Échantillonnage

Chapitre 7 : L'allocation optimale et l'effet de plan

Jaouad Madkour

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

Plan du chapitre

Le problème de l'allocation

L'allocation de Neyman

Coûts et dispersions inconnues

L'allocation égale et les domaines

L'effet de plan

Proportions et plusieurs variables

S'entraîner

Ce qu'il faut retenir

Le problème de l’allocation

La variance dépend de la répartition

La taille n et les strates sont fixées. Reste à choisir $n_1, \ldots, n_H$, avec $\sum_h n_h = n$. Comme $W_h/N_h = 1/N$:

$$V(\overline{y}_{st}) = \sum_h W_h^2 \,(1 - f_h)\, \frac{S_h^2}{n_h} = \underbrace{\sum_h \frac{W_h^2\, S_h^2}{n_h}}_{\text{dépend des } n_h} - \underbrace{\frac{1}{N} \sum_h W_h\, S_h^2}_{\text{fixe}}$$

Sur les provinces, le terme fixe vaut $15\,142\,148/2\,000 = 7\,571{,}1$.

Trois objectifs, trois répartitions

1. La **précision de la moyenne d'ensemble** : l'allocation de Neyman.
2. Le **coût** : tirer moins là où une interview coûte cher.
3. La **précision dans chaque strate** : l'allocation égale, ou un compromis.

Strates A, B, C et D, E : trois ménages à répartir

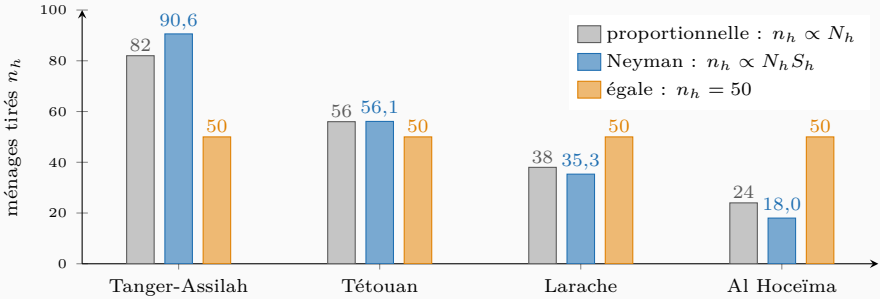
Répartition (n_1, n_2)	$(2,1)$	$(1,2)$
Strate 1 : $0{,}36\,(1-f_1)\,S_1^2/n_1$	$0{,}14$	$0{,}56$
Strate 2 : $0{,}16\,(1-f_2)\,S_2^2/n_2$	1	0
$V(\overline{y}_{st})$	$1{,}14$	$0{,}56$

Dépenses $4, 6, 7$ puis $9, 14$: $S_1^2 = 7/3$ et $S_2^2 = 12{,}5$; poids $W_1 = 0{,}6$ et $W_2 = 0{,}4$.

Tirer davantage là où cela varie

L'allocation proportionnelle ($1{,}8$ et $1{,}2$) conduit à $(2,1)$. Or la strate D, E est petite mais **très dispersée** : en la recensant, on supprime sa variance, et V est divisée par deux. Vérification : les trois échantillons possibles donnent 7 ; $8{,}2$; $8{,}8$, de variance $0{,}56$. Un sondage aléatoire simple de trois ménages : $1{,}93$.

Trois répartitions de 200 ménages



Même taille, trois précisions

Variance de $\overline{y}_{st}$: $68\,139{,}7$ (proportionnelle), $67\,172{,}3$ (Neyman), $89\,027{,}4$ (égale), contre $71\,435{,}5$ pour un sondage aléatoire simple. La meilleure répartition **suit la population** en l'inclinant vers les provinces les plus dispersées ; la répartition égale fait **moins bien** que le hasard simple.

L’allocation de Neyman

L'allocation de Neyman (1934)

Pour une taille n fixée, la variance de $\overline{y}_{st}$ est minimale pour

$$n_h = n \, \frac{N_h \, S_h}{\sum_k N_k \, S_k}$$

Le taux de sondage $f_h = n_h / N_h$ est alors **proportionnel à** S_h.

Deux raisons de tirer davantage dans une strate

1. Elle est **grande** (N_h) : elle pèse lourd dans la moyenne d'ensemble.
2. Elle est **dispersée** (S_h) : la même précision y demande plus d'unités.

Si tous les S_h sont égaux, on retrouve l'allocation proportionnelle $n_h = n \, W_h$.

Par l'inégalité de Cauchy-Schwarz

On minimise $\sum_h W_h^2 S_h^2 / n_h$ sous la contrainte $\sum_h n_h = n$. Pour des nombres positifs, $\left(\sum a_h^2\right)\left(\sum b_h^2\right) \geq \left(\sum a_h b_h\right)^2$; avec $a_h = \sqrt{n_h}$ et $b_h = W_h S_h / \sqrt{n_h}$:

$$n \times \sum_h \frac{W_h^2\, S_h^2}{n_h} \;\geq\; \Big(\sum_h W_h\, S_h\Big)^2$$

L'égalité a lieu quand $b_h / a_h = W_h S_h / n_h$ est constant : $n_h \propto W_h S_h$, c'est-à-dire $n_h \propto N_h S_h$.

La variance minimale

$$V_{Ney} = \frac{\overline{S}^2}{n} - \frac{1}{N} \sum_h W_h\, S_h^2 \qquad \text{où} \qquad \overline{S} = \sum_h W_h\, S_h$$

$\overline{S}$ est la **moyenne pondérée des écarts-types** des strates.

Deux cents ménages répartis selon Neyman

Province	$N_h\,S_h$	part	n_h	f_h
Tanger-Assilah	3 502 379	45,3 %	90,6	0,110
Tétouan	2 168 943	28,0 %	56,1	0,100
Larache	1 365 406	17,7 %	35,3	0,093
Al Hoceïma	695 975	9,0 %	18,0	0,075

La variance

$$\overline{S} = 0{,}41\times 4\,271{,}19 + 0{,}28\times 3\,873{,}11 + 0{,}19\times 3\,593{,}17 + 0{,}12\times 2\,899{,}90 = 3\,866{,}35.$$

$$V_{Ney} = \frac{3\,866{,}35^2}{200} - 7\,571{,}1 = 74\,743{,}4 - 7\,571{,}1 = 67\,172{,}3$$

Erreur-type $259{,}2$ dh ; $deff = 67\,172{,}3/71\,435{,}5 = 0{,}940$. Arrondie à $91, 56, 35, 18$: $V = 67\,174{,}1$, la même.

Ce que Neyman gagne sur la proportionnelle

Comme $V_{prop} = \sum_h W_h S_h^2/n - \sum_h W_h S_h^2/N$:

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

Le gain est la **variance des écarts-types** des strates, divisée par n.

Les provinces

$\sum_h W_h\,(S_h - \overline{S})^2 = 193\,474$, d'où un gain de $193\,474/200 = 967{,}4 = 68\,139{,}7 - 67\,172{,}3$. La variance ne baisse que de $1{,}4\,\%$: les S_h vont de $2\,900$ à $4\,271$ dh, ils se ressemblent trop pour que les répartir autrement change beaucoup.

Le chiffre d'affaires de 1 200 entreprises, chiffres inventés

Taille	N_h	S_h (millions de dh)	proportionnelle	Neyman
Petites	1 000	1	83,3	20
Moyennes	150	10	12,5	30
Grandes	50	60	4,2	50

Pour $n = 100$, la règle donne $54{,}5$ grandes entreprises, plus qu'il n'y en a : on les prend **toutes**, et l'on répartit les 50 autres selon Neyman. Variance : $1{,}497$, puis $0{,}076$, divisée par 20.

La strate exhaustive

Quand $n_h > N_h$, on recense la strate ($f_h = 1$: aucune erreur d'échantillonnage) et l'on répartit le reste. Dans les enquêtes auprès des entreprises, les plus grandes sont **toujours** interrogées.

Coûts et dispersions inconnues

L'allocation optimale avec coûts

Le coût de l'enquête est $C = c_0 + \sum_h c_h\, n_h$: frais fixes, plus c_h par unité de la strate h. À budget donné, la variance est minimale pour

$$n_h \propto \frac{N_h\, S_h}{\sqrt{c_h}}$$

La démonstration par le lagrangien

On minimise $\sum_h W_h^2 S_h^2 / n_h + \lambda \left(\sum_h c_h n_h - C \right)$. La dérivée en n_h s'annule :

$$-\frac{W_h^2\, S_h^2}{n_h^2} + \lambda\, c_h = 0 \quad \Longrightarrow \quad n_h = \frac{W_h\, S_h}{\sqrt{\lambda}\,\sqrt{c_h}}$$

La constante $\sqrt{\lambda}$ s'obtient en épuisant le budget. Si les c_h sont égaux, on retrouve Neyman.

Un ménage rural coûte deux fois plus, budget de 200 unités

Milieu	N_h	S_h	c_h	$N_h\,S_h/\sqrt{c_h}$	n_h
Rural	562	3 215,03	2	1 277 634	30,1
Urbain	1 438	4 134,11	1	5 944 846	139,9

$$n_h = 200 \times \frac{N_h\,S_h/\sqrt{c_h}}{\sum_k N_k\,S_k\,\sqrt{c_k}} \qquad \text{avec} \qquad \sum_k N_k\,S_k\,\sqrt{c_k} = 8\,500\,114$$

Coût : $2 \times 30{,}06 + 139{,}88 = 200$.

Moins de ruraux, mais plus de ménages

Taux de sondage : $5{,}3\,\%$ à la campagne, $9{,}7\,\%$ en ville. Le budget achète $169{,}9$ ménages ; la variance vaut $82\,718{,}5$, l'erreur-type $287{,}6$ dh.

Comparer à l'allocation proportionnelle de même coût

Un ménage coûte en moyenne $0{,}281 \times 2 + 0{,}719 \times 1 = 1{,}281$: le budget en paie $200/1{,}281 = 156{,}1$.

Allocation de même coût	ruraux	urbains	V	erreur-type
Optimale	$30{,}1$	$139{,}9$	$82\,718{,}5$	$287{,}6$
Proportionnelle	$43{,}9$	$112{,}3$	$89\,713{,}8$	$299{,}5$

À coût égal, la variance baisse de $7{,}8\,\%$. Un sondage aléatoire simple de 156 ménages ferait pire : $93\,739{,}2$.

Le prix de l'économie

La moyenne **rurale** perd en précision : erreur-type $570{,}5$ dh contre $466{,}1$. Si elle doit être publiée, il faut un plancher. Et le surcoût rural vient surtout des **déplacements** : le chapitre 9 le réduira en regroupant les interviews.

Où trouver les écarts-types des strates

1. Une enquête antérieure, ou une enquête pilote (chapitre 5).
2. Une variable liée, connue par le recensement.
3. Un **coefficient de variation** supposé commun : $S_h \approx CV \times \overline{Y}_h$, d'où $n_h \propto N_h \overline{Y}_h$, le **total** de la strate.

Une erreur sur les écarts-types se paie

Dans les provinces, $S_h/\overline{Y}_h$ va de $0{,}475$ à $0{,}495$: la répartition au prorata des totaux ($89{,}4$; $57{,}7$; $34{,}8$; $18{,}1$) a un $deff$ de $0{,}941$, presque celui de Neyman. Mais avec S_h surestimé de $30\,\%$ à Tanger-Assilah et sous-estimé de $30\,\%$ à Al Hoceïma, $deff = 0{,}978$: **moins bien que la proportionnelle** ($0{,}954$).

L’allocation égale et les domaines

Une allocation qui fait perdre

$V = 89\,027{,}4$, soit $deff = 1{,}246$: ce plan stratifié est **moins précis** qu'un sondage aléatoire simple de même taille. Les poids $d_k = N_h/50$ vont de $16{,}4$ à Tanger-Assilah à $4{,}8$ à Al Hoceïma (chapitre 1).

Des poids inégaux sans raison

Si toutes les strates avaient la même dispersion, sans correction de population finie :

$$deff = n \sum_h \frac{W_h^2}{n_h} = 4 \sum_h W_h^2 = 4 \times 0{,}297 = 1{,}188 = 1 + CV^2(d_k)$$

C'est la formule de **Kish** : des poids inégaux coûtent $1 + CV^2$ des poids. Ils ne se justifient que par des S_h ou des coûts différents, ou par un autre objectif.

Erreur-type de la moyenne provinciale, en dh

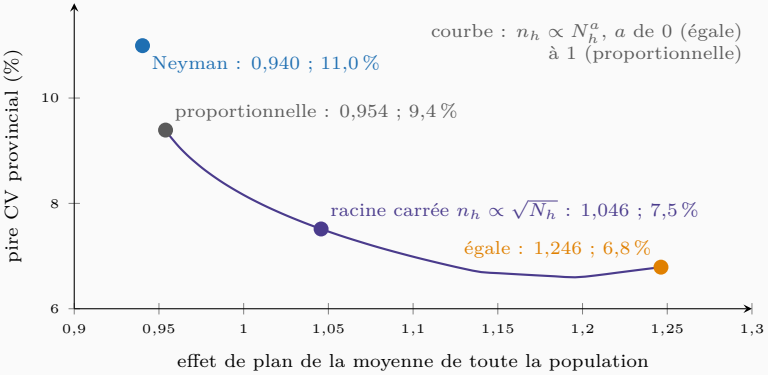
$\sqrt{(1-f_h)\,S_h^2/n_h}$ pour chaque province ; dernière ligne : la moyenne d'ensemble.

Province	proportionnelle	Neyman	égale
Tanger-Assilah	447,5	423,3	585,3
Tétouan	491,0	490,5	522,7
Larache	553,0	575,9	473,5
Al Hoceïma	561,6	657,4	364,9
Ensemble	261,0	259,2	298,4

Chaque domaine est une petite enquête

Pour publier la dépense d'Al Hoceïma, Neyman est le pire choix : 18 ménages, une marge de $\pm 1\,288$ dh. L'allocation égale la ramène à ± 715 dh. Ce qui est optimal pour la région ne l'est pas pour chaque province.

Le compromis



Lire la courbe

En abscisse, l'effet de plan de la moyenne régionale ; en ordonnée, le pire coefficient de variation provincial. La règle $n_h \propto \sqrt{N_h}$ $(65{,}6\,;\,54{,}2\,;\,44{,}7\,;\,35{,}5)$ abaisse ce pire CV de $9{,}4\,\%$ à $7{,}5\,\%$, pour un *deff* de $1{,}046$ au lieu de $0{,}954$.

Un plancher de 30 ménages à Al Hoceïma

On donne 30 ménages à Al Hoceïma et l'on répartit les 170 autres au prorata : $79{,}2$; $54{,}1$; $36{,}7$. Poids : $10{,}35$ dans les trois grandes provinces, 8 à Al Hoceïma. Résultat : $deff = 0{,}975$, et la marge d'Al Hoceïma passe de $\pm 1\,101$ à ± 971 dh.

Suréchantillonner, puis repondérer

On **suréchantillonne** les domaines trop petits pour être bien estimés, puis l'on rend à chaque ménage son poids $d_k = N_h / n_h$ dans les estimations d'ensemble. L'effet de plan, un peu plus élevé, est le prix de la précision locale : un choix de diffusion, pas une erreur.

L’effet de plan

Effet de plan, taille effective

Pour un estimateur $\hat{\theta}$ et un plan de taille n :

$$deff = \frac{V_{plan}(\hat{\theta})}{V_{SAS}(\hat{\theta})} \qquad n_{eff} = \frac{n}{deff} \qquad deft = \sqrt{deff}$$

où V_{SAS} est la variance d'un sondage aléatoire simple de **même taille** n.

Trois lectures d'un même nombre

Allocation égale, $deff = 1{,}246$:

- la **variance** est supérieure de $24{,}6\,\%$ à celle du sondage aléatoire simple ;
- les 200 ménages **valent** $n_{eff} = 160{,}5$ ménages tirés au hasard simple ;
- la **marge** est plus large de $deft - 1 = 11{,}6\,\%$: $\pm 584{,}8$ dh au lieu de $\pm 523{,}9$.

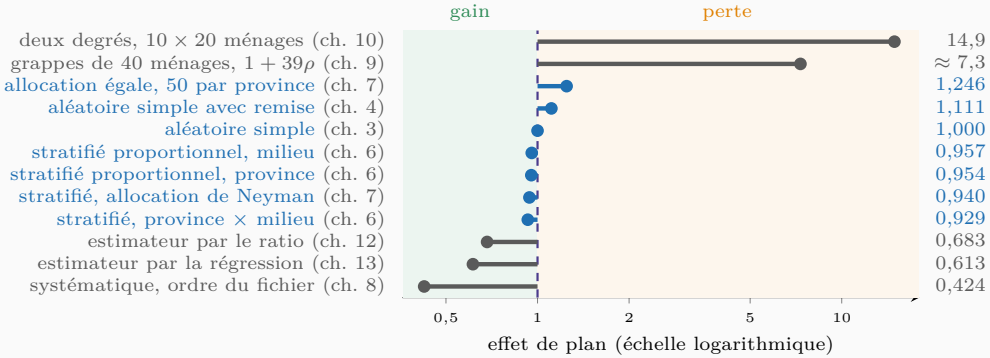
Les allocations de 200 ménages, dépense

Plan	$deff$	n_{eff}	marge à $95\,\%$
Aléatoire simple	1	200	$\pm 523{,}9$
Stratifié, proportionnelle	$0{,}954$	$209{,}7$	$\pm 511{,}6$
Stratifié, Neyman	$0{,}940$	$212{,}7$	$\pm 508{,}0$
Stratifié, égale	$1{,}246$	$160{,}5$	$\pm 584{,}8$

L'effet de plan multiplie le budget

Pour une marge de ± 300 dh, le chapitre 5 demandait 507 ménages en aléatoire simple. Le calcul exact donne 483 avec Neyman et 624 avec l'allocation égale. L'allocation égale exige $23\,\%$ d'interviews de plus que le hasard simple : l'effet de plan se lit comme un **coût**.

L'échelle des effets de plan



Ce que montre l'échelle

En bleu, les plans déjà vus : entre $0{,}93$ et $1{,}25$, **stratifier rapporte peu** sur cette population.

En gris, la suite : les grappes multiplient la variance par 7 à 15 ; ratio, régression ou ordre du fichier la divisent par $1{,}5$ à $2{,}4$.

Proportions et plusieurs variables

Neyman pour une proportion

$$S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h) \qquad \text{d'où, à peu près,} \qquad n_h \propto N_h \sqrt{P_h (1 - P_h)}$$

Or $\sqrt{P(1-P)}$ varie peu : $0{,}50$ pour $P = 0{,}5$; $0{,}46$ pour $0{,}3$; $0{,}40$ pour $0{,}2$. Il ne tombe vite que pour les proportions **rares** : $0{,}14$ pour $0{,}02$.

Les ménages connectés, stratifiés par milieu

$P_h = 0{,}434$ (rural) et $0{,}768$ (urbain), d'où $S_h = 0{,}496$ et $0{,}422$. Neyman : $62{,}9$ ruraux et $137{,}1$ urbains, contre $56{,}2$ et $143{,}8$ en proportionnelle. Effet de plan : $0{,}893$ contre $0{,}898$.

Pour une proportion, **Neyman se confond presque avec la proportionnelle**.

Effet de plan pour la dépense et pour internet, 200 ménages

Allocation	n_h	dépense	internet
Proportionnelle	$82, 56, 38, 24$	$0{,}954$	$0{,}966$
Neyman dépense	$91, 56, 35, 18$	$0{,}940$	$0{,}997$
Neyman internet	$78, 55, 41, 26$	$0{,}968$	$0{,}963$
Égale	50 partout	$1{,}246$	$1{,}122$

La proportionnelle, un compromis robuste

L'allocation optimale pour une variable est médiocre pour une autre : Neyman pour la dépense ne gagne **rien** sur internet. La proportionnelle n'est jamais loin de la meilleure. En pratique : on optimise pour la variable principale, on contrôle les autres, et l'on ajoute des planchers pour les domaines.

S’entraîner

Trois catégories d'exploitations agricoles

Une province compte $1\,000$ exploitations. On veut estimer leur superficie cultivée moyenne, en hectares, avec $n = 140$ exploitations.

Strate	Petites	Moyennes	Grandes
N_h	600	300	100
S_h (hectares)	2	5	15

1. Calculer l'allocation proportionnelle et l'allocation de Neyman.
2. Calculer les deux variances de $\overline{y}_{st}$.
3. Retrouver leur différence par $\frac{1}{n}\sum_h W_h\,(S_h - \overline{S})^2$.
4. Quels sont le taux de sondage et le poids d'une grande exploitation ?

Exercice 1 : solution

Les deux allocations

1. Proportionnelle : $84, 42, 14$. $N_h S_h = 1\,200, 1\,500, 1\,500$, de somme $4\,200$: Neyman donne $40, 50, 50$.
2. $\sum_h W_h S_h^2 = 2{,}4 + 7{,}5 + 22{,}5 = 32{,}4$; $V_{prop} = (1 - 0{,}14) \times 32{,}4/140 = 0{,}1990$. $\overline{S} = 1{,}2 + 1{,}5 + 1{,}5 = 4{,}2$; $V_{Ney} = 4{,}2^2/140 - 32{,}4/1\,000 = 0{,}126 - 0{,}0324 = 0{,}0936$: la variance est **divisée par** $2{,}1$.

Le gain et les poids

3. $0{,}6 \times 2{,}2^2 + 0{,}3 \times 0{,}8^2 + 0{,}1 \times 10{,}8^2 = 2{,}904 + 0{,}192 + 11{,}664 = 14{,}76$; $14{,}76/140 = 0{,}1054 = 0{,}1990 - 0{,}0936$. Les S_h sont très inégaux : le gain est fort.
4. $f_h = 50/100 = 0{,}5$ et $d_k = 2$ pour une grande exploitation, contre $f_h = 0{,}067$ et $d_k = 15$ pour une petite.

Ville et campagne

Une enquête sur la dépense porte sur $6\,000$ ménages urbains ($S_h = 4\,000$ dh) et $4\,000$ ruraux ($S_h = 3\,000$ dh). Un questionnaire coûte 100 dh en ville et 400 dh à la campagne ; le budget des questionnaires est de $60\,000$ dh.

1. Calculer l'allocation optimale.
2. Calculer l'allocation proportionnelle de même coût.
3. Comparer les variances de $\overline{y}_{st}$.
4. Que perd-on sur la moyenne rurale ?

Exercice 2 : solution

Les deux allocations

1. $N_h S_h / \sqrt{c_h}$: $6\,000 \times 4\,000/10 = 2\,400\,000$ et $4\,000 \times 3\,000/20 = 600\,000$, soit $n_u = 4\,n_r$. Le budget : $100 \times 4\,n_r + 400\,n_r = 800\,n_r = 60\,000$, d'où $n_r = 75$ et $n_u = 300$.
2. Coût moyen : $0{,}6 \times 100 + 0{,}4 \times 400 = 220$ dh, donc $n = 272{,}7$: $163{,}6$ urbains et $109{,}1$ ruraux.

Les variances

3. Optimale : $0{,}36 \times 0{,}95 \times 16\,000\,000/300 = 18\,240$ en ville, $0{,}16 \times 0{,}981 \times 9\,000\,000/75 = 18\,840$ à la campagne, d'où $V_{opt} = 37\,080$. Proportionnelle : $V_{prop} = (1 - 0{,}0273) \times 13\,200\,000/272{,}7 = 47\,080$. Soit $21\,\%$ de variance en moins, au même prix.
4. L'erreur-type de la moyenne rurale passe de $283{,}3$ à $343{,}1$ dh. Si elle doit être publiée, on impose un plancher.

Cent ménages par milieu, le retour

On reprend le plan des chapitres 1 et 6 : 100 ménages urbains et 100 ruraux. Sa variance vaut $88\,918{,}3$, contre $71\,435{,}5$ pour un sondage aléatoire simple de 200 ménages.

1. Calculer $deff$, n_{eff} et $deft$. Interpréter.
2. Calculer les poids, puis l'effet de plan de Kish $n \sum_h n_h d_h^2 / \left(\sum_h n_h d_h\right)^2$. Pourquoi est-il plus faible que le vrai ?
3. Répartir 200 ménages selon Neyman, avec $S_h = 4\,134{,}11$ (urbain) et $3\,215{,}03$ (rural). Son effet de plan vaut $0{,}945$: conclure.

Le plan à 100 et 100

1. $deff = 88\,918{,}3 / 71\,435{,}5 = 1{,}245$; $n_{eff} = 200 / 1{,}2447 = 160{,}7$; $deft = 1{,}116$. Les 200 ménages en valent 161 tirés au hasard simple ; la marge est $11{,}6\,\%$ plus large.
2. Poids $14{,}38$ et $5{,}62$; Kish : $200 \times 100 \times (14{,}38^2 + 5{,}62^2) / 2\,000^2 = 1{,}192$. Le vrai est plus fort, car la strate **sous-représentée**, l'urbaine, est aussi la **plus dispersée**.

Neyman par milieu

3. $N_h S_h = 5\,944\,846$ (urbain) et $1\,806\,848$ (rural) : $76{,}7\,\%$ et $23{,}3\,\%$, soit $153{,}4$ urbains et $46{,}6$ ruraux. C'est l'inverse du plan à 100 et 100 : moins de ruraux que la proportionnelle $(56{,}2)$, et un gain au lieu d'une perte.

Ce qu’il faut retenir

Le résumé du chapitre

Sept points

1. À n fixé, seul $\sum_h W_h^2 S_h^2 / n_h$ dépend de la répartition de l'échantillon.
2. **Neyman** : $n_h \propto N_h S_h$ minimise la variance ; f_h est proportionnel à S_h.
3. Son gain sur la proportionnelle vaut $\frac{1}{n}\sum_h W_h (S_h - \overline{S})^2$: faible si les S_h se ressemblent, énorme sinon (strate exhaustive).
4. Avec des coûts c_h : $n_h \propto N_h S_h / \sqrt{c_h}$.
5. L'allocation égale sert les **domaines**, au prix d'un $deff$ de $1{,}246$ ici.
6. $deff = V_{plan}/V_{SAS}$, $n_{eff} = n/deff$; des poids inégaux coûtent $1 + CV^2$ (Kish).
7. Pour les proportions et quand il y a plusieurs variables, la proportionnelle est un compromis robuste.

La suite

Le chapitre 8 tire un ménage tous les k dans une liste ordonnée : le **sondage systématique**, une stratification implicite qui ne dit pas son nom.