Échantillonnage

Chapitre 4 : Proportions, totaux et tirage avec remise

Jaouad Madkour

31 octobre 2026

Plan du chapitre

Estimer une proportion **2**
- Une proportion est une moyenne . . . 2
- La variance de la proportion estimée . . . 2
- Estimer cette variance sur l'échantillon . . . 3
- L'intervalle de confiance d'une proportion . . . 4
- La précision dépend de la proportion elle-même . . . 4

Estimer un total **5**
- Du total à la moyenne, et retour . . . 5
- Le nombre de ménages connectés . . . 5
- Même erreur absolue, autre erreur relative . . . 6
- Quand N est inconnu . . . 6

Les sous-populations **7**
- Estimer pour un domaine . . . 7
- Une taille d'échantillon aléatoire . . . 8
- La précision d'une moyenne de domaine . . . 8
- Le total d'un domaine . . . 9
- Les petits domaines . . . 10

Le tirage avec remise **10**
- Tirer avec remise . . . 10
- La variance d'un tirage avec remise . . . 11
- Avec ou sans remise . . . 12
- Combien de ménages distincts ? . . . 12
- Pourquoi tirer avec remise ? . . . 13

Tout estimer sur un échantillon **13**
- Six estimations sur l'échantillon fil rouge . . . 13
- Ce que montrent les intervalles . . . 14
- Les intervalles tiennent-ils leurs promesses ? . . . 14

S'entraîner **14**
- Exercice 1 : une proportion et un total . . . 14

Exercice 1 : solution 15
Exercice 2 : tirer avec remise 15
Exercice 2 : solution 15
Exercice 3 : un domaine 15
Exercice 3 : solution 16

Ce qu'il faut retenir **16**
Le résumé du chapitre 16

Estimer une proportion

Une proportion est une moyenne

L'estimateur d'une proportion

Avec $y_k = 1$ si le ménage k possède la caractéristique, 0 sinon, $P = \overline{Y}$ (chapitre 1). Un sondage aléatoire simple l'estime par la proportion observée :

$$\widehat{P} = \overline{y} = \frac{1}{n}\sum_{k\in S} y_k = \frac{\text{nombre de « oui » dans l'échantillon}}{n}$$

Moyenne d'échantillon, elle est **sans biais** (chapitre 3) : $E(\widehat{P}) = P$.

Sur la mini-population

Ménages dépensant plus de 8 : D et E, donc $P = 2/5 = 0{,}4$. Parmi les 10 échantillons de 2 ménages, AB, AC et BC donnent $\widehat{P} = 0$, DE donne $\widehat{P} = 1$, les 6 autres $\widehat{P} = 0{,}5$:

$$E(\widehat{P}) = (3 \times 0 + 6 \times 0{,}5 + 1 \times 1)/10 = 0{,}4 = P$$

La variance de la proportion estimée

Une variance qui ne dépend que de P

On remplace $S^2 = \frac{N}{N-1} P(1-P)$ (chapitre 1) dans $V(\overline{y}) = (1-f)\, S^2/n$ (chapitre 3) :

$$V(\widehat{P}) = (1-f)\,\frac{N}{N-1}\,\frac{P(1-P)}{n} = \frac{N-n}{N-1}\,\frac{P(1-P)}{n}$$

Vérification sur la mini-population

$S^2 = \frac{5}{4} \times 0{,}4 \times 0{,}6 = 0{,}3$, d'où $V(\widehat{P}) = (1-0{,}4) \times 0{,}3/2 = 0{,}09$. Par énumération des dix échantillons : $(3 \times 0{,}4^2 + 6 \times 0{,}1^2 + 1 \times 0{,}6^2)/10 = 0{,}9/10 = 0{,}09$.

Une loi déjà connue

Le nombre de « oui » $n\widehat{P}$ suit une loi **hypergéométrique** (cours de Probabilités), de variance $n\,P(1-P)\,\frac{N-n}{N-1}$. Avec des tirages indépendants, ce serait la binomiale, de variance $n\,P(1-P)$: le facteur $\frac{N-n}{N-1}$ est le gain du tirage sans remise.

Estimer cette variance sur l'échantillon

Remplacer P par son estimation

Pour une variable 0/1, $y_k^2 = y_k$, donc $\sum_{k\in S}(y_k - \widehat{P})^2 = n\widehat{P} - n\widehat{P}^2 = n\widehat{P}(1-\widehat{P})$, et $s^2 = \frac{n}{n-1}\widehat{P}(1-\widehat{P})$. D'où :

$$\widehat{V}(\widehat{P}) = (1-f)\,\frac{s^2}{n} = (1-f)\,\frac{\widehat{P}(1-\widehat{P})}{n-1}$$

Elle est sans biais, puisque s^2 l'est pour S^2 (chapitre 3).

Le piège du dénominateur

On rencontre souvent $\widehat{P}(1-\widehat{P})/n$. Pour $n = 200$, l'écart est négligeable : l'erreur-type est multipliée par $\sqrt{200/199} = 1{,}0025$. Pour $n = 20$, il atteint 2,6 %. L'oubli plus grave est celui de $(1-f)$ quand le taux de sondage est élevé.

Démonstration : Pourquoi n-1 au dénominateur pour une proportion

La variance d'échantillon d'une variable 0/1. Dans l'échantillon, $y_k^2 = y_k$ et $\sum_{k\in S} y_k = n\widehat{P}$. En développant le carré comme au chapitre 1,

$$\sum_{k\in S}(y_k - \widehat{P})^2 = \sum_{k\in S} y_k^2 - n\widehat{P}^2 = n\widehat{P} - n\widehat{P}^2 = n\widehat{P}\,(1-\widehat{P}),$$

donc $s^2 = \dfrac{n}{n-1}\,\widehat{P}(1-\widehat{P})$ et $\widehat{V}(\widehat{P}) = (1-f)\,\dfrac{s^2}{n} = (1-f)\,\dfrac{\widehat{P}(1-\widehat{P})}{n-1}$.

Vérification directe de l'absence de biais. Calculons l'espérance de $\widehat{P}(1-\widehat{P})$, avec $E(\widehat{P}^2) = V(\widehat{P}) + P^2$:

$$E\big[\widehat{P}(1-\widehat{P})\big] = P - V(\widehat{P}) - P^2 = P(1-P)\left[1 - \frac{N-n}{n\,(N-1)}\right] = P(1-P)\,\frac{N\,(n-1)}{n\,(N-1)} = \frac{n-1}{n}\,S^2.$$

Le produit $\widehat{P}(1-\widehat{P})$ sous-estime donc S^2 du facteur $\frac{n-1}{n}$: diviser par $n-1$ au lieu de n le corrige exactement, et $E\big[\widehat{V}(\widehat{P})\big] = (1-f)\,\frac{S^2}{n} = V(\widehat{P})$.

Ce que cela dit. Le piège du dénominateur est un biais : $\widehat{P}(1-\widehat{P})/n$ donne une variance trop petite d'un facteur $\frac{n-1}{n}$, donc une erreur-type trop petite d'un facteur $\sqrt{\frac{n-1}{n}}$. Pour $n = 200$, c'est $\sqrt{200/199} = 1{,}0025$; pour $n = 20$, $\sqrt{20/19} = 1{,}026$, soit 2,6 % d'erreur-type en moins.

■

L'intervalle de confiance d'une proportion

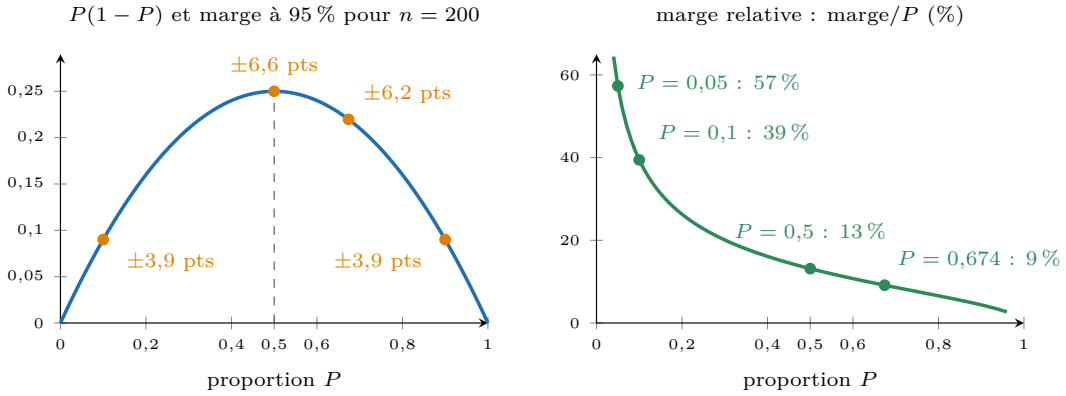
L'approximation normale

Pour n grand, $\widehat{P}$ est approximativement normale (théorème central limite, Probabilités, chapitre 20) : l'intervalle $\widehat{P} \pm 1{,}96\sqrt{\widehat{V}(\widehat{P})}$ contient P pour environ 95 % des échantillons, s'il y a assez de « oui » et de « non » dans l'échantillon (Statistique inférentielle, chapitres 8 à 10).

Les ménages connectés, n = 200

- **Sur tous les échantillons** : $V(\widehat{P}) = 0{,}9 \times \frac{2\,000}{1\,999} \times \frac{0{,}674 \times 0{,}326}{200} = 0{,}000989$; erreur-type 0,0315, marge ±0,0616, soit ±6,2 points.
- **Sur l'échantillon fil rouge** : 134 connectés sur 200, $\widehat{P} = 0{,}67$; $\widehat{V} = 0{,}9 \times \frac{0{,}67 \times 0{,}33}{199} = 0{,}00100$, erreur-type 0,0316.
- **Intervalle** : $0{,}67 \pm 0{,}062$, soit $[0{,}608\,;0{,}732]$, qui contient $P = 0{,}674$.

La précision dépend de la proportion elle-même



Marge absolue, marge relative

La marge est maximale en $P = 0{,}5$ (±6,6 points pour $n = 200$) et varie peu entre 0,3 et 0,7. Une proportion **rare** est mal estimée relativement à sa valeur : ±2,9 points pour $P = 0{,}05$, soit 57 % de P (chapitre 5).

Démonstration : Marge absolue et marge relative d'une proportion

La marge à 95 % de $\widehat{P}$ est $e(P) = 1{,}96\sqrt{(1-f)\,\frac{N}{N-1}\,\frac{P(1-P)}{n}}$; elle est proportionnelle à $\sqrt{P(1-P)}$.

Le maximum. Comme $P(1-P) = \frac{1}{4} - \left(P - \frac{1}{2}\right)^2$, la marge est maximale pour $P = \frac{1}{2}$, et symétrique autour de ce point. Pour $n = 200$: ±6,6 points en $P = 0{,}5$, mais déjà $1{,}96\sqrt{0{,}9 \times 0{,}21/200} = 6{,}0$ points en $P = 0{,}3$: la courbe est plate autour du maximum.

La marge relative. Divisons par P :

$$\frac{e(P)}{P} = 1{,}96\sqrt{\frac{(1-f)\,N}{(N-1)\,n}\cdot\frac{1-P}{P}}.$$

Le rapport $\frac{1-P}{P}$ décroît de $+\infty$ à 0 quand P croît de 0 à 1 : plus la proportion est rare, plus la marge relative est grande. Pour $P = 0{,}05$, la marge absolue n'est que de $\pm 2{,}9$ points, mais elle vaut 57 % de P.

Ce que cela dit. La marge **absolue** est maximale à mi-chemin, la marge **relative** explose pour les petites proportions : on ne peut pas juger la précision d'une proportion rare en points, seulement en pourcentage d'elle-même.

■

Estimer un total

Du total à la moyenne, et retour

L'estimateur du total

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

Chaque ménage tiré compte pour son poids $d_k = N/n$: c'est l'estimateur de Horvitz-Thompson (chapitre 2). Il est sans biais, et

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

Le total des dépenses

- Mini-population, $n = 2$: $V(\widehat{Y}) = 5^2 \times 4{,}35 = 108{,}75$, erreur-type 10,4 milliers de dh.
- Échantillon fil rouge : $\widehat{Y} = 2\,000 \times 7\,843{,}05 = 15\,686\,100$ dh ; erreur-type $2\,000 \times 256{,}87 = 513\,736$ dh ; intervalle $[14{,}68\,;16{,}69]$ millions de dh, qui contient $Y = 15{,}82$ millions.

Le nombre de ménages connectés

Compter, c'est estimer un total

Le nombre de ménages ayant une caractéristique est le total d'une variable 0/1 : NP. On l'estime par $N\widehat{P}$, de variance $N^2\,V(\widehat{P})$, estimée par

$$\widehat{V}(N\widehat{P}) = N^2(1-f)\,\frac{\widehat{P}(1-\widehat{P})}{n-1}$$

Combien de ménages ont accès à internet ?

- **La vérité** : 1 348 ménages. Pour $n = 200$, l'erreur-type de $N\widehat{P}$ vaut $2\,000 \times 0{,}0315 = 62{,}9$ ménages, la marge ± 123.

- **L'échantillon fil rouge** : $2\,000 \times 0{,}67 = 1\,340$ ménages ; erreur-type estimée $2\,000 \times 0{,}0316 = 63{,}2$; intervalle $[1\,216\,;1\,464]$.

Même erreur absolue, autre erreur relative

Le coefficient de variation

$CV = \sqrt{V(\hat{\theta})}/\theta$ mesure l'erreur **relative**. Multiplier par N ne le change pas : $CV(\widehat{Y}) = CV(\overline{y}) = 267{,}27/7\,911{,}44 = 3{,}4\,\%$ pour la dépense.

Connectés et non connectés

Les 652 ménages non connectés s'estiment par $N(1-\widehat{P})$, avec **la même** erreur-type, 62,9 ménages. Mais le CV double :

Nombre à estimer	Vraie valeur	Erreur-type	CV
Ménages connectés	1 348	62,9	4,7 %
Ménages non connectés	652	62,9	9,6 %

Plus le groupe compté est petit, moins son effectif est connu avec précision, en termes relatifs.

Quand N est inconnu

Tirer sans connaître la taille de la liste

$N\overline{y}$ exige de connaître N. Sans liste complète (commerçants ambulants, exploitations non recensées), on peut garder chaque unité rencontrée avec une probabilité fixée π : c'est le tirage de Bernoulli (chapitre 1). On estime alors $\widehat{Y} = \frac{1}{\pi}\sum_{k\in S} y_k$, $\widehat{N} = n/\pi$, et la moyenne par $\widehat{Y}/\widehat{N} = \overline{y}$.

Un essai sur la population du cours

Avec $\pi = 0{,}1$ (graine fixée), on obtient $n = 202$ ménages qui dépensent 1 601 800 dh : $\widehat{Y} = 16\,018\,000$ dh (+1,2 %), $\widehat{N} = 2\,020$, moyenne 7 929,7 dh. Mais l'erreur-type de $\widehat{Y}$, $\sqrt{\sum_U \frac{1-\pi}{\pi} y_k^2} = 1{,}19$ million de dh (chapitre 2), est plus du double de celle de $N\overline{y}$ (0,53 million) : la taille aléatoire ajoute de l'incertitude. La moyenne, elle, n'en souffre pas (chapitre 12).

Démonstration : Le tirage de Bernoulli, variance du total estimé

Chaque unité est gardée avec la probabilité π, **indépendamment** des autres : $\pi_k = \pi$ et, pour $k \neq l$, $\pi_{kl} = \pi^2 = \pi_k\pi_l$, donc $\Delta_{kl} = 0$. La taille n est aléatoire, et l'estimateur de Horvitz-Thompson du total est $\widehat{Y} = \frac{1}{\pi}\sum_{k\in S} y_k$, sans biais d'après le chapitre 2.

Variance. Dans la formule générale du chapitre 2, seuls les termes diagonaux subsistent :

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

Effectif et moyenne. $\widehat{N} = n/\pi$ est le total estimé de la variable égale à 1, sans biais pour N, et $\widehat{Y}/\widehat{N} = \frac{\sum_S y_k}{n} = \overline{y}$.

Pour le cours. Avec $\pi = 0{,}1$, l'erreur-type est $\sqrt{V(\widehat{Y})} = \sqrt{9 \times \sum_k y_k^2} = 1{,}19$ million de dh. Le sondage aléatoire simple de 200 ménages donne $N\sqrt{(1-f)S^2/n} = 2\,000 \times 267{,}27 = 0{,}53$ million : le tirage de Bernoulli a une erreur-type plus de deux fois plus grande.

Ce que cela dit. La taille de l'échantillon est une source de variation en plus : quand on garde plus de ménages que prévu, le total estimé augmente avec elle. Un plan de taille fixe écarte cette source ; la moyenne $\overline{y}$, qui divise par n, n'en souffre pas.

■

Les sous-populations

Estimer pour un domaine

Domaine

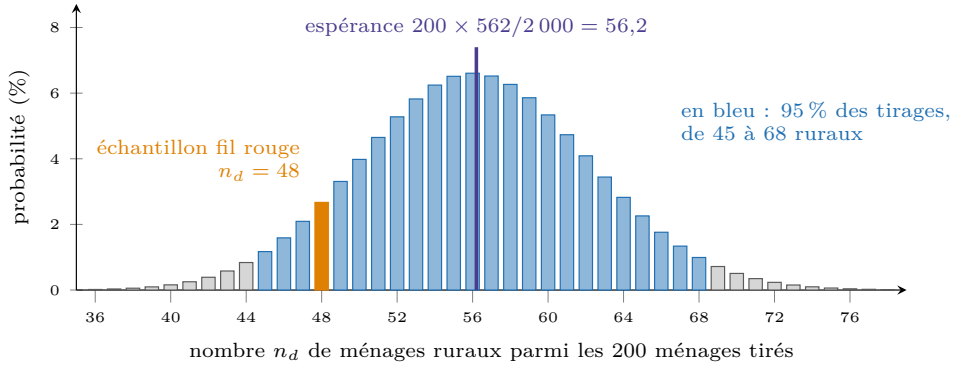
Un **domaine** est une sous-population U_d de taille N_d : les ruraux, une province, les ménages dirigés par une femme. On veut sa moyenne $\overline{Y}_d$, son total Y_d, une proportion en son sein. On utilise les n_d ménages tirés qui en font partie :

$$\overline{y}_d = \frac{1}{n_d} \sum_{k \in S \cap U_d} y_k$$

Les ruraux de l'échantillon fil rouge

L'échantillon compte 48 ruraux et 152 urbains. Dépense moyenne des 48 ruraux : $\overline{y}_d = 6\,916{,}67$ dh ($s_d = 3\,375{,}32$), contre $\overline{Y}_d = 6\,585{,}14$ dh pour les 562 ruraux de la population. Les 152 urbains dépensent en moyenne 8 135,59 dh.

Une taille d'échantillon aléatoire



Le nombre de ruraux tirés n'est pas choisi

n_d dépend du tirage : il suit une loi hypergéométrique d'espérance $200 \times 0{,}281 = 56{,}2$ et d'écart-type 6,0. 95 % des échantillons comptent entre 45 et 68 ruraux. Avec 48, l'échantillon fil rouge en a un peu moins que prévu : $P(n_d \leq 48) = 0{,}10$.

La précision d'une moyenne de domaine

Raisonner à taille de domaine fixée

Sachant n_d, les ménages du domaine tirés forment un sondage aléatoire simple de n_d ménages parmi N_d. Donc $\overline{y}_d$ est sans biais, et

$$\widehat{V}(\overline{y}_d) = \left(1 - \frac{n_d}{N_d}\right) \frac{s_d^2}{n_d}$$

Si N_d est inconnu, on l'estime par $\widehat{N}_d = N\, n_d/n$, ce qui revient à remplacer n_d/N_d par f.

Démonstration : Sachant la taille, le domaine est un sondage aléatoire simple

Fixons m et cherchons la probabilité que la partie de l'échantillon située dans le domaine soit un ensemble donné s_d de m unités de U_d. Les échantillons s tels que $s \cap U_d = s_d$ sont formés de s_d et de $n - m$ unités choisies **hors du domaine**, parmi les $N - N_d$ restantes : il y en a $\binom{N-N_d}{n-m}$, **le même nombre quel que soit** s_d. Tous les échantillons ayant la même probabilité $1/\binom{N}{n}$, tous les ensembles s_d de m unités ont la même probabilité, qui vaut $\binom{N-N_d}{n-m} \Big/ \binom{N}{n}$: sachant $n_d = m$, la partie de l'échantillon dans le domaine est un sondage aléatoire simple de m unités parmi N_d.

Conséquences. Les résultats du chapitre 3 s'appliquent à $\overline{y}_d$, avec N_d pour population et m pour taille : pour chaque valeur de $n_d \geq 1$, $E(\overline{y}_d \mid n_d) = \overline{Y}_d$, donc $\overline{y}_d$ est sans biais ; et $V(\overline{y}_d \mid n_d) = \left(1 - \frac{n_d}{N_d}\right) \frac{S_d^2}{n_d}$, estimée sans biais par $\widehat{V}(\overline{y}_d)$ dès que $n_d \geq 2$.

Effectif du domaine inconnu. L'espérance de n_d est $\sum_{k \in U_d} \pi_k = N_d \frac{n}{N}$, donc $\widehat{N}_d = N \frac{n_d}{n}$ est sans biais pour N_d, et $\dfrac{n_d}{\widehat{N}_d} = \dfrac{n}{N} = f$.

Ce que cela dit. La taille n_d n'est pas choisie, mais elle n'empêche pas de raisonner : on se place **sachant** n_d, et le plan redevient un plan simple. Pour les ruraux, $E(n_d) = 200 \times \frac{562}{2\,000} = 56{,}2$.

■

La dépense des ruraux

$\widehat{V}(\overline{y}_d) = (1 - 48/562) \times 3\,375{,}32^2/48$, erreur-type 465,9 dh. Intervalle : $6\,916{,}67 \pm 913{,}2$, soit $[6\,003{,}5\,;7\,829{,}9]$, qui contient 6 585,14. L'erreur-type est 1,8 fois celle de la moyenne générale (256,87) : l'estimation repose sur 48 ménages, pas sur 200.

Le total d'un domaine

Deux estimateurs du total des ruraux

- N_d **connu** (par le recensement) : $N_d\,\overline{y}_d$.
- N_d **inconnu** : $\widehat{Y}_d = \frac{N}{n}\sum_{k\in S\cap U_d} y_k$, l'estimateur du total de la variable $y_k^{(d)}$ = « y_k si k est rural, 0 sinon ». Sans biais, de variance $N^2(1-f)\,S^2_{(d)}/n$, où $S^2_{(d)}$ est la variance de $y^{(d)}$.

Le rapport $\widehat{Y}_d/\widehat{N}_d$ redonne $\overline{y}_d$: c'est un **ratio** de deux totaux estimés (chapitre 12).

Le total des dépenses rurales

Vrai total : 3 700 850 dh. Échantillon fil rouge : $562 \times 6\,916{,}67 = 3\,887\,167$ dh si N_d est connu, $10 \times 48 \times 6\,916{,}67 = 3\,320\,000$ dh sinon. Sur tous les échantillons, les erreurs-types valent environ 230 000 dh et 458 000 dh : **connaître** N_d **divise la variance par** 4.

Démonstration : Pourquoi connaître la taille du domaine divise la variance par quatre

Notons $W_d = N_d/N$, $\overline{Y}_d$ et S_d^2 la moyenne et la variance du domaine, et $y_k^{(d)} = y_k$ si $k \in U_d$, 0 sinon.

N_d inconnu. Le total estimé est $\widehat{Y}_d = N\,\overline{y^{(d)}}$, de variance $N^2(1-f)\,S^2_{(d)}/n$. La moyenne de $y^{(d)}$ est $W_d\overline{Y}_d$ et sa somme des carrés est $\sum_{U_d} y_k^2 = (N_d-1)S_d^2 + N_d\overline{Y}_d^{\,2}$ (chapitre 1), donc

$$(N-1)\,S^2_{(d)} = (N_d-1)\,S_d^2 + N_d\overline{Y}_d^{\,2} - N\,W_d^2\,\overline{Y}_d^{\,2} = (N_d-1)\,S_d^2 + N_d\,(1-W_d)\,\overline{Y}_d^{\,2}.$$

Pour N et N_d grands, $S^2_{(d)} \approx W_d\left[S_d^2 + (1-W_d)\overline{Y}_d^{\,2}\right]$ et $V(\widehat{Y}_d) \approx \frac{N^2(1-f)\,W_d}{n}\left[S_d^2 + (1-W_d)\overline{Y}_d^{\,2}\right]$.

N_d connu. On estime $N_d\,\overline{y}_d$, de variance $N_d^2\left(1-\frac{n_d}{N_d}\right)\frac{S_d^2}{n_d}$. Avec $n_d \approx nW_d$ et $\frac{n_d}{N_d} \approx f$, cela donne $V \approx \frac{N^2(1-f)\,W_d}{n}\,S_d^2$.

Le rapport.

$$\frac{V(\widehat{Y}_d)}{V(N_d\overline{y}_d)} \approx 1 + (1 - W_d)\,\frac{\overline{Y}_d^{\,2}}{S_d^2}.$$

Pour les ruraux, $W_d = 0{,}281$, $\overline{Y}_d = 6\,585$ dh et $S_d = 3\,215$ dh : $1 + 0{,}719 \times 4{,}20 = 4{,}0$.

Ce que cela dit. Quand N_d est inconnu, l'estimateur doit aussi estimer **combien** de ménages sont ruraux, et ce nombre varie d'un échantillon à l'autre : cette variation, multipliée par le niveau $\overline{Y}_d$ des dépenses, s'ajoute à la dispersion S_d^2. Plus le domaine a une moyenne élevée relativement à sa dispersion, plus connaître N_d rapporte : ici, les erreurs-types passent de 458 000 à 230 000 dh.

■

Les petits domaines

Al Hoceïma : 24 ménages attendus

Dans un sondage aléatoire simple de 200 ménages, la province d'Al Hoceïma (240 ménages) n'en reçoit que 24 en moyenne. Marge à 95 % : environ $\pm 1\,101$ dh sur une moyenne de 5 979 dh, soit ± 18 %. Dans l'échantillon fil rouge : 28 ménages, dépense moyenne 5 115 dh, 14 % sous la vraie valeur.

Fixer la taille des domaines

Quand les domaines figurent dans la base de sondage, mieux vaut tirer **séparément** dans chacun : n_d est alors fixé à l'avance. C'est le sondage stratifié (chapitre 6) ; une répartition égale entre domaines (chapitre 7) sert les estimations par province.

Le tirage avec remise

Tirer avec remise

Des tirages indépendants

On effectue n tirages ; à chacun, chaque unité a la probabilité $1/N$ d'être choisie, qu'elle l'ait déjà été ou non. Une unité peut donc apparaître plusieurs fois. Estimateur : la moyenne des n valeurs tirées, répétitions comprises,

$$\overline{y}_{AR} = \frac{1}{n}\sum_{i=1}^{n} y_{(i)} \qquad (y_{(i)} : \text{valeur du } i\text{-ème tirage})$$

Chaque tirage a pour espérance $\sum_k \frac{1}{N}\, y_k = \overline{Y}$: $\overline{y}_{AR}$ est sans biais.

Mini-population, deux tirages avec remise

$5 \times 5 = 25$ suites ordonnées équiprobables, dont 5 répètent le même ménage (AA, BB, …). La moyenne des 25 moyennes vaut $8 = \overline{Y}$; leur variance vaut 5,8, contre 4,35 sans remise.

La variance d'un tirage avec remise

La démonstration

Les n tirages sont indépendants et de même loi : $y_{(i)}$ vaut y_k avec la probabilité $1/N$, d'espérance $\overline{Y}$ et de variance $\sigma^2 = \frac{1}{N}\sum_U (y_k - \overline{Y})^2$. La variance d'une moyenne de n variables indépendantes de variance σ^2 est σ^2/n.

Les formules

$$V(\overline{y}_{AR}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\,\frac{S^2}{n} \qquad \widehat{V}(\overline{y}_{AR}) = \frac{s^2}{n}$$

où s^2, calculée sur les n tirages, est sans biais pour σ^2. Total : $N\,\overline{y}_{AR}$, de variance $N^2\sigma^2/n$. Pas de facteur $(1-f)$.

Vérification

Mini-population : $\sigma^2 = 58/5 = 11{,}6$, et $11{,}6/2 = 5{,}8$. Population du cours : $\sigma^2/200 = 3\,983{,}29^2/200 = 79\,333{,}1$.

Démonstration : La variance du tirage avec remise et son estimation

Loi d'un tirage. À chaque tirage, $y_{(i)}$ vaut y_k avec la probabilité $1/N$: $E(y_{(i)}) = \frac{1}{N}\sum_k y_k = \overline{Y}$ et $V(y_{(i)}) = \frac{1}{N}\sum_k (y_k - \overline{Y})^2 = \sigma^2$, la variance **divisée par** N. Les n tirages étant indépendants,

$$V(\overline{y}_{AR}) = \frac{1}{n^2}\sum_{i=1}^{n} V(y_{(i)}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\,\frac{S^2}{n}.$$

Estimation. Comme au chapitre 3, $\sum_i (y_{(i)} - \overline{y}_{AR})^2 = \sum_i (y_{(i)} - \overline{Y})^2 - n(\overline{y}_{AR} - \overline{Y})^2$. L'espérance du premier terme est $n\sigma^2$, celle du second $n\,V(\overline{y}_{AR}) = \sigma^2$: l'écart vaut $(n-1)\,\sigma^2$, donc $E(s^2) = \sigma^2$ et $E(s^2/n) = V(\overline{y}_{AR})$: sans facteur $(1-f)$, la variance estimée est s^2/n.

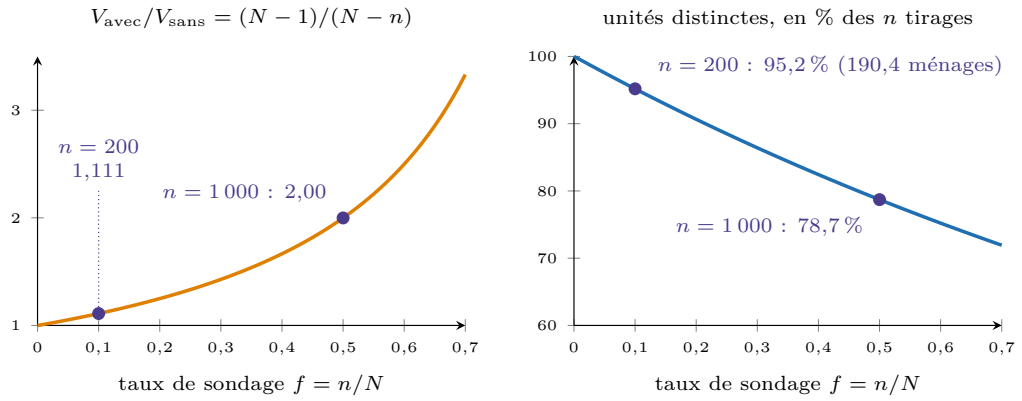
Comparaison avec le tirage sans remise.

$$\frac{V_{AR}}{V_{SR}} = \frac{(N-1)\,S^2/(N\,n)}{(N-n)\,S^2/(N\,n)} = \frac{N-1}{N-n} > 1 \quad \text{pour } n \geq 2.$$

Ce que cela dit. Le tirage avec remise est moins précis, puisqu'il peut interroger deux fois la même unité. Sur la mini-population, $\frac{4}{3}$; sur la population du cours avec $n = 200$, $\frac{1\,999}{1\,800} = 1{,}1106$, et le rapport tombe à 1,002 pour 18 000 ménages parmi 9 275 038.

■

Avec ou sans remise



Le rapport des variances

$$\frac{V_{AR}}{V_{SR}} = \frac{(N-1)\,S^2/(Nn)}{(1-f)\,S^2/n} = \frac{N-1}{N-n} \approx \frac{1}{1-f}$$

Pour $n = 200$: 79 333,1/71 435,5 = 1,1106. Sur la mini-population : 5,8/4,35 = 4/3.

Combien de ménages distincts ?

Le nombre attendu d'unités distinctes

Une unité n'est jamais tirée avec la probabilité $(1 - 1/N)^n$. Avec $I_k = 1$ si k est tirée au moins une fois, $\pi_k = 1 - (1 - 1/N)^n$ et le nombre ν d'unités distinctes vérifie

$$E(\nu) = \sum_{k \in U} \pi_k = N \left[1 - \left(1 - \frac{1}{N} \right)^n \right]$$

Deux cents tirages parmi 2 000 ménages

$\pi_k = 1 - 0{,}9995^{200} = 0{,}0952$ (contre 0,1 sans remise) et $E(\nu) = 190{,}4$: en moyenne 9,6 tirages servent à réinterroger un ménage déjà vu. Il faut $200 \times 1{,}1106 = 222$ tirages avec remise pour la précision de 200 sans remise. Mini-population : $E(\nu) = 5 \times (1 - 0{,}8^2) = 1{,}8$.

Démonstration : Le nombre attendu d'unités distinctes

À chaque tirage, l'unité k n'est pas choisie avec la probabilité $1 - \frac{1}{N}$; les n tirages étant indépendants, elle n'est **jamais** choisie avec la probabilité $\left(1 - \frac{1}{N}\right)^n$. L'indicatrice I_k « k est tirée au moins une fois » a donc pour espérance

$$\pi_k = 1 - \left(1 - \frac{1}{N} \right)^n .$$

Le nombre d'unités distinctes est $\nu = \sum_{k \in U} I_k$, et par linéarité $E(\nu) = N\pi_k$: la formule annoncée.

Pour le cours. $\pi_k = 1 - 0{,}9995^{200} = 0{,}0952$ et $E(\nu) = 2\,000 \times 0{,}0952 = 190{,}4$: en moyenne 9,6 des 200 tirages retombent sur un ménage déjà interrogé. Pour égaler la précision de 200

ménages sans remise, il faut n' tirages avec $\frac{\sigma^2}{n'} = (1-f)\frac{S^2}{200}$, soit $n' = 200 \times \frac{N-1}{N-200} = 222{,}1$, environ 222 tirages.

Ce que cela dit. Le tirage avec remise gaspille environ 5 % des tirages à $f = 0{,}1$: c'est exactement ce que mesure le rapport 1,1106 des variances.

■

Pourquoi tirer avec remise ?

Quand le taux de sondage est faible

Pour 18 000 ménages parmi 9 275 038 (l'échelle de l'enquête sur le niveau de vie de 2022), $f = 0{,}0019$ et le rapport des variances vaut 1,002 : la différence est invisible. On calcule alors souvent la variance **comme si** le tirage était avec remise. Sur l'échantillon fil rouge, où $f = 0{,}1$, cette approximation donne $s/\sqrt{n} = 270{,}76$ au lieu de 256,87 : elle est **prudente**.

La simplicité de l'indépendance

Des tirages indépendants donnent des formules simples, même lorsque les unités ont des probabilités de tirage **inégales** : c'est l'estimateur de Hansen-Hurwitz, au cœur des tirages proportionnels à la taille (chapitre 11).

Tout estimer sur un échantillon

Six estimations sur l'échantillon fil rouge

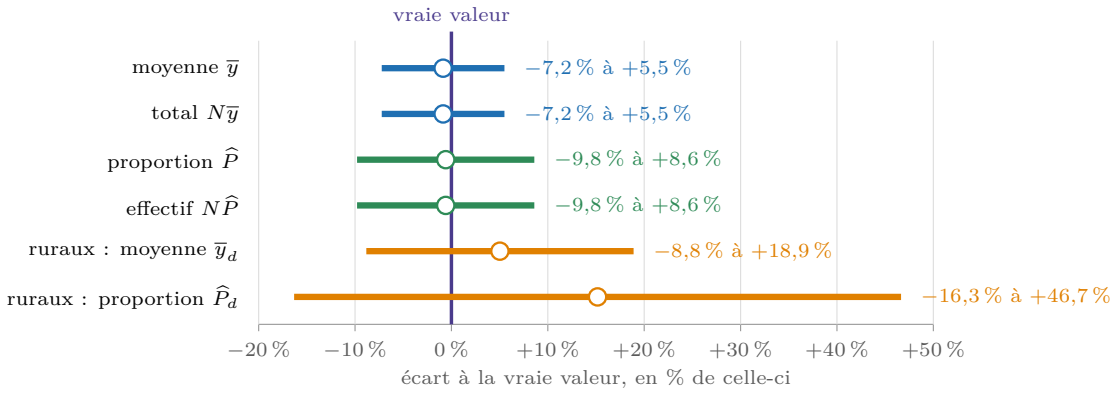
Les mêmes 200 ménages, six paramètres

Paramètre	Estimation	Erreur-type	Vérité
Dépense moyenne (dh)	7 843,05	256,87	7 911,44
Dépense totale (millions dh)	15,686	0,514	15,823
Part de ménages connectés	0,670	0,0316	0,674
Nombre de ménages connectés	1 340	63,2	1 348
Moyenne des ruraux (dh)	6 916,67	465,92	6 585,14
Part de connectés, ruraux	0,500	0,0697	0,434

Un seul tirage, toutes les questions

Une enquête sert à estimer de nombreux paramètres, sur toute la population et sur des domaines, avec les **mêmes** ménages. Chez les ruraux : 24 connectés sur 48, d'où $\widehat{P}_d = 0{,}5$, avec la formule du domaine.

Ce que montrent les intervalles



Lire la figure

Les six intervalles à 95 % contiennent la vraie valeur. Moyenne et total ont le **même** intervalle relatif, de −7,2 % à +5,5 % : multiplier par N ne change rien. Les estimations sur les 48 ruraux sont nettement moins précises.

Les intervalles tiennent-ils leurs promesses ?

Vingt mille échantillons de 200 ménages

Part des intervalles à 95 % qui contiennent la vraie valeur, sur 20 000 tirages (graine fixée) :

Paramètre	Ménages utilisés	Couverture
Part de ménages connectés	200	95,3 %
Dépense moyenne des ruraux	56 en moyenne	94,4 %
Dépense moyenne d'Al Hoceïma	24 en moyenne	92,3 %

Quand l'approximation normale faiblit

Plus le domaine est petit, plus l'approximation normale est fragile : l'intervalle « à 95 % » couvre moins souvent. L'estimateur reste sans biais ; c'est l'**annonce de sa précision** qui devient trop optimiste.

S'entraîner

Exercice 1 : une proportion et un total

Le microcrédit

Une association de microcrédit compte 2 500 clientes. On en tire 250 par sondage aléatoire simple : 40 ont un retard de remboursement.

1. Estimer la proportion P de clientes en retard et son erreur-type.
2. Donner un intervalle de confiance à 95 % pour P.
3. Estimer le nombre de clientes en retard, avec son intervalle de confiance.
4. Calculer le coefficient de variation de ces deux estimations. Commenter.

Exercice 1 : solution

La proportion

1. $\widehat{P} = 40/250 = 0{,}16$ et $f = 0{,}1$: $\widehat{V}(\widehat{P}) = 0{,}9 \times \dfrac{0{,}16 \times 0{,}84}{249} = 0{,}000486$, erreur-type $0{,}0220$.
2. $0{,}16 \pm 1{,}96 \times 0{,}0220 = 0{,}16 \pm 0{,}043$, soit $[0{,}117\,;0{,}203]$.

Le total

3. $2\,500 \times 0{,}16 = 400$ clientes ; erreur-type $2\,500 \times 0{,}0220 = 55{,}1$; intervalle $[292\,;508]$.
4. $CV = 0{,}0220/0{,}16 = 13{,}8\,\%$, le même pour la proportion et pour le total. Une proportion faible est estimée avec une erreur **relative** élevée, même si l'erreur absolue paraît petite.

Exercice 2 : tirer avec remise

Quatre boutiques, deux tirages

Les quatre boutiques de l'exercice 1 du chapitre 1 ont un chiffre d'affaires de 2, 5, 8 et 9 milliers de dh. On tire deux boutiques **avec remise**.

1. Combien de suites de tirages sont possibles ? Avec quelle probabilité chacune ?
2. Calculer σ^2 et $V(\overline{y}_{AR})$.
3. Comparer à la variance sans remise ($S^2 = 10$). Retrouver le rapport $(N-1)/(N-n)$.
4. Quel est le nombre attendu de boutiques distinctes ?

Exercice 2 : solution

Seize suites

1. $4 \times 4 = 16$ suites ordonnées, de probabilité $1/16$ chacune ; 4 d'entre elles répètent la même boutique.
2. Écarts à 6 : -4, -1, 2, 3 ; $\sigma^2 = 30/4 = 7{,}5$ et $V(\overline{y}_{AR}) = 7{,}5/2 = 3{,}75$. Les 16 moyennes ont bien pour moyenne 6 et pour variance 3,75.

La comparaison

3. Sans remise : $(1 - 0{,}5) \times 10/2 = 2{,}5$. Rapport : $3{,}75/2{,}5 = 1{,}5 = (4-1)/(4-2)$.
4. $E(\nu) = 4 \times [1 - (3/4)^2] = 1{,}75$: une fois sur quatre, on interroge deux fois la même boutique.

Exercice 3 : un domaine

Les ménages dirigés par une femme

Dans une ville de 8 000 ménages, un sondage aléatoire simple de 400 ménages en compte 96 dirigés par une femme. Leur dépense moyenne est de 6 200 dh, avec $s_d = 2\,800$ dh.

1. Estimer la part et le nombre de ménages dirigés par une femme, avec leur erreur-type.
2. Pourquoi $n_d = 96$ est-il aléatoire ? Si la vraie part était 0,25, quelles seraient son espérance et son écart-type ?
3. Estimer la dépense moyenne de ces ménages et son intervalle de confiance à 95 %.

4. Estimer leur dépense totale.

Exercice 3 : solution

La taille du domaine

1. $\widehat{P} = 96/400 = 0{,}24$, erreur-type $\sqrt{0{,}95 \times 0{,}24 \times 0{,}76/399} = 0{,}0208$. Nombre : $8\,000 \times 0{,}24 = 1\,920$ ménages, erreur-type 167.
2. n_d dépend des ménages tirés : loi hypergéométrique, d'espérance $400 \times 0{,}25 = 100$ et d'écart-type $\sqrt{400 \times 0{,}25 \times 0{,}75 \times 7\,600/7\,999} = 8{,}4$.

La moyenne et le total du domaine

3. N_d est inconnu : on remplace n_d/N_d par $f = 0{,}05$. Erreur-type $\sqrt{0{,}95 \times 2\,800^2/96} = 278{,}5$ dh ; intervalle $6\,200 \pm 546$, soit $[5\,654\,;6\,746]$.
4. $\widehat{N}_d\,\overline{y}_d = 1\,920 \times 6\,200 = 11\,904\,000$ dh, qui est aussi $\frac{N}{n}\sum_{k\in S\cap U_d} y_k = 20 \times 96 \times 6\,200$.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Une proportion est la moyenne d'une variable 0/1 : $\widehat{P} = \overline{y}$, sans biais.
2. $V(\widehat{P}) = \frac{N-n}{N-1}\,\frac{P(1-P)}{n}$, estimée par $(1-f)\,\frac{\widehat{P}(1-\widehat{P})}{n-1}$; maximale en $P = 0{,}5$.
3. Total : $N\overline{y}$, de variance $N^2V(\overline{y})$ et de même CV que la moyenne.
4. Un effectif est le total d'une variable 0/1 : on l'estime par $N\widehat{P}$.
5. Dans un domaine, n_d est aléatoire ; à n_d fixé, $\overline{y}_d$ se traite comme un sondage aléatoire simple.
6. Avec remise, $V(\overline{y}_{AR}) = \sigma^2/n$, soit $\frac{N-1}{N-n}$ fois la variance sans remise.
7. On attend $N\,[1-(1-1/N)^n]$ unités distinctes ; l'indépendance des tirages simplifie les calculs (chapitre 11).

La suite

Le chapitre 5 renverse la question : quelle taille n faut-il pour atteindre une précision fixée à l'avance ?