Échantillonnage

Chapitre 6 : Le sondage stratifié

Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 21 octobre 2026

Plan du chapitre

Le principe

L'estimateur et sa variance

L'allocation proportionnelle

D'où vient le gain

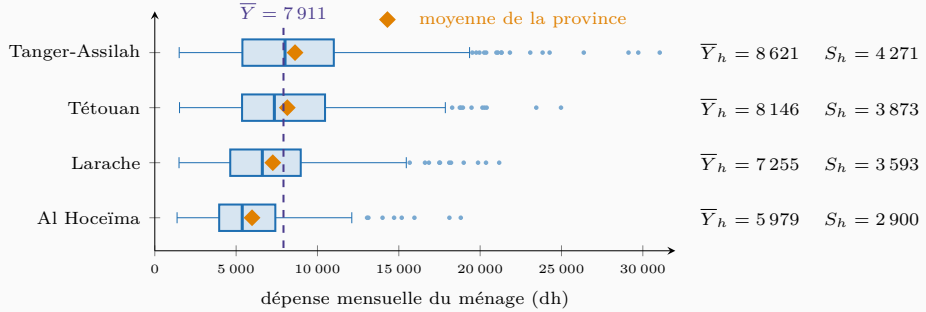
Stratifier en pratique

S'entraîner

Ce qu'il faut retenir

Le principe

Des provinces qui ne se ressemblent pas



Des moyennes différentes, des ménages très dispersés

Les moyennes vont de $5\,979$ dh (Al Hoceïma) à $8\,621$ dh (Tanger-Assilah). Mais **à l'intérieur** de chaque province, les dépenses vont de moins de $1\,600$ dh à plus de $18\,000$ dh : les boîtes se chevauchent. Tirer séparément dans chaque province rend-il l'enquête plus précise ?

Tirer séparément dans chaque strate

Le plan stratifié

On partage la population U en H **strates** $U_1, \ldots, U_H$, disjointes et qui la recouvrent entièrement. Dans chaque strate h, on tire un sondage aléatoire simple de n_h unités, **indépendamment** des autres strates. La taille de l'échantillon est $n = \sum_h n_h$.

Trois raisons de stratifier

1. **La précision** : si les strates sont homogènes, l'estimation varie moins d'un échantillon à l'autre.
2. **Des résultats par strate** : chaque province est sûre d'avoir des ménages dans l'échantillon, en nombre choisi à l'avance.
3. **Le terrain** : une équipe d'enquêteurs et un budget par province.

Les notations, strate par strate

Population : N_h unités, poids $W_h = N_h/N$, moyenne $\overline{Y}_h$, variance S_h^2 (en $N_h - 1$).
Échantillon : n_h unités, taux $f_h = n_h/N_h$, moyenne $\overline{y}_h$, variance s_h^2 (en $n_h - 1$).

Les quatre provinces

Province	N_h	W_h	$\overline{Y}_h$	S_h
Tanger-Assilah	820	0,41	8 621,04	4 271,19
Tétouan	560	0,28	8 145,93	3 873,11
Larache	380	0,19	7 254,97	3 593,17
Al Hoceïma	240	0,12	5 979,21	2 899,90
Ensemble	2 000	1	7 911,44	3 984,29

La moyenne générale est la moyenne des $\overline{Y}_h$ **pondérée** par les W_h : $\overline{Y} = \sum_h W_h \overline{Y}_h$.

La mini-population en deux strates

Au chapitre 1, on tirait un ménage parmi A, B, C (dépenses 4, 6, 7 milliers de dh) et un ménage parmi D, E (9, 14). C'est un plan stratifié :

$$H = 2, \qquad N_1 = 3,\ N_2 = 2, \qquad n_1 = n_2 = 1, \qquad W_1 = \tfrac{3}{5},\ W_2 = \tfrac{2}{5}$$

La moyenne pondérée était déjà l'estimateur stratifié

La moyenne pondérée $\frac{3}{5}\,y_1 + \frac{2}{5}\,y_2$ du chapitre 1 est exactement $W_1\,\overline{y}_1 + W_2\,\overline{y}_2$. Ce chapitre en calcule la variance, et montre pourquoi elle est **plus petite** que celle du sondage aléatoire simple de deux ménages.

L’estimateur et sa variance

Moyenne et total

$$\overline{y}_{st} = \sum_{h=1}^{H} W_h \, \overline{y}_h \qquad \widehat{Y}_{st} = N \, \overline{y}_{st} = \sum_{h=1}^{H} N_h \, \overline{y}_h$$

Sans biais, en une ligne

Dans chaque strate, $\overline{y}_h$ est la moyenne d'un sondage aléatoire simple : $E(\overline{y}_h) = \overline{Y}_h$ (chapitre 3). Donc

$$E(\overline{y}_{st}) = \sum_h W_h \, \overline{Y}_h = \overline{Y}$$

C'est aussi l'estimateur de Horvitz-Thompson (chapitre 2) : $\pi_k = n_h / N_h$ dans la strate h, d'où $\sum_{k \in S} y_k / \pi_k = \sum_h \frac{N_h}{n_h} \sum_{k \in S_h} y_k = \sum_h N_h \, \overline{y}_h$.

Une somme de variances de strates

$$V(\overline{y}_{st}) = \sum_{h=1}^{H} W_h^2 \, (1 - f_h) \, \frac{S_h^2}{n_h}$$

La démonstration

Les tirages des strates sont indépendants : les covariances entre les $\overline{y}_h$ sont nulles, et

$$V\Big(\sum_h W_h \, \overline{y}_h \Big) = \sum_h W_h^2 \, V(\overline{y}_h), \qquad V(\overline{y}_h) = (1 - f_h) \, \frac{S_h^2}{n_h} \text{ (chapitre 3)}$$

Seules les dispersions à l'intérieur des strates apparaissent : les écarts entre les moyennes $\overline{Y}_h$ ne comptent plus.

La formule

- Strate 1 $(4, 6, 7)$: $\overline{Y}_1 = 17/3$, $S_1^2 = 7/3$, $f_1 = 1/3$, d'où $V(\overline{y}_1) = \frac{2}{3} \times \frac{7}{3} = 1{,}556$.
- Strate 2 $(9, 14)$: $\overline{Y}_2 = 11{,}5$, $S_2^2 = 12{,}5$, $f_2 = 1/2$, d'où $V(\overline{y}_2) = 6{,}25$.

$$V(\overline{y}_{st}) = 0{,}36 \times 1{,}556 + 0{,}16 \times 6{,}25 = 0{,}56 + 1 = 1{,}56$$

La vérification par énumération

Les six moyennes pondérées du chapitre 1 valent 6 ; 8 ; $7{,}2$; $9{,}2$; $7{,}8$; $9{,}8$. Leur moyenne est 8, et leurs carrés d'écarts à 8 somment à $9{,}36$: $V = 9{,}36/6 = 1{,}56$. Le sondage aléatoire simple de deux ménages avait une variance de $4{,}35$: **stratifier l'a divisée par $2{,}8$**.

La variance estimée

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(1 - f_h\right) \frac{s_h^2}{n_h}$$

Elle est sans biais, car s_h^2 est sans biais pour S_h^2 dans chaque strate (chapitre 3). Intervalle de confiance à $95\,\% : \overline{y}_{st} \pm 1{,}96 \sqrt{\widehat{V}(\overline{y}_{st})}$.

Au moins deux unités par strate

Avec $n_h = 1$, s_h^2 ne peut pas être calculée. Le plan de la mini-population a une variance, $1{,}56$, mais son échantillon ne permet pas de l'**estimer**. En pratique, on tire au moins deux unités dans chaque strate, ou l'on regroupe les strates trop petites.

L’allocation proportionnelle

L'allocation proportionnelle

Chaque strate reçoit, dans l'échantillon, sa part dans la population :

$$n_h = n\,\frac{N_h}{N} = n\,W_h \qquad \text{d'où} \qquad f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

200 ménages répartis entre les provinces

Province	Tanger-Assilah	Tétouan	Larache	Al Hoceïma
N_h	820	560	380	240
$n_h = 200\,W_h$	82	56	38	24

Ici les n_h tombent juste. Par milieu, $200 \times 0{,}719 = 143{,}8$ et $200 \times 0{,}281 = 56{,}2$ s'arrondissent à 144 urbains et 56 ruraux.

Le même poids pour tous

Avec l'allocation proportionnelle, $\pi_k = n_h/N_h = f$ dans toutes les strates : chaque ménage tiré a le poids $d_k = N/n = 10$. Et comme $\frac{N_h}{N} \cdot \frac{1}{n_h} = \frac{1}{n}$,

$$\overline{y}_{st} = \sum_h \frac{N_h}{N} \frac{1}{n_h} \sum_{k \in S_h} y_k = \frac{1}{n} \sum_{k \in S} y_k = \overline{y}$$

Commode pour l'utilisateur

L'échantillon s'exploite comme un sondage aléatoire simple : moyennes et proportions **non pondérées** sont sans biais. Les arrondis l'écartent à peine de l'idéal : par milieu, les poids valent $1\,438/144 = 9{,}99$ et $562/56 = 10{,}04$. Mais attention : la **variance**, elle, n'est pas celle d'un sondage aléatoire simple.

Une formule simple

$$V_{prop} = \frac{1-f}{n} \sum_h W_h \, S_h^2 \qquad \text{à comparer à} \qquad V_{SAS} = \frac{1-f}{n} \, S^2$$

Il suffit de remplacer $n_h = n \, W_h$ et $f_h = f$ dans la formule générale :

$$W_h^2 \, (1-f) \, \frac{S_h^2}{n \, W_h} = \frac{1-f}{n} \, W_h \, S_h^2.$$

Les provinces, 200 ménages

$$\sum_h W_h S_h^2 = 7\,479\,671 + 4\,200\,279 + 2\,453\,070 + 1\,009\,128 = 15\,142\,148$$

$V_{prop} = \frac{0{,}9}{200} \times 15\,142\,148 = 68\,139{,}7$, soit une erreur-type de $261{,}0$ dh, contre $V_{SAS} = 71\,435{,}5$ et $267{,}3$ dh.

Comparer un plan au sondage aléatoire simple

L’**effet de plan** rapporte la variance d’un plan à celle d’un sondage aléatoire simple de **même taille** :

$$deff = \frac{V_{plan}}{V_{SAS}} \qquad \text{ici} \qquad deff = \frac{\sum_h W_h \, S_h^2}{S^2}$$

Moins de 1 : le plan est plus précis. Le chapitre 7 en fera l’outil de comparaison de tous les plans.

Stratifier par province

$deff = 15\,142\,148 / 15\,874\,548 = 0{,}954$: la variance baisse de $4{,}6\,\%$, l’erreur-type passe de $267{,}3$ à $261{,}0$ dh. Un gain réel, mais **modeste**, alors que les moyennes provinciales diffèrent de plus de $2\,600$ dh. La section suivante dit pourquoi.

D’où vient le gain

Variance totale, variance intra, variance inter

$$(N-1)\,S^2 = \sum_h (N_h - 1)\,S_h^2 + \sum_h N_h\,(\overline{Y}_h - \overline{Y})^2$$

En divisant par $N-1$, avec $N_h - 1 \approx N_h$ et $N - 1 \approx N$:

$$S^2 \approx \underbrace{\sum_h W_h\,S_h^2}_{\text{intra-strates}} + \underbrace{\sum_h W_h\,(\overline{Y}_h - \overline{Y})^2}_{\text{inter-strates}}$$

La démonstration

Dans la strate h, on écrit $y_k - \overline{Y} = (y_k - \overline{Y}_h) + (\overline{Y}_h - \overline{Y})$ et l'on élève au carré. Le double produit disparaît, car $\sum_{k \in U_h} (y_k - \overline{Y}_h) = 0$; il reste $(N_h - 1)\,S_h^2 + N_h\,(\overline{Y}_h - \overline{Y})^2$. On additionne sur les strates. C'est la décomposition de la variance de la statistique descriptive : variance **intra** plus variance **inter**.

Ce que la stratification retire

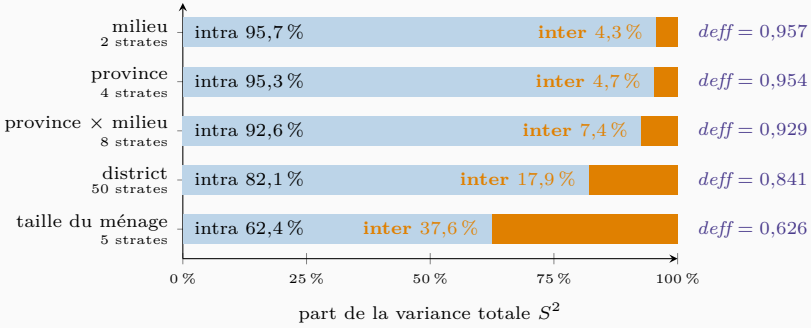
$$V_{SAS} - V_{prop} = \frac{1-f}{n}\Big(S^2 - \sum_h W_h\, S_h^2\Big) \approx \frac{1-f}{n}\sum_h W_h\,(\overline{Y}_h - \overline{Y})^2 \;\geq\; 0$$

Stratifier retire de la variance la part due aux écarts **entre** strates : $deff \approx 1 -$ part inter.

Les provinces

Part inter : $206\,449 + 15\,396 + 81\,879 + 448\,020 = 751\,744$, soit $4{,}7\,\%$ de S^2. Al Hoceïma en fournit $60\,\%$ à elle seule : petite, mais très différente. Gain : $V_{SAS} - V_{prop} = 3\,295{,}8$, proche de $0{,}0045 \times 751\,744 = 3\,382{,}8$.

Cinq façons de stratifier



Ce que montre la figure

- Province et milieu captent à peine $5\,\%$ de la variance ; les croiser, $7{,}4\,\%$ ($deff = 0{,}929$).
- La **taille du ménage** en capte $37{,}6\,\%$: la variance baisserait de plus d'un tiers.
- Le **district**, $17{,}9\,\%$: un atout pour stratifier, un handicap pour tirer des districts entiers (chapitre 9).

Les ménages connectés, stratifiés par milieu

Milieu	W_h	P_h	$S_h^2 = \frac{N_h}{N_h - 1} P_h(1 - P_h)$
Rural	$0{,}281$	$0{,}434$	$0{,}24610$
Urbain	$0{,}719$	$0{,}768$	$0{,}17844$

$\sum_h W_h S_h^2 = 0{,}06916 + 0{,}12830 = 0{,}19746$, contre $S^2 = 0{,}21983$: $deff = 0{,}898$.
L'erreur-type passe de $0{,}0315$ à $0{,}0298$.

Une stratification sert plus ou moins chaque variable

Le taux de connexion diffère beaucoup entre milieux : la part inter atteint $10{,}2\,\%$, contre $4{,}3\,\%$ pour la dépense. À l'inverse, la province sert peu pour internet ($deff = 0{,}966$). Une enquête mesure plusieurs variables : la stratification est un **compromis**.

Stratifier en pratique

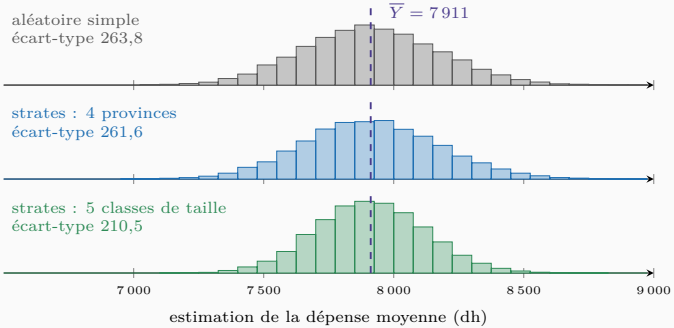
Trois conditions

1. La variable de stratification est **connue pour chaque unité** de la base de sondage, avant le tirage.
2. Elle est **liée à la variable d'intérêt** : strates homogènes, moyennes éloignées.
3. Elle sert la **diffusion** et le **terrain** : résultats par province ou par milieu, équipes par zone.

La province ou la taille du ménage

Supposons que la base donne la taille de chaque ménage. Pour la dépense, stratifier par taille fait bien mieux que par province ($deff$ de $0{,}626$ contre $0{,}954$), mais ne garantit aucun effectif par province. On **croise** souvent les deux : province $\times$ taille, 20 strates, $deff = 0{,}553$.

Ce que montre la simulation



20 000 enquêtes de 200 ménages pour chaque plan

Les trois histogrammes sont centrés sur $\overline{Y}$: pas de biais. Stratifier par province se voit à peine ($261{,}6$ contre $263{,}8$) ; par taille, l'écart-type baisse de $20\,\%$. À moins de 300 dh de $\overline{Y}$: $85\,\%$ des estimations stratifiées par taille, $74\,\%$ en aléatoire simple.

82, 56, 38 et 24 ménages tirés dans les provinces

Province	n_h	$\overline{y}_h$	s_h	vraie $\overline{Y}_h$
Tanger-Assilah	82	8 264,88	3 580	8 621,04
Tétouan	56	7 644,29	3 993	8 145,93
Larache	38	6 498,42	4 075	7 254,97
Al Hoceïma	24	5 643,75	2 585	5 979,21

Comment on a tiré

Province après province, on tire sans remise n_h ménages parmi les N_h, avec le générateur `numpy.random.default_rng(2026)`. La graine fixée rend le tirage **reproductible** : chacun peut retrouver ces 200 ménages.

Les calculs

$\overline{y}_{st} = 0{,}41 \times 8\,264{,}88 + \cdots + 0{,}12 \times 5\,643{,}75 = 7\,440{,}95$ dh : c’est la moyenne simple des 200 ménages, car l’échantillon est autopondéré.

$$\widehat{V} = \frac{0{,}9}{200} \sum_{h} W_h \, s_h^2 = 0{,}0045 \times 13\,676\,976 = 61\,546{,}4$$

Erreur-type $248{,}1$ dh ; intervalle de confiance à $95\,\%$: $[6\,954{,}7\,;\,7\,927{,}2]$.

Un échantillon malchanceux

L’estimation se trompe de -470 dh, contre -68 dh pour l’échantillon aléatoire simple fil rouge : les quatre provinces sont sous-estimées à la fois. Un tel écart, $1{,}8$ erreur-type, arrive environ 7 fois sur 100 ; l’intervalle contient $\overline{Y} = 7\,911{,}44$, de justesse. **Un échantillon ne juge pas un plan** : seule la distribution sur tous les tirages le fait.

Ce qu'il faut respecter

- Les strates sont **disjointes** et couvrent **toute** la base : chaque ménage appartient à une strate et une seule.
- On tire au moins **deux unités** par strate, pour pouvoir estimer la variance.
- Si l'allocation n'est pas proportionnelle, chaque unité tirée garde son poids $d_k = N_h/n_h$: il faut **pondérer** (chapitre 1).

Ce que la stratification ne fait pas

Avec l'allocation proportionnelle, stratifier ne fait pratiquement jamais perdre de précision : si les moyennes des strates sont égales, on retrouve le sondage aléatoire simple. Mais la stratification ne corrige ni une base incomplète, ni la non-réponse (chapitre 14). Et une autre répartition de l'échantillon peut faire mieux, ou moins bien : c'est l'objet du chapitre 7.

S’entraîner

Cinq boutiques en deux strates

Cinq boutiques réalisent un chiffre d'affaires mensuel de 3, 5 et 7 milliers de dh (les petites) et de 12 et 16 milliers de dh (les grandes). On tire 2 petites et 1 grande, par sondage aléatoire simple dans chaque strate.

1. Calculer W_1, W_2, $\overline{Y}$, S_1^2 et S_2^2.
2. Énumérer les 6 échantillons possibles et la valeur de $\overline{y}_{st}$ pour chacun. Vérifier que $E(\overline{y}_{st}) = \overline{Y}$.
3. Calculer $V(\overline{y}_{st})$ par énumération, puis par la formule.
4. Comparer à un sondage aléatoire simple de 3 boutiques, sachant que $S^2 = 28{,}3$.

Exercice 1 : solution

Les six échantillons

1. $W_1 = 3/5, W_2 = 2/5, \overline{Y} = 43/5 = 8{,}6$; $S_1^2 = (4+0+4)/2 = 4$; $S_2^2 = (4+4)/1 = 8$.
2. Les paires de petites ont pour moyennes 4, 5 et 6. $\overline{y}_{st} = 0{,}6\,\overline{y}_1 + 0{,}4\,y_2$ vaut $7{,}2$; $8{,}8$; $7{,}8$; $9{,}4$; $8{,}4$; 10. Moyenne : $51{,}6/6 = 8{,}6 = \overline{Y}$.

La variance

3. Écarts à $8{,}6$: $\pm 1{,}4$, $\pm 0{,}8$ et $\pm 0{,}2$; $V = (2 \times 1{,}96 + 2 \times 0{,}64 + 2 \times 0{,}04)/6 = 5{,}28/6 = 0{,}88$. Formule : $0{,}36 \times \frac{1}{3} \times \frac{4}{2} + 0{,}16 \times \frac{1}{2} \times \frac{8}{1} = 0{,}24 + 0{,}64 = 0{,}88$.
4. $V_{SAS} = (1 - 3/5) \times 28{,}3/3 = 3{,}77$: stratifier divise la variance par plus de quatre. Les deux strates sont très différentes : la variance inter domine.

Urbain et rural, 200 ménages

Milieu	N_h	$\overline{Y}_h$	S_h
Urbain	1 438	8 429,78	4 134,11
Rural	562	6 585,14	3 215,03

On tire $n = 200$ ménages avec allocation proportionnelle. On rappelle $S^2 = 15\,874\,548$.

1. Retrouver $\overline{Y}$. Calculer la part inter-strates $\sum_h W_h(\overline{Y}_h - \overline{Y})^2$ et la rapporter à S^2.
2. Calculer V_{prop} et l'effet de plan. Comparer à la stratification par province ($0{,}954$).
3. Pour la proportion de ménages connectés, l'effet de plan vaut $0{,}898$. Pourquoi est-il plus faible ?

La décomposition

1. $W_h = 0{,}719$ et $0{,}281$; $\overline{Y} = 0{,}719 \times 8\,429{,}78 + 0{,}281 \times 6\,585{,}14 = 7\,911{,}44$. Inter : $0{,}719 \times 518{,}34^2 + 0{,}281 \times 1\,326{,}29^2 = 193\,180 + 494\,294 = 687\,474$, soit $4{,}3\,\%$ de S^2.
2. $\sum_h W_h S_h^2 = 12\,288\,314 + 2\,904\,536 = 15\,192\,850$; $V_{prop} = 0{,}0045 \times 15\,192\,850 = 68\,367{,}8$; $deff = 15\,192\,850 / 15\,874\,548 = 0{,}957$, à peine moins bon que la province.

Internet

3. Les taux de connexion diffèrent bien plus entre milieux ($0{,}434$ contre $0{,}768$) que les dépenses : la part inter de la variance est plus grande ($10{,}2\,\%$ contre $4{,}3\,\%$), donc le gain aussi.

Un échantillon qui n'est pas proportionnel

Comme à l'exercice 3 du chapitre 1, on tire 100 ménages urbains et 100 ruraux (graine 2026) :

Milieu	N_h	n_h	$\overline{y}_h$	s_h
Urbain	$1\,438$	100	$8\,280{,}30$	$3\,527{,}90$
Rural	562	100	$6\,463{,}60$	$3\,250{,}92$

1. Calculer $\overline{y}_{st}$, puis la moyenne non pondérée des 200 ménages. Laquelle retenir ?
2. Calculer $\widehat{V}(\overline{y}_{st})$ et l'intervalle de confiance à $95\,\%$.
3. L'échantillon est-il autopondéré ?

L'estimation

1. $\overline{y}_{st} = 0{,}719 \times 8\,280{,}30 + 0{,}281 \times 6\,463{,}60 = 5\,953{,}5 + 1\,816{,}3 = 7\,769{,}8$ dh. Non pondérée : $(8\,280{,}30 + 6\,463{,}60)/2 = 7\,371{,}95$ dh, biaisée : les ruraux font la moitié de l'échantillon, mais $28\,\%$ de la population. On retient $\overline{y}_{st}$.

La précision

2. $\widehat{V} = 0{,}719^2 \left(1 - \frac{100}{1\,438}\right) \frac{3\,527{,}90^2}{100} + 0{,}281^2 \left(1 - \frac{100}{562}\right) \frac{3\,250{,}92^2}{100} = 59\,866{,}9 + 6\,860{,}1 = 66\,727{,}0$; erreur-type $258{,}3$; intervalle $[7\,263{,}5\,;\,8\,276{,}1]$.
3. Non : poids $14{,}38$ en ville, $5{,}62$ à la campagne. La variance exacte de ce plan, $88\,918$, dépasse celle de l'allocation proportionnelle, $68\,368$: le chapitre 7 dira comment bien répartir.

Ce qu’il faut retenir

Sept points

1. Un plan **stratifié** tire un sondage aléatoire simple indépendant dans chaque strate.
2. $\overline{y}_{st} = \sum_h W_h \, \overline{y}_h$ est sans biais : c'est l'estimateur de Horvitz-Thompson.
3. $V(\overline{y}_{st}) = \sum_h W_h^2 \, (1 - f_h) \, S_h^2 / n_h$, estimée avec les s_h^2 : $n_h \geq 2$.
4. Allocation **proportionnelle** : échantillon autopondéré, $V_{prop} = \frac{1-f}{n} \sum_h W_h S_h^2$.
5. $S^2 \approx$ intra + inter : stratifier supprime la variance **inter-strates**.
6. $deff = V_{plan} / V_{SAS}$: $0{,}954$ par province, $0{,}626$ par taille du ménage.
7. On stratifie sur une variable connue dans la base, liée à y, utile à la diffusion.

La suite

L'allocation proportionnelle n'est pas la meilleure : le chapitre 7 répartit l'échantillon selon la dispersion et le coût des strates (Neyman), et compare tous les plans par leur effet de plan.