Échantillonnage

Chapitre 2 : Estimateur, biais et précision

Jaouad Madkour

30 octobre 2026

Plan du chapitre

L'estimateur de Horvitz-Thompson **2**
Pondérer par l'inverse de la probabilité . . . 2
Sans biais, quel que soit le plan . . . 2
La moyenne, et l'aléatoire simple . . . 3
Le plan un ménage par groupe . . . 4
Sur la population du cours . . . 4

La variance de l'estimateur **4**
La formule générale . . . 4
La forme de Sen-Yates-Grundy . . . 5
Vérification sur l'aléatoire simple . . . 6
Vérification sur le plan par groupes . . . 7
Deux plans, deux précisions . . . 7

Estimer la variance **8**
L'estimateur de la variance . . . 8
Vérification sur la mini-population . . . 8
Quand la variance ne s'estime pas . . . 9

Biais, variance et erreur quadratique **10**
Trois mesures de la qualité . . . 10
Les cibles . . . 11
Brute ou pondérée sous le plan par groupes . . . 11
Sur la population du cours . . . 12
Le biais ne diminue pas avec n . . . 12

L'intervalle de confiance **13**
L'approximation normale . . . 13
Cent échantillons, cent intervalles . . . 14
La lecture correcte . . . 14
La couverture simulée . . . 14
Le coefficient de variation . . . 15

S'entraîner **15**

Exercice 1 : un plan à probabilités inégales 15
Exercice 1 : solution 15
Exercice 2 : variance et estimation de la variance 16
Exercice 2 : solution 16
Exercice 3 : un sondage de 400 ménages 16
Exercice 3 : solution 16

Ce qu'il faut retenir **17**
Le résumé du chapitre 17

L'estimateur de Horvitz-Thompson

Pondérer par l'inverse de la probabilité

L'estimateur de Horvitz-Thompson du total

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k \, y_k = \sum_{k \in U} I_k \, \frac{y_k}{\pi_k}$$

Chaque unité tirée compte pour son **poids** $d_k = 1/\pi_k$ (chapitre 1). La dernière écriture fait apparaître les indicatrices I_k : seules les unités tirées contribuent.

Un ménage tiré en représente plusieurs

Un ménage tiré avec la probabilité 0,1 représente 10 ménages : sa dépense est comptée 10 fois. Un ménage tiré avec la probabilité 1/2 n'en représente que 2. Les unités **rarement** tirées comptent **davantage** quand elles le sont : c'est ce qui compense leur rareté.

Sans biais, quel que soit le plan

La démonstration en une ligne

Les y_k/π_k sont des nombres fixes, et $E(I_k) = \pi_k$:

$$E(\widehat{Y}_{HT}) = \sum_{k \in U} E(I_k) \, \frac{y_k}{\pi_k} = \sum_{k \in U} \pi_k \, \frac{y_k}{\pi_k} = \sum_{k \in U} y_k = Y$$

Seule l'espérance des I_k intervient : le résultat vaut pour **tout** plan, à une condition, $\pi_k > 0$ pour **chaque** unité de U.

Une unité oubliée ne se rattrape pas

Si $\pi_k = 0$, l'unité k disparaît de la somme : $E(\widehat{Y}_{HT}) = Y - \sum_{\pi_k = 0} y_k$. Une base de sondage qui omet les 134 ménages des quatre districts ruraux d'Al Hoceïma fait estimer, en moyenne, $Y - 739\,640$ dh : 4,7 % de moins que le vrai total, **quelle que soit** la taille de l'échantillon.

Démonstration : L'estimateur de Horvitz-Thompson est sans biais

L'écriture avec les indicatrices. Dans $\sum_{k \in S} y_k/\pi_k$, une unité figure si elle est tirée, et pas

sinon : c'est exactement ce que dit l'indicatrice I_k, qui vaut 1 si $k \in S$ et 0 sinon. Donc

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} I_k \frac{y_k}{\pi_k}.$$

L'espérance. Les y_k/π_k sont des **nombres fixes** : seul I_k est aléatoire. L'espérance d'une somme finie est la somme des espérances, et $E(I_k) = \pi_k$ (chapitre 1) :

$$E(\widehat{Y}_{HT}) = \sum_{k \in U} E(I_k) \frac{y_k}{\pi_k} = \sum_{k \in U} \pi_k \frac{y_k}{\pi_k} = \sum_{k \in U} y_k = Y.$$

La simplification de π_k par π_k n'est possible que si $\pi_k > 0$: c'est la condition du théorème. Si l'unité k a $\pi_k = 0$, elle ne figure jamais dans l'échantillon, son terme n'apparaît pas dans $\widehat{Y}_{HT}$, et

$$E(\widehat{Y}_{HT}) = Y - \sum_{k\,:\,\pi_k = 0} y_k.$$

Remarquons que rien n'a été supposé sur le plan : ni la taille de l'échantillon, ni l'indépendance des tirages, ni l'égalité des probabilités.

Le plan aléatoire simple. Ici $\pi_k = n/N$ pour toute unité, donc $\widehat{Y}_{HT} = \frac{N}{n} \sum_{k \in S} y_k = N\,\overline{y}$, et $\widehat{\overline{Y}}_{HT} = \widehat{Y}_{HT}/N = \overline{y}$. La moyenne d'échantillon est sans biais pour la moyenne de la population.

Ce que cela dit. Le poids $d_k = 1/\pi_k$ compense exactement la rareté de l'unité k : elle est tirée une fois sur $1/\pi_k$ en moyenne, et compte alors pour $1/\pi_k$. Et la condition $\pi_k > 0$ n'est pas un détail : c'est le défaut de la base de sondage. Avec les 134 ménages des quatre districts ruraux d'Al Hoceïma absents de la liste, l'estimateur manque en moyenne 739 640 dh, soit 4,7 % du total, quelle que soit la taille de l'échantillon.

■

La moyenne, et l'aléatoire simple

Estimer la moyenne

Quand N est connu, $\widehat{\overline{Y}}_{HT} = \widehat{Y}_{HT}/N$ est sans biais pour $\overline{Y}$. Pour un sondage aléatoire simple, $\pi_k = n/N$ et

$$\widehat{Y}_{HT} = \frac{N}{n} \sum_{k \in S} y_k = N\,\overline{y}, \qquad \widehat{\overline{Y}}_{HT} = \overline{y}$$

La moyenne d'échantillon est l'estimateur de Horvitz-Thompson de ce plan : elle est **sans biais**.

Mini-population : deux ménages sur cinq

$\pi_k = 0{,}4$ et $d_k = 2{,}5$. Échantillon $\{A, E\}$: $\widehat{Y}_{HT} = 2{,}5 \times (4 + 14) = 45$ milliers de dh, soit une moyenne de 9. Les dix échantillons donnent 25 ; 27,5 ; 32,5 ; 45 ; 32,5 ; 37,5 ; 50 ; 40 ; 52,5 ; 57,5, de moyenne exactement $Y = 40$.

Le plan un ménage par groupe

Un ménage parmi A, B, C, un ménage parmi D, E

$\pi_A = \pi_B = \pi_C = 1/3$ et $\pi_D = \pi_E = 1/2$, d'où $\widehat{Y}_{HT} = 3\,y_1 + 2\,y_2$, où y_1 vient du premier groupe.

Échantillon	AD	AE	BD	BE	CD	CE
$\widehat{Y}_{HT}$	30	40	36	46	39	49

Chaque échantillon a la probabilité 1/6 : $E(\widehat{Y}_{HT}) = 240/6 = 40 = Y$.

La moyenne pondérée du chapitre 1

Divisé par $N = 5$: $\widehat{\overline{Y}}_{HT} = \frac{3}{5}\,y_1 + \frac{2}{5}\,y_2$. C'est la moyenne pondérée que le chapitre 1 avait construite « à la main ». L'estimateur de Horvitz-Thompson la retrouve **automatiquement**, à partir des seuls π_k.

Sur la population du cours

Un tirage de 50 ménages par province

Province	$d_k = N_h/50$	$W_h = N_h/N$	Moyenne $\overline{y}_h$
Tanger-Assilah	16,4	0,41	9 030,4
Tétouan	11,2	0,28	7 121,6
Larache	7,6	0,19	7 551,0
Al Hoceïma	4,8	0,12	6 137,4

$\widehat{Y}_{HT} = \sum_{k\in S} d_k\,y_k = \sum_h N_h\,\overline{y}_h = 15\,735\,380$ dh, soit $\widehat{\overline{Y}}_{HT} = \sum_h W_h\,\overline{y}_h = 7\,867{,}69$ dh.

Les poids font le travail

Moyenne **brute** des 200 ménages : 7 460,1 dh, contre $\overline{Y} = 7\,911{,}44$. L'écart de l'estimation pondérée, −44 dh, vient du hasard ; celui de la brute est surtout **systématique** : on le mesurera plus loin.

La variance de l'estimateur

La formule générale

Variance de l'estimateur de Horvitz-Thompson

$$V(\widehat{Y}_{HT}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\,\frac{y_k}{\pi_k}\,\frac{y_l}{\pi_l}, \qquad \Delta_{kl} = \pi_{kl} - \pi_k\,\pi_l, \quad \Delta_{kk} = \pi_k\,(1-\pi_k)$$

La démonstration

Posons $a_k = y_k/\pi_k$, un nombre fixe. $\widehat{Y}_{HT} = \sum_k I_k\, a_k$ est une somme de variables aléatoires, dont la variance est la somme de toutes les covariances :

$$V\Big(\sum_k I_k\, a_k\Big) = \sum_k \sum_l a_k\, a_l\, \mathrm{Cov}(I_k, I_l) \quad \text{et} \quad \mathrm{Cov}(I_k, I_l) = \Delta_{kl}$$

d'après le chapitre 1 ; pour $k = l$, $\mathrm{Cov}(I_k, I_k) = V(I_k) = \pi_k(1-\pi_k)$.

Démonstration : La variance de l'estimateur de Horvitz-Thompson

Posons $a_k = y_k/\pi_k$ (un nombre fixe), de sorte que $\widehat{Y}_{HT} = \sum_k a_k\, I_k$ et, d'après la démonstration précédente, $E(\widehat{Y}_{HT}) = \sum_k a_k\, \pi_k$. L'écart à l'espérance s'écrit

$$\widehat{Y}_{HT} - E(\widehat{Y}_{HT}) = \sum_{k\in U} a_k\, (I_k - \pi_k).$$

Élevons au carré : le carré d'une somme est la somme de **tous** les produits deux à deux, k et l parcourant chacun U (le produit (k,l) et le produit (l,k) sont comptés tous les deux) :

$$\Big(\sum_k a_k\, (I_k - \pi_k)\Big)^2 = \sum_{k\in U} \sum_{l\in U} a_k\, a_l\, (I_k - \pi_k)(I_l - \pi_l).$$

Prenons l'espérance. Les a_k sont fixes, et $E[(I_k - \pi_k)(I_l - \pi_l)] = \mathrm{Cov}(I_k, I_l)$:

$$V(\widehat{Y}_{HT}) = \sum_{k\in U} \sum_{l\in U} a_k\, a_l\, \mathrm{Cov}(I_k, I_l).$$

Il reste à nommer la covariance. Pour $k \neq l$, c'est $\pi_{kl} - \pi_k\pi_l = \Delta_{kl}$; pour $k = l$, c'est $V(I_k) = \pi_k(1-\pi_k)$, c'est-à-dire Δ_{kk} avec la convention $\pi_{kk} = \pi_k$. D'où la formule annoncée.

Ce que cela dit. La variance ne dépend du plan que par les nombres Δ_{kl} : les probabilités d'inclusion simples et doubles résument tout ce que le plan a d'utile pour la précision. Les y_k n'interviennent que par les rapports y_k/π_k.

■

La forme de Sen-Yates-Grundy

Pour un plan de taille fixe

$$V(\widehat{Y}_{HT}) = \frac{1}{2} \sum_{k\in U} \sum_{l\neq k} (\pi_k\, \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

Elle est **nulle** si y_k/π_k est le même pour toutes les unités, c'est-à-dire si π_k est proportionnel à y_k : tirer plus souvent les grandes unités peut rendre l'estimateur très précis (chapitre 11).

D'où elle vient

Si $n = \sum_l I_l$ est fixe, $\mathrm{Cov}(I_k, n) = 0$, donc $\sum_l \Delta_{kl} = 0$ et $\Delta_{kk} = -\sum_{l\neq k} \Delta_{kl}$. En remplaçant les termes diagonaux :

$$V = -\sum_k \sum_{l\neq k} \Delta_{kl}\, a_k^2 + \sum_k \sum_{l\neq k} \Delta_{kl}\, a_k\, a_l = -\frac{1}{2} \sum_k \sum_{l\neq k} \Delta_{kl}\, (a_k - a_l)^2$$

Démonstration : La forme de Sen-Yates-Grundy

Première étape : les covariances d'une unité s'annulent. Si la taille $n = \sum_l I_l$ est fixe, c'est une constante, donc $\mathrm{Cov}(I_k, n) = 0$. Par bilinéarité de la covariance, $\mathrm{Cov}(I_k, n) = \sum_{l\in U} \mathrm{Cov}(I_k, I_l) = \sum_{l\in U} \Delta_{kl}$. Ainsi

$$\sum_{l\in U} \Delta_{kl} = 0 \qquad \text{donc} \qquad \Delta_{kk} = -\sum_{l\neq k} \Delta_{kl}.$$

Deuxième étape : on remplace les termes diagonaux. Dans la formule générale, séparons $l = k$ et $l \neq k$, puis remplaçons Δ_{kk} :

$$V = \sum_k \Delta_{kk}\, a_k^2 + \sum_k \sum_{l\neq k} \Delta_{kl}\, a_k\, a_l = -\sum_k \sum_{l\neq k} \Delta_{kl}\, a_k^2 + \sum_k \sum_{l\neq k} \Delta_{kl}\, a_k\, a_l.$$

Troisième étape : on symétrise. Comme $\Delta_{kl} = \Delta_{lk}$, échanger les noms k et l dans la première somme donne $\sum_k \sum_{l\neq k} \Delta_{kl}\, a_k^2 = \sum_k \sum_{l\neq k} \Delta_{kl}\, a_l^2$. Elle est donc aussi égale à la moyenne des deux écritures, $\frac{1}{2} \sum_k \sum_{l\neq k} \Delta_{kl}\, (a_k^2 + a_l^2)$, et

$$V = -\frac{1}{2} \sum_k \sum_{l\neq k} \Delta_{kl}\, (a_k^2 + a_l^2 - 2\, a_k\, a_l) = -\frac{1}{2} \sum_k \sum_{l\neq k} \Delta_{kl}\, (a_k - a_l)^2.$$

Comme $-\Delta_{kl} = \pi_k \pi_l - \pi_{kl}$, c'est la formule de Sen-Yates-Grundy.

Ce que cela dit. La variance est une somme de carrés d'écarts **entre unités** : elle s'annule si tous les y_k/π_k sont égaux, c'est-à-dire si π_k est proportionnel à y_k. L'hypothèse de taille fixe a servi une seule fois, à la première étape ; sans elle, il faut la formule générale.

■

Vérification sur l'aléatoire simple

Deux ménages sur cinq

$\pi_k = 0{,}4$, $\pi_{kl} = 0{,}1$, donc $\pi_k\, \pi_l - \pi_{kl} = 0{,}16 - 0{,}1 = 0{,}06$. Les y_k/π_k valent 10 ; 15 ; 17,5 ; 22,5 ; 35. Chaque paire apparaît deux fois dans la double somme, d'où

$$V(\widehat{Y}_{HT}) = 0{,}06 \sum_{k<l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 = 0{,}06 \times 1\,812{,}5 = 108{,}75$$

Trois calculs, un seul résultat

- Variance des dix totaux de l'énumération : 108,75.
- Formule du chapitre 3 : $N^2\,(1-f)\,S^2/n = 25 \times 4{,}35 = 108{,}75$.
- Pour la moyenne, on divise par $N^2 = 25$: $V(\overline{y}) = 4{,}35$.

Vérification sur le plan par groupes

Deux groupes, des probabilités doubles nulles

- Même groupe : $\pi_{AB} = 0$, donc $\Delta_{AB} = -\frac{1}{9}$; $\pi_{DE} = 0$, donc $\Delta_{DE} = -\frac{1}{4}$.
- Groupes différents : $\pi_{AD} = \frac{1}{6} = \frac{1}{3} \times \frac{1}{2}$, donc $\Delta_{AD} = 0$.

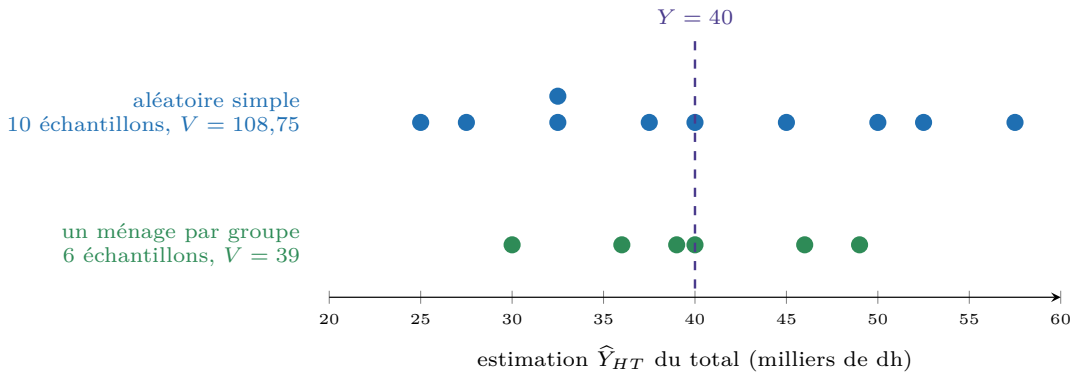
Les y_k/π_k valent 12 ; 18 ; 21 dans le premier groupe, 18 ; 28 dans le second :

$$V(\widehat{Y}_{HT}) = \tfrac{1}{9}\left[6^2 + 9^2 + 3^2\right] + \tfrac{1}{4} \times 10^2 = 14 + 25 = 39$$

L'énumération confirme

Les six totaux 30 ; 40 ; 36 ; 46 ; 39 ; 49 s'écartent de 40 de -10 ; 0 ; -4 ; 6 ; -1 ; 9. Somme des carrés : 234, et $234/6 = 39$.

Deux plans, deux précisions



Même taille, variance presque trois fois plus faible

Avec deux ménages dans les deux cas, la variance tombe de 108,75 à 39. Le plan par groupes **interdit** les échantillons extrêmes, $\{A, B\}$ ou $\{D, E\}$: il associe toujours un petit budget à un grand. C'est le principe de la **stratification** (chapitre 6).

Estimer la variance

L'estimateur de la variance

Deux estimateurs sans biais

$$\widehat{V}(\widehat{Y}_{HT}) = \sum_{k\in S}\sum_{l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \qquad (\pi_{kk} = \pi_k)$$

et, pour un plan de taille fixe, la forme de Sen-Yates-Grundy :

$$\widehat{V}_{SYG} = \frac{1}{2} \sum_{k\in S} \sum_{l\in S,\, l\neq k} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Le même argument que pour le total

Chaque terme de V concerne une **paire** (k, l) ; la paire est observée avec la probabilité $E(I_k I_l) = \pi_{kl}$. On divise donc chaque terme par π_{kl}, comme on divisait y_k par π_k. Condition : $\pi_{kl} > 0$ pour **toutes** les paires.

Démonstration : L'estimateur de la variance est sans biais

Posons encore $a_k = y_k/\pi_k$. Comme pour le total, on écrit la somme sur l'échantillon comme une somme sur la population, en faisant apparaître l'indicatrice de chaque paire. L'indicatrice de « k et l sont tous deux tirés » est $I_k\, I_l$ (et $I_k\, I_k = I_k$ pour $l = k$), d'où

$$\widehat{V}(\widehat{Y}_{HT}) = \sum_{k\in S}\sum_{l\in S} \frac{\Delta_{kl}}{\pi_{kl}}\, a_k\, a_l = \sum_{k\in U}\sum_{l\in U} I_k\, I_l\, \frac{\Delta_{kl}}{\pi_{kl}}\, a_k\, a_l.$$

Les nombres $\Delta_{kl}\, a_k\, a_l/\pi_{kl}$ sont fixes, et $E(I_k\, I_l) = \pi_{kl}$ (avec $\pi_{kk} = \pi_k$). Donc

$$E(\widehat{V}) = \sum_{k\in U}\sum_{l\in U} \pi_{kl}\, \frac{\Delta_{kl}}{\pi_{kl}}\, a_k\, a_l = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\, a_k\, a_l = V(\widehat{Y}_{HT}).$$

Le même argument s'applique à la forme de Sen-Yates-Grundy : chaque terme $\frac{1}{2}(\pi_k\pi_l - \pi_{kl})(a_k - a_l)^2$ de la variance est divisé par π_{kl} et multiplié par $I_k I_l$, dont l'espérance est π_{kl}.

Ce que cela dit. La division par π_{kl} joue pour les paires le rôle que la division par π_k jouait pour les unités : elle compense la probabilité d'observer la paire. Elle exige $\pi_{kl} > 0$: une paire qui ne peut jamais être tirée ne peut pas être « remontée » par un poids.

■

Vérification sur la mini-population

Les dix estimations de la variance

Pour l'aléatoire simple, $(\pi_k\pi_l - \pi_{kl})/\pi_{kl} = 0{,}6$: $\widehat{V} = 0{,}6 \times (2{,}5\, y_k - 2{,}5\, y_l)^2 = 3{,}75\, (y_k - y_l)^2$.

Échantillon	AB	AC	AD	AE	BC
$\widehat{V}$	15	33,75	93,75	375	3,75

Échantillon	BD	BE	CD	CE	DE
$\widehat{V}$	33,75	240	15	183,75	93,75

Moyenne des dix : $1\,087{,}5/10 = 108{,}75 = V(\widehat{Y}_{HT})$.

Sans biais, mais très dispersé

Les estimations vont de 3,75 à 375 : avec deux unités, la variance est mal connue. Pour l'aléatoire simple, la formule se réduit à $N^2\,(1 - f)\,s^2/n$; le chapitre 3 l'exploitera avec $n = 200$.

Quand la variance ne s'estime pas

Le plan par groupes ne voit jamais deux ménages d'un même groupe

$\pi_{AB} = \pi_{AC} = \pi_{BC} = \pi_{DE} = 0$: aucun échantillon ne renseigne sur la dispersion **à l'intérieur** d'un groupe, celle qui fait la variance. Les deux estimateurs échouent : $\widehat{V}$ vaut 479 en moyenne, $\widehat{V}_{SYG}$ vaut toujours 0, contre $V = 39$.

Démonstration : Pourquoi les deux estimateurs échouent sous le plan par groupes

Les $a_k = y_k/\pi_k$ valent 12, 18, 21 pour A, B, C et 18, 28 pour D, E. Un échantillon contient toujours un ménage de chaque groupe : deux ménages de groupes différents.

Les paires observées ont $\Delta_{kl} = 0$. Pour k et l de groupes différents, $\pi_{kl} = \frac{1}{3} \times \frac{1}{2} = \pi_k \pi_l$, donc $\Delta_{kl} = 0$ et $\pi_k \pi_l - \pi_{kl} = 0$. Dans l'estimateur général, il ne reste que les termes diagonaux $\Delta_{kk}/\pi_k = 1 - \pi_k$; dans la forme de Sen-Yates-Grundy, tous les termes sont nuls :

$$\widehat{V}_{SYG} = 0 \text{ pour tout échantillon,} \qquad \widehat{V} = \sum_{k \in S} (1 - \pi_k)\, a_k^2.$$

L'espérance de $\widehat{V}$. Chaque ménage est tiré avec la probabilité π_k, donc

$$E(\widehat{V}) = \sum_{k \in U} \pi_k (1 - \pi_k)\, a_k^2 = \tfrac{2}{9}\,(144 + 324 + 441) + \tfrac{1}{4}\,(324 + 784) = 202 + 277 = 479.$$

Ce qui manque. La vraie variance contient aussi les paires du **même** groupe, avec $\Delta_{AB} = \Delta_{AC} = \Delta_{BC} = -\frac{1}{9}$ et $\Delta_{DE} = -\frac{1}{4}$; chaque paire compte deux fois dans la double somme :

$$2 \times \left(-\tfrac{1}{9}\right)(12 \times 18 + 12 \times 21 + 18 \times 21) + 2 \times \left(-\tfrac{1}{4}\right)(18 \times 28) = -188 - 252 = -440.$$

D'où $V = 479 - 440 = 39$. Ces paires ont $\pi_{kl} = 0$: on ne les voit jamais, et aucune pondération ne remplace ce qu'on n'observe pas.

Ce que cela dit. Le plan est excellent pour estimer le total (variance 39 contre 108,75), mais il ne contient aucun échantillon qui informe sur la dispersion **à l'intérieur** d'un groupe : celle-là seule fait la variance.

■

Un problème qui reviendra

Le tirage systématique (chapitre 8) et les plans à une seule unité par strate ont des π_{kl} nuls : leur estimateur reste sans biais, mais sa variance ne peut qu'être **approchée**. Un bon plan permet de mesurer sa propre précision.

Biais, variance et erreur quadratique

Trois mesures de la qualité

Biais, variance, erreur quadratique moyenne

Pour un estimateur $\hat{\theta}$ d'un paramètre θ :

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta \qquad V(\hat{\theta}) = E\big[(\hat{\theta} - E\hat{\theta})^2\big]$$

$$EQM(\hat{\theta}) = E\big[(\hat{\theta} - \theta)^2\big] = V(\hat{\theta}) + B(\hat{\theta})^2$$

La démonstration

On écrit $\hat{\theta} - \theta = (\hat{\theta} - E\hat{\theta}) + B$, où B est un nombre fixe. En élevant au carré puis en prenant l'espérance, le double produit $2B\,E(\hat{\theta} - E\hat{\theta})$ est nul, car $E(\hat{\theta} - E\hat{\theta}) = 0$. Il reste $V + B^2$.

Démonstration : Erreur quadratique moyenne égale variance plus biais au carré

Notons $\mu = E(\hat{\theta})$ et $B = \mu - \theta$ (un nombre fixe). On ajoute et retranche μ :

$$\hat{\theta} - \theta = (\hat{\theta} - \mu) + (\mu - \theta) = (\hat{\theta} - \mu) + B.$$

En élevant au carré :

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mu)^2 + 2B\,(\hat{\theta} - \mu) + B^2.$$

Prenons l'espérance, terme à terme. $E\big[(\hat{\theta} - \mu)^2\big] = V(\hat{\theta})$ par définition. B est une constante, donc $E\big[2B(\hat{\theta} - \mu)\big] = 2B\,(E(\hat{\theta}) - \mu) = 0$. Il reste

$$EQM(\hat{\theta}) = V(\hat{\theta}) + B^2.$$

Ce que cela dit. Biais et variance s'additionnent sans terme croisé : on peut les étudier séparément, puis les sommer. Un estimateur sans biais a pour EQM sa variance ; un estimateur biaisé n'est pas pour autant condamné, si sa variance est assez faible (voir ci-dessous).

■

Les cibles



sans biais, précis | sans biais, imprécis | biaisé, précis | biaisé, imprécis

Lire les quatre cibles

Le centre est le paramètre ; chaque point est l'estimation d'un échantillon. Le **biais** est le décalage du nuage, la **variance** sa dispersion, l'**EQM** la distance moyenne au centre, au carré. Un estimateur biaisé mais précis peut tomber plus près, en moyenne, qu'un estimateur sans biais très dispersé.

Brute ou pondérée sous le plan par groupes

Trois estimateurs de la moyenne, n = 2

Estimateur	Espérance	Biais	Variance	EQM
Aléatoire simple, $\overline{y}$	8	0	4,35	4,35
Groupes, moyenne brute	8,58	0,58	1,95	2,29
Groupes, HT pondérée	8	0	1,56	1,56

$\overline{Y} = 8$. Brute : $E = 51{,}5/6$; $EQM = 1{,}95 + 0{,}58^2 = 2{,}29$. Pondérée : $V = 39/25 = 1{,}56$.

Deux leçons

Sous le plan par groupes, la pondération supprime le biais **et** réduit la variance. Et la moyenne brute, pourtant biaisée, fait mieux que l'aléatoire simple sans biais : 2,29 contre 4,35. Le biais n'est pas tout ; c'est l'EQM qui mesure l'erreur.

Démonstration : Les trois lignes du tableau

On a $\overline{Y} = 8$. Chaque plan est énuméré : dix échantillons de probabilité $\frac{1}{10}$ pour l'aléatoire simple, six de probabilité $\frac{1}{6}$ pour le plan par groupes.

Aléatoire simple, $\overline{y}$. Sans biais, donc $EQM = V$. La variance du total est 108,75 (démontrée plus haut par la formule générale, retrouvée sur les dix échantillons) ; celle de la moyenne s'obtient en divisant par $N^2 = 25$: $108{,}75/25 = 4{,}35$.

Groupes, moyenne brute. Les six moyennes sont 6,5 ; 9 ; 7,5 ; 10 ; 8 ; 10,5. Leur espérance est $\frac{51{,}5}{6} = 8{,}583$, d'où un biais de 0,583. La moyenne de leurs carrés est $\frac{453{,}75}{6} = 75{,}625$, donc

$$V = 75{,}625 - 8{,}583^2 = 1{,}951, \qquad EQM = V + B^2 = 1{,}951 + 0{,}340 = 2{,}292.$$

Groupes, moyenne pondérée de Horvitz-Thompson. Les six valeurs sont 6 ; 8 ; 7,2 ; 9,2 ; 7,8 ; 9,8. Leur espérance est 8 (aucun biais) et les écarts à 8 sont

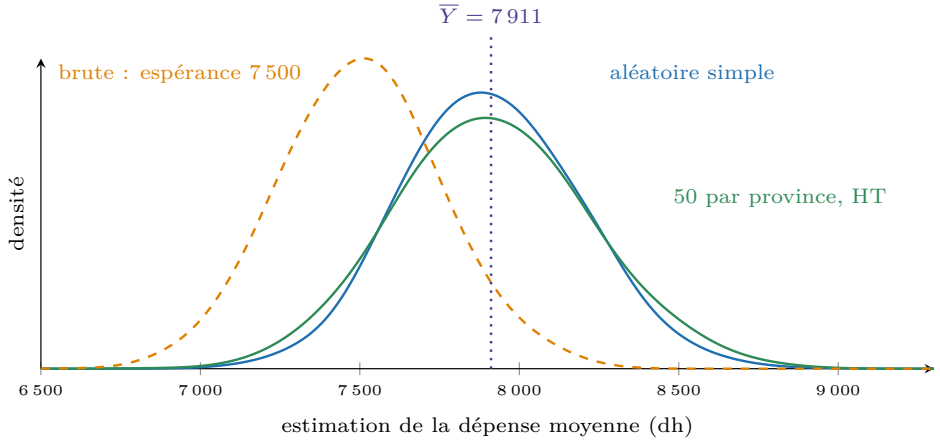
$-2\,;\ 0\,;\ -0{,}8\,;\ 1{,}2\,;\ -0{,}2\,;\ 1{,}8$, de carrés $4+0+0{,}64+1{,}44+0{,}04+3{,}24=9{,}36$:

$$V = EQM = \frac{9{,}36}{6} = 1{,}56 = \frac{39}{25}.$$

Ce que cela dit. Le classement par EQM est $1{,}56 < 2{,}29 < 4{,}35$. La moyenne brute, biaisée, bat l'aléatoire simple parce que le plan par groupes interdit les échantillons extrêmes ; la pondérée garde cet avantage et supprime le biais.

■

Sur la population du cours



Quatre mille tirages de 200 ménages pour chaque estimateur

Les deux estimateurs de Horvitz-Thompson sont centrés sur $\overline{Y}$. La moyenne brute du plan à 50 ménages par province est la plus resserrée, mais centrée sur 7 500 dh.

Le biais ne diminue pas avec n

Valeurs exactes, en dh

Plan et estimateur	n	Biais	Écart-type	$\sqrt{EQM}$
Aléatoire simple	200	0	267,3	267,3
50 par province, HT	200	0	298,4	298,4
50 par province, brute	200	−411,1	246,6	479,4
200 par province, brute	800	−411,1	98,4	422,7

Plus de ménages ne corrige pas un biais

Quadrupler l'échantillon divise l'écart-type de la moyenne brute par 2,5, mais laisse le biais intact : il représente 74 % de l'EQM avec 200 ménages, 95 % avec 800. Au-delà d'une certaine taille, **seul le biais compte** : c'est la leçon du Literary Digest.

Démonstration : Le biais ne dépend pas de la taille de l'échantillon

Soit m ménages tirés par sondage aléatoire simple dans chacune des H provinces, indépendamment. D'après la démonstration du chapitre 1, l'espérance de la moyenne brute est $\frac{1}{H}\sum_h \overline{Y}_h$, **quel que soit** m : le biais $B = \frac{1}{H}\sum_h \overline{Y}_h - \overline{Y}$ ne contient pas m. Pour les provinces du cours, $B = 7\,500{,}29 - 7\,911{,}44 = -411{,}1$ dh, que l'on tire 50 ou 200 ménages par province.

La variance, elle, dépend de m. Dans chaque province, la moyenne $\overline{y}_h$ est celle d'un sondage aléatoire simple de m ménages parmi N_h : sa variance est $(1 - m/N_h)\, S_h^2/m$ (chapitre 3). Les provinces étant tirées indépendamment, les variances s'ajoutent, et chaque moyenne $\overline{y}_h$ pèse $\frac{1}{H}$:

$$V(\overline{y}_{\text{brute}}) = \frac{1}{H^2}\sum_{h=1}^{H}\left(1 - \frac{m}{N_h}\right)\frac{S_h^2}{m}.$$

Avec $S_h = 4\,271$, $3\,873$, $3\,593$ et $2\,900$ dh, cela donne un écart-type de 246,6 dh pour $m = 50$ et de 98,4 dh pour $m = 200$.

L'EQM. En utilisant $EQM = V + B^2$:

$$\sqrt{EQM} = \sqrt{246{,}6^2 + 411{,}1^2} = 479{,}4 \quad (m = 50), \qquad \sqrt{98{,}4^2 + 411{,}1^2} = 422{,}7 \quad (m = 200).$$

Le carré du biais représente $\dfrac{411{,}1^2}{479{,}4^2} = 74\,\%$ de l'EQM pour $m = 50$ et $\dfrac{411{,}1^2}{422{,}7^2} = 95\,\%$ pour $m = 200$.

Ce que cela dit. Quadrupler m divise la variance par un facteur voisin de 4 (un peu plus, par la correction de population finie) mais ne touche pas au biais : l'EQM tend vers B^2, et il n'y a plus rien à gagner à interroger davantage de ménages avec une mauvaise pondération.

■

L'intervalle de confiance

L'approximation normale

Intervalle de confiance à 95 pour cent

Quand n est grand, $\widehat{Y}_{HT}$ suit approximativement une loi normale (théorème central limite, chapitre 20 de Probabilités ; Statistique inférentielle, chapitres 8 à 10) :

$$\left[\,\widehat{Y}_{HT} - 1{,}96\,\sqrt{\widehat{V}}\;;\;\widehat{Y}_{HT} + 1{,}96\,\sqrt{\widehat{V}}\,\right]$$

Pour la moyenne, on divise les bornes par N.

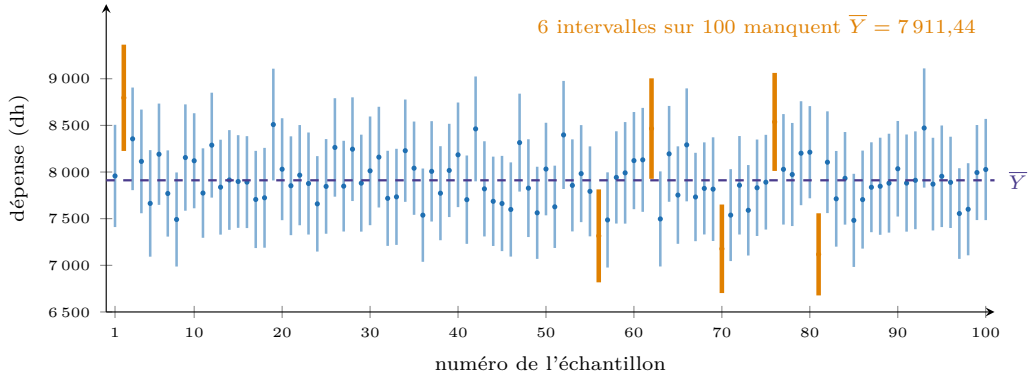
L'échantillon fil rouge : aléatoire simple de 200 ménages

$\overline{y} = 7\,843{,}05$ dh, $s = 3\,829{,}16$ dh, erreur-type estimée $\sqrt{(1-f)\,s^2/n} = 256{,}87$ dh :

$$\text{IC}_{95\,\%}(\overline{Y}) = 7\,843{,}05 \pm 1{,}96 \times 256{,}87 = [\,7\,339{,}6\;;\;8\,346{,}5\,]$$

Total : $\widehat{Y}_{HT} = 2\,000 \times 7\,843{,}05 = 15\,686\,100$ dh, intervalle $[\,14\,679\,178\;;\;16\,693\,022\,]$.

Cent échantillons, cent intervalles



L'intervalle bouge, le paramètre reste

Chaque segment est l'intervalle à 95 % d'un échantillon de 200 ménages. Six d'entre eux, en orange, manquent $\overline{Y}$: on en attendait environ cinq.

La lecture correcte

Ce que garantit le niveau de confiance

Avant le tirage, l'intervalle est **aléatoire** : la probabilité qu'il contienne $\overline{Y}$ vaut environ 0,95. Le niveau de confiance qualifie la **méthode** : sur un grand nombre d'échantillons, 95 % des intervalles contiennent $\overline{Y}$.

Une phrase à ne pas écrire

« Il y a 95 % de chances que $\overline{Y}$ soit entre 7 339,6 et 8 346,5 » est **faux** : une fois l'échantillon tiré, l'intervalle et $\overline{Y}$ sont deux choses fixes. Ou bien il le contient, ou bien non. Ici il le contient (7 911,44) ; l'enquêteur, lui, ne le saura jamais.

La couverture simulée

Cinq mille sondages aléatoires simples

Taille n	50	200	800
Couverture	0,9420	0,9424	0,9512

La **couverture** est la part des intervalles qui contiennent $\overline{Y}$. Avec $n = 200$: 288 échecs, dont 188 intervalles **trop bas** et 100 trop hauts. L'écart à 0,95 vaut 2,5 erreurs-types de simulation (0,0031) : il est réel.

Pourquoi un peu moins de 95

La dépense est **asymétrique** : quelques ménages dépensent beaucoup. Un échantillon qui en contient peu a une moyenne basse **et** un s faible : son intervalle est trop court précisément quand il est décalé. L'approximation normale s'améliore quand n grandit.

Le coefficient de variation

Une précision relative

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{\theta}, \qquad \text{estimé par } \frac{\sqrt{\widehat{V}}}{\hat{\theta}}$$

Sans unité, il est le même pour le total et pour la moyenne ($\widehat{Y}_{HT} = N\,\widehat{\overline{Y}}_{HT}$). La marge à 95 % vaut environ $2\,CV$ en valeur relative.

Dépense moyenne, aléatoire simple

Taille n	50	200	1 000
Erreur-type de $\overline{y}$	556,38	267,27	89,09
CV	7,0 %	3,4 %	1,1 %

Fil rouge : CV estimé $= 256{,}87/7\,843{,}05 = 3{,}3\,\%$, marge relative $\pm 6{,}4\,\%$.

S'entraîner

Exercice 1 : un plan à probabilités inégales

Trois unités, trois échantillons

Une population compte trois unités, de valeurs $y_1 = 10$, $y_2 = 21$ et $y_3 = 40$ ($Y = 71$). Un plan de taille 2 donne aux échantillons $\{1,2\}$, $\{1,3\}$ et $\{2,3\}$ les probabilités 0,2, 0,3 et 0,5.

1. Calculer les π_k et vérifier que leur somme vaut n.
2. Calculer $\widehat{Y}_{HT}$ pour chaque échantillon, puis son espérance.
3. Calculer $V(\widehat{Y}_{HT})$.
4. On propose plutôt $N\,\overline{y} = 3\,\overline{y}$. Calculer son biais, sa variance et son EQM. Conclure.

Exercice 1 : solution

Horvitz-Thompson

1. $\pi_1 = 0{,}2 + 0{,}3 = 0{,}5$; $\pi_2 = 0{,}2 + 0{,}5 = 0{,}7$; $\pi_3 = 0{,}3 + 0{,}5 = 0{,}8$; somme $2 = n$.
2. y_k/π_k : 20 ; 30 ; 50. D'où $\widehat{Y}_{HT} = 50$; 70 ; 80 et $E = 0{,}2\times 50 + 0{,}3\times 70 + 0{,}5\times 80 = 71 = Y$.
3. $E(\widehat{Y}_{HT}^2) = 0{,}2 \times 2\,500 + 0{,}3 \times 4\,900 + 0{,}5 \times 6\,400 = 5\,170$; $V = 5\,170 - 71^2 = 129$.

La moyenne brute extrapolée

4. $3\,\overline{y}$ vaut 46,5 ; 75 ; 91,5. $E = 77{,}55$, biais $+6{,}55$: l'unité 3, la plus grande, est tirée quatre fois sur cinq. $V = 6\,306{,}075 - 77{,}55^2 = 292{,}07$ et $EQM = 292{,}07 + 6{,}55^2 = 334{,}98$, contre 129 : Horvitz-Thompson l'emporte sur les deux tableaux.

Exercice 2 : variance et estimation de la variance

Les quatre boutiques du chapitre 1

Les chiffres d'affaires sont 2, 5, 8 et 9 milliers de dh ($S^2 = 10$) ; on en tire deux par sondage aléatoire simple.

1. Calculer π_k, π_{kl}, $\pi_k\pi_l - \pi_{kl}$ et les y_k/π_k.
2. Calculer $V(\widehat{Y}_{HT})$ par la forme de Sen-Yates-Grundy ; comparer à $N^2(1-f)\,S^2/n$.
3. Calculer $\widehat{V}_{SYG}$ pour chacun des six échantillons. Vérifier qu'il est sans biais.

Exercice 2 : solution

La variance

1. $\pi_k = 0{,}5$, $\pi_{kl} = 1/6$, $\pi_k\pi_l - \pi_{kl} = \frac{1}{4} - \frac{1}{6} = \frac{1}{12}$; y_k/π_k : 4 ; 10 ; 16 ; 18.
2. Carrés des écarts sur les six paires : 36 ; 144 ; 196 ; 36 ; 64 ; 4, de somme 480. $V = \frac{1}{12} \times 480 = 40$, et $16 \times 0{,}5 \times 10/2 = 40$.

Son estimation

3. Une seule paire par échantillon, et $\frac{1/12}{1/6} = 0{,}5$: $\widehat{V}_{SYG} = 0{,}5\,(a_k - a_l)^2$. Pour $\{2,5\}$, $\{2,8\}$, $\{2,9\}$, $\{5,8\}$, $\{5,9\}$, $\{8,9\}$: 18 ; 72 ; 98 ; 18 ; 32 ; 2. Moyenne : $240/6 = 40 = V$.

Exercice 3 : un sondage de 400 ménages

Intervalle, coefficient de variation et EQM

Un sondage aléatoire simple de 400 ménages de la population du cours donne $\overline{y} = 7\,952$ dh et $s = 4\,074$ dh.

1. Calculer l'erreur-type estimée $\sqrt{(1-f)\,s^2/n}$, l'intervalle de confiance à 95 % et le CV.
2. Vrai ou faux : « il y a 95 % de chances que $\overline{Y}$ soit dans cet intervalle » ?
3. Un plan moins coûteux donne un estimateur de biais 200 dh et d'écart-type 120 dh. Comparer son EQM à celle du sondage ($V = 31\,749{,}1$).

Exercice 3 : solution

Précision et lecture

1. $f = 0{,}2$; $\sqrt{0{,}8 \times 4\,074^2/400} = 182{,}19$ dh ; IC : $7\,952 \pm 357{,}1 = [\,7\,594{,}9\ ;\ 8\,309{,}1\,]$; $CV = 182{,}19/7\,952 = 2{,}3\,\%$.
2. Faux : l'intervalle tiré contient $\overline{Y}$ ou non (ici oui : 7 911,44). Les 95 % décrivent la méthode, sur l'ensemble des échantillons possibles.

Le biais, un plancher

3. $EQM = 120^2 + 200^2 = 54\,400$ (racine 233,2 dh), contre 31 749,1 (racine 178,2 dh). Même avec un écart-type réduit à 60 dh, $EQM = 43\,600$: la racine de l'EQM ne descend jamais sous le biais, 200 dh.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. L'estimateur de **Horvitz-Thompson** $\widehat{Y}_{HT} = \sum_{k \in S} y_k/\pi_k$ compte chaque unité tirée pour son poids d_k.
2. Il est **sans biais** pour tout plan, pourvu que $\pi_k > 0$ pour toute unité.
3. $V(\widehat{Y}_{HT}) = \sum_k \sum_l \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$; à taille fixe, elle est nulle si y_k/π_k est constant.
4. La variance s'estime sur l'échantillon si $\pi_{kl} > 0$ pour toutes les paires.
5. $EQM = V + B^2$: la variance baisse avec n, **le biais ne baisse pas**.
6. L'intervalle $\hat{\theta} \pm 1{,}96\sqrt{\widehat{V}}$ couvre le paramètre dans environ 95 % des échantillons (0,9424 simulé pour $n = 200$).
7. Le $CV = \sqrt{V}/\theta$ mesure la précision relative.

La suite

Le chapitre 3 applique ces résultats au sondage aléatoire simple : il démontre $V(\overline{y}) = (1 - f)\, S^2/n$ et montre comment l'estimer.