Échantillonnage

Chapitre 1 : Population, paramètres et plan de sondage

Exercices corrigés

Jaouad Madkour
Faculté d'Économie et de Gestion de Tanger
Échantillonnage

Table des matières

Partie 1 — Énoncés 1

Partie 2 — Corrections 3

1 Exercices du chapitre

Exercice 1 — Énumérer un sondage aléatoire simple

Cinq petites entreprises emploient 2, 3, 5, 6 et 9 salariés. On en tire trois par sondage aléatoire simple sans remise, pour estimer le nombre moyen de salariés.

(a) Calculer $\overline{Y}$, σ^2 et S^2.

(b) Énumérer les échantillons possibles et la moyenne $\overline{y}$ de chacun.

(c) Calculer l'espérance de $\overline{y}$ sur les échantillons, puis sa variance (en divisant par le nombre d'échantillons). Comparer cette variance à $(1-f)\,S^2/n$, où $f = n/N$.

(d) Calculer π_k et π_{kl}, et vérifier que $\sum_k \pi_k = n$.

Exercice 2 — Un plan à probabilités inégales

Dans une population de quatre unités, un plan de taille 2 attribue aux six échantillons possibles les probabilités suivantes :

Échantillon	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$p(s)$	0,1	0,2	0,3	0,1	0,1	0,2

(a) Vérifier qu'il s'agit bien d'un plan de sondage.

(b) Calculer les quatre probabilités d'inclusion, et vérifier leur somme.

(c) Calculer π_{12}, puis $\mathrm{Cov}(I_1, I_2)$. Interpréter son signe.

(d) Quels poids d_k faudrait-il attribuer aux unités tirées ? Laquelle « représente » le plus d'unités de la population ?

Exercice 3 — Les indicatrices d'appartenance

On tire $n = 4$ unités parmi $N = 10$ par sondage aléatoire simple.

(a) Calculer $E(I_k)$, $V(I_k)$ et π_{kl} pour $k \neq l$.

(b) Calculer $\mathrm{Cov}(I_k, I_l)$ de deux façons : par $\pi_{kl} - \pi_k \pi_l$, puis par la formule $-n(N-n)/(N^2(N-1))$.

(c) La taille de l'échantillon s'écrit $n = \sum_{k \in U} I_k$. Calculer $V\left(\sum_k I_k\right)$ à partir des variances et des covariances des I_k. Le résultat était-il prévisible ?

2 Exercice cumulatif (introduction et chapitre 1)

Exercice 4 — Un plan provincial et une base incomplète

On reprend la population du cours : 2 000 ménages répartis en quatre provinces de 820, 560, 380 et 240 ménages, de dépenses moyennes respectives 8 621, 8 146, 7 255 et 5 979 dh ($\overline{Y} = 7\,911{,}44$ dh).

(a) Un enquêteur tire 50 ménages au hasard dans chaque province. Calculer les probabilités d'inclusion, les poids, et vérifier que $\sum_k \pi_k = 200$.

(b) Que vaut, en espérance, la moyenne non pondérée des 200 ménages ? Quelle erreur commet-elle ?

(c) La base de sondage disponible omet les quatre districts ruraux d'Al Hoceïma, soit 134 ménages dont la dépense moyenne est de 5 520 dh. Quelle est la probabilité d'inclusion de ces ménages ? Quelle est la moyenne des 1 866 ménages couverts par la base ? Quel type d'erreur cela introduit-il, et un échantillon plus grand le corrigerait-il ?

3 Applications en sciences économiques

Exercice 5 — Une enquête auprès des entreprises

Une chambre de commerce compte 2 400 très petites entreprises (TPE), 500 PME et 100 grandes entreprises. Leur chiffre d'affaires annuel moyen est de 0,8, 12 et 150 millions de dh. Elle interroge 60 TPE, 50 PME et 50 grandes entreprises, tirées au hasard dans chaque catégorie.

(a) Calculer la probabilité d'inclusion et le poids d'une entreprise de chaque catégorie.

(b) Calculer le chiffre d'affaires moyen des 3 000 entreprises. Que vaut, en espérance, la moyenne non pondérée des 160 entreprises interrogées ?

(c) Pourquoi a-t-on interrogé la moitié des grandes entreprises et seulement 2,5 % des TPE ?

Exercice 6 — L'échantillon de l'enquête emploi

L'enquête nationale sur l'emploi interroge chaque année environ 90 000 ménages, dont 30 000 ruraux. Le recensement de 2024 compte 6 173 930 ménages urbains et 3 101 108 ménages ruraux.

(a) Calculer la probabilité d'inclusion moyenne d'un ménage urbain et d'un ménage rural sur une année, et les poids correspondants.

(b) Un échantillon est dit « autopondéré » lorsque toutes ses unités ont le même poids. L'est-il ici, approximativement ?

(c) Quel avantage pratique présente un échantillon autopondéré pour celui qui exploite les données ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $\overline{Y} = 25/5 = 5$. Écarts : $-3, -2, 0, 1, 4$; somme des carrés $9+4+0+1+16 = 30$. D'où $\sigma^2 = 30/5 = 6$ et $S^2 = 30/4 = 7{,}5$.
(b) $\binom{5}{3} = 10$ échantillons, chacun de probabilité $1/10$:

s	2,3,5	2,3,6	2,3,9	2,5,6	2,5,9	2,6,9	3,5,6	3,5,9	3,6,9	5,6,9
$\overline{y}$	3,33	3,67	4,67	4,33	5,33	5,67	4,67	5,67	6	6,67

(c) La somme des dix moyennes vaut 50 : $E(\overline{y}) = 5 = \overline{Y}$. Les écarts à 5 sont $-\frac{5}{3}, -\frac{4}{3}, -\frac{1}{3}, -\frac{2}{3}, \frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, \frac{2}{3}$, $1, \frac{5}{3}$; leurs carrés somment à $90/9 = 10$, d'où $V(\overline{y}) = 10/10 = 1$. Et $(1-f)\,S^2/n = (1-0{,}6)\times 7{,}5/3 = 1$: la formule du chapitre 3 donne exactement le résultat de l'énumération.
(d) Chaque entreprise figure dans $\binom{4}{2} = 6$ échantillons sur 10 : $\pi_k = 0{,}6 = n/N$, et $\sum_k \pi_k = 5 \times 0{,}6 = 3 = n$. Chaque paire figure dans $\binom{3}{1} = 3$ échantillons : $\pi_{kl} = 0{,}3 = \dfrac{3\times 2}{5\times 4}$.

Correction — Exercice 2

(a) Les six probabilités sont positives et leur somme vaut $0{,}1 + 0{,}2 + 0{,}3 + 0{,}1 + 0{,}1 + 0{,}2 = 1$.
(b) $\pi_1 = 0{,}1+0{,}2+0{,}3 = 0{,}6$; $\pi_2 = 0{,}1+0{,}1+0{,}1 = 0{,}3$; $\pi_3 = 0{,}2+0{,}1+0{,}2 = 0{,}5$; $\pi_4 = 0{,}3+0{,}1+0{,}2 = 0{,}6$. Somme : $2 = n$.
(c) $\pi_{12} = p(\{1,2\}) = 0{,}1$ et $\mathrm{Cov}(I_1, I_2) = 0{,}1 - 0{,}6\times 0{,}3 = -0{,}08$. Négative : l'échantillon n'ayant que deux places, la présence de l'unité 1 réduit la chance que l'unité 2 soit tirée avec elle.
(d) $d_1 = 1/0{,}6 = 1{,}67$; $d_2 = 1/0{,}3 = 3{,}33$; $d_3 = 2$; $d_4 = 1{,}67$. L'unité 2, la moins souvent tirée, représente le plus d'unités : quand elle est tirée, elle doit compter davantage.

Correction — Exercice 3

(a) $E(I_k) = \pi_k = 0{,}4$; $V(I_k) = 0{,}4\times 0{,}6 = 0{,}24$; $\pi_{kl} = \dfrac{4\times 3}{10\times 9} = 0{,}1333$.
(b) $0{,}1333 - 0{,}4^2 = -0{,}0267$, et $-\dfrac{4\times 6}{100\times 9} = -0{,}0267$: les deux calculs concordent.
(c) $V\left(\sum_k I_k\right) = \sum_k V(I_k) + \sum_{k\neq l} \mathrm{Cov}(I_k, I_l) = 10\times 0{,}24 + 90\times(-0{,}0267) = 2{,}4 - 2{,}4 = 0$. C'était prévisible : la taille d'un sondage aléatoire simple vaut toujours 4, elle n'a aucune variabilité. Les covariances négatives compensent exactement les variances.

2 Exercice cumulatif (introduction et chapitre 1)

Correction — Exercice 4

(a) $\pi_k = 50/N_h$: 0,061 (Tanger-Assilah), 0,089 (Tétouan), 0,132 (Larache), 0,208 (Al Hoceïma) ; poids 16,4, 11,2, 7,6 et 4,8. $\sum_k \pi_k = 820\times\frac{50}{820} + 560\times\frac{50}{560} + 380\times\frac{50}{380} + 240\times\frac{50}{240} = 4\times 50 = 200$.
(b) Chaque province fournit un quart de l'échantillon : l'espérance de la moyenne non pondérée est la moyenne simple des quatre moyennes, $\frac{1}{4}(8\,621 + 8\,146 + 7\,255 + 5\,979) = 7\,500$ dh. Elle sous-estime $\overline{Y}$ de 411 dh, soit 5,2 %.
(c) Ces 134 ménages ont une probabilité d'inclusion **nulle**. La moyenne des ménages couverts vaut $(15\,822\,870 - 134\times 5\,519{,}70)/1\,866 = 8\,083$ dh : une erreur de **couverture** de +172 dh. Elle ne dépend pas de la taille de l'échantillon : seule une base complète la supprime.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) TPE : $\pi = 60/2\,400 = 0{,}025$, poids 40. PME : $\pi = 50/500 = 0{,}1$, poids 10. Grandes : $\pi = 50/100 = 0{,}5$, poids 2.
(**b**) Moyenne vraie : $(2\,400 \times 0{,}8 + 500 \times 12 + 100 \times 150)/3\,000 = 22\,920/3\,000 = 7{,}64$ millions de dh. Moyenne non pondérée attendue : $(60 \times 0{,}8 + 50 \times 12 + 50 \times 150)/160 = 8\,148/160 = 50{,}9$ millions : près de sept fois trop. Les grandes entreprises forment 31 % de l'échantillon, mais 3 % de la population.
(**c**) Les 100 grandes entreprises réalisent 15 000 millions sur 22 920, soit 65 % du chiffre d'affaires total : une erreur sur elles fausse tout le total. On les interroge donc beaucoup plus, quitte à corriger ensuite par les poids. C'est l'idée de l'allocation optimale (chapitre 7).

Correction — Exercice 6

(**a**) Urbain : $60\,000/6\,173\,930 = 0{,}00972$, poids 102,9. Rural : $30\,000/3\,101\,108 = 0{,}00967$, poids 103,4.
(**b**) Oui, à 0,5 % près : chaque ménage interrogé représente environ 103 ménages, qu'il soit urbain ou rural. La répartition de l'échantillon entre milieux est proportionnelle à celle de la population.
(**c**) Dans un échantillon autopondéré, les moyennes et proportions non pondérées sont sans biais : un utilisateur qui oublie les poids ne se trompe pas. Les tableaux s'interprètent directement.