Échantillonnage

Chapitre 12 : L'estimateur par le ratio

Jaouad Madkour

30 octobre 2026

Plan du chapitre

L'information auxiliaire **2**
Ce que l'on sait avant l'enquête . . . 2
La dépense suit la taille . . . 2
L'estimateur par le ratio . . . 3
Sur la mini-population . . . 3
Dix échantillons, dix estimations . . . 3

Le biais **4**
Pourquoi le ratio est biaisé . . . 4
Un biais d'ordre 1/n . . . 5
Négligeable devant l'erreur-type . . . 6

La précision **6**
La variance approchée . . . 6
Sur la population du cours . . . 7
Quand le ratio gagne-t-il ? . . . 7
Le gain selon la corrélation . . . 8

L'échantillon fil rouge **9**
L'estimation . . . 9
Le ratio a trop corrigé . . . 9
La précision estimée sur l'échantillon . . . 9
En moyenne, le ratio fait mieux . . . 10
Juger une méthode, pas un résultat . . . 10

Ratios et domaines **11**
La dépense par personne . . . 11
Un ratio n'est pas une moyenne de ratios . . . 11
La moyenne d'un domaine est un ratio . . . 11
Quand la taille de l'échantillon est aléatoire . . . 12

Bien utiliser le ratio **14**
Ratio, strates ou régression . . . 14
Une variable auxiliaire mal choisie . . . 14

Quand l'employer 14

S'entraîner **15**
Exercice 1 : quatre exploitations 15
Exercice 1 : solution 15
Exercice 2 : la masse salariale 15
Exercice 2 : solution 16
Exercice 3 : choisir la variable auxiliaire 16
Exercice 3 : solution 16

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

L'information auxiliaire

Ce que l'on sait avant l'enquête

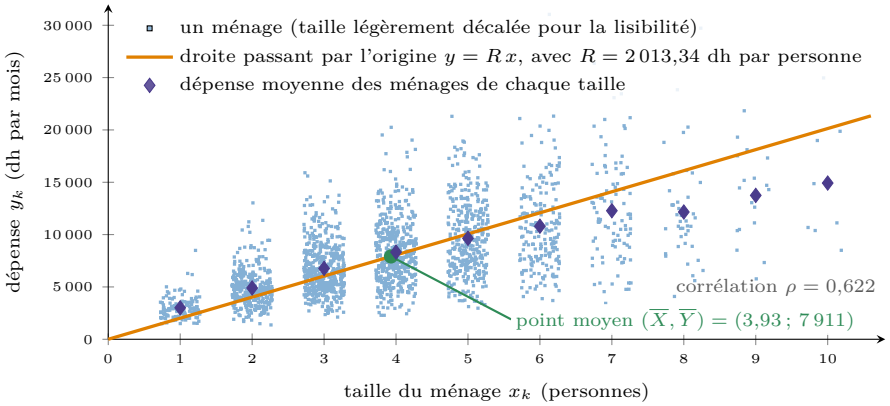
Une variable auxiliaire

Une **variable auxiliaire** x est mesurée sur chaque unité tirée, et son **total** $X = \sum_{k \in U} x_k$ est connu pour toute la population, par un recensement ou un fichier administratif. On note $\overline{X} = X/N$ sa moyenne, $\overline{x}$ sa moyenne dans l'échantillon.

Le nombre de personnes

Le recensement donne le nombre de personnes de la population du cours : $X = 7\,859$ pour $N = 2\,000$ ménages, soit $\overline{X} = 3{,}93$ personnes par ménage (le RGPH 2024 en compte 3,9 pour tout le pays). L'enquête demande la taille de chaque ménage tiré : l'échantillon fil rouge en compte $\overline{x} = 3{,}805$ en moyenne.

La dépense suit la taille



Lire le nuage

Les grands ménages dépensent plus : $\rho = 0{,}622$. La droite issue de l'origine, de pente $R = Y/X = 2\,013{,}34$ dh par personne, résume ce lien. Des ménages tirés **plus petits** que la moyenne donnent une dépense moyenne **trop basse**.

L'estimateur par le ratio

Corriger la moyenne par la taille

On estime le ratio $R = Y/X$ par $\widehat{R} = \overline{y}/\overline{x}$, puis on le multiplie par le total connu :

$$\widehat{Y}_R = X\,\widehat{R} = X\,\frac{\overline{y}}{\overline{x}}, \qquad \widehat{\overline{Y}}_R = \overline{X}\,\frac{\overline{y}}{\overline{x}} = \overline{y} \times \frac{\overline{X}}{\overline{x}}$$

Un facteur de correction

Si les ménages tirés sont plus petits que l'ensemble ($\overline{x} < \overline{X}$), le facteur $\overline{X}/\overline{x}$ dépasse 1 et **relève** la moyenne ; s'ils sont plus grands, il l'**abaisse**. Si $\overline{x} = \overline{X}$, l'estimateur vaut $\overline{y}$. Il ne demande qu'une chose de plus que $\overline{y}$: le total X.

Sur la mini-population

Dépense et taille des cinq ménages, sondage de deux ménages

Tailles 2, 3, 3, 4, 6 : $X = 18$, $\overline{X} = 3{,}6$, $R = 40/18 = 2{,}22$. Estimateur : $3{,}6\,\overline{y}/\overline{x}$.

Échantillon	AB	AC	AD	AE	BC
$\overline{y}$	5	5,5	6,5	9	6,5
$\overline{x}$	2,5	2,5	3	4	3
Ratio	7,2	7,92	7,8	8,1	7,8

Échantillon	BD	BE	CD	CE	DE
$\overline{y}$	7,5	10	8	10,5	11,5
$\overline{x}$	3,5	4,5	3,5	4,5	5
Ratio	7,71	8	8,23	8,4	8,28

Dix échantillons, dix estimations

Les deux estimateurs sur les dix échantillons

	Moyenne $\overline{y}$	Ratio
Espérance	8	7,944
Biais	0	−0,056
Variance	4,35	0,108
Valeurs extrêmes	5 et 11,5	7,2 et 8,4

Un petit biais, une variance divisée par quarante

La dépense est presque proportionnelle à la taille : y_k/x_k ne varie que de 2 à 2,33 milliers de dh par personne. Le ratio efface l'essentiel des écarts entre échantillons : la variance tombe de 4,35 à 0,108. Il n'est plus exactement sans biais, mais son erreur quadratique moyenne,

0,111, reste 39 fois plus faible que celle de $\overline{y}$.

Le biais

Pourquoi le ratio est biaisé

L'espérance d'un quotient

$\overline{y}$ et $\overline{x}$ sont sans biais, mais $E(\overline{y}/\overline{x}) \neq E(\overline{y})/E(\overline{x})$: l'espérance d'un quotient n'est pas le quotient des espérances. Sur la mini-population, l'estimateur vaut 7,944 en moyenne au lieu de 8.

L'erreur s'écrit avec des résidus

Posons $e_k = y_k - R\,x_k$: l'écart du ménage k à la droite de pente R. Leur moyenne dans la population est **nulle**, car $\sum_{k\in U} e_k = Y - R\,X = 0$. Comme $R = \overline{Y}/\overline{X}$:

$$\widehat{\overline{Y}}_R - \overline{Y} = \frac{\overline{X}}{\overline{x}}\,(\overline{y} - R\,\overline{x}) = \frac{\overline{X}}{\overline{x}}\,\overline{e}$$

où $\overline{e}$ est la moyenne des e_k des unités tirées.

Démonstration : L'erreur du ratio s'écrit avec les résidus

Les résidus ont une moyenne nulle. Avec $e_k = y_k - R\,x_k$ et $R = Y/X$, la somme sur la population est $\sum_{k\in U} e_k = Y - R\,X = 0$.

L'identité exacte. Dans l'échantillon, $\overline{e} = \overline{y} - R\,\overline{x}$, donc $\overline{y} = R\,\overline{x} + \overline{e}$. En reportant dans l'estimateur et en utilisant $R\,\overline{X} = \overline{Y}$,

$$\widehat{\overline{Y}}_R = \overline{X}\,\frac{\overline{y}}{\overline{x}} = \overline{X}\,\frac{R\,\overline{x} + \overline{e}}{\overline{x}} = R\,\overline{X} + \frac{\overline{X}}{\overline{x}}\,\overline{e} = \overline{Y} + \frac{\overline{X}}{\overline{x}}\,\overline{e}.$$

L'erreur du ratio est donc **exactement** $\dfrac{\overline{X}}{\overline{x}}\,\overline{e}$, sans approximation.

Ce que l'on en tire. Si $y_k = R\,x_k$ pour tous les ménages, tous les e_k sont nuls et le ratio vaut $\overline{Y}$ sur **tout** échantillon. En général, l'erreur est la moyenne d'échantillon $\overline{e}$ d'une variable de moyenne nulle, multipliée par un facteur proche de 1. Et le ratio est biaisé : $E(\overline{e}) = 0$, mais le facteur $\overline{X}/\overline{x}$ est aléatoire et lié à $\overline{e}$, donc $E(\frac{\overline{X}}{\overline{x}}\,\overline{e}) \neq E(\frac{\overline{X}}{\overline{x}})\,E(\overline{e}) = 0$ en général.

Ce que cela dit. Sur la mini-population, $R = 40/18$ et les résidus valent $-\frac{4}{9}$, $-\frac{2}{3}$, $\frac{1}{3}$, $\frac{1}{9}$, $\frac{2}{3}$, de somme nulle ; l'identité tient sur chacun des dix échantillons de deux ménages. Le ratio y a pour espérance 7,944 (biais $-0{,}056$), pour variance 0,108 contre 4,35 pour $\overline{y}$, et son erreur quadratique 0,111 est 39 fois plus faible. Sur le fil rouge, $\overline{y} - \overline{Y} = R\,(\overline{x} - \overline{X}) + \overline{e} = -250{,}7 + 182{,}3 = -68{,}4$, tandis que l'erreur du ratio est $1{,}0327 \times 182{,}28 = 188{,}24$: il retire la première erreur et garde la seconde.

■

Un biais d'ordre 1/n

Le biais approché

$$B(\widehat{\overline{Y}}_R) \approx \frac{1-f}{n} \cdot \frac{R\,S_x^2 - S_{xy}}{\overline{X}} = \overline{Y}\,\frac{1-f}{n}\left(CV_x^2 - \rho\,CV_x\,CV_y\right)$$

S_{xy} est la covariance de x et y dans la population (en $N-1$), $CV_x = S_x/\overline{X}$ et $CV_y = S/\overline{Y}$ leurs coefficients de variation.

La démonstration courte

Pour $\overline{x}$ proche de $\overline{X}$, $\overline{X}/\overline{x} \approx 1 - (\overline{x} - \overline{X})/\overline{X}$, d'où

$$\widehat{\overline{Y}}_R - \overline{Y} \approx \overline{e} - \frac{\overline{e}\,(\overline{x} - \overline{X})}{\overline{X}}$$

Le premier terme est d'espérance nulle (sondage aléatoire simple des e_k, de moyenne 0). Le second a pour espérance $-\mathrm{Cov}(\overline{e},\overline{x})/\overline{X} = -(1-f)\,S_{ex}/(n\,\overline{X})$, avec $S_{ex} = S_{xy} - R\,S_x^2$.

Démonstration : Le biais du ratio, d'ordre 1/n

La covariance de deux moyennes. Pour deux variables a et b et un sondage aléatoire simple, $\mathrm{Cov}(\overline{a},\overline{b}) = (1-f)\,S_{ab}/n$. En effet, la moyenne de $a+b$ a pour variance $(1-f)\,S_{a+b}^2/n$ (chapitre 3), avec $S_{a+b}^2 = S_a^2 + S_b^2 + 2\,S_{ab}$; or $V(\overline{a}+\overline{b}) = V(\overline{a}) + V(\overline{b}) + 2\,\mathrm{Cov}(\overline{a},\overline{b})$. En retranchant $V(\overline{a})$ et $V(\overline{b})$, il reste la covariance.

Le développement. Avec l'identité du résultat précédent, $\dfrac{\overline{X}}{\overline{x}} = 1 - \dfrac{\overline{x} - \overline{X}}{\overline{x}}$, donc exactement

$$\widehat{\overline{Y}}_R - \overline{Y} = \overline{e} - \overline{e}\,\frac{\overline{x} - \overline{X}}{\overline{x}}.$$

Pour n grand, $\overline{x} - \overline{X}$ est de l'ordre de $1/\sqrt{n}$; remplacer $\overline{x}$ par $\overline{X}$ au dénominateur change le second terme d'un ordre $1/n^{3/2}$, négligeable.

L'espérance. Le premier terme est d'espérance nulle. Dans le second, $E(\overline{x}) = \overline{X}$, donc $E\left[\overline{e}\,(\overline{x} - \overline{X})\right] = \mathrm{Cov}(\overline{e},\overline{x}) = (1-f)\,S_{ex}/n$, avec $S_{ex} = S_{xy} - R\,S_x^2$. D'où

$$B \approx -\frac{1-f}{n} \cdot \frac{S_{ex}}{\overline{X}} = \frac{1-f}{n} \cdot \frac{R\,S_x^2 - S_{xy}}{\overline{X}} = \overline{Y}\,\frac{1-f}{n}\left(CV_x^2 - \rho\,CV_x\,CV_y\right),$$

car $R\,S_x^2/\overline{X} = \overline{Y}\,S_x^2/\overline{X}^2 = \overline{Y}\,CV_x^2$ et $S_{xy}/\overline{X} = \rho\,S_x\,S/\overline{X} = \overline{Y}\,\rho\,CV_x\,CV_y$.

Un biais qui s'efface plus vite que l'erreur-type. B varie comme $1/n$, l'erreur-type comme $1/\sqrt{n}$: leur rapport varie comme $1/\sqrt{n}$ et tend vers 0.

Ce que cela dit. Sur la mini-population, la formule donne $-0{,}053$ pour un biais exact de $-0{,}056$; l'énumération des échantillons de deux et de trois ménages confirme $\mathrm{Cov}(\overline{a},\overline{b}) = (1-f)\,S_{ab}/n$ pour les couples (y,x), (x,x), (y,y) et (e,x). Sur la population du cours, $CV_x^2 = 0{,}1989$ et $\rho\,CV_x\,CV_y = 0{,}1397$: pour $n = 200$, le biais relatif vaut $0{,}0045 \times 0{,}0592 = 0{,}027\,\%$, soit 2,11 dh, pour une erreur-type de 220,84 dh ; pour $n = 20$,

23,2 dh contre 732, soit 0,032 ; le carré de ce rapport, 0,1 %, est la part du biais dans l'erreur quadratique moyenne.

■

Négligeable devant l'erreur-type

Sur la population du cours

$CV_x^2 = 0{,}1989$ et $\rho\, CV_x\, CV_y = 0{,}1397$. Pour $n = 200$: biais relatif $(0{,}9/200) \times 0{,}0592 = 0{,}027\,\%$, soit 2,11 dh.

Taille de l'échantillon n	20	50	200	800
Biais (dh)	23,2	9,1	2,1	0,35
Erreur-type (dh)	732	460	221	90
Biais / erreur-type	0,032	0,020	0,010	0,004

Le biais s'efface plus vite que l'erreur-type

Le biais décroît comme $1/n$, l'erreur-type comme $1/\sqrt{n}$: leur rapport tend vers 0. Dès $n = 20$, le carré du biais fait 0,1 % de l'erreur quadratique moyenne. La formule est fiable même sur la mini-population : $-0{,}053$, pour un biais exact de $-0{,}056$.

La précision

La variance approchée

Variance de l'estimateur par le ratio

$$V(\widehat{\overline{Y}}_R) \approx \frac{1-f}{n}\, S_e^2, \qquad S_e^2 = \frac{1}{N-1}\sum_{k\in U} (y_k - R\, x_k)^2 = S^2 - 2R\, S_{xy} + R^2 S_x^2$$

Pour le total : $V(\widehat{Y}_R) \approx N^2\,(1-f)\, S_e^2/n$.

La démonstration courte

Au premier ordre, $\widehat{\overline{Y}}_R - \overline{Y} \approx \overline{e}$, la moyenne d'un sondage aléatoire simple de la variable e_k. Le chapitre 3 donne sa variance : $(1-f)\, S_e^2/n$. Le ratio se comporte comme $\overline{y}$, **avec les résidus e_k à la place des y_k**. Sur la mini-population, l'approximation donne 0,091 pour une variance exacte de 0,108 : grossière avec $n = 2$, mais du bon ordre de grandeur.

Démonstration : La variance du ratio, les résidus d'un sondage simple

Au premier ordre, un sondage simple des résidus. D'après le résultat précédent, $\widehat{\overline{Y}}_R - \overline{Y} \approx \overline{e}$: le terme omis est de l'ordre de $1/n$, devant $1/\sqrt{n}$. Or $\overline{e}$ est la moyenne d'un sondage aléatoire simple de la variable e_k, de moyenne nulle dans la population et de variance

$S_e^2 = \frac{1}{N-1}\sum_{k\in U} e_k^2$. Le chapitre 3 donne

$$V(\widehat{\overline{Y}}_R) \approx \frac{1-f}{n}\, S_e^2.$$

La variance des résidus. Comme $R\,\overline{X} = \overline{Y}$, on a $e_k = (y_k - \overline{Y}) - R\,(x_k - \overline{X})$. En élevant au carré et en sommant,

$$S_e^2 = S^2 - 2R\,S_{xy} + R^2\,S_x^2.$$

L'estimation. On remplace R par $\widehat{R}$ et la somme sur la population par la somme sur l'échantillon : $s_e^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \widehat{R}\,x_k)^2$, et $\widehat{V} = (1-f)\,s_e^2/n$.

Ce que cela dit. Sur la population du cours, $S = 3\,984{,}29$ mais $S_e = 3\,292{,}09$: $V \approx 0{,}0045 \times 3\,292{,}09^2 = 48\,770{,}3$, soit une erreur-type de 220,84 dh contre 267,27, un effet de plan de 0,683 et une taille effective de $200/0{,}683 \approx 293$. Sur la mini-population ($n = 2$), l'approximation donne 0,0907 pour une variance exacte de 0,1078 : grossière avec si peu d'unités, du bon ordre de grandeur.

■

Sur la population du cours

Moyenne simple et ratio, 200 ménages

	Moyenne simple	Ratio
Dispersion de la variable	$S = 3\,984{,}29$	$S_e = 3\,292{,}09$
Variance de l'estimateur	71 435,5	48 770,3
Erreur-type (dh)	267,27	220,84
Effet de plan	1	0,683

Ce que vaut l'information

L'erreur-type baisse de 17 %. Avec la taille du ménage, 200 ménages valent $200/0{,}683 = 293$ ménages interrogés sans elle : c'est la **taille effective** (chapitre 7). Il a suffi d'une question de plus dans l'enquête, et du total X fourni par le recensement.

Quand le ratio gagne-t-il ?

La condition de gain

Le ratio est plus précis que $\overline{y}$ si et seulement si $S_e^2 < S^2$, c'est-à-dire

$$\rho > \frac{CV_x}{2\,CV_y}$$

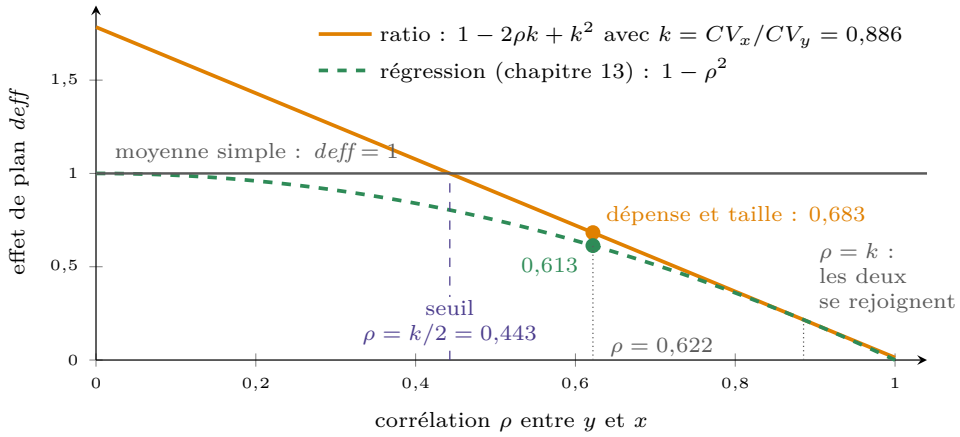
La démonstration

$S_e^2 < S^2 \iff R^2 S_x^2 < 2R\,S_{xy} = 2R\,\rho\,S_x\,S$. On divise par $2R\,S_x\,S > 0$: $\rho > \dfrac{R\,S_x}{2S} = \dfrac{S_x/\overline{X}}{2\,S/\overline{Y}}$, puisque $R = \overline{Y}/\overline{X}$.

Dépense et taille

$CV_x = 0{,}446$ et $CV_y = 0{,}504$: il faut $\rho > 0{,}443$. Avec $\rho = 0{,}622$, la condition est remplie.

Le gain selon la corrélation



Lire la figure

L'effet de plan du ratio vaut $S_e^2/S^2 = 1 - 2\rho k + k^2$, avec $k = CV_x/CV_y = 0{,}886$. Sous le seuil $\rho = 0{,}443$, le ratio fait **pire** que $\overline{y}$. La régression (chapitre 13) fait toujours au moins aussi bien ; les deux coïncident quand $\rho = k$.

Démonstration : L'effet de plan du ratio selon la corrélation

La réduction. Posons $k = CV_x/CV_y$. Comme $R = \overline{Y}/\overline{X}$, $R\,S_x = \overline{Y}\,CV_x$ et $S = \overline{Y}\,CV_y$, donc $R\,S_x/S = k$. Avec $S_{xy} = \rho\,S_x\,S$, on divise $S_e^2 = S^2 - 2R\,S_{xy} + R^2 S_x^2$ par S^2 :

$$deff = \frac{S_e^2}{S^2} = 1 - 2\rho\,\frac{R\,S_x}{S} + \left(\frac{R\,S_x}{S}\right)^2 = 1 - 2\rho\,k + k^2.$$

La condition de gain. $1 - 2\rho k + k^2 < 1 \iff k^2 < 2\rho k \iff \rho > k/2$, puisque $k > 0$.

Un carré parfait. $1 - 2\rho k + k^2 = (1 - \rho^2) + (\rho - k)^2$. Le premier terme est l'effet de plan de la régression (chapitre 13) ; le second est ≥ 0, et nul si et seulement si $\rho = k$: le ratio ne fait jamais mieux que la régression, et autant seulement si $\rho = k$.

Ce que cela dit. Dépense et taille : $k = 0{,}886$, $\rho = 0{,}622$, et $0{,}683 = 0{,}613 + 0{,}070$; le seuil $k/2 = 0{,}443$ est dépassé. Avec l'accès à internet comme variable auxiliaire, $\rho = 0{,}276$ ne dépasse pas le seuil de 0,691 : effet de plan 2,145 ; avec le milieu urbain, $\rho = 0{,}208$ pour un seuil de 0,621 : 2,025. Un ratio mal choisi **double** la variance.

■

L'échantillon fil rouge

L'estimation

Les 200 ménages du chapitre 3

	Échantillon	Population
Taille moyenne (personnes)	3,805	3,9295
Dépense moyenne (dh)	7 843,05	7 911,44
Dépense par personne (dh)	2 061,25	2 013,34

L'enquêteur ne connaît, dans la seconde colonne, que $\overline{X} = 7\,859/2\,000$.

Des ménages tirés un peu petits

Les ménages tirés comptent 0,12 personne de moins que l'ensemble : le facteur $\overline{X}/\overline{x} = 1{,}0327$ relève la moyenne de 3,3 %.

$$\widehat{\overline{Y}}_R = \overline{X}\,\widehat{R} = 3{,}9295 \times 2\,061{,}25 = 8\,099{,}68 \text{ dh}$$

Total estimé : $X\,\widehat{R} = 16\,199\,351$ dh, pour un vrai total de 15 822 870 dh.

Le ratio a trop corrigé

Plus loin de la vérité que la moyenne simple

Erreur de $\overline{y}$: $7\,843{,}05 - 7\,911{,}44 = -68{,}39$ dh. Erreur du ratio : $8\,099{,}68 - 7\,911{,}44 = +188{,}24$ dh. **Sur cet échantillon, le ratio fait moins bien que la moyenne simple.**

Deux erreurs qui se compensaient

L'erreur de $\overline{y}$ se décompose : $\overline{y} - \overline{Y} = R\,(\overline{x} - \overline{X}) + \overline{e} = -250{,}7 + 182{,}3 = -68{,}4$.

- $R\,(\overline{x} - \overline{X}) = 2\,013{,}34 \times (-0{,}1245)$: les ménages tirés sont **trop petits** ;
- $\overline{e} = +182{,}3$: à taille égale, ils **dépensent plus** que la moyenne.

$\overline{y}$ cumule les deux erreurs, qui se compensent presque. Le ratio retire la première et garde la seconde : $\frac{\overline{X}}{\overline{x}}\,\overline{e} = 1{,}0327 \times 182{,}28 = 188{,}24$.

La précision estimée sur l'échantillon

Estimer la variance

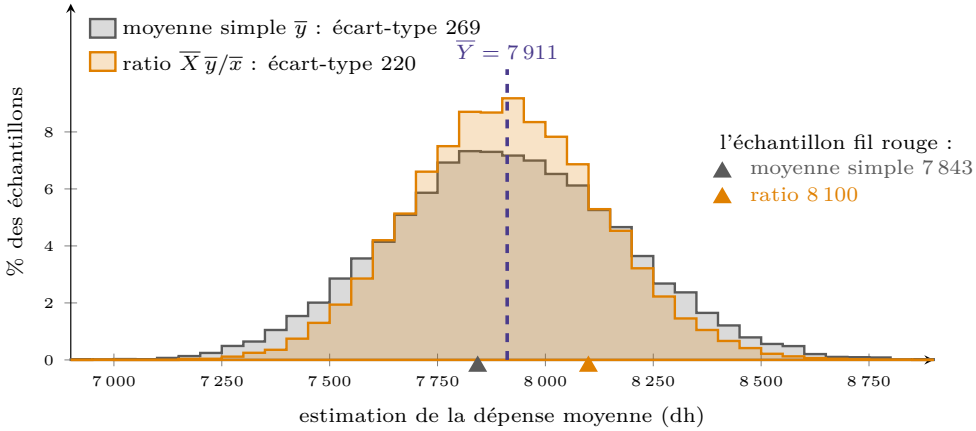
On remplace R par $\widehat{R}$ et S_e^2 par sa version d'échantillon :

$$\widehat{V}(\widehat{\overline{Y}}_R) = \frac{1-f}{n}\,s_e^2, \qquad s_e^2 = \frac{1}{n-1}\sum_{k\in S}\left(y_k - \widehat{R}\,x_k\right)^2 = s_y^2 - 2\widehat{R}\,s_{xy} + \widehat{R}^2 s_x^2$$

L'échantillon fil rouge

$s_y = 3\,829{,}16$, $s_x = 1{,}7812$, $s_{xy} = 4\,630{,}25$ (corrélation 0,679), d'où $s_e = 3\,009{,}13$ et une erreur-type estimée de $\sqrt{0{,}0045} \times 3\,009{,}13 = 201{,}86$ dh. Intervalle à 95 % : $8\,099{,}68 \pm 395{,}6$, soit $[7\,704{,}0\,;8\,495{,}3]$, qui contient $\overline{Y}$. La moyenne simple avait une erreur-type estimée de 256,87 dh.

En moyenne, le ratio fait mieux



Vingt mille sondages de 200 ménages

Les deux distributions sont centrées sur $\overline{Y}$: le biais du ratio, 2 dh, est invisible. Celle du ratio est plus **resserrée** : écart-type 220 dh, comme le prévoyait la formule (220,84), contre 269 pour $\overline{y}$. Le fil rouge tombe à gauche pour $\overline{y}$, à droite pour le ratio.

Juger une méthode, pas un résultat

Ce que disent les simulations

Sur 20 000 échantillons	Moyenne simple	Ratio
Erreur de moins de 300 dh	73,4 %	82,8 %
Racine de l'erreur quadratique moyenne	269,3	220,1
Couverture de l'intervalle à 95 %	94,6 %	94,7 %

Le ratio est le plus proche de $\overline{Y}$ dans 57,3 % des échantillons ; dans les 42,7 % restants, comme le fil rouge, c'est $\overline{y}$.

Choisir l'estimateur avant de voir le résultat

On choisit un estimateur pour sa **loi** sur tous les échantillons, pas pour sa valeur sur celui qu'on a tiré : $\overline{Y}$ étant inconnu, on ne sait jamais lequel est tombé le plus près. Retenir après coup celui qui « arrange » fausserait toutes les marges d'erreur.

Ratios et domaines

La dépense par personne

Un ratio comme paramètre

$R = Y/X$ s'estime par $\widehat{R} = \overline{y}/\overline{x}$, **sans** avoir besoin de X. Même linéarisation : $\widehat{R} - R = \overline{e}/\overline{x} \approx \overline{e}/\overline{X}$, d'où

$$V(\widehat{R}) \approx \frac{1-f}{n} \cdot \frac{S_e^2}{\overline{X}^2}, \qquad \widehat{V}(\widehat{R}) = \frac{1-f}{n} \cdot \frac{s_e^2}{\overline{x}^2}$$

Sur l'échantillon fil rouge

$\widehat{R} = 2\,061{,}25$ dh par personne, d'erreur-type estimée $201{,}86/3{,}805 = 53{,}05$ dh. Intervalle à 95 % : [1 957,3 ; 2 165,2], qui contient $R = 2\,013{,}34$. La vraie erreur-type vaut $220{,}84/3{,}9295 = 56{,}2$ dh.

Un ratio n'est pas une moyenne de ratios

Dépense par personne selon la taille du ménage

Taille du ménage	Ménages	Dépense moyenne	Par personne
1 personne	119	3 005	3 005
2 personnes	318	4 899	2 450
3 personnes	454	6 769	2 256
4 personnes	419	8 326	2 082
5 personnes	338	9 645	1 929
6 personnes et plus	352	11 606	1 719

Des économies d'échelle

Logement, équipement et cuisine se partagent : la dépense par personne baisse avec la taille. La moyenne des y_k/x_k des 2 000 ménages, 2 147,75 dh, compte autant un ménage d'une personne qu'un ménage de huit ; $R = 2\,013{,}34$ dh compte chaque **personne** une fois (chapitre 1).

La moyenne d'un domaine est un ratio

Un domaine, deux totaux

Avec $\delta_k = 1$ si le ménage est rural et 0 sinon, $\overline{y}_d = \sum_S \delta_k\, y_k \big/ \sum_S \delta_k$ est le ratio de la variable $\delta_k\, y_k$ à la variable δ_k. La linéarisation, avec $e_k = \delta_k\,(y_k - \overline{y}_d)$ et $\overline{x} = n_d/n$, donne

$$\widehat{V}(\overline{y}_d) = \frac{1-f}{n} \cdot \frac{s_e^2}{(n_d/n)^2}$$

Les ruraux de l'échantillon fil rouge

$n_d = 48$, $\overline{y}_d = 6\,916{,}67$ dh, erreur-type 458,49 dh. Le chapitre 4, à n_d fixé, donnait 465,9 dh ; en y remplaçant n_d/N_d par f, on obtient 462,18 dh. Le rapport de ces deux dernières variances, $(n_d - 1)\,n/[n_d\,(n-1)] = 0{,}984$, tend vers 1 : pour un domaine de quelques dizaines de ménages, les deux approches **coïncident**.

Démonstration : Un ratio, et la moyenne d'un domaine

Un ratio sans total connu. $\widehat{R} - R = \dfrac{\overline{y} - R\,\overline{x}}{\overline{x}} = \dfrac{\overline{e}}{\overline{x}} \approx \dfrac{\overline{e}}{\overline{X}}$: même linéarisation, d'où $V(\widehat{R}) \approx \dfrac{1-f}{n} \cdot \dfrac{S_e^2}{\overline{X}^2}$. On n'a pas besoin de X pour **estimer** R ; la variance, elle, a besoin de $\overline{x}$, que l'on met à la place de $\overline{X}$.

Un domaine est un ratio. Soit $\delta_k = 1$ si k est dans le domaine, 0 sinon. Alors $\overline{y}_d = \sum_S \delta_k\,y_k \big/ \sum_S \delta_k$ est le ratio de la variable $\delta_k\,y_k$ à la variable δ_k. Le ratio de population est $\overline{Y}_d$, les résidus sont $e_k = \delta_k\,y_k - \overline{Y}_d\,\delta_k = \delta_k\,(y_k - \overline{Y}_d)$, estimés par $\delta_k\,(y_k - \overline{y}_d)$, de somme nulle sur l'échantillon, et $\overline{x} = n_d/n$. D'où

$$\widehat{V}(\overline{y}_d) = \frac{1-f}{n} \cdot \frac{s_e^2}{(n_d/n)^2}.$$

Le lien avec le chapitre 4. Les résidus sont nuls hors du domaine : $s_e^2 = \dfrac{1}{n-1}\sum_{k\in S_d}(y_k - \overline{y}_d)^2 = \dfrac{n_d - 1}{n-1}\,s_d^2$. En remplaçant,

$$\widehat{V}(\overline{y}_d) = \frac{1-f}{n_d}\,s_d^2 \times \frac{(n_d-1)\,n}{n_d\,(n-1)}.$$

Le dernier facteur, $\dfrac{1 - 1/n_d}{1 - 1/n}$, tend vers 1 dès que le domaine compte quelques dizaines de ménages : on retrouve la formule d'un sondage simple de n_d ménages.

Ce que cela dit. Sur le fil rouge, $\widehat{R} = 2\,061{,}25$ dh par personne, d'erreur-type estimée $201{,}86/3{,}805 = 53{,}05$ dh. Pour les 48 ruraux, $\overline{y}_d = 6\,916{,}67$ et l'erreur-type est 458,49 dh ; le chapitre 4, avec f à la place de n_d/N_d, donne 462,18, et le rapport des variances vaut $47 \times 200/(48 \times 199) = 0{,}984$.

■

Quand la taille de l'échantillon est aléatoire

Le tirage de Bernoulli du chapitre 4

Avec $\pi = 0{,}1$, le tirage a donné $n = 202$ ménages : $\widehat{Y} = 16\,018\,000$ dh et $\widehat{N} = 2\,020$.

Estimateur de la moyenne	Valeur	Erreur-type
$\widehat{Y}/N$ (Horvitz-Thompson)	8 009,0	594,2
$\widehat{Y}/\widehat{N} = \overline{y}$	7 929,7	267,2

Le ratio absorbe la taille aléatoire

$\widehat{Y}$ et $\widehat{N}$ varient ensemble : un tirage qui ramène trop de ménages gonfle les deux. Leur rapport efface cet aléa et retrouve la précision d'un sondage aléatoire simple (267,3) : c'était l'annonce du chapitre 4. Même mécanisme pour les grappes de tailles inégales (chapitre 9) : effet de plan 32,2 pour Horvitz-Thompson, 8,4 pour le ratio.

Démonstration : Une taille d'échantillon aléatoire

Le tirage de Bernoulli. Chaque ménage est tiré avec la probabilité π, indépendamment des autres : les indicatrices I_k sont indépendantes, de moyenne π et de variance $\pi\,(1 - \pi)$, et la taille $n = \sum_k I_k$ est aléatoire, d'espérance $N\pi$.

Horvitz-Thompson. $\widehat{Y} = \frac{1}{\pi}\sum_k I_k\, y_k$ est sans biais, et par indépendance

$$V(\widehat{Y}) = \frac{1}{\pi^2}\sum_k y_k^2\,\pi\,(1-\pi) = \frac{1-\pi}{\pi}\sum_{k\in U} y_k^2.$$

Cette variance contient $\sum y_k^2$: elle porte le **niveau** de la dépense, et pas seulement sa dispersion.

Le ratio. La moyenne $\overline{y} = \widehat{Y}/\widehat{N}$, avec $\widehat{N} = n/\pi$, est le ratio de $\widehat{Y}$ à $\widehat{N}$. Sachant la taille $n \geq 1$, tous les échantillons de n ménages ont la même probabilité $\pi^n\,(1 - \pi)^{N-n}$: c'est un sondage aléatoire simple de taille n. Donc $E(\overline{y} \mid n) = \overline{Y}$ et $V(\overline{y} \mid n) = (\frac{1}{n} - \frac{1}{N})S^2$, et

$$V(\overline{y}) = E\Big(\frac{1}{n} - \frac{1}{N}\Big)S^2 \approx \Big(\frac{1}{N\pi} - \frac{1}{N}\Big)S^2 = (1-\pi)\,\frac{S^2}{N\pi} = (1-f)\,\frac{S^2}{\overline{n}},$$

avec $\overline{n} = N\pi$ et $f = \pi$: la variance d'un sondage aléatoire simple de taille $\overline{n}$. Elle ne contient que la dispersion S^2 : l'aléa de la taille a été absorbé.

Ce que cela dit. Pour $\pi = 0{,}1$ sur la population du cours, l'erreur-type exacte de $\widehat{Y}/N$ est 594,2 dh, celle de $\overline{y}$ est 267,95 dh, à 0,3 % de celle du sondage aléatoire simple (267,27 ; l'écart vient de $E(1/n)$, un peu supérieur à $1/E(n)$). Sur la mini-population avec $\pi = 2/5$, l'énumération des 32 sous-ensembles confirme $V(\widehat{Y}) = \frac{1-\pi}{\pi}\sum y_k^2$, et celle des 31 sous-ensembles non vides $E(\overline{y}) = 8 = \overline{Y}$ et $V(\overline{y}) = 5{,}43 = E(\frac{1}{n} - \frac{1}{5})S^2$.

■

Trois façons d'utiliser la taille, 200 ménages

Méthode	Effet de plan
Moyenne simple	1
Strates de taille (1–2, 3, 4, 5, 6+), proportionnelle	0,626
Estimateur par le ratio	0,683
Estimateur par la régression (chapitre 13)	0,613

Stratifier exige la taille de **chaque** ménage de la base ; le ratio, le seul total X.

Ce que coûte la droite par l'origine

$1-2\rho k+k^2 = (1-\rho^2)+(\rho-k)^2 = 0{,}613+0{,}070$. La régression ajuste la meilleure droite ; le ratio l'impose **par l'origine**. Or les petits ménages dépensent plus que $R\,x_k$, les grands moins.

Trois variables auxiliaires pour la dépense

Variable auxiliaire x	ρ	Seuil	Effet de plan
Taille du ménage	0,622	0,443	0,683
Accès à internet (0/1)	0,276	0,691	2,145
Milieu urbain (0/1)	0,208	0,621	2,025

Le seuil est $CV_x/(2\,CV_y)$; l'effet de plan, S_e^2/S^2.

Un ratio peut doubler la variance

Les ménages connectés dépensent plus, mais la dépense n'est pas **proportionnelle** à l'indicatrice : un ménage non connecté ($x_k = 0$) dépense quand même. Sous le seuil, le ratio fait pire que $\overline{y}$: ici, deux fois pire. Avant d'employer un ratio, on vérifie la condition sur une enquête antérieure.

Trois conditions

1. x_k est mesuré sur chaque unité tirée, et X est **connu** : recensement, fichier administratif.
2. y est à peu près **proportionnel** à x : un nuage allongé le long d'une droite qui passe près de l'origine.
3. $\rho > CV_x/(2\,CV_y)$, vérifié sur une enquête antérieure ou une enquête pilote.

Des usages courants en économie

- La production agricole et la **superficie** cultivée.
- Le chiffre d'affaires ou la masse salariale et l'**effectif** de l'entreprise.
- La valeur actuelle d'une variable et sa valeur au **dernier recensement** : le ratio estime alors une évolution.

Production et superficie

Quatre exploitations ont une superficie x_k de 2, 3, 5 et 6 ha, et une production y_k de 5, 8, 11 et 16 tonnes. On en tire deux par sondage aléatoire simple ; la superficie totale est connue.

1. Calculer $\overline{Y}$, $\overline{X}$ et R.
2. Pour chacun des six échantillons, calculer $\overline{y}$, $\overline{x}$ et l'estimateur par le ratio $\overline{X}\,\overline{y}/\overline{x}$.
3. Calculer l'espérance et la variance des deux estimateurs. Le ratio est-il sans biais ? Lequel préférer ?
4. Comparer la variance du ratio à l'approximation $(1-f)\,S_e^2/n$.

Les six échantillons

1. $\overline{Y} = 40/4 = 10$ t, $\overline{X} = 16/4 = 4$ ha, $R = 40/16 = 2{,}5$ t par ha.
2. L'estimateur vaut $4\,\overline{y}/\overline{x}$:

Exploitations	1, 2	1, 3	1, 4	2, 3	2, 4	3, 4
$\overline{y}$	6,5	8	10,5	9,5	12	13,5
$\overline{x}$	2,5	3,5	4	4	4,5	5,5
Ratio	10,4	9,14	10,5	9,5	10,67	9,82

Espérance et variance

3. $E(\overline{y}) = 10$, $V(\overline{y}) = 0{,}5 \times 22/2 = 5{,}5$. Ratio : $E = 60{,}03/6 = 10{,}005$ (biais $+0{,}005$), $V = 1{,}872/6 = 0{,}312$, 18 fois plus faible : on le préfère.
4. $e_k = y_k - 2{,}5\,x_k$ vaut 0 ; 0,5 ; $-1{,}5$; 1, d'où $S_e^2 = 3{,}5/3 = 1{,}167$ et $0{,}5 \times 1{,}167/2 = 0{,}292$, proche de 0,312.

Une zone industrielle

Une zone industrielle compte $N = 500$ entreprises qui emploient, selon un fichier administratif, $X = 12\,000$ salariés. Un sondage aléatoire simple de $n = 50$ entreprises donne, pour la masse salariale mensuelle y (milliers de dh) et l'effectif x :

$$\overline{y} = 110, \quad \overline{x} = 25, \quad s_y = 80, \quad s_x = 15, \quad r = 0{,}9$$

1. Estimer la masse salariale moyenne par $\overline{y}$, puis par le ratio. Pourquoi diffèrent-elles ?

2. Calculer l'erreur-type et l'intervalle à 95 % des deux estimations.
3. Estimer la masse salariale totale par le ratio, et l'effet de plan.

Les deux estimations

1. $\overline{X} = 24$; $\widehat{R} = 110/25 = 4{,}4$ milliers de dh par salarié, d'où $24 \times 4{,}4 = 105{,}6$. Les entreprises tirées sont un peu plus grandes que la moyenne (25 salariés contre 24) : $\overline{y}$ surestime, le ratio corrige.
2. Avec $s_{xy} = 0{,}9 \times 15 \times 80 = 1\,080$, on obtient $s_e^2 = 80^2 - 2 \times 4{,}4 \times 1\,080 + 4{,}4^2 \times 15^2 = 1\,252$. Avec $(1-f)/n = 0{,}018$: ratio, $\sqrt{0{,}018 \times 1\,252} = 4{,}75$, intervalle $[96{,}3\,;114{,}9]$; moyenne simple, $\sqrt{0{,}018 \times 6\,400} = 10{,}73$, intervalle $[89{,}0\,;131{,}0]$.

Le total

3. $X\,\widehat{R} = 12\,000 \times 4{,}4 = 52\,800$ milliers de dh, soit 52,8 millions de dh par mois, d'erreur-type $500 \times 4{,}75 = 2\,374$ milliers de dh. Effet de plan estimé : $22{,}54/115{,}2 = 0{,}196$. L'effectif explique l'essentiel des écarts de masse salariale.

Une enquête sur la production de blé

On veut estimer la production moyenne de blé de 15 000 exploitations ($CV_y = 0{,}8$). Un recensement agricole fournit trois variables auxiliaires :

Variable x	Corrélation ρ avec y	CV_x
Superficie totale	0,85	0,7
Superficie en céréales	0,95	0,9
Nombre d'actifs familiaux	0,4	0,5

1. Pour chaque variable, vérifier la condition de gain et calculer l'effet de plan du ratio.
2. Laquelle choisir ? Pour $n = 100$, quelle est la taille effective (on néglige f) ?
3. Comparer à l'effet de plan $1 - \rho^2$ de la régression.

Seuils et effets de plan

1. Avec $k = CV_x/CV_y$, seuil $k/2$ et $deff = 1 - 2\rho k + k^2$:

Variable	k	Seuil	*deff* du ratio
Superficie totale	0,875	0,4375	0,2781
Superficie en céréales	1,125	0,5625	0,1281
Actifs familiaux	0,625	0,3125	0,8906

Les trois corrélations dépassent leur seuil.

Le choix

2. La superficie en céréales. 100 exploitations valent alors $100/0{,}1281 \approx 780$ exploitations tirées sans information auxiliaire.
3. Régression : 0,2775 ; 0,0975 ; 0,84. Superficie totale : $\rho \approx k$, le ratio fait presque aussi bien. Ailleurs, l'écart $(\rho - k)^2$ pèse.

Sept points

1. Une variable auxiliaire de total X connu améliore l'estimation : $\widehat{Y}_R = X\,\overline{y}/\overline{x}$.
2. Le ratio est biaisé, mais d'ordre $1/n$: négligeable devant l'erreur-type.
3. $V \approx (1-f)\,S_e^2/n$ avec $e_k = y_k - R\,x_k$, estimée par $(1-f)\,s_e^2/n$.
4. Il gagne si $\rho > CV_x/(2\,CV_y)$; dépense et taille : effet de plan 0,683.
5. Sur le fil rouge, il fait moins bien que $\overline{y}$; en moyenne, mieux : on juge une **méthode**.
6. $\overline{y}/\overline{x}$ estime aussi un ratio (dépense par personne) et une moyenne de domaine.
7. Le ratio neutralise une taille d'échantillon aléatoire (Bernoulli, grappes).

La suite

Le chapitre 13 remplace la droite par l'origine par la meilleure droite : l'estimateur par la **régression**. Il combine ensuite plusieurs informations auxiliaires : post-stratification et **calage**.