# Échantillonnage

Chapitre 11 : Les probabilités inégales

Jaouad Madkour

31 octobre 2026

## Plan du chapitre

**Tirer selon la taille** **2**
Des districts de tailles inégales . . . 2
La dépense d'un district suit sa taille . . . 2
Une probabilité proportionnelle à la taille . . . 3

**Avec remise : Hansen-Hurwitz** **3**
La méthode des tailles cumulées . . . 3
L'estimateur de Hansen-Hurwitz . . . 3
Sa variance et son estimation . . . 4
Sur la mini-population . . . 5

**Cinq districts tirés selon leur taille** **5**
La moyenne des moyennes de district . . . 5
Un tirage de cinq districts . . . 6
Estimer la précision . . . 7
Le gain . . . 7
Vingt mille enquêtes . . . 8

**Sans remise : Horvitz-Thompson** **8**
Éviter les répétitions . . . 8
Le tirage systématique proportionnel à la taille . . . 9
Cinq districts sur l'axe des ménages . . . 10
Sur la mini-population . . . 10
La variance de Sen-Yates-Grundy . . . 10

**Deux degrés autopondéré** **12**
Districts selon leur taille, ménages en nombre fixe . . . 12
Pourquoi les instituts l'aiment . . . 12
La variance et son estimation . . . 12
Vingt districts de dix ménages sur la population . . . 14

**Le risque d'une mauvaise mesure de taille** **14**
Une mesure de taille, plusieurs variables . . . 14
Quand la taille va à rebours . . . 14

Une taille périmée . . . . . 15

**S'entraîner** **16**
Exercice 1 : quatre villages . . . . . 16
Exercice 1 : solution . . . . . 16
Exercice 2 : huit douars . . . . . 16
Exercice 2 : solution . . . . . 17
Exercice 3 : vingt-cinq districts de huit ménages . . . . . 17
Exercice 3 : solution . . . . . 17

**Ce qu'il faut retenir** **18**
Le résumé du chapitre . . . . . 18

# Tirer selon la taille

## Des districts de tailles inégales

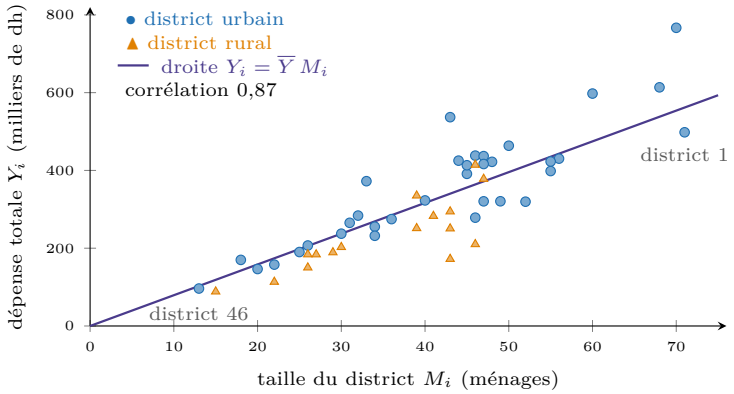
**Cinq districts tirés à probabilités égales (chapitre 9)**

| Estimateur de la dépense moyenne | Écart-type | *deff* |
|---|---|---|
| Aléatoire simple de 200 ménages | 267,3 | 1 |
| 5 districts, Horvitz-Thompson | 1 516,6 | 32,2 |
| 5 districts, ratio | 773,8 | 8,4 |

**Ce qui fait varier le total d'un district**

Les districts comptent de 13 à 71 ménages. Le total $Y_i$ d'un district dépend d'abord de sa **taille** $M_i$, ensuite seulement de la dépense de ses ménages. Tirés à probabilités égales, cinq districts héritent de toute la variation des tailles : c'est elle qui gonfle la variance de Horvitz-Thompson.

## La dépense d'un district suit sa taille



**Une corrélation de 0,87**

Les points s'alignent autour de la droite $Y_i = \overline{Y}\, M_i$ : la moyenne $\overline{Y}_i = Y_i/M_i$ varie beaucoup moins que le total $Y_i$. Coefficient de variation des totaux : 0,45 ; des moyennes : 0,21.

### Une probabilité proportionnelle à la taille

**Le tirage proportionnel à la taille**

On connaît pour chaque unité $i$, avant l'enquête, une **mesure de taille** $x_i > 0$, de total $X$. On tire l'unité $i$ avec la probabilité

$$p_i = \frac{x_i}{X}, \qquad \sum_i p_i = 1$$

On parle de tirage **PPT** (probabilité proportionnelle à la taille). Pour les districts : $x_i = M_i$, $X = N = 2\,000$ et $p_i = M_i/2\,000$.

**Le plus grand et le plus petit district**

District 1 : $p_1 = 71/2\,000 = 0{,}0355$. District 46 : $p_{46} = 13/2\,000 = 0{,}0065$. Le premier a $71/13 = 5{,}46$ fois plus de chances d'être tiré que le second.

## Avec remise : Hansen-Hurwitz

### La méthode des tailles cumulées

**Tirer selon la taille avec un seul nombre**

1. Calculer les tailles cumulées $C_i = x_1 + \cdots + x_i$, avec $C_0 = 0$.
2. Tirer un entier $a$ au hasard entre 1 et $X$.
3. Retenir l'unité $i$ telle que $C_{i-1} < a \leq C_i$.

L'unité $i$ occupe $x_i$ entiers sur $X$ : elle sort avec la probabilité $x_i/X$. On répète $n$ fois, **indépendamment** : c'est un tirage avec remise.

**La mini-population, avec la taille pour mesure**

Tailles 2, 3, 3, 4, 6 ; cumuls 2, 5, 8, 12, 18. On tire $a$ entre 1 et 18 : A si $a \leq 2$, B de 3 à 5, C de 6 à 8, D de 9 à 12, E de 13 à 18. Ainsi $p_A = 2/18 = 1/9$ et $p_E = 6/18 = 1/3$.

**Tirer un ménage, garder son district**

Pour les districts, $X = N$ : tirer $a$, c'est tirer **un ménage** au hasard et retenir son district.

### L'estimateur de Hansen-Hurwitz

**Une estimation par tirage, puis leur moyenne**

Au tirage $j$ sort une unité de valeur $y_{(j)}$ et de probabilité $p_{(j)}$. Ce tirage fournit à lui seul

l'estimation $z_j = y_{(j)}/p_{(j)}$ du total, et

$$\widehat{Y}_{HH} = \frac{1}{n}\sum_{j=1}^{n} z_j = \frac{1}{n}\sum_{j=1}^{n} \frac{y_{(j)}}{p_{(j)}}$$

**Sans biais, en une ligne**

$z_j$ vaut $y_i/p_i$ avec la probabilité $p_i$, donc

$$E(z_j) = \sum_i p_i \frac{y_i}{p_i} = \sum_i y_i = Y$$

La moyenne de $n$ estimations sans biais est sans biais. Avec $p_i = 1/N$, on retrouve $N\,\overline{y}_{AR}$, l'estimateur avec remise du chapitre 4.

## Sa variance et son estimation

**Des tirages indépendants et de même loi**

$$V(\widehat{Y}_{HH}) = \frac{1}{n}\sum_i p_i \left(\frac{y_i}{p_i} - Y\right)^2 \qquad \widehat{V}(\widehat{Y}_{HH}) = \frac{s_z^2}{n} = \frac{1}{n\,(n-1)}\sum_{j=1}^{n}\left(z_j - \widehat{Y}_{HH}\right)^2$$

**La démonstration**

Les $z_j$ sont indépendants, d'espérance $Y$ et de variance $V(z) = \sum_i p_i\,(y_i/p_i - Y)^2$ : leur moyenne a la variance $V(z)/n$. La variance $s_z^2$ des $n$ valeurs $z_j$ est sans biais pour $V(z)$ (Statistique inférentielle, chapitre 3), quels que soient les $p_i$.

**Démonstration : Hansen-Hurwitz, sans biais, sa variance et son estimation**

*Une estimation par tirage.* Les tirages se font avec remise : à chaque tirage, l'unité $i$ sort avec la probabilité $p_i$, quoi qu'il soit sorti avant. La variable $z_j = y_{(j)}/p_{(j)}$ prend donc la valeur $y_i/p_i$ avec la probabilité $p_i$, et $z_1, \ldots, z_n$ sont **indépendantes et de même loi**.

*Pas de biais.*

$$E(z_j) = \sum_i p_i \frac{y_i}{p_i} = \sum_i y_i = Y, \qquad E(\widehat{Y}_{HH}) = \frac{1}{n}\sum_{j=1}^{n} E(z_j) = Y.$$

*La variance de $z$.* En développant le carré, avec $\sum_i p_i = 1$ et $\sum_i p_i \frac{y_i}{p_i} = Y$,

$$V(z) = \sum_i p_i \left(\frac{y_i}{p_i} - Y\right)^2 = \sum_i \frac{y_i^2}{p_i} - 2\,Y\sum_i y_i + Y^2 \sum_i p_i = \sum_i \frac{y_i^2}{p_i} - Y^2.$$

Les $z_j$ étant indépendantes, $V(\widehat{Y}_{HH}) = V(z)/n$.

*L'estimation de la variance.* Soit $s_z^2 = \frac{1}{n-1}\sum_j (z_j - \overline{z})^2$. En écrivant $z_j - \overline{z} = (z_j - Y) - (\overline{z} - Y)$,

$$\sum_j (z_j - \overline{z})^2 = \sum_j (z_j - Y)^2 - n\,(\overline{z} - Y)^2,$$

dont l'espérance est $n\,V(z) - n\,\dfrac{V(z)}{n} = (n-1)\,V(z)$. Donc $E(s_z^2) = V(z)$ et $E(s_z^2/n) = V(\widehat{Y}_{HH})$ : l'estimateur de variance est sans biais, **quels que soient les** $p_i$.

**Ce que cela dit.** Sur la mini-population ($n = 2$, tailles 2, 3, 3, 4, 6), $z = 36, 36, 42, 40{,}5, 42$ : $E(z) = 40 = Y$ et $V(z) = 117/18 = 6{,}5$, d'où $V(\widehat{Y}_{HH}) = 3{,}25$. L'énumération des 25 couples ordonnés de tirages, de probabilité $p_i\,p_j$, retrouve exactement $E(\widehat{Y}_{HH}) = 40$, $V(\widehat{Y}_{HH}) = 3{,}25$ et $E(s_z^2/n) = 3{,}25$ ; l'énumération des 125 triplets donne de même $V(z)/3$. À probabilités égales, la même population donne 145 avec remise : le tirage selon la taille divise la variance par 45 environ ; par 33 environ face aux 108,75 du sondage aléatoire simple sans remise.

■

**Une variance nulle**

Si $y_i/p_i$ est le même pour toutes les unités, chaque tirage donne exactement $Y$ : il faudrait des $p_i$ **proportionnels aux** $y_i$. Comme les $y_i$ sont inconnus, on prend une taille $x_i$ qui leur est à peu près proportionnelle.

### Sur la mini-population

**Deux tirages proportionnels à la taille**

| Ménage | A | B | C | D | E |
|---|---|---|---|---|---|
| $y_i$ (milliers de dh) | 4 | 6 | 7 | 9 | 14 |
| $p_i = x_i/18$ | 1/9 | 1/6 | 1/6 | 2/9 | 1/3 |
| $z_i = y_i/p_i$ | 36 | 36 | 42 | 40,5 | 42 |

$V(z) = \frac{1}{18}(2\times 4^2 + 3\times 4^2 + 3\times 2^2 + 4\times 0{,}5^2 + 6\times 2^2) = \frac{117}{18} = 6{,}5$, d'où $V(\widehat{Y}_{HH}) = 6{,}5/2 = 3{,}25$.

**Trente-trois fois moins que le sondage aléatoire simple**

À probabilités égales, $V(\widehat{Y}) = 25 \times 4{,}35 = 108{,}75$ sans remise (chapitre 2), $25 \times 5{,}8 = 145$ avec remise (chapitre 4). Ici 3,25 : la dépense par personne $y_i/x_i$ ne va que de 2 à 2,33, et chaque $z_i$ reste proche de $Y = 40$.

## Cinq districts tirés selon leur taille

### La moyenne des moyennes de district

**Hansen-Hurwitz pour des grappes**

Avec $p_i = M_i/N$ et $y_i = Y_i$, chaque tirage donne $z_j = Y_i/p_i = N\,\overline{Y}_i$. La dépense moyenne s'estime par la **moyenne simple** des moyennes des districts tirés :

$$\frac{\widehat{Y}_{HH}}{N} = \frac{1}{n_I}\sum_{j=1}^{n_I}\overline{Y}_{(j)} \qquad V\Big(\frac{\widehat{Y}_{HH}}{N}\Big) = \frac{1}{n_I}\sum_{i=1}^{N_I}\frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2$$

**Les tailles ont disparu**

La variance ne dépend plus que des écarts entre **moyennes** de district : $\sum_i \frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2 = 2\,844\,644$, soit 18 % seulement de $S^2$. Pour $n_I = 5$ : $2\,844\,644/5 = 568\,929$, et l'écart-type vaut 754,3 dh.

**Démonstration : Hansen-Hurwitz avec des grappes, la moyenne des moyennes**

*Les tailles disparaissent.* Le district $i$ est tiré avec la probabilité $p_i = M_i/N$ : tirer le ménage $a$ au hasard parmi les $N$ et retenir son district. Pour la variable $y_i = Y_i = M_i\overline{Y}_i$, le total du district,

$$z = \frac{Y_i}{p_i} = \frac{M_i\overline{Y}_i}{M_i/N} = N\overline{Y}_i.$$

L'estimateur de la moyenne est donc la moyenne simple des moyennes de district tirées :

$$\frac{\widehat{Y}_{HH}}{N} = \frac{1}{n_I}\sum_{j=1}^{n_I}\frac{z_j}{N} = \frac{1}{n_I}\sum_{j=1}^{n_I}\overline{Y}_{(j)}.$$

*Sans biais.* $E(\overline{Y}_{(j)}) = \sum_i \frac{M_i}{N}\overline{Y}_i = \frac{1}{N}\sum_i Y_i = \overline{Y}$.

*La variance.* $\overline{Y}_{(j)}$ prend la valeur $\overline{Y}_i$ avec la probabilité $M_i/N$ et a pour espérance $\overline{Y}$ ; sa variance est $\sum_i \frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2$. Les $n_I$ tirages sont indépendants : on divise par $n_I$. Seuls les écarts entre **moyennes** interviennent.

*Le coût en ménages.* Le nombre de ménages enquêtés dans un tirage est $M_i$, d'espérance $\sum_i \frac{M_i}{N}M_i = \sum_i M_i^2/N$.

**Ce que cela dit.** Sur la population du cours, $\sum_i \frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2 = 2\,844\,644$, soit 17,9 % de $S^2$ ; pour $n_I = 5$ districts, 568 929 et un écart-type de 754,3 dh. À probabilités égales, les cinq districts héritent des écarts de **taille** : 1 516,6 dh avec Horvitz-Thompson. Le prix du tirage selon la taille est le nombre de ménages : $\sum_i M_i^2/N = 44{,}5$ par tirage, soit 222,6 pour cinq tirages. Pour la suite 1704, 358, 53, 1280, 731, les tailles cumulées désignent les districts 42, 7, 1, 30, 16, d'où $\frac{1}{5}\sum_j \overline{Y}_{(j)} = 7\,986{,}74$ dh.

■

## Un tirage de cinq districts

### Cinq nombres entre 1 et 2 000, graine 2026

| Nombre $a$ | District | Province, milieu | $M_i$ | $\overline{Y}_i$ |
|---:|---:|---|---:|---:|
| 1 704 | 42 | Larache, rural | 43 | 6 845,35 |
| 358 | 7 | Tanger-Assilah, urbain | 70 | 10 953,29 |
| 53 | 1 | Tanger-Assilah, urbain | 71 | 7 014,65 |
| 1 280 | 30 | Tétouan, rural | 26 | 7 074,23 |
| 731 | 16 | Tanger-Assilah, rural | 47 | 8 046,17 |

### L'estimation

$\frac{1}{5}(6\,845{,}35 + 10\,953{,}29 + 7\,014{,}65 + 7\,074{,}23 + 8\,046{,}17) = 7\,986{,}74$ dh, pour $\overline{Y} = 7\,911{,}44$. On a enquêté 257 ménages. Les deux plus grands districts, 1 et 7, sont tirés tous les deux : le plan les favorise.

### Cinq moyennes, une variance

Écarts des $\overline{Y}_i$ à leur moyenne : $-1\,141{,}4$ ; $2\,966{,}5$ ; $-972{,}1$ ; $-912{,}5$ ; $59{,}4$. Somme des carrés $11\,884\,335$, d'où $s^2 = 11\,884\,335/4 = 2\,971\,084$ et

$$\widehat{V} = \frac{2\,971\,084}{5} = 594\,217 \qquad \text{erreur-type estimée } 770{,}9 \text{ dh}$$

Intervalle à 95 %, Student à 4 degrés de liberté : $7\,986{,}7 \pm 2{,}776 \times 770{,}9 = [5\,846{,}8\,;\,10\,126{,}6]$.

### Cinq observations, pas 257

La variance se calcule sur les **cinq** moyennes de district : c'est le tirage des districts qui porte l'aléa, et les ménages d'un même district se ressemblent. L'erreur-type estimée, 770,9, est proche de la vraie, 754,3 ; l'erreur commise, +75 dh, est bien couverte.

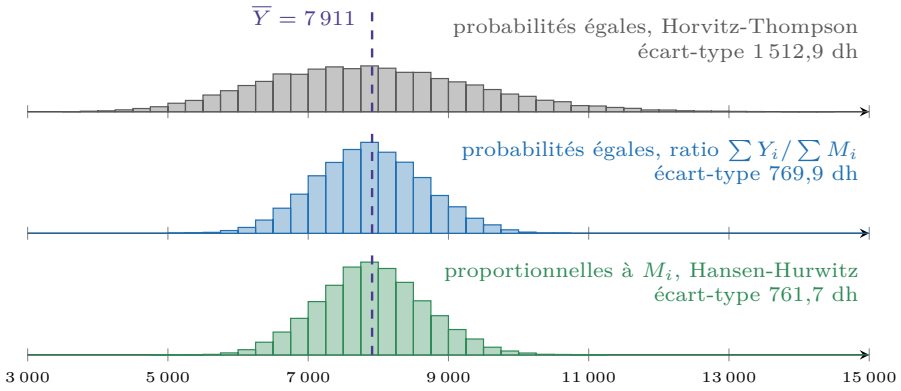
### Cinq districts, trois façons de faire

| Cinq districts, estimateur | Ménages | Écart-type | *deff* |
|---|---:|---:|---:|
| (aléatoire simple de 200 ménages) | 200 | 267,3 | 1 |
| Égales, Horvitz-Thompson | 200 | 1 516,6 | 32,2 |
| Égales, ratio | 200 | 773,8 | 8,4 |
| PPT, Hansen-Hurwitz | 222,6 | 754,3 | 8,0 |

Ménages : nombre moyen de ménages enquêtés. Effet de plan rapporté à l'aléatoire simple de 200 ménages.

### Aussi bon que le ratio, sans ses défauts

Le tirage PPT fait aussi bien que le ratio, mais son estimateur est **exactement sans biais**, et sa variance s'estime par une formule exacte. Le prix : les grands districts, plus souvent tirés, font enquêter en moyenne $\sum_i M_i^2/N \;=\; 44{,}5$ ménages par tirage, soit 222,6 au lieu de 200.



estimation de la dépense moyenne (dh), 5 districts

### Trois estimateurs, cinq districts, graine 11

Horvitz-Thompson à probabilités égales se trompe de plus de 2 000 dh dans 18,6 % des enquêtes, le ratio dans 0,7 %, Hansen-Hurwitz dans 0,8 %. Écarts-types simulés : 1 512,9, 769,9 et 761,7 dh.

### Un district tiré deux fois

Avec remise, cinq tirages donnent cinq districts **distincts** avec la probabilité 0,795 seulement : une enquête sur cinq tire deux fois le même district, et n'apprend rien de plus la seconde fois.

### Des probabilités d'inclusion proportionnelles à la taille

Sans remise, on vise $\pi_i \;=\; n\,x_i/X$, ce qui exige $n\,x_i \;\leq\; X$, et l'on emploie l'estimateur de Horvitz-Thompson (chapitre 2), $\widehat{Y}_{HT} \;=\; \sum_{i\in S} y_i/\pi_i$. Pour 5 districts, $\pi_i \;=\; 5\,M_i/2\,000 \;=\; M_i/400$, de 0,0325 à 0,1775 : $\widehat{Y}_{HT}/N$ est encore la moyenne simple des $\overline{Y}_i$ tirés.

### Les unités trop grandes

Si $n\,x_i \;>\; X$, l'unité est prise d'office ($\pi_i \;=\; 1$) et l'on tire les autres parmi le reste. Avec 40 districts, $\pi_i = M_i/50$ dépasserait 1 pour les 8 districts de plus de 50 ménages.

### Tailles cumulées et pas constant

1. Ranger les unités dans un ordre choisi ; calculer les tailles cumulées $C_i$.
2. Fixer le pas $k = X/n$ ; tirer un départ $r$ au hasard entre 1 et $k$.
3. Retenir les unités dont l'intervalle $]C_{i-1}\,;C_i]$ contient l'un des points $r$, $r+k$, ..., $r+(n-1)\,k$.

Si $x_i \leq k$, l'intervalle de l'unité $i$ reçoit un point pour exactement $x_i$ des $k$ départs : $\pi_i = x_i/k = n\,x_i/X$.

### Cinq districts dans l'ordre du fichier

$k = 2\,000/5 = 400$ ; départ tiré (graine 2026) : $r = 341$. Les points 341, 741, 1 141, 1 541 et 1 941 tombent dans les districts 7, 17, 26, 37 et 49.

### Démonstration : Le tirage systématique proportionnel à la taille

*Un point par départ.* Posons $k = X/n$ et tirons le départ $r$ parmi $1, \dots, k$, chacun avec la probabilité $1/k$. Un entier $a$ de l'axe $\{1, \dots, X\}$ est l'un des points $r$, $r+k$, ..., $r+(n-1)k$ pour **un seul** départ : celui qui vaut $a$ modulo $k$.

*La probabilité d'inclusion.* L'unité $i$ occupe les $x_i$ entiers de $]C_{i-1}\,;C_i]$. Si $x_i \leq k$, deux de ces entiers diffèrent de moins de $k$ : ils ne sont pas congrus modulo $k$, donc ils correspondent à $x_i$ départs **distincts**. L'unité $i$ est tirée pour exactement $x_i$ des $k$ départs :

$$\pi_i = \frac{x_i}{k} = \frac{n\,x_i}{X}.$$

*Un échantillon de taille fixe.* Deux points d'un même départ sont distants de $k$, plus que la longueur $x_i$ d'un intervalle : ils tombent dans deux unités différentes. Chaque départ donne donc $n$ unités distinctes.

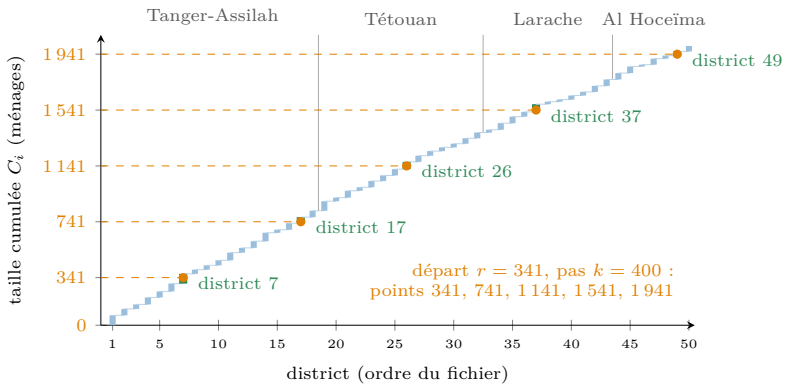
*Horvitz-Thompson.* $y_i/\pi_i = k\,y_i/x_i$. Pour les districts, $y_i = Y_i = M_i\,\overline{Y}_i$ et $x_i = M_i$ : $Y_i/\pi_i = k\,\overline{Y}_i$, avec $k = N/n_I$. Alors

$$\frac{\widehat{Y}_{HT}}{N} = \frac{1}{N} \cdot \frac{N}{n_I} \sum_{i \in S} \overline{Y}_i = \frac{1}{n_I} \sum_{i \in S} \overline{Y}_i,$$

la moyenne simple des moyennes de district, comme avec remise. L'estimateur est sans biais : $E(\widehat{Y}_{HT}) = \sum_i \pi_i \, \dfrac{y_i}{\pi_i} = Y$.

**Ce que cela dit.** Sur la mini-population ($n = 2$, $k = 9$), les 9 départs donnent cinq échantillons : AD, BD, BE, CE et DE, de probabilités 2/9, 1/9, 2/9, 3/9 et 1/9 ; chaque ménage a bien $\pi_i = 2\,x_i/18$. Sur les districts ($k = 400$), chacun reçoit un départ pour $M_i$ des 400 départs, chaque départ donne cinq districts distincts, et la moyenne des 400 estimations vaut exactement $\overline{Y} = 7\,911{,}44$. Le départ 341 donne 7 475,0 dh.

■




## L'estimation et sa précision

$\frac{1}{5}\,(10\,953{,}29 + 5\,828{,}14 + 6\,146{,}35 + 7\,688{,}91 + 6\,758{,}33) = 7\,475{,}0$ dh, soit 436 dh de moins que $\overline{Y}$. Sur les 400 départs possibles, l'écart-type exact vaut 668,3 dh, contre 754,3 avec remise : chaque province reçoit sa part de points, une **stratification implicite** (chapitre 8).

## Deux ménages, pas de 9, ordre A à E

$\pi_i = 2\,x_i/18$ : $\frac{2}{9}$, $\frac{1}{3}$, $\frac{1}{3}$, $\frac{4}{9}$, $\frac{2}{3}$ ; $y_i/\pi_i$ : 18, 18, 21, 20,25, 21. Les 9 départs donnent cinq échantillons :

| Échantillon | AD | BD | BE | CE | DE |
|---|---|---|---|---|---|
| Probabilité | 2/9 | 1/9 | 2/9 | 3/9 | 1/9 |
| $\widehat{Y}_{HT}$ | 38,25 | 38,25 | 39 | 42 | 41,25 |

## Sans remise, un peu mieux

$E(\widehat{Y}_{HT}) = 360/9 = 40 = Y$ et $V(\widehat{Y}_{HT}) = 2{,}75$, contre 3,25 avec remise : aucun tirage n'est perdu à reprendre un ménage déjà tiré. Mais AB, AC, AE, BC et CD ne sont **jamais** tirés ensemble : $\pi_{kl} = 0$ pour cinq paires sur dix.

## La forme du chapitre 2, pour un plan de taille fixe

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

Mini-population : les dix paires, multipliées par 81, donnent 54 (AC), −50,625 (AD), 108 (AE), 81 (BC), 15,1875 (BD), 6,75 (CD), 8,4375 (DE), 0 pour les autres. Total : 222,75/81 = 2,75, la variance de l'énumération.

**Pas d'estimation sans biais**

$\widehat{V}_{SYG}$ divise chaque paire par $\pi_{kl}$ : il exige $\pi_{kl} > 0$. Or 1 028 paires de districts sur 1 225 (84 %) ne sont jamais tirées ensemble. On calcule alors la variance **comme avec remise**, $s^2/n_I$ sur les $\overline{Y}_i$ tirés : erreur-type 925,6 dh ici, $\sqrt{599\,512} = 774{,}3$ dh en moyenne sur les 400 départs, pour un écart-type vrai de 668,3. C'est **prudent**.

**Démonstration : Des couples jamais tirés ensemble, et la variance prudente**

*Pourquoi* $\pi_{kl} = 0$. Deux unités sont tirées ensemble seulement si un même départ place un point dans chacune : il faut deux entiers $a \in ]C_{k-1}\,;C_k]$ et $b \in ]C_{l-1}\,;C_l]$ dont l'écart $b-a$ est un multiple de $k$. Si tous les écarts possibles tombent strictement entre deux multiples consécutifs de $k$, la paire n'est jamais tirée. Sur la mini-population ($k = 9$), A occupe 1-2, B 3-5, C 6-8, D 9-12, E 13-18 : les écarts entre A et B vont de 1 à 4, entre 0 et 9 ; aucun n'est un multiple de 9, donc $\pi_{AB} = 0$. Même chose pour AC (4 à 7), AE (11 à 17), BC (1 à 5) et CD (1 à 6) ; les écarts entre A et D, de 7 à 11, atteignent 9 : $\pi_{AD} > 0$.

*La variance exacte.* Pour un plan de taille fixe, la forme de Sen-Yates-Grundy est

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

En développant, $V(\widehat{Y}_{HT}) = \sum_k \frac{1-\pi_k}{\pi_k}\, y_k^2 + \sum_{k\neq l} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\,\pi_l}\, y_k\, y_l$. Pour un couple de $\pi_{kl} = 0$, le produit $y_k\, y_l$ a pour coefficient $-1$ : un estimateur sans biais devrait le contenir. Mais aucun échantillon ne contient à la fois $k$ et $l$ : **le produit n'est jamais observé**, et aucun estimateur ne peut être sans biais pour la variance. L'estimateur de Sen-Yates-Grundy divise par $\pi_{kl}$ : il n'existe pas.

*Que faire.* On remplace le plan par un tirage avec remise de même probabilité $p_i = \pi_i/n$ et l'on estime $s_z^2/n$, ici $s^2/n_I$ sur les $\overline{Y}_i$ tirés. L'estimateur est sans biais pour la variance du tirage **avec** remise. Ici elle dépasse celle du plan systématique, qui stratifie implicitement : l'erreur est du côté de la prudence. Ce n'est pas une règle générale, qui dépend de l'ordre de la liste.

**Ce que cela dit.** Mini-population : les dix paires, multipliées par 81, donnent 54 (AC), −50,625 (AD), 108 (AE), 81 (BC), 15,1875 (BD), 6,75 (CD), 8,4375 (DE) et 0 pour AB, BE et CE : leur somme est 222,75, d'où $V(\widehat{Y}_{HT}) = 222{,}75/81 = 2{,}75$ ; l'énumération des cinq échantillons le confirme, et $E(\widehat{Y}_{HT}) = 40$. Pour les districts, 1 028 des 1 225 paires ne sont jamais tirées ensemble ; la variance prudente vaut $774{,}3^2 = 599\,512$ en moyenne sur les 400 départs, pour un écart-type exact de 668,3 dh.

■

### Le plan

1. Tirer $n_I$ districts proportionnellement à leur taille : $\pi_i = n_I\, M_i/N$.
2. Dans chaque district tiré, tirer $m$ ménages par sondage aléatoire simple, **le même nombre partout**.

Un ménage du district $i$ est tiré avec la probabilité

$$\pi_k = \frac{n_I\, M_i}{N} \times \frac{m}{M_i} = \frac{n_I\, m}{N}$$

la même pour tous : le plan est **autopondéré**.

### Vingt districts de dix ménages

$\pi_i = 20\, M_i/2\,000 = M_i/100$, de 0,13 à 0,71 ; $\pi_k = 200/2\,000 = 0{,}1$ et $d_k = 10$ pour chacun des 200 ménages. L'estimateur est la simple moyenne $\overline{y}$ des 200 ménages.

### Trois avantages pratiques

1. **Une charge fixe par district** : chaque équipe réalise $m$ entretiens, quel que soit le district.
2. **Une taille d'échantillon fixée** à $n_I\, m$ ménages, donc un budget connu.
3. **Des poids égaux** : moyennes et proportions se calculent sans pondération.

### Ce que le chapitre 10 ne réunissait pas

À probabilités égales, $m$ ménages par district donnent des poids $d_k = N_I\, M_i/(n_I\, m)$ qui varient de 3,25 à 17,75 ; des $m_i$ proportionnels à $M_i$ rendent le plan autopondéré, mais la charge devient inégale. Le tirage selon la taille réunit les deux avantages.

### Districts tirés avec remise

Chaque tirage donne la moyenne $\overline{y}_{(j)}$ de $m$ ménages d'un district ; l'estimateur est $\overline{y} = \frac{1}{n_I}\sum_j \overline{y}_{(j)}$, de variance

$$V(\overline{y}) = \frac{1}{n_I}\Big[\underbrace{\sum_i \frac{M_i}{N}\big(\overline{Y}_i - \overline{Y}\big)^2}_{\text{entre districts}} + \underbrace{\sum_i \frac{M_i}{N}\Big(1 - \frac{m}{M_i}\Big)\frac{S_i^2}{m}}_{\text{dans les districts}}\Big]$$

où $S_i^2$ est la variance des dépenses dans le district $i$.

**Une estimation qui voit les deux degrés**

Les $n_I$ moyennes $\overline{y}_{(j)}$ sont indépendantes et de même loi, d'espérance $\sum_i \frac{M_i}{N}\overline{Y}_i = \overline{Y}$. Donc $\widehat{V}(\overline{y}) = s_b^2/n_I$, où $s_b^2$ est la variance de ces $n_I$ moyennes : les deux degrés y sont, sans calculer un seul $S_i^2$.

**Démonstration : Le plan autopondéré à deux degrés et sa variance**

*La probabilité d'inclusion.* Un ménage $k$ du district $i$ est tiré si son district l'est, puis s'il est tiré dans son district :

$$\pi_k = \frac{n_I\, M_i}{N} \times \frac{m}{M_i} = \frac{n_I\, m}{N}.$$

Le poids $d_k = N/(n_I\, m)$ est le même pour tous : $\widehat{Y} = d\sum_{k\in S} y_k$ et l'estimateur de la moyenne est la moyenne simple $\overline{y}$ des $n_I\, m$ ménages.

*La variance.* Au tirage $j$, on tire le district $i$ avec la probabilité $p_i = M_i/N$, puis $m$ ménages par sondage aléatoire simple : on obtient la moyenne $\overline{y}_{(j)}$. Les $n_I$ tirages sont indépendants et de même loi, donc $V(\overline{y}) = V(\overline{y}_{(j)})/n_I$. La loi de la variance totale (chapitre 10) donne

$$V(\overline{y}_{(j)}) = V\big[E(\overline{y}_{(j)} \mid i)\big] + E\big[V(\overline{y}_{(j)} \mid i)\big].$$

Sachant $i$, $\overline{y}_{(j)}$ est la moyenne d'un sondage aléatoire simple de $m$ ménages : $E(\overline{y}_{(j)} \mid i) = \overline{Y}_i$ et $V(\overline{y}_{(j)} \mid i) = (1 - \frac{m}{M_i})\dfrac{S_i^2}{m}$. Avec la loi $p_i$,

$$V(\overline{y}_{(j)}) = \underbrace{\sum_i \frac{M_i}{N}\big(\overline{Y}_i - \overline{Y}\big)^2}_{\text{entre districts}} + \underbrace{\sum_i \frac{M_i}{N}\Big(1 - \frac{m}{M_i}\Big)\frac{S_i^2}{m}}_{\text{dans les districts}}, \qquad \text{car } E(\overline{y}_{(j)}) = \sum_i \frac{M_i}{N}\overline{Y}_i = \overline{Y}.$$

*L'estimation.* Les $n_I$ moyennes $\overline{y}_{(j)}$ sont indépendantes, d'espérance $\overline{Y}$ et de variance $V(\overline{y}_{(j)})$ : par le même calcul que pour Hansen-Hurwitz, $s_b^2/n_I$ est sans biais pour $V(\overline{y})$, les **deux** termes inclus.

**Ce que cela dit.** Sur trois districts de 3, 2 et 4 ménages et $m = 2$, l'énumération exacte des tirages donne $V(\overline{y}_{(j)}) = 13{,}85 = 9{,}47 + 4{,}38$ ; pour deux tirages, l'énumération des tirages et des échantillons confirme $E(s_b^2/n_I) = V(\overline{y})$. Sur la population du cours ($m = 10$) : $\sum_i \frac{M_i}{N}(1 - \frac{10}{M_i})S_i^2 = 10\,105\,597$, d'où, pour $n_I = 20$, $2\,844\,644/20 = 142\,232$ entre les districts et $10\,105\,597/200 = 50\,528$ dans les districts : variance 192 760, écart-type 439,0 dh. Les districts tirés à probabilités égales donnent 383 368 entre districts, 2,7 fois plus. Avec 25 districts de 8 ménages, $113\,786 + 53\,768 = 167\,554$.

■

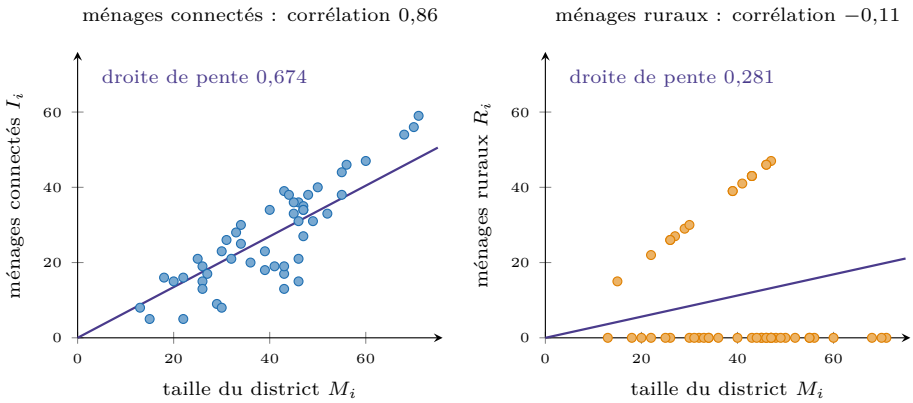
## Deux plans de 200 ménages

| Districts tirés | Entre | Dans | Écart-type | *deff* |
|---|---|---|---|---|
| Probabilités égales | 383 368 | 59 169 | 665,2 | 6,19 |
| Selon leur taille | 142 232 | 50 528 | 439,0 | 2,70 |

Variance de $\overline{y}$ **entre** districts et **dans** les districts. PPT : $2\,844\,644/20 = 142\,232$ et $10\,105\,597/200 = 50\,528$, où $10\,105\,597 = \sum_i \frac{M_i}{N}(1 - \frac{10}{M_i})S_i^2$.

## Le premier degré reste le plus lourd

Le tirage selon la taille divise par 2,7 la variance entre districts et ramène l'effet de plan de 6,19 à 2,70. Le premier degré fait encore 74 % de la variance. Pour le plan $10 \times 20$ du chapitre 10 ($deff = 14{,}9$), l'effet de plan tomberait à 4,5 environ.




## Le total estimé avec cinq districts

Ménages connectés : coefficient de variation de 11,6 % en PPT, contre 20,8 % à probabilités égales. Ménages ruraux : 71,5 % en PPT, contre 65,9 % : le tirage selon la taille fait **moins bien**.

## La mini-population, tailles inversées

Avec les tailles 6, 4, 3, 3, 2 pour A à E : $z_i = y_i/p_i = 12, 27, 42, 54, 126$, et $V(z) = 20\,772/18 = 1\,154$. Avec les vraies tailles, $V(z) = 6{,}5$ ; à probabilités égales, 290, soit quatre fois mieux.

**Une petite probabilité sur une grande valeur**

$V(z) = \sum_i p_i\,(y_i/p_i - Y)^2$ explose quand une unité a une **grande valeur** $y_i$ et une **petite probabilité** $p_i$ : E, qui dépense le plus, n'a plus que $p_E = 2/18$ et compte, s'il est tiré, pour 126 au lieu de 40. Les ruraux sont un cas voisin : $R_i/p_i$ vaut $N$ pour un district rural, 0 pour un district urbain.

**Démonstration : Une mesure de taille mal choisie fait exploser la variance**

*Ce qui fait la variance.* Pour le tirage avec remise, $V(z) = \sum_i y_i^2/p_i - Y^2$ : le terme de l'unité $i$ est grand quand sa valeur $y_i$ est grande et sa probabilité $p_i$ petite.

*Le minimum.* Par l'inégalité de Cauchy-Schwarz,

$$Y^2 = \Big(\sum_i \frac{y_i}{\sqrt{p_i}}\cdot\sqrt{p_i}\Big)^2 \le \Big(\sum_i \frac{y_i^2}{p_i}\Big)\Big(\sum_i p_i\Big) = \sum_i \frac{y_i^2}{p_i},$$

donc $V(z) \ge 0$, avec égalité si et seulement si $y_i/\sqrt{p_i}$ est proportionnel à $\sqrt{p_i}$, c'est-à-dire $p_i \propto y_i$. La variance est nulle pour des probabilités **proportionnelles aux valeurs**, inconnues avant l'enquête ; la mesure de taille $x_i$ est un substitut, d'autant meilleur que $y_i/x_i$ varie peu.

*Un exemple à rebours.* Avec les tailles 6, 4, 3, 3, 2 au lieu de 2, 3, 3, 4, 6, les probabilités sont $p_i = 6/18, \dots, 2/18$ et $z_i = y_i/p_i = 12$, 27, 42, 54, 126. Alors

$$V(z) = \sum_i p_i\,(z_i - 40)^2 = \frac{20\,772}{18} = 1\,154 = \sum_i \frac{y_i^2}{p_i} - 40^2.$$

**Ce que cela dit.** Trois mesures de taille pour la même population : $V(z) = 6{,}5$ avec les vraies tailles, 290 à probabilités égales, 1 154 à rebours, soit quatre fois plus que sans information. Le ménage E, qui dépense le plus, n'a plus que $p_E = 2/18$ : s'il sort, il compte pour 126 au lieu de 40.

■

**Un district qui a quadruplé**

Un district compté 13 ménages au recensement en abrite aujourd'hui 52, après la construction d'un lotissement. Tiré avec $p_i = 13/2\,000$, il fournit à lui seul $z_j/2\,000 = 52\,\overline{Y}_i/13 = 4\,\overline{Y}_i$. L'estimateur reste sans biais, car $p_i$ est la vraie probabilité de tirage, mais sa variance s'envole ; en deux degrés, ses $m$ ménages devraient peser **quatre fois** plus que les autres : le plan n'est plus autopondéré.

**Bien choisir la mesure de taille**

1. **Liée à la variable principale** : $y_i/x_i$ doit varier peu d'une unité à l'autre.
2. **À jour** : recensement récent, liste actualisée, ou dénombrement des districts tirés avant l'enquête.
3. **Adaptée à toutes les variables** : une taille en ménages convient aux totaux qui

croissent avec la taille ; pour un groupe concentré dans certains districts, comme les ruraux, mieux vaut **stratifier** (chapitre 6).

### Hansen-Hurwitz sur quatre villages

Quatre villages comptent 10, 20, 30 et 40 ménages ; leurs dépenses totales sont 60, 140, 180 et 320 milliers de dh. On tire deux villages avec remise, proportionnellement au nombre de ménages.

1. Calculer les $p_i$, les $z_i = Y_i/p_i$, et vérifier que $E(z) = Y$.
2. Calculer $V(\widehat{Y}_{HH})$. Comparer à deux tirages avec remise à probabilités égales.
3. On tire les villages 2 et 4. Calculer $\widehat{Y}_{HH}$, son erreur-type estimée, et la dépense moyenne par ménage estimée.

### Probabilités et variance

1. $p_i = 0{,}1$ ; $0{,}2$ ; $0{,}3$ ; $0{,}4$ et $z_i = 600, 700, 600, 800$. $E(z) = 60 + 140 + 180 + 320 = 700 = Y$.
2. $V(z) = 0{,}1 \times 100^2 + 0 + 0{,}3 \times 100^2 + 0{,}4 \times 100^2 = 8\,000$, d'où $V(\widehat{Y}_{HH}) = 4\,000$. À probabilités égales, $z_i = 4\,Y_i = 240, 560, 720, 1\,280$ : $V(z) = 142\,000$ et $V = 71\,000$, soit 17,75 fois plus.

### Une estimation

3. $\widehat{Y}_{HH} = (700 + 800)/2 = 750$ milliers de dh ; $\widehat{V} = \frac{1}{2\times 1}(50^2 + 50^2) = 2\,500$, erreur-type 50. Par ménage : $750/100 = 7{,}5$ milliers de dh, pour une vraie valeur de 7.

### Un tirage systématique proportionnel à la taille

Huit douars comptent, dans l'ordre de la liste, 12, 30, 18, 25, 40, 15, 35 et 25 ménages. On en tire 4 par tirage systématique proportionnel à la taille, et l'on enquête tous leurs ménages.

1. Calculer les tailles cumulées, le pas $k$ et les $\pi_i$. Tous les $\pi_i$ sont-ils inférieurs à 1 ?
2. Avec le départ $r = 37$, quels douars sont tirés ?
3. Leurs dépenses moyennes sont 6,2 ; 7,5 ; 5,8 et 6,9 milliers de dh. Estimer la dépense moyenne et le total.
4. Peut-on estimer la variance sans biais ? Que faire ?

### Le tirage

1. Cumuls : 12, 42, 60, 85, 125, 140, 175, 200 ; $k = 200/4 = 50$ ; $\pi_i = 4\,M_i/200 = M_i/50$, de 0,24 à 0,8 : tous les douars ont $M_i \leq 50$.
2. Points 37, 87, 137, 187 : douars 2 (]12 ; 42]), 5 (]85 ; 125]), 6 (]125 ; 140]) et 8 (]175 ; 200]).

### L'estimation

3. $Y_i/\pi_i = M_i\overline{Y}_i/(M_i/50) = 50\,\overline{Y}_i$ : $\widehat{Y}_{HT} = 50 \times 26{,}4 = 1\,320$ milliers de dh, et la moyenne vaut $1\,320/200 = 6{,}6$, la moyenne simple des quatre $\overline{Y}_i$.
4. Non : sept paires, dont les douars 1 et 2, ne sont jamais tirées ensemble ($\pi_{kl} = 0$). On calcule comme avec remise : $s^2 = 1{,}7/3 = 0{,}5667$, $\widehat{V} = 0{,}5667/4 = 0{,}1417$, erreur-type 0,376 millier de dh, une valeur prudente.

### Un plan autopondéré sur la population du cours

On tire 25 districts par tirage systématique proportionnel à la taille, puis 8 ménages dans chacun ; la variance se calcule comme si les districts étaient tirés avec remise.

1. Calculer $\pi_k$ et le poids des ménages, puis $\pi_i$ pour les districts de 71 et de 13 ménages.
2. Avec $\sum_i \frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2 = 2\,844\,644$ et $\sum_i \frac{M_i}{N}(1 - \frac{8}{M_i})S_i^2 = 10\,753\,679$, calculer $V(\overline{y})$, l'écart-type et l'effet de plan ($V_{SAS} = 71\,435{,}5$).
3. Pourquoi ne pourrait-on pas tirer ainsi 40 districts de 5 ménages ?

### Probabilités et variance

1. $\pi_k = 25 \times 8/2\,000 = 0{,}1$ et $d_k = 10$ pour tous. $\pi_i = 25\,M_i/2\,000 = M_i/80$ : 0,8875 et 0,1625.
2. $2\,844\,644/25 = 113\,786$ et $10\,753\,679/(25 \times 8) = 53\,768$ : $V = 167\,554$, écart-type 409,3 dh, *deff* $= 167\,554/71\,435{,}5 = 2{,}35$, contre 4,45 si les districts sont tirés à probabilités égales.

### Les grands districts

3. $\pi_i = 40\,M_i/2\,000 = M_i/50$ dépasserait 1 pour les 8 districts de plus de 50 ménages. Il faudrait les prendre d'office, et y tirer $M_i/10$ ménages pour garder $\pi_k = 0{,}1$.

**Sept points**

1. Tirer **proportionnellement à une taille** : $p_i = x_i/X$, par les tailles cumulées.
2. Hansen-Hurwitz : $\widehat{Y}_{HH} = \frac{1}{n}\sum_j y_{(j)}/p_{(j)}$, sans biais, et $\widehat{V} = s_z^2/n$.
3. Sa variance est faible quand $y_i/x_i$ varie peu, nulle si $y_i$ est proportionnel à $x_i$.
4. Cinq districts selon leur taille : écart-type 754,3 dh, contre 1 516,6 à probabilités égales.
5. Sans remise : $\pi_i = n\,x_i/X$ et Horvitz-Thompson ; des $\pi_{kl}$ nuls, une variance prudente.
6. Districts selon leur taille, puis $m$ ménages : un plan **autopondéré** à charge fixe.
7. Une taille mal liée à $y$, ou périmée, peut faire **pire** que les probabilités égales.

**La suite**

Le chapitre 12 se sert de l'information auxiliaire pour **estimer**, non plus pour tirer : connaissant le nombre total de personnes, $X = 7\,859$, l'estimateur par le ratio corrige la moyenne de l'échantillon.