# Échantillonnage

Chapitre 1 : Population, paramètres et plan de sondage

Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 21 octobre 2026

# Plan du chapitre

Ce que l'on cherche

La dispersion

Le plan de sondage

Les probabilités d'inclusion

Sur la population du cours

S'entraîner

Ce qu'il faut retenir

# Ce que l’on cherche

## Quatre paramètres

Une variable $y$ prend la valeur $y_k$ pour chaque unité $k$ de la population $U = \{1, \dots, N\}$.

| Paramètre | Définition | Exemple |
|---|---|---|
| Total | $Y = \sum_{k \in U} y_k$ | dépense de tous les ménages |
| Moyenne | $\overline{Y} = Y/N$ | dépense par ménage |
| Proportion | $P$ = moyenne d'une variable 0/1 | part des ménages connectés |
| Ratio | $R = Y/X$ | dépense par personne |

## Le total d'abord

La comptabilité nationale, la fiscalité ou la politique sociale ont besoin de **totaux** : masse des dépenses, nombre de chômeurs, nombre de ménages pauvres. La moyenne s'en déduit en divisant par $N$, quand $N$ est connu.

**La variable indicatrice**

Si $y_k = 1$ quand le ménage $k$ a accès à internet et $y_k = 0$ sinon, alors

$$Y = \sum_{k \in U} y_k = \text{nombre de ménages connectés}, \qquad \overline{Y} = \frac{Y}{N} = P$$

**Une seule théorie pour deux questions**

Tout ce que le cours établit pour une moyenne vaut donc pour une proportion : il suffit de coder la caractéristique en 0/1. Le chapitre 4 en tirera les formules particulières.

## Cinq ménages : dépense en milliers de dh, taille en personnes

| Ménage | A | B | C | D | E | Total |
|---|---|---|---|---|---|---|
| Dépense $y_k$ | 4 | 6 | 7 | 9 | 14 | 40 |
| Taille $x_k$ | 2 | 3 | 3 | 4 | 6 | 18 |

## Les quatre paramètres

Total $Y = 40$ ; moyenne $\overline{Y} = 40/5 = 8$ ; proportion des ménages dépensant plus de $8$ : $P = 2/5 = 0{,}4$ (D et E) ; dépense par personne $R = 40/18 = 2{,}22$ milliers de dh.

## Les paramètres que l'enquêteur cherche

| Paramètre | Valeur |
|---|---|
| Total des dépenses $Y$ | $15\,822\,870$ dh |
| Dépense moyenne $\overline{Y}$ | $7\,911{,}44$ dh |
| Proportion de ménages connectés $P$ | $0{,}674$ |
| Total des personnes $X$ | $7\,859$ |
| Dépense par personne $R = Y/X$ | $2\,013{,}34$ dh |

## Un ratio n'est pas une moyenne de ratios

$R$ divise la dépense **totale** par le nombre **total** de personnes. La moyenne des dépenses par personne de chaque ménage donnerait autre chose, car elle compte autant un ménage d'une personne qu'un ménage de dix.

# La dispersion

## Diviser par N ou par N − 1

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

## La convention des sondages : S²

La statistique descriptive divise par $N$. La théorie des sondages préfère $S^2$, divisée par $N-1$, pour une seule raison : avec elle, la variance d'un sondage aléatoire simple s'écrit $(1-f)\,S^2/n$, une formule **exacte** et simple. Le chapitre 3 le démontrera.

**Sur la mini-population**

Écarts à la moyenne : $-4, -2, -1, 1, 6$ ; somme des carrés : $16 + 4 + 1 + 1 + 36 = 58$.

$$\sigma^2 = 58/5 = 11{,}6 \qquad S^2 = 58/4 = 14{,}5$$

**Sur la population du cours**

$\sigma = 3\,983{,}29$ dh et $S = 3\,984{,}29$ dh : un dirham d'écart sur $N = 2\,000$ ménages. La convention ne compte que sur de petites populations.

## Une formule fermée

Pour une variable 0/1, $y_k^2 = y_k$, d'où $\sigma^2 = P - P^2 = P(1-P)$ et

$$S^2 = \frac{N}{N-1} P\,(1-P)$$

## Les ménages connectés

$P = 0{,}674$, donc $S^2 = \dfrac{2\,000}{1\,999} \times 0{,}674 \times 0{,}326 = 0{,}2198$. La dispersion d'une proportion se déduit de la proportion elle-même : c'est ce qui permettra de dimensionner une enquête **avant** de la faire (chapitre 5).

# Le plan de sondage

# Combien d'échantillons possibles ?

## Un échantillon est une partie de U

Un échantillon sans répétition $s$ est un sous-ensemble de $U$. Il y a $\binom{N}{n}$ échantillons de taille $n$.

## De dix à un nombre astronomique

| Population | Taille | Nombre d'échantillons |
|---|---|---|
| Mini-population, $N = 5$ | $n = 2$ | $\binom{5}{2} = 10$ |
| Population du cours, $N = 2\,000$ | $n = 100$ | environ $10^{171}$ |
| Population du cours, $N = 2\,000$ | $n = 200$ | environ $10^{281}$ |

## On ne les énumère jamais

Les formules du cours remplacent l'énumération : elles donnent la moyenne et la variance d'un estimateur sur **tous** les échantillons possibles, sans en tirer un seul.

**Plan de sondage**

Un **plan de sondage** est une loi de probabilité $p(\cdot)$ sur l'ensemble des échantillons possibles :

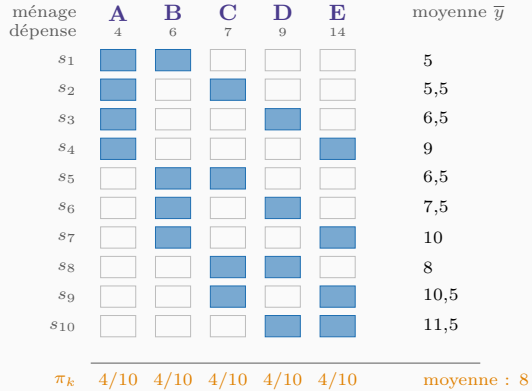
$$p(s) \geq 0 \text{ pour tout } s, \qquad \sum_{s} p(s) = 1$$

L'échantillon effectivement tiré, $S$, est une variable aléatoire de loi $p$.

**Ce que le plan contient**

Le plan dit **tout** : quels échantillons sont possibles, lesquels sont plus probables, quelle taille ils ont. Il est choisi par l'enquêteur, **avant** le tirage, et c'est lui qui rend l'erreur calculable.

# Le plan aléatoire simple sur la mini-population




## Dix échantillons de même probabilité

Chaque ligne est un échantillon de deux ménages, de probabilité $p(s) = 1/10$. La moyenne des dix moyennes vaut exactement $8 = \overline{Y}$.

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

On sépare les petits budgets (A, B, C) des grands (D, E), et l'on tire un ménage dans chaque groupe. Il n'y a plus que $3 \times 2 = 6$ échantillons possibles, chacun de probabilité $1/6$ :

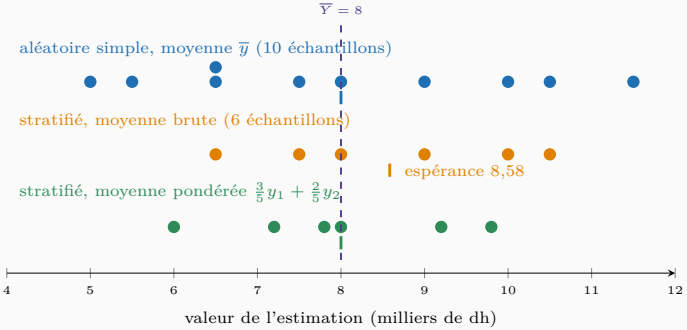
| Échantillon | AD | AE | BD | BE | CD | CE |
|---|---|---|---|---|---|---|
| Moyenne brute | $6{,}5$ | $9$ | $7{,}5$ | $10$ | $8$ | $10{,}5$ |
| Moyenne pondérée | $6$ | $8$ | $7{,}2$ | $9{,}2$ | $7{,}8$ | $9{,}8$ |

Brute : $\frac{1}{2}(y_1 + y_2)$. Pondérée : $\frac{3}{5}\, y_1 + \frac{2}{5}\, y_2$, où $y_1$ vient du premier groupe.

## Le plan a changé, l'estimateur doit suivre

Moyenne des six moyennes brutes : $51{,}5/6 = 8{,}58 \neq 8$. Le premier groupe compte trois ménages sur cinq, mais n'en fournit qu'un sur deux : il faut lui redonner son **poids**, $3/5$.

# Ce que montre la comparaison




## Un estimateur se juge avec son plan

La même moyenne $\overline{y}$ est juste en moyenne sous le plan aléatoire simple, faussée sous le second plan. Un estimateur n'est bon ou mauvais que **relativement au plan** qui a produit l'échantillon.

## Trois distinctions

- **Sans remise** : une unité figure au plus une fois. **Avec remise** : elle peut être tirée plusieurs fois (chapitre 4).
- **Taille fixe** : tous les échantillons possibles ont $n$ unités. **Taille aléatoire** : $n$ dépend du tirage.
- **Probabilités égales** ou **inégales** pour les unités (chapitre 11).

## Un plan de taille aléatoire : le tirage de Bernoulli

On décide pour chaque ménage, **indépendamment**, de le garder avec probabilité $0{,}4$. La taille $n$ suit une loi binomiale $\mathcal{B}(5\,;0{,}4)$ : en moyenne $2$, mais $P(n=0) = 0{,}6^5 = 0{,}078$. Une fois sur treize, l'enquête n'interroge personne.

# Les probabilités d’inclusion

# La chance qu'a chaque unité d'être tirée

## Probabilité d'inclusion

$$\pi_k = P(k \in S) = \sum_{s \ni k} p(s)$$

On additionne les probabilités des échantillons qui **contiennent** $k$.

## Sous les deux plans de la mini-population

- Aléatoire simple : chaque ménage figure dans $4$ échantillons sur $10$, donc $\pi_k = 0{,}4 = n/N$.
- Un ménage par groupe : $\pi_A = \pi_B = \pi_C = 1/3$ et $\pi_D = \pi_E = 1/2$.

## Lire la matrice des échantillons

$\pi_k$ est la part des lignes où la colonne $k$ est cochée, chaque ligne étant pondérée par sa probabilité.

# Une propriété utile : la somme des probabilités

## La somme des probabilités d'inclusion

$$\sum_{k \in U} \pi_k = E(n) \qquad \text{et donc} \qquad \sum_{k \in U} \pi_k = n \text{ pour un plan de taille fixe}$$

## La démonstration en une ligne

$\sum_k \pi_k = \sum_k \sum_{s \ni k} p(s) = \sum_s p(s) \times (\text{nombre d'unités de } s) = \sum_s p(s)\, n(s) = E(n)$ : on intervertit les deux sommes.

## Vérification

Aléatoire simple : $5 \times 0{,}4 = 2$. Un ménage par groupe : $3 \times \frac{1}{3} + 2 \times \frac{1}{2} = 2$. Bernoulli : $5 \times 0{,}4 = 2 = E(n)$.

## Être tirés ensemble

$$\pi_{kl} = P(k \in S \text{ et } l \in S) = \sum_{s \ni k,\, l} p(s), \qquad k \neq l$$

Pour un plan aléatoire simple : $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$.

## Sur la mini-population

- Aléatoire simple : chaque paire forme exactement un échantillon, $\pi_{kl} = 1/10 = \dfrac{2 \times 1}{5 \times 4}$.
- Un ménage par groupe : $\pi_{AD} = 1/6$, mais $\pi_{AB} = 0$ et $\pi_{DE} = 0$ : deux ménages d'un même groupe ne sont **jamais** tirés ensemble.

La variance d'un estimateur dépend des $\pi_{kl}$ (chapitre 2). Quand certaines sont nulles, elle ne peut plus être estimée à partir de l'échantillon seul : ce sera le cas du tirage systématique (chapitre 8).

## Une variable aléatoire par unité

$I_k = 1$ si l'unité $k$ est tirée, $0$ sinon. C'est une variable de Bernoulli :

$$E(I_k) = \pi_k \qquad V(I_k) = \pi_k(1-\pi_k) \qquad \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\,\pi_l$$

## Le hasard est dans les I, pas dans les y

Tout estimateur s'écrit avec les $I_k$ : par exemple $\sum_{k\in S} y_k = \sum_{k\in U} I_k\, y_k$. Les $y_k$ sont des nombres fixes, les $I_k$ portent tout l'aléa. Les calculs d'espérance et de variance du cours se ramènent à ceux des $I_k$.

## Aléatoire simple sur la mini-population

$V(I_k) = 0{,}4 \times 0{,}6 = 0{,}24$ et

$$\mathrm{Cov}(I_k, I_l) = 0{,}1 - 0{,}4 \times 0{,}4 = -0{,}06$$

## Une covariance négative

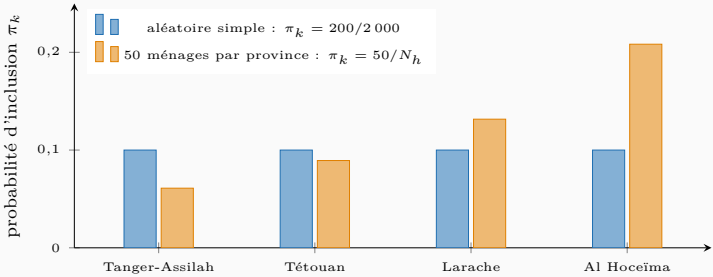
L'échantillon n'a que deux places : si A est tiré, il en reste une seule pour les quatre autres.

Être tiré **diminue** la chance des autres de l'être. En général, pour un plan aléatoire simple :

$$\mathrm{Cov}(I_k, I_l) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n\,(N-n)}{N^2\,(N-1)} < 0$$

# Sur la population du cours

# Deux plans de 200 ménages




## Même taille, autres probabilités

- **Aléatoire simple** de $200$ ménages : $\pi_k = 200/2\,000 = 0{,}1$ pour tous.
- **50 ménages par province** : $\pi_k = 50/N_h$, de $0{,}061$ à Tanger-Assilah ($N_h = 820$) à $0{,}208$ à Al Hoceïma ($N_h = 240$).

## Combien de ménages représente un ménage tiré

Le **poids** de l'unité $k$ est $d_k = 1/\pi_k$. Un ménage tiré avec la probabilité $0{,}1$ « représente » $10$ ménages de la population, lui compris.

## Les poids du plan à 50 ménages par province

| Province | Tanger-Assilah | Tétouan | Larache | Al Hoceïma |
|---|---|---|---|---|
| $N_h$ | $820$ | $560$ | $380$ | $240$ |
| Poids $d_k = N_h/50$ | $16{,}4$ | $11{,}2$ | $7{,}6$ | $4{,}8$ |

La somme des poids des $200$ ménages tirés vaut

$$50 \times (16{,}4 + 11{,}2 + 7{,}6 + 4{,}8) = 2\,000 = N.$$

## Une erreur de 411 dirhams

Avec $50$ ménages par province, la moyenne **non pondérée** des $200$ ménages vaut en espérance la moyenne des quatre moyennes provinciales :

$$\tfrac{1}{4}\,(8\,621 + 8\,146 + 7\,255 + 5\,979) = 7\,500{,}29 \text{ dh, contre } \overline{Y} = 7\,911{,}44$$

Al Hoceïma, la province la moins dépensière, pèse un quart dans l'échantillon mais $12\,\%$ dans la population.

## La règle

Dès que les $\pi_k$ ne sont pas tous égaux, il faut **pondérer**. Le chapitre 2 donne l'estimateur qui le fait automatiquement : chaque unité tirée y compte pour $d_k = 1/\pi_k$.

# S’entraîner

**Quatre boutiques**

Quatre boutiques ont un chiffre d'affaires mensuel de $2$, $5$, $8$ et $9$ milliers de dh. On en tire deux par sondage aléatoire simple.

1. Calculer $\overline{Y}$ et $S^2$.
2. Énumérer les échantillons possibles, leur probabilité et la moyenne de chacun.
3. Calculer la moyenne des moyennes. Conclure.
4. Calculer $\pi_k$ et $\pi_{kl}$ ; vérifier que $\sum_k \pi_k = n$.

## Paramètres et échantillons

1. $\overline{Y} = 24/4 = 6$ ; écarts $-4, -1, 2, 3$ ; $S^2 = (16 + 1 + 4 + 9)/3 = 10$.
2. $\binom{4}{2} = 6$ échantillons, chacun de probabilité $1/6$ : $\{2,5\} : 3{,}5$ ; $\{2,8\} : 5$ ; $\{2,9\} : 5{,}5$ ; $\{5,8\} : 6{,}5$ ; $\{5,9\} : 7$ ; $\{8,9\} : 8{,}5$.

## Espérance et probabilités d'inclusion

3. $(3{,}5 + 5 + 5{,}5 + 6{,}5 + 7 + 8{,}5)/6 = 36/6 = 6 = \overline{Y}$ : en moyenne sur les tirages, la moyenne d'échantillon retrouve la vraie moyenne.
4. Chaque boutique figure dans $3$ échantillons sur $6$ : $\pi_k = 0{,}5 = n/N$, et $\sum_k \pi_k = 4 \times 0{,}5 = 2 = n$. Chaque paire forme un seul échantillon : $\pi_{kl} = 1/6 = \dfrac{2 \times 1}{4 \times 3}$.

**Le tirage de Bernoulli**

Dans une population de $N = 6$ ménages, chaque ménage est retenu indépendamment avec la probabilité $\pi = 0{,}5$.

1. Quelle est la loi de la taille $n$ de l'échantillon ? Son espérance ?
2. Calculer $P(n = 0)$ et $P(n = 3)$.
3. Quelle est la probabilité d'inclusion de chaque ménage ? Vérifier la relation $\sum_k \pi_k = E(n)$.
4. Pourquoi les instituts préfèrent-ils les plans de taille fixe ?

## Une binomiale

1. $n$ compte les succès de $6$ épreuves indépendantes : $n \sim \mathcal{B}(6\,;0{,}5)$, d'espérance $3$.
2. $P(n=0) = 0{,}5^6 = 1/64 = 0{,}016$ ; $P(n=3) = \binom{6}{3}\,0{,}5^6 = 20/64 = 0{,}3125$.
3. $\pi_k = 0{,}5$ pour chacun ; $\sum_k \pi_k = 6 \times 0{,}5 = 3 = E(n)$.

## Le budget

4. Le coût d'une enquête est proportionnel au nombre de questionnaires : avec une taille aléatoire, ni le budget ni la précision ne sont connus à l'avance. Ici, deux fois sur trois, la taille n'est pas celle prévue : $P(n=3) = 0{,}31$ seulement.

**Cent ménages dans chaque milieu**

La population du cours compte $1\,438$ ménages urbains (dépense moyenne $8\,430$ dh) et $562$ ruraux ($6\,585$ dh). On tire $100$ ménages au hasard dans chaque milieu.

1. Calculer les probabilités d'inclusion et les poids de chaque milieu.
2. Vérifier que $\sum_k \pi_k = 200$ et que la somme des poids des ménages tirés vaut $N$.
3. Que vaut, en espérance, la moyenne non pondérée des $200$ ménages ? Quelle erreur commet-elle ?

## Probabilités et poids

1. Urbain : $\pi_k = 100/1\,438 = 0{,}0695$, poids $14{,}38$. Rural : $\pi_k = 100/562 = 0{,}178$, poids $5{,}62$.
2. $1\,438 \times \frac{100}{1\,438} + 562 \times \frac{100}{562} = 200$ ; $100 \times 14{,}38 + 100 \times 5{,}62 = 2\,000$.

## Le biais de la moyenne brute

3. Elle vaut en espérance $\frac{1}{2}(8\,430 + 6\,585) = 7\,507$ dh, contre $7\,911$ : elle sous-estime de $404$ dh. Les ruraux, $28\,\%$ de la population, font la moitié de l'échantillon. La moyenne pondérée $\frac{1\,438}{2\,000}\,\overline{y}_{\text{urb}} + \frac{562}{2\,000}\,\overline{y}_{\text{rur}}$ corrige ce déséquilibre.

# Ce qu’il faut retenir

## Sept points

1. On cherche des **paramètres** de la population : total, moyenne, proportion, ratio.
2. Une proportion est la moyenne d'une variable 0/1, de variance $S^2 = \frac{N}{N-1}P(1-P)$.
3. Les sondages emploient $S^2$, divisée par $N-1$.
4. Un **plan de sondage** est une loi de probabilité sur les échantillons possibles.
5. $\pi_k = P(k \in S)$, et $\sum_k \pi_k = n$ pour un plan de taille fixe.
6. Les indicatrices $I_k$ portent tout l'aléa : $E(I_k) = \pi_k$, $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$.
7. Quand les $\pi_k$ diffèrent, chaque unité tirée doit compter pour son **poids** $d_k = 1/\pi_k$.

## La suite

Le chapitre 2 construit, pour n'importe quel plan, un estimateur sans biais du total, et calcule sa précision.