Échantillonnage

Chapitre 10 : Le sondage à deux degrés

Exercices corrigés

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

Table des matières

Partie 1 — Énoncés 1

Partie 2 — Corrections 4

1 Exercices du chapitre

Exercice 1 — Six villages

Une commune rurale compte $N_I = 6$ villages de 30, 50, 20, 60, 40 et 40 ménages ($N = 240$). On tire trois villages par sondage aléatoire simple, puis quatre ménages par sondage aléatoire simple dans chaque village tiré. Les villages 2, 4 et 5 ont été tirés ; la dépense mensuelle des ménages interrogés vaut, en milliers de dh :

Village tiré	Ménages M_i	Dépenses des quatre ménages interrogés
2	50	9, 11, 8, 12
4	60	6, 8, 5, 9
5	40	7, 5, 8, 4

(a) Calculer la probabilité d'inclusion et le poids d'un ménage de chacun des six villages. Vérifier que $\sum_{k\in U} \pi_k$ est égale à la taille de l'échantillon.

(b) Estimer le total Y des dépenses de deux façons : par les totaux estimés $\widehat{Y}_i = M_i \overline{y}_i$, puis par les poids d_k. En déduire l'estimation $\widehat{Y}/N$ de la dépense moyenne.

(c) Calculer $\widehat{N}$ et l'estimateur par le ratio $\widehat{Y}/\widehat{N}$. Pourquoi diffère-t-il autant de $\widehat{Y}/N$? La moyenne simple des douze ménages interrogés est-elle un estimateur sans biais ?

(d) Estimer la variance de $\widehat{Y}$ par ses deux termes, puis son écart-type et son coefficient de variation. Quelle part de la variance estimée vient du premier degré ?

(e) Estimer l'écart-type de l'estimateur par le ratio de la moyenne, et le comparer à celui de $\widehat{Y}/N$.

Exercice 2 — Comment répartir six entretiens

Un lotissement compte $N_I = 5$ quartiers de $M = 6$ ménages ($N = 30$). Les dépenses moyennes des quartiers valent 5, 7, 8, 9 et 11 milliers de dh, et la variance (en $M-1$) des dépenses à l'intérieur de chaque quartier vaut $S_i^2 = 4$. On veut interroger six ménages et estimer la dépense moyenne par $\widehat{Y}/N$.

(a) Calculer $\overline{Y}$, la variance S_T^2 des cinq totaux de quartier, et S^2 (décomposer la somme des carrés en une partie interne aux quartiers et une partie entre quartiers).

(b) Calculer la variance de $\widehat{Y}/N$ et l'effet de plan pour trois plans à deux degrés : un quartier entier (1×6), deux quartiers et trois ménages dans chacun (2×3), trois quartiers et deux ménages dans chacun (3×2). Comparer à un sondage aléatoire simple de six ménages.

(c) Visiter un quartier coûte $c_1 = 100$ dh, interroger un ménage $c_2 = 20$ dh. Un sondage aléatoire simple de six ménages touche en moyenne 3,87 quartiers. Calculer le coût de chaque plan, puis le produit de sa variance par son coût. Pourquoi ce produit permet-il de comparer les plans ? Lequel choisir ?

(d) Expliquer pourquoi le plan 3×2 l'emporte ici sur le sondage aléatoire simple, alors qu'il est moins précis.

Exercice 3 — De l'effet de plan au nombre de ménages par grappe

Une enquête antérieure a interrogé des grappes entières de 25 ménages ; l'effet de plan de la dépense y valait 3,4. Une nouvelle enquête, à deux degrés, est prévue dans la même province : visiter une grappe coûte $c_1 = 2\,000$ dh, interroger un ménage $c_2 = 80$ dh, et le budget est de 128 000 dh. On néglige $1-f$.

(a) Déduire de l'enquête antérieure la corrélation intra-grappe ρ, par la formule $deff = 1 + (M-1)\rho$.

(b) Calculer le nombre optimal m^* de ménages par grappe, puis le nombre de grappes, le nombre de ménages, l'effet de plan et la taille effective.

(c) Mêmes calculs si l'on interroge, comme avant, les 25 ménages de chaque grappe. De combien la variance augmente-t-elle ? Commenter.

(d) Dans une ville de la province, visiter une grappe ne coûte que $c_1 = 500$ dh. Que devient m^* ?

Interpréter.

2 Exercice cumulatif (chapitres 10, 9, 8, 7 et 6)

Exercice 4 — Sept plans, deux terrains

Sur la population du cours ($N = 2\,000$ ménages, 50 districts), les chapitres précédents ont calculé la variance de l'estimation de la dépense moyenne pour plusieurs plans d'environ 200 ménages :

Plan (chapitre)	V
Aléatoire simple (3)	71 435,5
Stratifié par province, allocation proportionnelle (6)	68 139,7
Stratifié par province, allocation de Neyman (7)	67 172,3
Systématique dans l'ordre du fichier (8)	30 287,0
Cinq districts entiers, estimateur par le ratio (9)	598 829,9
Dix districts et vingt ménages par district, ratio (10)	308 755,1
Vingt districts et dix ménages par district, ratio (10)	158 974,1

(a) Calculer pour chaque plan l'effet de plan, la taille effective et la demi-largeur de l'intervalle de confiance à 95 % ($1{,}96\sqrt{V}$).

(b) Visiter un district coûte $c_1 = 800$ dh, interroger un ménage $c_2 = 60$ dh. Les 200 ménages des quatre premiers plans se répartissent en moyenne sur 48,2 districts. Calculer le coût de chaque plan, puis le produit $V \times C$, et classer les plans.

(c) Dans une grande région, les 200 ménages des quatre premiers plans seraient presque tous dans des districts différents. Refaire le classement avec 200 districts visités pour ces plans, en supposant les effets de plan inchangés.

(d) Conclure : quand le sondage à deux degrés est-il avantageux, et comment répartir alors les entretiens ?

3 Applications en sciences économiques

Exercice 5 — Le bénéfice des épiceries de quartier

Une ville compte $N_I = 80$ quartiers et, d'après un recensement des commerces, $N = 9\,600$ épiceries (chiffres fictifs). Pour estimer leur bénéfice mensuel total, on tire huit quartiers par sondage aléatoire simple, on dénombre les épiceries de chacun, puis on interroge dix épiceries tirées au hasard dans chaque quartier tiré. Bénéfice mensuel, en milliers de dh :

Quartier tiré	1	2	3	4	5	6	7	8
Épiceries M_i	140	95	210	60	125	180	75	110
Moyenne $\overline{y}_i$	4,2	3,1	5,6	2,4	3,8	4,9	2,7	3,5
Écart-type s_i	1,8	1,2	2,4	1,0	1,5	2,1	1,1	1,4

(a) Calculer la probabilité d'inclusion et le poids d'une épicerie du quartier 1, puis du quartier 4. Le plan est-il autopondéré ?

(b) Estimer le bénéfice total $\widehat{Y}$, le nombre d'épiceries $\widehat{N}$ et le bénéfice moyen par le ratio $\widehat{R} = \widehat{Y}/\widehat{N}$. Comparer $\widehat{N}$ à N, et en déduire un second estimateur du total, $N\widehat{R}$.

(c) Estimer la variance de $\widehat{Y}$ par ses deux termes. Quelle part vient du premier degré ? Donner le coefficient de variation et l'intervalle de confiance à 95 % ($t_7 = 2{,}365$).

(d) Pour $N\widehat{R}$, le premier terme se calcule sur les écarts $\widehat{Y}_i - \widehat{R}\,M_i$ au lieu des $\widehat{Y}_i$; le second ne change pas. Estimer son écart-type, son coefficient de variation et son intervalle de confiance. Pourquoi

est-il tellement plus précis ?

Exercice 6 — Mesurer le chômage d'une ville

Un observatoire régional prépare une enquête sur l'emploi dans une ville de 20 000 logements, répartis en $N_I = 400$ îlots d'environ 50 logements (chiffres fictifs). Il veut estimer la proportion P des logements où vit au moins un chômeur, voisine de 0,12. Visiter un îlot (cartographie et dénombrement des logements) coûte $c_1 = 600$ dh, interroger un logement $c_2 = 90$ dh, et la corrélation intra-îlot de la variable vaut $\rho = 0{,}04$. Le budget est de 90 000 dh. On néglige $1 - f$, et l'on prend pour variance $deff \times P(1-P)/n$.

(a) Calculer le nombre optimal m^* de logements à interroger par îlot.

(b) Avec $m = 12$: combien d'îlots le budget permet-il ? Calculer le nombre de logements interrogés, l'effet de plan, la taille effective et la marge d'erreur à 95 %. La comparer à la marge que l'on calculerait en oubliant l'effet de plan.

(c) Mêmes questions si l'on interroge tous les logements des îlots tirés ($m = 50$). Conclure.

(d) Les îlots sont tirés par sondage aléatoire simple, puis 12 logements dans chacun. Calculer la probabilité d'inclusion d'un logement d'un îlot de 30 logements, puis de 80 logements. Proposer deux façons de rendre le plan autopondéré, et comparer la charge de travail des enquêteurs dans ces deux îlots.

1 Exercices du chapitre

Correction — Exercice 1

(a) $\pi_k = \dfrac{n_I}{N_I}\cdot\dfrac{m}{M_i} = \dfrac{3}{6}\cdot\dfrac{4}{M_i} = \dfrac{2}{M_i}$:

Village	1	2	3	4	5	6
M_i	30	50	20	60	40	40
π_k	0,0667	0,04	0,1	0,0333	0,05	0,05
Poids $d_k = M_i/2$	15	25	10	30	20	20

Chaque village apporte $M_i \times 2/M_i = 2$ à la somme : $\sum_{k\in U}\pi_k = 6\times 2 = 12$, soit 3 villages de 4 ménages. Les ménages des grands villages ont moins de chances d'être tirés.
(b) Moyennes : $\overline{y}_2 = 40/4 = 10$, $\overline{y}_4 = 28/4 = 7$, $\overline{y}_5 = 24/4 = 6$. Totaux estimés : $\widehat{Y}_2 = 50\times 10 = 500$, $\widehat{Y}_4 = 60\times 7 = 420$, $\widehat{Y}_5 = 40\times 6 = 240$. D'où $\widehat{Y} = \frac{6}{3}\times(500+420+240) = 2\times 1\,160 = 2\,320$ milliers de dh. Par les poids : $25\times 40 + 30\times 28 + 20\times 24 = 1\,000 + 840 + 480 = 2\,320$: c'est le même estimateur. Dépense moyenne : $\widehat{Y}/N = 2\,320/240 = 9{,}667$ milliers de dh.
(c) $\widehat{N} = 2\times(50+60+40) = 300$ ménages, alors que $N = 240$; ratio : $1\,160/150 = 7{,}733$ milliers de dh. Le tirage a ramené trois grands villages (150 ménages, contre $3\times 40 = 120$ en moyenne) : $\widehat{Y}$ surestime le total parce qu'il surestime le nombre de ménages ; le ratio corrige cette taille. La moyenne simple vaut $92/12 = 7{,}667$; elle donne le même poids à chaque ménage interrogé, alors que les poids vont de 10 à 30 : elle n'est pas sans biais. Elle est proche du ratio ici, mais par hasard : le ratio pèse chaque village selon M_i, la moyenne simple selon le nombre d'entretiens, le même partout.
(d) La moyenne des trois $\widehat{Y}_i$ vaut 386,67 ; écarts 113,33, 33,33, −146,67 ; $s_1^2 = 35\,466{,}7/2 = 17\,733{,}3$. Premier terme : $6^2\times(1-\frac{1}{2})\times 17\,733{,}3/3 = 106\,400$. Dans chaque village tiré, les écarts à la moyenne sont ± 1 et ± 2 : $s_i^2 = 10/3$. Second terme : $\frac{6}{3}\times[50^2\times\frac{46}{50}+60^2\times\frac{56}{60}+40^2\times\frac{36}{40}]\times\dfrac{10/3}{4} = 2\times(1\,916{,}7+2\,800+1\,200) = 11\,833{,}3$. D'où $\widehat{V}(\widehat{Y}) = 118\,233{,}3$, un écart-type de 343,9 milliers de dh et un coefficient de variation de $343{,}9/2\,320 = 14{,}8\,\%$. Le premier degré fait 90 % de la variance estimée.
(e) Écarts $\widehat{Y}_i - 7{,}733\,M_i$: 113,33 ; −44 ; −69,33 ; leur variance vaut 9 793,8. Premier terme : $36\times 0{,}5\times 9\,793{,}8/3 = 58\,762{,}7$; avec le second, inchangé : 70 596. Écart-type du ratio : $\sqrt{70\,596}/240 = 1{,}107$ millier de dh, contre $343{,}9/240 = 1{,}433$ pour $\widehat{Y}/N$. Le ratio est plus précis : il ne subit pas la variation des tailles des villages.

Correction — Exercice 2

(a) Quartiers de même taille : $\overline{Y} = (5+7+8+9+11)/5 = 8$. Totaux : 30, 42, 48, 54, 66, de moyenne 48 ; écarts −18, −6, 0, 6, 18 ; $S_T^2 = 720/4 = 180$. Somme des carrés interne : $5\times(6-1)\times 4 = 100$; entre quartiers : $6\times(9+1+0+1+9) = 120$. D'où $S^2 = 220/29 = 7{,}586$.
(b) $V(\widehat{Y}) = N_I^2(1-f_1)\dfrac{S_T^2}{n_I} + \dfrac{N_I}{n_I}N_I M^2\Big(1-\dfrac{m}{M}\Big)\dfrac{4}{m}$, puis on divise par $N^2 = 900$:

Plan	V_1	V_2	$V(\widehat{Y}/N)$	*deff*
1×6	$25\times 0{,}8\times 180 = 3\,600$	0	4	3,955
2×3	$25\times 0{,}6\times 90 = 1\,350$	$2{,}5\times 5\times 36\times 0{,}5\times\frac{4}{3} = 300$	1,833	1,812
3×2	$25\times 0{,}4\times 60 = 600$	$\frac{5}{3}\times 5\times 36\times\frac{2}{3}\times 2 = 400$	1,111	1,098
Aléatoire simple			1,0115	1

Pour le sondage aléatoire simple de six ménages, $V(\overline{y}) = (1-6/30)\times 7{,}586/6 = 0{,}8\times 7{,}586/6 = 1{,}0115$. À six entretiens fixés, plus on visite de quartiers, plus la variance baisse : V_1 diminue vite, V_2 augmente peu.
(c) Coûts : 1×6 : $100+6\times 20 = 220$ dh ; 2×3 : 320 dh ; 3×2 : 420 dh ; aléatoire simple : $3{,}87\times 100+120 = 506{,}7$ dh. Produits $V\times C$: 880 ; 586,7 ; 466,7 ; 512,5. Avec un budget B, on peut répéter le plan B/C fois

(ou l'agrandir dans cette proportion) ; la variance est alors divisée par B/C et vaut $V \times C/B$: à budget égal, le meilleur plan est celui dont $V \times C$ est le plus petit. On choisit 3×2.
(**d**) Le sondage aléatoire simple est un peu plus précis (1,0115 contre 1,111), mais il coûte 21 % de plus, car il visite presque quatre quartiers pour six entretiens. Le plan 3×2 garde l'essentiel de la dispersion entre quartiers dans l'échantillon (trois quartiers sur cinq) tout en économisant les déplacements. Le plan 1×6 est le pire : toute l'estimation repose sur un seul quartier, alors que les moyennes vont de 5 à 11.

Correction — Exercice 3

(**a**) $3{,}4 = 1 + 24\rho$, d'où $\rho = 2{,}4/24 = 0{,}1$.
(**b**) $m^* = \sqrt{\dfrac{2\,000}{80} \times \dfrac{0{,}9}{0{,}1}} = \sqrt{25 \times 9} = 15$ ménages. Une grappe coûte $2\,000 + 15 \times 80 = 3\,200$ dh : $128\,000/3\,200 = 40$ grappes et 600 ménages. $deff = 1 + 14 \times 0{,}1 = 2{,}4$; taille effective $600/2{,}4 = 250$.
(**c**) Une grappe entière coûte $2\,000 + 25 \times 80 = 4\,000$ dh : 32 grappes, 800 ménages, $deff = 3{,}4$, taille effective $800/3{,}4 = 235{,}3$. La variance est multipliée par $250/235{,}3 = 1{,}0625$: 6,25 % de plus, pour 200 entretiens supplémentaires. L'optimum est plat : interroger les 25 ménages ne coûte pas très cher en précision, mais le budget des entretiens en trop aurait pu payer des grappes.
(**d**) $m^* = \sqrt{\dfrac{500}{80} \times 9} = \sqrt{56{,}25} = 7{,}5$, soit 7 ou 8 ménages. Quand le déplacement coûte moins cher, on multiplie les grappes et l'on interroge moins de ménages dans chacune : m^* varie comme la racine carrée de c_1/c_2, et il est divisé par 2 quand c_1 est divisé par 4.

2 Exercice cumulatif (chapitres 10, 9, 8, 7 et 6)

Correction — Exercice 4

(**a**) $deff = V/71\,435{,}5$, taille effective $200/deff$, demi-largeur $1{,}96\sqrt{V}$:

Plan	*deff*	Taille effective	Demi-largeur (dh)
Aléatoire simple	1	200	523,9
Stratifié, proportionnelle	0,954	209,7	511,6
Stratifié, Neyman	0,940	212,7	508,0
Systématique	0,424	471,7	341,1
Cinq districts entiers, ratio	8,383	23,9	1 516,7
10×20, ratio	4,322	46,3	1 089,1
20×10, ratio	2,225	89,9	781,5

(**b**) Les quatre premiers plans coûtent $48{,}24 \times 800 + 200 \times 60 = 50\,594$ dh ; cinq districts entiers : $5 \times 800 + 12\,000 = 16\,000$ dh ; 10×20 : 20 000 dh ; 20×10 : 28 000 dh. Produits $V \times C$, en milliards : systématique 1,53 ; Neyman 3,40 ; proportionnelle 3,45 ; aléatoire simple 3,61 ; 20×10 : 4,45 ; 10×20 : 6,18 ; cinq districts entiers : 9,58. Sur la population du cours, les plans à un seul degré l'emportent : avec 50 districts seulement, un échantillon de 200 ménages en touche de toute façon 48, et les déplacements économisés ne compensent pas l'effet de grappe.
(**c**) Les quatre premiers plans coûtent maintenant $200 \times (800 + 60) = 172\,000$ dh. Produits $V \times C$: aléatoire simple 12,29 ; proportionnelle 11,72 ; Neyman 11,55 ; systématique 5,21. Le classement s'inverse : 20×10 (4,45) passe en tête, devant le systématique (5,21) et 10×20 (6,18) ; même les cinq districts entiers (9,58) battent l'aléatoire simple.
(**d**) Le sondage à deux degrés est avantageux quand le territoire est vaste et que chaque lieu visité coûte cher : il paie l'effet de grappe par l'économie des déplacements. Il faut alors répartir les entretiens sur assez de districts, avec peu de ménages dans chacun : 20×10 fait mieux que 10×20, conformément à la règle $m^* = \sqrt{\frac{c_1}{c_2} \cdot \frac{1-\rho}{\rho}} = 8{,}3$ pour ces coûts et $\rho = 0{,}162$. Et l'on estime la moyenne par le ratio.

3 Applications en sciences économiques

Correction — Exercice 5

(a) $f_1 = 8/80 = 0{,}1$. Quartier 1 : $\pi_k = 0{,}1 \times 10/140 = 0{,}00714$, poids 140. Quartier 4 : $\pi_k = 0{,}1 \times 10/60 = 0{,}0167$, poids 60. Le poids d'une épicerie interrogée vaut $10 \times M_i/10 = M_i$: il varie d'un quartier à l'autre, le plan n'est pas autopondéré.
(b) $\widehat{Y}_i = M_i \overline{y}_i$: 588 ; 294,5 ; 1 176 ; 144 ; 475 ; 882 ; 202,5 ; 385, de somme 4 147. $\widehat{Y} = 10 \times 4\,147 = 41\,470$ milliers de dh, soit 41,5 millions de dh par mois. $\widehat{N} = 10 \times 995 = 9\,950$ épiceries, pour $N = 9\,600$: les quartiers tirés sont un peu plus grands que la moyenne (124,4 épiceries contre 120). $\widehat{R} = 4\,147/995 = 4{,}168$ milliers de dh par épicerie ; $N\widehat{R} = 9\,600 \times 4{,}1678 = 40\,011$ milliers de dh.
(c) Écart-type des huit $\widehat{Y}_i$: 354,4. Premier terme : $80^2 \times 0{,}9 \times 125\,609{,}3/8 = 90\,438\,724$. Second terme : $10 \times \sum M_i^2 (1 - \frac{10}{M_i}) \frac{s_i^2}{10} = 10 \times 51\,026{,}4 = 510\,264$ (le quartier 3 y contribue pour près de la moitié). $\widehat{V}(\widehat{Y}) = 90\,948\,989$; écart-type 9 536,7 ; le premier degré fait 99,4 % de la variance. Coefficient de variation : $9\,536{,}7/41\,470 = 23\,\%$. Intervalle : $41\,470 \pm 2{,}365 \times 9\,536{,}7 = 41\,470 \pm 22\,554$, soit [18 916 ; 64 024] milliers de dh.
(d) Écarts $\widehat{Y}_i - 4{,}1678\, M_i$: 4,5 ; −101,4 ; 300,8 ; −106,1 ; −46,0 ; 131,8 ; −110,1 ; −73,5 ; variance 21 287,7. Premier terme : $80^2 \times 0{,}9 \times 21\,287{,}7/8 = 15\,327\,113$; avec le second : 15 837 377, écart-type 3 979,6, coefficient de variation 10 %. Intervalle : $40\,011 \pm 2{,}365 \times 3\,979{,}6 = 40\,011 \pm 9\,412$, soit [30 599 ; 49 423]. La variance est divisée par 5,7. Le total d'un quartier dépend surtout de son nombre d'épiceries (corrélation 0,991 entre $\widehat{Y}_i$ et M_i) ; $\widehat{Y}$ subit cette variation des tailles, alors que $N\widehat{R}$ la neutralise grâce au nombre total d'épiceries, connu par le recensement.

Correction — Exercice 6

(a) $m^* = \sqrt{\dfrac{600}{90} \times \dfrac{0{,}96}{0{,}04}} = \sqrt{6{,}67 \times 24} = \sqrt{160} = 12{,}6$: 12 ou 13 logements par îlot.
(b) Un îlot coûte $600 + 12 \times 90 = 1\,680$ dh ; $90\,000/1\,680 = 53{,}6$, donc 53 îlots (89 040 dh) et 636 logements. $deff = 1 + 11 \times 0{,}04 = 1{,}44$; taille effective $636/1{,}44 = 441{,}7$. Marge : $1{,}96\sqrt{1{,}44 \times 0{,}12 \times 0{,}88/636} = 1{,}96 \times 0{,}0155 = 0{,}030$, soit $\pm 3{,}0$ points. En oubliant l'effet de plan : $1{,}96\sqrt{0{,}12 \times 0{,}88/636} = 0{,}025$, soit $\pm 2{,}5$ points : une précision annoncée trop optimiste de 20 % ($\sqrt{1{,}44} = 1{,}2$).
(c) Un îlot entier coûte $600 + 50 \times 90 = 5\,100$ dh : 17 îlots (86 700 dh), 850 logements, $deff = 1 + 49 \times 0{,}04 = 2{,}96$, taille effective 287,2, marge $1{,}96\sqrt{2{,}96 \times 0{,}1056/850} = 0{,}038$, soit $\pm 3{,}8$ points. On interroge 214 logements de plus, mais la précision est moindre : les logements d'un même îlot se ressemblent, même avec un ρ de 0,04, et 49ρ n'est plus négligeable.
(d) $\pi_k = \dfrac{53}{400} \cdot \dfrac{12}{M_i}$: 0,053 (poids 18,9) pour un îlot de 30 logements, 0,0199 (poids 50,3) pour un îlot de 80. Le plan n'est pas autopondéré. Deux façons de l'être :

- tirer dans chaque îlot la même **fraction** $f_2 = 12/50 = 0{,}24$ de ses logements : $\pi_k = 0{,}1325 \times 0{,}24 = 0{,}0318$, poids 31,4 ; mais l'enquêteur interroge 7 logements dans l'îlot de 30 et 19 dans celui de 80, et la taille totale de l'échantillon dépend des îlots tirés ;
- tirer les îlots avec des probabilités **proportionnelles à leur taille**, puis 12 logements dans chacun : $\pi_k = 53 \times 12/20\,000 = 0{,}0318$, le même poids 31,4, et 12 entretiens dans chaque îlot. C'est le plan du chapitre 11 : autopondéré et à charge de travail égale.