\frac{c_1}{c_2} \cdot \frac{1-\rho}{\rho}}$$

puis $n_I = C/(c_1 + c_2\, m^*)$.

La démonstration

À budget fixé, $n_I = C/(c_1 + c_2 m)$, donc

$$V \approx \frac{S^2}{C}\,(c_1 + c_2 m)\Big(\rho + \frac{1-\rho}{m}\Big) = \frac{S^2}{C}\Big[c_1\rho + c_2(1-\rho) + c_2\rho\, m + \frac{c_1(1-\rho)}{m}\Big]$$

La dérivée en m s'annule quand $c_2\,\rho = c_1(1-\rho)/m^2$. Le terme $c_2\rho\, m$, croissant, est le prix de la ressemblance ; le terme $c_1(1-\rho)/m$, décroissant, celui des déplacements.

Démonstration : Le nombre optimal de ménages par district, sans dérivée

Le coût fixe le nombre de districts. À budget C, $n_I = C/(c_1 + c_2\, m)$. Avec $\frac{1}{m}[1 + (m-1)\rho] = \rho + \frac{1-\rho}{m}$, la variance est $V \approx \frac{S^2}{C}\, g(m)$ avec

$$g(m) = (c_1 + c_2\, m)\Big(\rho + \frac{1-\rho}{m}\Big).$$

Une borne. L'inégalité de Cauchy-Schwarz (démontrée au chapitre 7) avec $a = (\sqrt{c_1},\, \sqrt{c_2\, m})$ et $b = (\sqrt{\rho},\, \sqrt{(1-\rho)/m})$ donne, pour tout $m > 0$,

$$g(m) = \sum a_h^2 \sum b_h^2 \;\geq\; \Big(\sum a_h b_h\Big)^2 = \Big(\sqrt{c_1\,\rho} + \sqrt{c_2\,(1-\rho)}\Big)^2.$$

Le cas d'égalité. Les deux vecteurs doivent être proportionnels : $\dfrac{\sqrt{c_1}}{\sqrt{\rho}} = \dfrac{\sqrt{c_2\, m}}{\sqrt{(1-\rho)/m}}$, soit $\sqrt{c_1\,(1-\rho)} = m\,\sqrt{c_2\,\rho}$:

$$m^* = \sqrt{\frac{c_1}{c_2} \cdot \frac{1-\rho}{\rho}}.$$

La borne est atteinte : c'est un minimum global, ce que l'annulation de la dérivée ne garantit pas à elle seule. La variance minimale vaut $\dfrac{S^2}{C}\Big(\sqrt{c_1\,\rho} + \sqrt{c_2\,(1-\rho)}\Big)^2$.

Ce que cela dit. Le nombre optimal de ménages par district grandit avec le rapport c_1/c_2 des coûts de déplacement et d'entretien, et diminue avec ρ. Pour $c_1 = 800$, $c_2 = 60$ et $\rho = 0{,}162$: $m^* = 8{,}3$. L'optimum est plat : $g(m)$ dépasse son minimum de moins de 5 % pour m de 6 à 13, de 19,5 % pour $m = 20$ et de 71,5 % pour $m = 40$.

■

Sur la population du cours

Budget de 25 000 dh

$m^* = \sqrt{(800/60) \times (0{,}838/0{,}162)} = \sqrt{13{,}33 \times 5{,}17} = 8{,}3$.

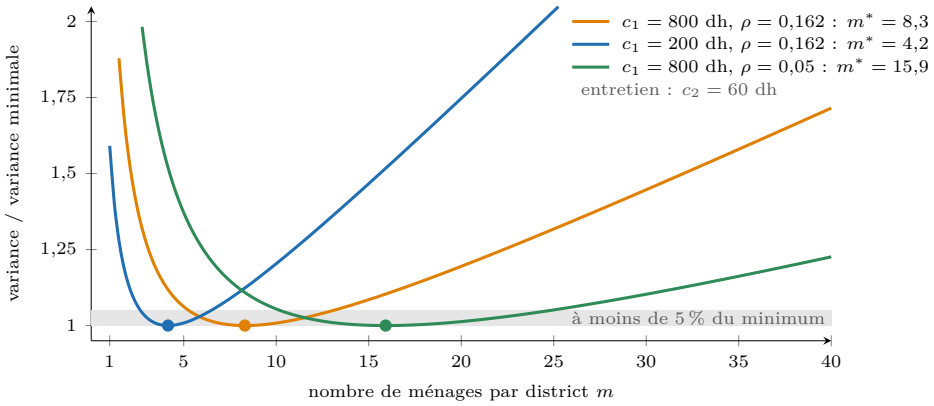
m	n_I	n	Erreur-type, modèle	Erreur-type, population
2	27,2	54	582,6	568,1

4	24,0	96	495,3	464,5
8	19,5	156	465,6	428,7
12	16,4	197	473,0	438,5
20	12,5	250	508,9	483,4
40	7,8	313	609,7	606,7

Le modèle et le calcul exact s'accordent

La dernière colonne est calculée sur les 50 districts, avec l'estimateur par le ratio. Les deux calculs placent l'optimum en $m = 8$: moins de ménages, mais plus de districts.

Un optimum plat



Lecture

Avec $c_1 = 800$ dh, tout m entre 6 et 13 reste à moins de 5 % de la variance minimale ; $m = 20$ coûte 19,5 % de variance en plus, $m = 40$, 71,5 %. En ville ($c_1 = 200$ dh), $m^* = 4{,}2$; si les ménages d'un district se ressemblent moins ($\rho = 0{,}05$), $m^* = 15{,}9$.

Ce qui fait varier le nombre optimal

Le nombre optimal selon les coûts et la ressemblance

ρ	$c_1/c_2 = 5$	$c_1/c_2 = 10$	$c_1/c_2 = 20$
0,02	15,7	22,1	31,3
0,05	9,7	13,8	19,5
0,162 (dépense)	5,1	7,2	10,2
0,3	3,4	4,8	6,8

Lecture économique

On interroge beaucoup de ménages par district quand le **déplacement coûte cher** et que les ménages d'un même district **diffèrent** (ρ petit). ρ dépend de la variable : 0,162 pour la dépense, 0,120 pour l'accès à internet ($m^* = 9{,}9$ avec les coûts ruraux). Une enquête à

plusieurs objectifs choisit m pour sa variable principale ; l'optimum étant plat, l'erreur coûte peu.

Le plan autopondéré

Des poids tous égaux

Autopondération

Un plan est **autopondéré** si tous les ménages ont la même probabilité d'inclusion, donc le même poids. Comme $\pi_k = f_1 \times m_i/M_i$, deux façons d'y parvenir :

1. Tirer dans chaque district la **même fraction** f_2 de ses ménages : $\pi_k = f_1\, f_2$.
2. Tirer les districts **proportionnellement à leur taille**, puis m ménages dans chacun : $\pi_k = n_I \dfrac{M_i}{N} \cdot \dfrac{m}{M_i} = \dfrac{n_I\, m}{N}$ (chapitre 11).

Sur la population du cours

1. $f_1 = 0{,}2$ et $f_2 = 0{,}5$: $\pi_k = 0{,}1$, poids 10. Mais la taille de l'échantillon dépend des districts tirés : environ 107 ménages pour les dix plus petits, 293 pour les dix plus grands.
2. 10 districts proportionnellement à leur taille, 20 ménages dans chacun : $\pi_k = 10 \times 20/2\,000 = 0{,}1$, toujours 200 ménages, et la même charge de travail dans chaque district.

Le piège de l'autopondération

Une estimation simple

Dans un plan autopondéré, l'estimateur par le ratio est la **moyenne simple** des ménages interrogés : $\widehat{Y}/\widehat{N} = \sum_{k\in S} y_k/n$. Avec la fraction $f_2 = 0{,}5$, son erreur-type vaut 547 dh, effet de plan 4,19, proche du ratio 10×20 (4,32).

Mais pas une variance simple

Sur 20 000 tirages du second plan (graine 2026), on estime $\overline{Y}$ par la moyenne des dix moyennes de district. L'intervalle calculé comme pour un sondage aléatoire simple de 200 ménages a une demi-largeur moyenne de 519 dh et ne contient $\overline{Y}$ que 64,6 % du temps. Calculé à partir de la dispersion entre les dix districts (t à 9 degrés de liberté), il a une demi-largeur de 1 238 dh et contient $\overline{Y}$ 95,0 % du temps.

S'entraîner

Exercice 1 : quatre douars

Deux douars, puis deux ménages dans chacun

Quatre douars de 4 ménages ont des dépenses, en milliers de dh : D1 : $3, 5, 6, 6$; D2 : $6, 7, 9, 10$; D3 : $8, 10, 11, 11$; D4 : $4, 6, 7, 7$. On tire 2 douars, puis 2 ménages dans chaque douar tiré, par sondages aléatoires simples.

1. Calculer la probabilité d'inclusion et le poids d'un ménage.
2. On a tiré D2, avec les ménages 7 et 9, et D3, avec 8 et 11. Estimer le total Y.
3. Calculer V_1 et V_2, sachant que $S_T^2 = 236/3$ et que les S_i^2 valent 2 ; 10/3 ; 2 ; 2. Quel terme domine ?
4. Comparer à un sondage aléatoire simple de 4 ménages parmi 16 ($S^2 = 5{,}8$).

Exercice 1 : solution

Probabilités et estimation

1. $\pi_k = \frac{2}{4} \times \frac{2}{4} = 0{,}25$, poids 4 : chaque ménage tiré en représente quatre.
2. $\widehat{Y}_2 = 4 \times 8 = 32$ et $\widehat{Y}_3 = 4 \times 9{,}5 = 38$; $\widehat{Y} = \frac{4}{2} \times 70 = 140$, soit $4 \times (7 + 9 + 8 + 11)$. La vraie valeur est $Y = 116$.

Les deux termes

3. $V_1 = 4^2 \times 0{,}5 \times \frac{236/3}{2} = 314{,}67$; $V_2 = \frac{4}{2} \times 16 \times 0{,}5 \times \frac{2+10/3+2+2}{2} = 74{,}67$. $V = 389{,}33$, dont 81 % pour le premier degré. L'énumération des 216 échantillons possibles donne la même variance.
4. $16^2 \times 0{,}75 \times 5{,}8/4 = 278{,}4$: effet de plan $389{,}33/278{,}4 = 1{,}40$.

Exercice 2 : le budget d'une enquête rurale

Combien de ménages par douar

Une enquête rurale coûte 1 500 dh par douar visité et 75 dh par ménage interrogé. Une enquête antérieure donne $\rho = 0{,}08$. Le budget est de 60 000 dh.

1. Calculer le nombre optimal m^* de ménages par douar.
2. En déduire le nombre de douars, le nombre de ménages, l'effet de plan $1+(m-1)\rho$ et la taille effective.
3. Mêmes questions avec $m = 40$. Conclure.

Exercice 2 : solution

L'optimum

1. $m^* = \sqrt{\frac{1\,500}{75} \times \frac{0{,}92}{0{,}08}} = \sqrt{20 \times 11{,}5} = \sqrt{230} = 15{,}2$, soit 15 ménages.
2. Un douar coûte $1\,500 + 15 \times 75 = 2\,625$ dh : $60\,000/2\,625 = 22{,}9$ douars, 343 ménages. Effet de plan $1 + 14 \times 0{,}08 = 2{,}12$; taille effective $343/2{,}12 = 162$ ménages.

Quarante ménages par douar

3. Un douar coûte 4 500 dh : 13,3 douars, 533 ménages. Effet de plan 4,12 ($1 + 39 \times 0{,}08$), taille effective 129. Plus de ménages interrogés, mais une précision moindre : la variance est 25 % plus forte. Interroger beaucoup de voisins coûte des douars.

Exercice 3 : vingt districts ou dix

Deux plans de 200 ménages

Sur la population du cours, l'estimateur d'expansion a pour composantes de variance, en 10^{12} : $V_1 = 4{,}089$ et $V_2 = 0{,}170$ pour 10×20 ; $V_1 = 1{,}533$ et $V_2 = 0{,}237$ pour 20×10. Un sondage aléatoire simple de 200 ménages donne $V(\overline{y}) = 71\,435$.

1. Calculer la probabilité d'inclusion d'un ménage du district 1 (71 ménages) sous chacun des deux plans.
2. Calculer l'erreur-type de la moyenne, l'effet de plan et la taille effective de chaque plan.
3. Avec $c_1 = 800$ dh et $c_2 = 60$ dh, calculer le coût de chaque plan. Commenter.

Exercice 3 : solution

Mêmes probabilités, autres variances

1. $0{,}2 \times 20/71 = 0{,}056$ et $0{,}4 \times 10/71 = 0{,}056$: **les mêmes**. Les π_k ne dépendent que du produit $n_I\, m$; ils ne disent rien de la précision.
2. 10×20 : $\sqrt{4{,}259 \cdot 10^{12}/2\,000} = 1\,032$ dh, effet de plan 14,9, taille effective 13. 20×10 : $\sqrt{1{,}770 \cdot 10^{12}/2\,000} = 665$ dh, effet de plan 6,19, taille effective 32.

Le coût

3. $10 \times 800 + 200 \times 60 = 20\,000$ dh contre $20 \times 800 + 200 \times 60 = 28\,000$ dh. Pour 40 % de dépense en plus, la variance est divisée par 2,4 : les districts supplémentaires valent leur prix, car le premier degré domine.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. On tire des districts, puis des ménages dans les districts tirés, seuls à dénombrer.
2. $\pi_k = \frac{n_I}{N_I} \cdot \frac{m_i}{M_i}$; le poids d'un ménage est le produit des poids des deux degrés.
3. $\widehat{Y} = \frac{N_I}{n_I} \sum_{i \in S_I} M_i\, \overline{y}_i$ est sans biais.
4. $V(\widehat{Y}) = V_1 + V_2$; le premier degré domine (96 % pour 10×20).
5. Le ratio $\widehat{Y}/\widehat{N}$ ramène l'effet de plan de 14,9 à 4,32.
6. $m^* = \sqrt{\frac{c_1}{c_2} \cdot \frac{1-\rho}{\rho}}$, un optimum plat (8,3 ménages pour l'enquête rurale).
7. Un plan autopondéré simplifie l'estimation ; la variance s'estime **entre districts**.

La suite

Le chapitre 11 tire les districts avec des probabilités **proportionnelles à leur taille** : un plan autopondéré et plus précis. Pour 5 districts entiers, l'erreur-type de la moyenne tombe de 1 516,6 à 754,3 dh.