Échantillonnage

Chapitre 3 : Le sondage aléatoire simple

Jaouad Madkour

30 octobre 2026

Plan du chapitre

Tirer au hasard, en pratique **2**
La définition . . . 2
Première méthode : tirer des numéros . . . 2
Deuxième méthode : trier selon une clé aléatoire . . . 3
La graine : un tirage que l'on peut refaire . . . 4

Probabilités d'inclusion et estimateur **4**
Les probabilités d'inclusion . . . 4
Un plan autopondéré . . . 5
La moyenne d'échantillon est sans biais . . . 5

La variance de la moyenne **5**
Le résultat . . . 5
Les ingrédients de la démonstration . . . 6
La démonstration . . . 6
Vérification sur la mini-population . . . 7
Le facteur de correction de population finie . . . 7

Estimer la précision **9**
Estimer la dispersion de la population . . . 9
Une estimation dispersée, mais juste en moyenne . . . 10
La variance estimée et l'intervalle de confiance . . . 10

La précision selon la taille de l'échantillon **10**
Le tableau de précision . . . 10
La loi en un sur racine de n . . . 11
Trois tailles d'échantillon . . . 12

L'échantillon fil rouge **12**
Deux cents ménages tirés au hasard . . . 12
Les estimations . . . 13
Ce que garantit la méthode . . . 13

S'entraîner **13**

Exercice 1 : les quatre boutiques, suite 13
Exercice 1 : solution 14
Exercice 2 : une enquête auprès des commerces 14
Exercice 2 : solution 14
Exercice 3 : un ménage sur cinq, un ménage sur cinq cents 14
Exercice 3 : solution 15

Ce qu'il faut retenir **15**
Le résumé du chapitre 15

Tirer au hasard, en pratique

La définition

Sondage aléatoire simple sans remise

Un plan de taille fixe n dans lequel **tous** les échantillons de n unités distinctes ont la même probabilité :

$$p(s) = \frac{1}{\binom{N}{n}} \text{ si } s \text{ compte } n \text{ unités}, \qquad p(s) = 0 \text{ sinon}$$

On l'abrège en **SAS**. C'est le plan de référence du cours : les autres plans lui seront comparés.

Deux ordres de grandeur

- Mini-population, $n = 2$: $\binom{5}{2} = 10$ échantillons, chacun de probabilité 1/10 (chapitre 1).
- Population du cours, $n = 200$: environ 10^{281} échantillons, chacun de probabilité 10^{-281} environ.

Personne ne les écrit : il faut une **procédure** de tirage qui réalise ce plan sans les énumérer.

Première méthode : tirer des numéros

Le tirage numéro par numéro

1. Numéroter les unités de la base de sondage de 1 à N.
2. Tirer un numéro au hasard entre 1 et N ; s'il est **déjà sorti**, l'ignorer.
3. Recommencer jusqu'à obtenir n numéros **distincts**.

Pour 200 ménages parmi 2 000, il faut tirer en moyenne 210,7 numéros : les doublons sont rares.

Pourquoi on obtient bien un SAS

À chaque étape, les unités non encore tirées ont toutes la même chance. Une suite ordonnée de n unités a donc la probabilité $\frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-n+1}$, et un échantillon peut sortir dans $n!$ ordres :

$$p(s) = n! \times \frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$$

Mini-population : $\{A, B\}$ sort par A puis B ou par B puis A, d'où $2 \times \frac{1}{5} \times \frac{1}{4} = \frac{1}{10}$.

Démonstration : Combien de numéros faut-il tirer ?

Appelons *tirage utile* un tirage qui donne un numéro **nouveau**. Supposons que i numéros distincts aient déjà été obtenus, avec $0 \leq i \leq n-1$. À chaque tirage, le numéro est uniforme entre 1 et N : il est nouveau avec la probabilité $p_i = \dfrac{N-i}{N}$, et les tirages sont indépendants. Le nombre de tirages nécessaires pour passer de i à $i+1$ numéros distincts suit donc une loi géométrique de paramètre p_i, d'espérance $1/p_i = \dfrac{N}{N-i}$.

Obtenir n numéros distincts demande de franchir successivement les paliers $i = 0, 1, \ldots, n-1$. Par linéarité de l'espérance, le nombre moyen de tirages est

$$E(T) = \sum_{i=0}^{n-1} \frac{N}{N-i} = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-n+1}\right).$$

Valeur pour le cours. Avec $N = 2\,000$ et $n = 200$, la somme compte 200 termes, tous compris entre $\frac{1}{2\,000}$ et $\frac{1}{1\,801}$: $E(T) = 210{,}7$ numéros.

Ce que cela dit. Il y a donc en moyenne 10,7 numéros tirés pour rien sur 210,7, soit 5 % : tant que n reste petit devant N, l'élimination des doublons est un détail. Elle devient lourde quand n approche de N : pour $n = N$, la somme est $N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right)$, de l'ordre de $N \ln N$ tirages.

■

Deuxième méthode : trier selon une clé aléatoire

Le tri aléatoire

1. Attribuer à chaque unité un nombre au hasard u_k entre 0 et 1 : sa **clé**.
2. Trier la base de sondage selon les clés, puis garder les n premières unités.

Tous les ordres des N unités sont équiprobables : les n premières forment un SAS.

Sur la mini-population, deux ménages

Ménage	A	B	C	D	E
Clé u_k	0,26	0,30	0,81	0,09	0,60
Rang	2	3	5	1	4

L'échantillon est $\{D, A\}$, de moyenne $\overline{y} = (9+4)/2 = 6{,}5$. La méthode convient bien à un fichier informatique : une colonne de clés, un tri, et aucun doublon à gérer.

Démonstration : Le tri aléatoire réalise un sondage aléatoire simple

Les clés $u_1, \ldots, u_N$ sont tirées indépendamment avec la même loi continue. Deux clés ne sont jamais égales (la probabilité d'une égalité est nulle), donc le tri est un **ordre strict** des N unités.

Tous les ordres sont équiprobables. Le vecteur des clés a la même loi quand on permute ses coordonnées (elles sont indépendantes et de même loi). Un ordre des unités est une permuta-

tion ; les $N!$ permutations ont donc la même probabilité, qui vaut $1/N!$.

Les n premières unités. Un échantillon donné s de n unités est l'ensemble des n premières si l'ordre commence par les éléments de s, dans un ordre quelconque ($n!$ possibilités), et se termine par les $N - n$ autres, dans un ordre quelconque ($(N - n)!$ possibilités). Il y a donc $n!\,(N-n)!$ ordres favorables et

$$p(s) = \frac{n!\,(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$$

C'est la définition du sondage aléatoire simple.

Ce que cela dit. Sur la mini-population, $n = 2$: $\frac{2!\,3!}{5!} = \frac{12}{120} = \frac{1}{10}$, la probabilité de chacun des dix échantillons. La méthode n'exige rien d'autre qu'un générateur de nombres aléatoires et un tri.

■

La graine : un tirage que l'on peut refaire

Un hasard reproductible

Un ordinateur produit des nombres **pseudo-aléatoires** : une suite entièrement déterminée par un nombre de départ, la **graine**. Même graine, même échantillon. L'échantillon fil rouge du cours est tiré ainsi en Python :

```
import numpy as np
rng = np.random.default_rng(2026)              # graine
s = rng.choice(2000, 200, replace=False) + 1   # 1 à 2000
```

Ménages tirés, dans l'ordre : 200, 769, 1 747, 1 821, 1 560, etc.

Pourquoi noter la graine

Une enquête doit pouvoir être **contrôlée** : avec la base de sondage et la graine, chacun retrouve les mêmes 200 ménages et vérifie qu'aucun n'a été choisi à la main. Attention : R emploie un autre générateur, la même graine y donne un autre échantillon (chapitre 15).

Probabilités d'inclusion et estimateur

Les probabilités d'inclusion

Un décompte d'échantillons

Parmi les $\binom{N}{n}$ échantillons, $\binom{N-1}{n-1}$ contiennent l'unité k, et $\binom{N-2}{n-2}$ contiennent à la fois k et l :

$$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} = f \qquad \pi_{kl} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

Mini-population, $n = 2$: $\pi_k = 4/10$ et $\pi_{kl} = 1/10$, comme au chapitre 1.

Pour l'échantillon de 200 ménages

$\pi_k = 200/2\,000 = 0{,}1$ et $\pi_{kl} = \dfrac{200 \times 199}{2\,000 \times 1\,999} = 0{,}009955$, à peine moins que $\pi_k \pi_l = 0{,}01$: $\Delta_{kl} = -0{,}000045$. Le poids de chaque ménage tiré est $d_k = 1/0{,}1 = 10$.

Un plan autopondéré

La moyenne simple est l'estimateur de Horvitz-Thompson

Tous les π_k valent n/N, donc

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

Chaque ménage tiré en représente dix

Tous les ménages ont le même poids $d_k = N/n$: le plan est **autopondéré**. Moyennes, proportions et tableaux se calculent sur l'échantillon comme sur n'importe quel fichier, sans poids. Le piège du chapitre 1, où il fallait repondérer chaque province, n'existe pas ici.

La moyenne d'échantillon est sans biais

La démonstration par les indicatrices

$\overline{y} = \frac{1}{n} \sum_{k \in U} I_k\, y_k$, où seules les I_k sont aléatoires, et $E(I_k) = \pi_k = n/N$:

$$E(\overline{y}) = \frac{1}{n} \sum_{k \in U} \frac{n}{N}\, y_k = \frac{1}{N} \sum_{k \in U} y_k = \overline{Y}$$

De même, $E(N\overline{y}) = Y$: le total estimé est sans biais. C'est un cas particulier du résultat général du chapitre 2.

Vérification sur la mini-population

Les dix moyennes possibles (chapitre 1) : 5 ; 5,5 ; 6,5 ; 9 ; 6,5 ; 7,5 ; 10 ; 8 ; 10,5 ; 11,5. Leur somme vaut 80, leur moyenne $8 = \overline{Y}$.

La variance de la moyenne

Le résultat

Variance de la moyenne d'un SAS

$$V(\overline{y}) = (1 - f)\,\frac{S^2}{n}, \qquad f = \frac{n}{N}, \qquad S^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \overline{Y})^2$$

Pour le total estimé : $V(N\overline{y}) = N^2 (1 - f)\, S^2/n$.

Trois leviers

- La **dispersion** S^2 : plus les ménages diffèrent, plus l'estimation varie d'un échantillon à l'autre.
- La **taille** n : la variance est divisée par n.
- Le **taux de sondage** f : le facteur $1-f$ réduit la variance quand on interroge une part notable de la population, et l'annule pour un recensement ($n = N$).

Les ingrédients de la démonstration

Ce que donne le chapitre 1

$\overline{y} = \frac{1}{n}\sum_{k\in U} I_k\, y_k$, avec, pour un SAS,

$$V(I_k) = f(1-f), \qquad \mathrm{Cov}(I_k, I_l) = -\frac{n(N-n)}{N^2(N-1)} = -\frac{f(1-f)}{N-1}$$

Une simplification : centrer les valeurs

Remplacer chaque y_k par $y_k \;-\; \overline{Y}$ décale $\overline{y}$ de la constante $\overline{Y}$ sans changer sa variance. On peut donc supposer $\sum_{k\in U} y_k = 0$. Alors

$$\sum_{k\neq l} y_k\, y_l = \Big(\sum_{k\in U} y_k\Big)^2 - \sum_{k\in U} y_k^2 = -\sum_{k\in U} y_k^2 \qquad \text{et} \qquad \sum_{k\in U} y_k^2 = (N-1)\, S^2$$

La démonstration

Trois lignes de calcul

$$\begin{aligned} V(\overline{y}) &= \frac{1}{n^2}\Big[\sum_k y_k^2\, V(I_k) + \sum_{k\neq l} y_k\, y_l\, \mathrm{Cov}(I_k, I_l)\Big] \\ &= \frac{f(1-f)}{n^2}\Big[\sum_k y_k^2 + \frac{1}{N-1}\sum_k y_k^2\Big] = \frac{f(1-f)}{n^2}\cdot\frac{N}{N-1}\sum_k y_k^2 \\ &= \frac{f(1-f)}{n^2}\, N\, S^2 = (1-f)\,\frac{S^2}{n} \qquad \text{car } fN = n \end{aligned}$$

Ce que la démonstration explique

- Le facteur $1-f$ vient de $V(I_k) = f(1-f)$: si $n = N$, chaque I_k vaut 1 et plus rien n'est aléatoire.
- Le résultat s'écrit **exactement** avec S^2, divisée par $N-1$: c'est la raison de la convention annoncée au chapitre 1.

Démonstration : La variance de la moyenne par les différences deux à deux

La démonstration précédente passe par les covariances des indicatrices. En voici une seconde, qui vérifie le résultat par un autre chemin et fait apparaître une écriture utile de S^2.

Une écriture de S^2. Développons $\sum_k \sum_l (y_k - y_l)^2$, où k et l parcourent chacun U :

$$\sum_k \sum_l (y_k - y_l)^2 = 2N \sum_k y_k^2 - 2\Big(\sum_k y_k\Big)^2 = 2N\Big(\sum_k y_k^2 - N\overline{Y}^2\Big) = 2N(N-1)\,S^2.$$

Application de la forme de Sen-Yates-Grundy. Le plan est de taille fixe, avec $a_k = y_k/\pi_k = \frac{N}{n}\,y_k$ et, d'après le chapitre 1,

$$\pi_k\pi_l - \pi_{kl} = \frac{n^2}{N^2} - \frac{n(n-1)}{N(N-1)} = \frac{n\,(N-n)}{N^2\,(N-1)}$$

pour toute paire $k \neq l$. Donc

$$V(N\overline{y}) = \frac{1}{2}\sum_k \sum_{l\neq k} \frac{n\,(N-n)}{N^2(N-1)} \cdot \frac{N^2}{n^2}\,(y_k - y_l)^2 = \frac{N-n}{2\,n\,(N-1)} \cdot 2N(N-1)\,S^2 = \frac{N\,(N-n)}{n}\,S^2.$$

(les termes $k = l$ sont nuls, on peut donc les rétablir dans la double somme). Divisons par N^2 :

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

Ce que cela dit. Les deux démonstrations donnent le même résultat, ce qui en confirme chaque étape. La seconde montre en plus que la variance mesure les **écarts entre unités** : S^2 est, à un facteur près, la moyenne des carrés des différences deux à deux.

■

Vérification sur la mini-population

Par énumération et par la formule

Écarts des dix moyennes à 8, au carré : 9 ; 6,25 ; 2,25 ; 1 ; 2,25 ; 0,25 ; 4 ; 0 ; 6,25 ; 12,25. Leur somme vaut 43,5, d'où

$$V(\overline{y}) = \frac{43{,}5}{10} = 4{,}35 \qquad \text{et} \qquad (1-f)\,\frac{S^2}{n} = 0{,}6 \times \frac{14{,}5}{2} = 4{,}35$$

Avec la variance divisée par N, le compte serait faux

$(1-f)\,\sigma^2/n = 0{,}6 \times 11{,}6/2 = 3{,}48 \neq 4{,}35$. Et $\sigma^2/n = 5{,}8$ ne convient pas non plus : c'est la variance d'un tirage **avec remise** (chapitre 4). Seule l'écriture avec S^2 donne une formule à la fois exacte et simple.

Le facteur de correction de population finie

Une autre écriture

$$V(\bar{y}) = (1-f)\,\frac{S^2}{n} = \frac{S^2}{n} - \frac{S^2}{N}$$

Le second terme ne dépend que de la population. Quand N est grand devant n, il disparaît : **seul n compte**. Le facteur $1-f$ est appelé **correction de population finie**.

Deux taux de sondage

- Population du cours, $n = 200$: $1 - f = 0{,}9$; erreur-type 267,27 dh au lieu de 281,73 sans correction, soit 5,1 % de moins.
- Enquête sur le niveau de vie 2022 : 18 000 ménages rapportés aux 9 275 038 du recensement de 2024, $f = 0{,}0019$ et $1 - f = 0{,}998$: la correction est négligeable.

La promesse de l'introduction

À n égal, un pays entier est estimé presque aussi précisément qu'une ville. Le facteur $1-f$ ne compte que si l'on interroge plus de quelques pour cent de la population.

Démonstration : La promesse de l'introduction, la marge ne dépend presque pas de N

Pour une proportion voisine de 0,5, la variance de la population est $S^2 = \frac{N}{N-1}P(1-P) \approx 0{,}25$ (chapitre 1). La marge à 95 % d'un sondage aléatoire simple de n unités est donc

$$e = 1{,}96\sqrt{(1-f)\,\frac{0{,}25}{n}} = \underbrace{1{,}96\sqrt{\frac{0{,}25}{n}}}_{\text{ne dépend pas de } N} \times\sqrt{1-\frac{n}{N}}.$$

Le premier facteur ne contient pas N ; le second est compris entre 0 et 1 et tend vers 1 quand N augmente. Pour $n = 1\,000$, le premier facteur vaut 0,0310, soit 3,1 points :

N	10 000	1 000 000	36 800 000
$\sqrt{1-n/N}$	0,949	0,9995	0,99999
marge e	2,9 points	3,1 points	3,1 points

Ce que cela dit. Multiplier la population par 3 680 ne fait gagner que deux dixièmes de point. Une ville et un pays sont connus avec la même précision par 1 000 personnes : c'est n qui commande, non la part de la population interrogée. La correction ne devient sensible que pour une petite population, comme les 2 000 ménages du cours ($\sqrt{0{,}5} = 0{,}707$, soit une marge de 2,2 points).

■

Estimer la précision

Estimer la dispersion de la population

La variance d'échantillon

$$s^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \overline{y})^2 \qquad \text{vérifie} \qquad E(s^2) = S^2$$

La justification courte

On décompose autour de $\overline{Y}$: $\sum_S (y_k - \overline{y})^2 = \sum_S (y_k - \overline{Y})^2 - n(\overline{y} - \overline{Y})^2$. En espérance :

- $E\big[\sum_S (y_k - \overline{Y})^2\big] = \sum_U \pi_k (y_k - \overline{Y})^2 = \frac{n}{N}(N-1)S^2$;
- $E\big[n(\overline{y} - \overline{Y})^2\big] = n\,V(\overline{y}) = (1-f)\,S^2 = \frac{N-n}{N}S^2$.

La différence vaut $\frac{nN-n-N+n}{N}S^2 = (n-1)S^2$: il suffit de diviser par $n-1$.

Démonstration : La variance d'échantillon est sans biais pour S²

La décomposition. Ajoutons et retranchons $\overline{Y}$ dans chaque écart : $y_k - \overline{y} = (y_k - \overline{Y}) - (\overline{y} - \overline{Y})$. En élevant au carré et en sommant sur les n unités de l'échantillon :

$$\sum_{k\in S}(y_k - \overline{y})^2 = \sum_{k\in S}(y_k - \overline{Y})^2 - 2(\overline{y} - \overline{Y})\sum_{k\in S}(y_k - \overline{Y}) + n(\overline{y} - \overline{Y})^2.$$

Or $\sum_{k\in S}(y_k - \overline{Y}) = n(\overline{y} - \overline{Y})$: le terme du milieu vaut $-2n(\overline{y} - \overline{Y})^2$, et il reste

$$\sum_{k\in S}(y_k - \overline{y})^2 = \sum_{k\in S}(y_k - \overline{Y})^2 - n(\overline{y} - \overline{Y})^2.$$

Le premier terme. On l'écrit avec les indicatrices, $\sum_{k\in S}(y_k - \overline{Y})^2 = \sum_{k\in U} I_k\,(y_k - \overline{Y})^2$, d'espérance $\frac{n}{N}\sum_{k\in U}(y_k - \overline{Y})^2 = \frac{n}{N}(N-1)S^2$.

Le second terme. $E\big[(\overline{y} - \overline{Y})^2\big] = V(\overline{y})$ car $\overline{y}$ est sans biais : $E\big[n(\overline{y} - \overline{Y})^2\big] = n(1-f)\dfrac{S^2}{n} = \dfrac{N-n}{N}S^2$.

La différence. $E\Big[\sum_{k\in S}(y_k - \overline{y})^2\Big] = \dfrac{n(N-1) - (N-n)}{N}S^2 = \dfrac{nN-N}{N}S^2 = (n-1)S^2$, donc $E(s^2) = S^2$.

La variance estimée. Par linéarité, $E\big[\widehat{V}(\overline{y})\big] = \dfrac{1-f}{n}E(s^2) = (1-f)\dfrac{S^2}{n} = V(\overline{y})$.

Ce que cela dit. Le diviseur $n-1$ compense le fait que $\overline{y}$, calculée sur le même échantillon, est plus proche des y_k que ne l'est $\overline{Y}$: c'est le terme $n(\overline{y} - \overline{Y})^2$. Sur la mini-population, les dix valeurs de s^2 ont pour moyenne $14{,}5 = S^2$, et celles de $\widehat{V}$, 4,35.

■

Une estimation dispersée, mais juste en moyenne

Les dix variances d'échantillon de la mini-population

Pour deux ménages, $s^2 = (y_1 - y_2)^2/2$. AB : 2 ; AC : 4,5 ; AD : 12,5 ; AE : 50 ; BC : 0,5 ; BD : 4,5 ; BE : 32 ; CD : 2 ; CE : 24,5 ; DE : 12,5.

$$E(s^2) = \frac{145}{10} = 14{,}5 = S^2$$

Juste en moyenne, très variable

D'un échantillon à l'autre, s^2 va de 0,5 à 50 : avec deux ménages, la dispersion est très mal connue. Mais elle est juste **en moyenne**, et $\widehat{V}(\overline{y}) = (1 - f)\, s^2/n = 0{,}3\, s^2$ aussi : ses dix valeurs ont pour moyenne 4,35 $= V(\overline{y})$.

La variance estimée et l'intervalle de confiance

Ce que l'on publie

$$\widehat{V}(\overline{y}) = (1 - f)\,\frac{s^2}{n} \qquad \overline{y} \pm 1{,}96\,\sqrt{\widehat{V}(\overline{y})}$$

La variance estimée est sans biais, puisque $E(s^2) = S^2$. Sa racine est l'**erreur-type estimée** ; l'intervalle est l'intervalle de confiance à 95 %.

Pourquoi 1,96

Pour n assez grand, et $N - n$ aussi, $\overline{y}$ suit approximativement une loi normale : c'est une version du théorème central limite (Probabilités, chapitre 20) adaptée au tirage sans remise. Les intervalles de confiance sont étudiés en Statistique inférentielle (chapitres 8 à 10) ; ce cours fournit la **bonne variance** à y mettre.

La précision selon la taille de l'échantillon

Le tableau de précision

Population du cours, dépense mensuelle

n	f	Erreur-type	Marge à 95 %	Sans correction
50	0,025	556,38	1 090,5	563,46
100	0,05	388,34	761,1	398,43
200	0,1	267,27	523,9	281,73
400	0,2	178,18	349,2	199,21
1 000	0,5	89,09	174,6	125,99

Erreur-type $\sqrt{(1 - f)\, S^2/n}$, $S = 3\,984{,}29$ dh ; marge : 1,96× erreur-type ; sans correction : $S/\sqrt{n}$.

Lire le tableau

De 50 à 200 ménages, l'erreur-type est divisée par 2,08 : **quatre fois** plus de ménages pour une erreur **deux fois** plus petite. À $n = 200$, la marge vaut 6,6 % de $\overline{Y}$.

Démonstration : Quatre fois plus de ménages pour une erreur deux fois plus petite

Écrivons la variance sous la forme $V(\overline{y}) = S^2\left(\frac{1}{n} - \frac{1}{N}\right)$ (voir plus bas). Pour un échantillon quatre fois plus grand,

$$\frac{V(\overline{y}_{4n})}{V(\overline{y}_n)} = \frac{\frac{1}{4n} - \frac{1}{N}}{\frac{1}{n} - \frac{1}{N}} = \frac{1}{4} \times \frac{1 - 4n/N}{1 - n/N}.$$

Sans correction de population finie (N très grand), le rapport est exactement $\frac{1}{4}$: l'erreur-type est divisée par 2. Avec la correction, le rapport est un peu **inférieur** à $\frac{1}{4}$ dès que $n/N > 0$, puisque $1 - 4f < 1 - f$: le gain est un peu plus que 2.

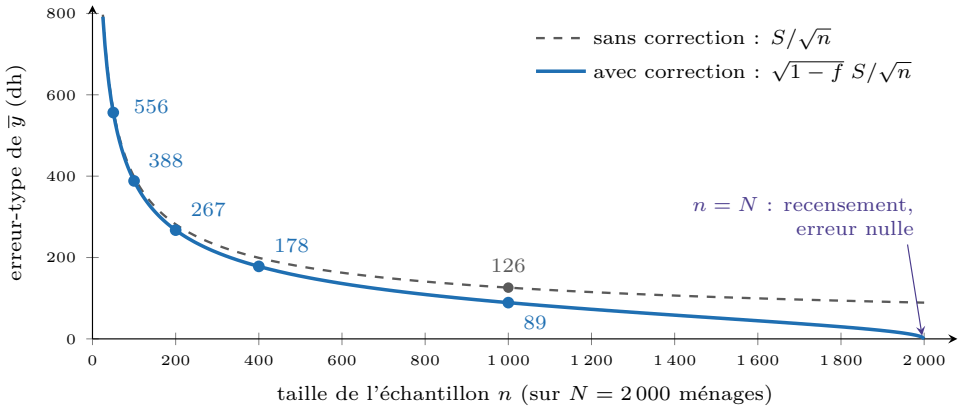
Pour le cours. De $n = 50$ à $n = 200$, f passe de 0,025 à 0,1 :

$$\frac{V_{200}}{V_{50}} = \frac{1}{4} \times \frac{0{,}9}{0{,}975} = 0{,}2308, \qquad \frac{\text{erreur-type}_{200}}{\text{erreur-type}_{50}} = \sqrt{0{,}2308} = 0{,}4804 = \frac{1}{2{,}08}.$$

Ce que cela dit. « Quatre fois plus pour deux fois mieux » est la règle des grands échantillons. Pour la population du cours, où f n'est pas négligeable, le gain est un peu meilleur (2,08 au lieu de 2) : le facteur $1 - f$ creuse l'écart d'autant plus que n est grand, et finit par annuler l'erreur quand $n = N$.

■

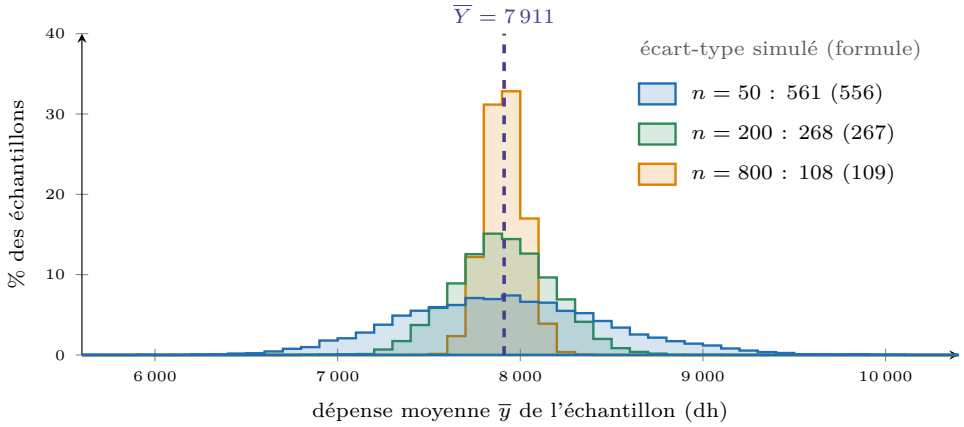
La loi en un sur racine de n



Une décroissance de plus en plus lente

Sans correction, l'erreur-type vaut $S/\sqrt{n}$: la diviser par 2 coûte 4 fois plus de questionnaires, la diviser par 10 en coûte 100 fois plus. Le facteur $\sqrt{1-f}$ ne creuse l'écart que lorsque n approche de N : 29,3 % de moins à $n = 1\,000$.

Trois tailles d'échantillon

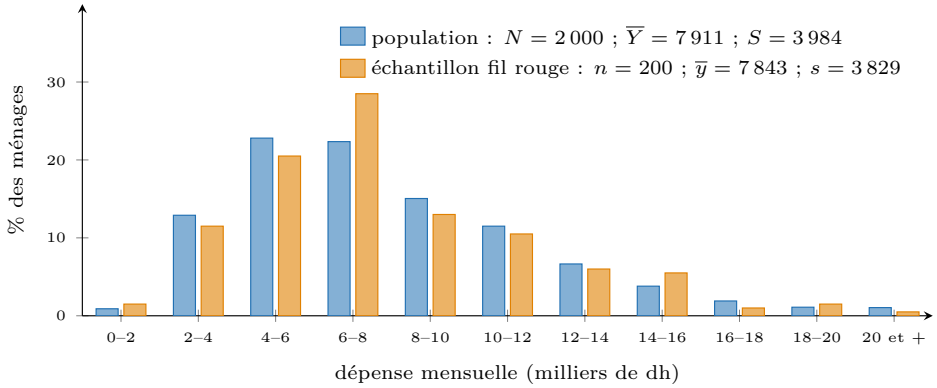


Dix mille sondages pour chaque taille

Les trois distributions sont **centrées** sur $\overline{Y}$ (moyennes simulées : 7 918 ; 7 912 ; 7 913) et leur écart-type suit la formule. Pour $n = 50$, la distribution garde un peu de l'**asymétrie** de la dépense ; dès $n = 200$, elle est très proche d'une loi normale.

L'échantillon fil rouge

Deux cents ménages tirés au hasard



Une population en miniature

L'échantillon reproduit la forme de la distribution, gros budgets compris. Il compte 152 urbains et 48 ruraux, pour 143,8 et 56,2 attendus : le hasard ne garantit pas une composition exacte, il la rend **probable**.

Les estimations

Ce que l'enquêteur calcule

Grandeur	Valeur
Moyenne $\overline{y}$	7 843,05 dh
Écart-type s	3 829,16 dh
Variance estimée $(1-f)\,s^2/n$	65 981,1
Erreur-type estimée	256,87 dh
Intervalle à 95 % : $\pm 503{,}46$	[7 339,6 ; 8 346,5]
Total estimé $N\,\overline{y}$	15 686 100 dh

Et la vérité

$\overline{Y} = 7\,911{,}44$ est dans l'intervalle : l'erreur, $-68{,}39$ dh, vaut 0,27 erreur-type. L'erreur-type estimée (256,87) est proche de la vraie (267,27) ; coefficient de variation estimé : 3,28 %. Le vrai total, 15 822 870, est dans [14 679 000 ; 16 693 000].

Ce que garantit la méthode

Cinq mille sondages de 200 ménages

On répète l'enquête 5 000 fois, chaque fois avec son intervalle $\overline{y} \pm 1{,}96\sqrt{\widehat{V}(\overline{y})}$: 94,24 % des intervalles contiennent $\overline{Y}$. Sur les 288 qui la manquent, 188 sont trop bas et 100 trop haut.

Un peu moins que promis, et plutôt du même côté

La dépense est asymétrique : un échantillon qui manque les gros budgets a **à la fois** une moyenne basse et un s faible. Son intervalle est trop bas **et** trop étroit. Avec 200 ménages l'approximation reste bonne ; elle se dégrade pour de petits échantillons.

La lecture correcte

« $\overline{Y}$ a 95 % de chances d'être dans [7 339,6 ; 8 346,5] » est **faux** : $\overline{Y}$ est fixe, et l'intervalle aussi une fois tiré. Ce qui vaut environ 95 %, c'est la part des échantillons dont l'intervalle contient $\overline{Y}$: la garantie porte sur la **méthode**.

S'entraîner

Exercice 1 : les quatre boutiques, suite

Vérifier les deux formules

On reprend les quatre boutiques du chapitre 1 (chiffres d'affaires 2, 5, 8 et 9 milliers de dh ; $\overline{Y} = 6$, $S^2 = 10$), dont on tire deux par SAS. Les six moyennes possibles sont 3,5 ; 5 ; 5,5 ; 6,5 ; 7 ; 8,5.

1. Calculer $V(\overline{y})$ par énumération, puis par la formule du chapitre.
2. Calculer s^2 pour chaque échantillon. Vérifier que $E(s^2) = S^2$.

3. En déduire l'espérance de $\widehat{V}(\overline{y})$. Que donnerait la formule sans le facteur $1-f$?

Exercice 1 : solution

La variance par les deux chemins

1. Écarts à 6, au carré : 6,25 ; 1 ; 0,25 ; 0,25 ; 1 ; 6,25, de somme 15 : $V(\overline{y}) = 15/6 = 2{,}5$. Formule : $(1-0{,}5)\times 10/2 = 2{,}5$.
2. $s^2 = (y_1-y_2)^2/2$: 4,5 ; 18 ; 24,5 ; 4,5 ; 8 ; 0,5. Somme 60, moyenne $10 = S^2$.

La variance estimée

3. $\widehat{V}(\overline{y}) = 0{,}5\times s^2/2 = s^2/4$, d'espérance $10/4 = 2{,}5 = V(\overline{y})$: elle est sans biais. Sans le facteur $1-f$, on obtiendrait $S^2/n = 5$, le **double** : on interroge ici la moitié de la population.

Exercice 2 : une enquête auprès des commerces

Le chiffre d'affaires des commerces

Une ville compte $N = 3\,000$ commerces. Un SAS de $n = 250$ commerces donne un chiffre d'affaires mensuel moyen $\overline{y} = 18{,}6$ milliers de dh, avec $s = 9{,}3$ (chiffres fictifs).

1. Calculer l'erreur-type estimée de $\overline{y}$ et l'intervalle de confiance à 95 %.
2. Estimer le chiffre d'affaires mensuel total des commerces, avec son intervalle.
3. Quelle erreur-type obtiendrait-on, avec le même s, pour $n = 1\,000$? Par combien est-elle divisée ?

Exercice 2 : solution

Moyenne et total

1. $f = 250/3\,000 = 0{,}0833$; $\widehat{V}(\overline{y}) = 0{,}9167\times 9{,}3^2/250 = 0{,}3171$; erreur-type 0,563. Intervalle : $18{,}6 \pm 1{,}104$, soit $[17{,}50\,;19{,}70]$ milliers de dh.
2. $\widehat{Y} = 3\,000\times 18{,}6 = 55\,800$ milliers de dh, soit 55,8 millions ; erreur-type $3\,000\times 0{,}563 = 1\,689$ milliers ; intervalle $[52{,}5\,;59{,}1]$ millions de dh.

Quatre fois plus de commerces

3. Avec $f = 1/3$: $\sqrt{0{,}6667\times 9{,}3^2/1\,000} = 0{,}240$. L'erreur-type est divisée par 2,35, un peu plus que $\sqrt{4} = 2$: le facteur $1-f$ passe de 0,917 à 0,667.

Exercice 3 : un ménage sur cinq, un ménage sur cinq cents

La taille de la population compte peu

On tire $n = 400$ ménages par SAS, d'une part dans la population du cours ($N = 2\,000$), d'autre part dans une ville de $N = 200\,000$ ménages dont la dépense a la même dispersion, $S = 3\,984{,}29$ dh.

1. Calculer l'erreur-type de $\overline{y}$ dans les deux cas.
2. Combien de ménages faudrait-il tirer dans la ville pour obtenir la même erreur-type

que dans la population du cours ?
3. Conclure.

Exercice 3 : solution

Deux populations

1. Population du cours : $f = 0{,}2$, erreur-type $3\,984{,}29 \times \sqrt{0{,}8/400} = 178{,}18$ dh. Ville : $f = 0{,}002$, erreur-type $3\,984{,}29 \times \sqrt{0{,}998/400} = 199{,}02$ dh.
2. La variance visée vaut $0{,}8\,S^2/400 = S^2/500$. Il faut donc $\frac{S^2}{n} - \frac{S^2}{200\,000} = \frac{S^2}{500}$, soit $\frac{1}{n} = \frac{1}{500} + \frac{1}{200\,000}$: $n = 498{,}75$, donc 499 ménages.

Ce qui compte

3. Une population **cent fois** plus grande demande 499 ménages au lieu de 400 : un quart de plus seulement. La précision dépend de n, et ne dépend de N que par le facteur $1 - f$.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Un **SAS** donne la même probabilité à tous les échantillons de taille n ; on le tire par numéros ou par clé aléatoire, en notant la **graine**.
2. $\pi_k = n/N$: le plan est autopondéré, et $\overline{y} = \widehat{Y}_{HT}/N$ est sans biais.
3. $V(\overline{y}) = (1 - f)\,S^2/n$, démontrée par les indicatrices I_k.
4. Si f est faible, $1 - f \approx 1$: la précision dépend de n, presque pas de N.
5. $E(s^2) = S^2$; $\widehat{V}(\overline{y}) = (1 - f)\,s^2/n$; intervalle $\overline{y} \pm 1{,}96\sqrt{\widehat{V}(\overline{y})}$.
6. L'erreur-type décroît en $1/\sqrt{n}$: deux fois plus précis coûte quatre fois plus cher.
7. Fil rouge : $\overline{y} = 7\,843{,}05$, intervalle $[7\,339{,}6\,;\,8\,346{,}5]$; la méthode couvre $\overline{Y}$ dans 94,2 % des tirages.

La suite

Le chapitre 4 traite les **proportions** et les **totaux**, et compare au tirage **avec remise**.