Échantillonnage

Chapitre 9 : Le sondage par grappes

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 — Quatre hameaux

Une commune rurale compte quatre hameaux. Le tableau donne, pour chacun, le nombre de ménages M_i et leur dépense mensuelle totale Y_i, en milliers de dh. On tire deux hameaux par sondage aléatoire simple, et l'on interroge tous leurs ménages.

Hameau	1	2	3	4
Ménages M_i	2	4	6	8
Dépense totale Y_i	10	24	42	64

(a) Calculer N, Y, $\overline{Y}$ et la variance S_T^2 des quatre totaux.

(b) Énumérer les six échantillons possibles ; pour chacun, calculer $\widehat{Y}_{HT}/N$ et l'estimateur par le ratio $\overline{y}_R = \sum Y_i / \sum M_i$.

(c) Calculer l'espérance et la variance de $\widehat{Y}_{HT}/N$, puis retrouver cette variance par la formule du cours :

$$V(\widehat{Y}_{HT}/N) = \frac{1-f_I}{n_I\,\overline{M}^2}\,S_T^2$$

(d) Calculer l'espérance, le biais et l'erreur quadratique moyenne de $\overline{y}_R$. Quel estimateur préférer ? Pourquoi ?

Exercice 2 — Des grappes homogènes ou hétérogènes

Douze ménages ont les dépenses suivantes, en milliers de dh : 2, 3, 4, 4, 5, 6, 6, 7, 8, 8, 9, 10. On les répartit en $N_I = 4$ grappes de $M = 3$ ménages, de deux façons :

- regroupement 1, par voisinage : $\{2,3,4\}$, $\{4,5,6\}$, $\{6,7,8\}$, $\{8,9,10\}$;
- regroupement 2 : $\{2,6,10\}$, $\{3,7,8\}$, $\{4,5,9\}$, $\{4,6,8\}$.

On tire $n_I = 2$ grappes par sondage aléatoire simple. Pour des grappes de même taille, on rappelle que $deff = M\,S_b^2/S^2$, où S_b^2 est la variance (en $N_I - 1$) des moyennes de grappes.

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

(b) Pour le regroupement 1, calculer les moyennes des grappes, S_b^2 et l'effet de plan. Que valent les six ménages interrogés, en ménages tirés au hasard ?

(c) Même question pour le regroupement 2. Que vaut $V(\overline{y})$? Interpréter.

(d) Calculer $\rho = \dfrac{CM_{\text{inter}} - CM_{\text{intra}}}{CM_{\text{inter}} + (M-1)\,CM_{\text{intra}}}$ pour chaque regroupement, avec $CM_{\text{inter}} = M\,S_b^2$ et CM_{intra} la moyenne des variances (en $M-1$) à l'intérieur des grappes. Commenter le signe et la valeur obtenue pour le regroupement 2.

(e) Le statisticien préférerait des grappes comme celles du regroupement 2. Pourquoi le terrain lui donne-t-il plutôt celles du regroupement 1 ?

Exercice 3 — Six districts d'une commune

Une commune urbaine compte $N_I = 120$ districts et $N = 4\,800$ ménages. On tire six districts par sondage aléatoire simple, et l'on relève dans chacun le nombre de ménages et leur dépense mensuelle totale, en milliers de dh :

District tiré	1	2	3	4	5	6
Ménages M_i	32	45	38	51	29	47
Dépense totale Y_i	230	380	290	440	190	360

(a) Calculer $\widehat{Y}_{HT}$, l'estimation $\widehat{Y}_{HT}/N$ de la dépense moyenne par ménage, et l'estimateur par le

ratio $\overline{y}_R$.

(b) Estimer l'écart-type de chacun des deux estimateurs de la moyenne.

(c) Construire l'intervalle de confiance à 95 % fondé sur le ratio ($t_5 = 2{,}571$).

(d) Pourquoi le ratio est-il ici beaucoup plus précis ?

2 Exercice cumulatif (chapitres 9, 8, 7, 6 et 5)

Exercice 4 — Cinq plans pour deux cents ménages

Sur la population du cours ($N = 2\,000$ ménages), les chapitres précédents ont calculé, pour $n = 200$ ménages et la dépense, les effets de plan suivants :

Plan	*deff*
Aléatoire simple (chapitre 3)	1
Stratifié par province, allocation proportionnelle (chapitre 6)	0,954
Stratifié par province, allocation de Neyman (chapitre 7)	0,940
Systématique dans l'ordre du fichier (chapitre 8)	0,424
Cinq districts entiers, estimateur par le ratio (chapitre 9)	8,383

(a) Calculer la taille effective $n/deff$ de chaque plan, et les classer.

(b) Expliquer en une phrase pourquoi chacun des quatre derniers plans gagne ou perd par rapport à l'aléatoire simple.

(c) Pour estimer la dépense moyenne à ± 500 dh près (niveau 95 %), le chapitre 5 a trouvé $n_0 = 243{,}9$, puis 218 ménages par aléatoire simple. Combien de ménages faut-il avec des districts entiers et l'estimateur par le ratio (multiplier n_0 par *deff*, puis corriger pour la population finie) ? Combien de districts de 40 ménages ? Le chapitre 5 avait trouvé 24 districts avec $deff \approx 7{,}32$: d'où vient l'écart ?

(d) Le plan systématique paraît le meilleur. Quel inconvénient le chapitre 8 lui a-t-il reconnu ?

3 Applications en sciences économiques

Exercice 5 — La production de blé d'une province

Une province compte $N_I = 80$ douars ; le nombre total de ses exploitations agricoles n'est pas connu. Pour estimer la production de blé de la campagne, on tire huit douars par sondage aléatoire simple et l'on recense, dans chacun, le nombre d'exploitations M_i et leur production totale Y_i en tonnes (chiffres fictifs) :

Douar tiré	1	2	3	4	5	6	7	8
Exploitations M_i	24	41	18	35	52	29	33	20
Production Y_i (t)	55	98	31	77	130	60	81	40

(a) Estimer la production totale de la province, $\widehat{Y}_{HT}$, et le nombre total d'exploitations, $\widehat{N}_{HT}$.

(b) Estimer la production moyenne par exploitation. Pourquoi ne peut-on pas utiliser $\widehat{Y}_{HT}/N$?

(c) Estimer l'écart-type de $\widehat{Y}_{HT}$, puis celui de l'estimateur par le ratio (dans la formule, remplacer $\overline{M}$, inconnu, par la taille moyenne $\overline{m}$ des douars tirés). Comparer les coefficients de variation.

(d) Donner l'intervalle de confiance à 95 % de la production totale ($t_7 = 2{,}365$). Le ministère souhaite une précision de ± 10 % sur le total : est-elle atteinte ? Quelle information permettrait de l'améliorer fortement ?

Exercice 6 — Les commerçants des souks hebdomadaires

Une enquête régionale veut connaître le revenu des commerçants des souks hebdomadaires (chiffres fictifs). Atteindre un souk coûte $c_1 = 1\,200$ dh (déplacement et installation d'une équipe), interroger un commerçant $c_2 = 40$ dh. Un souk compte environ 60 commerçants, répartis en quatre allées de 15. La corrélation intra-grappe du revenu vaut $\rho = 0{,}1$; on admet qu'elle est la même pour un souk entier et pour une allée. Le budget est de 72 000 dh, et la région compte beaucoup de souks : on néglige $1 - f$.

(a) Calculer l'effet de plan si l'on interroge tout le souk, puis si l'on n'interroge qu'une allée par souk visité.

(b) Calculer, dans les deux cas, le coût par commerçant interrogé, puis le coût d'un commerçant « effectif » (coût par commerçant multiplié par l'effet de plan).

(c) Combien de commerçants effectifs le budget permet-il dans chaque cas ? Et avec un sondage aléatoire simple de commerçants, où chaque entretien demande son propre déplacement ?

(d) Conclure. Quelle idée du chapitre suivant cet exercice annonce-t-il ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $N = 2+4+6+8 = 20$ ménages ; $Y = 10+24+42+64 = 140$ milliers de dh ; $\overline{Y} = 140/20 = 7$ milliers de dh. La moyenne des quatre totaux vaut 35 ; écarts $-25, -11, 7, 29$; somme des carrés $625+121+49+841 = 1\,636$, d'où $S_T^2 = 1\,636/3 = 545{,}33$.
(b) $\binom{4}{2} = 6$ échantillons, chacun de probabilité $1/6$. Avec $N_I/n_I = 2$, $\widehat{Y}_{HT}/N = 2\sum Y_i/20 = \sum Y_i/10$:

Hameaux tirés	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$\sum Y_i$	34	52	74	66	88	106
$\sum M_i$	6	8	10	10	12	14
$\widehat{Y}_{HT}/N$	3,4	5,2	7,4	6,6	8,8	10,6
$\overline{y}_R$	5,67	6,5	7,4	6,6	7,33	7,57

(c) $E(\widehat{Y}_{HT}/N) = 42/6 = 7 = \overline{Y}$: sans biais. Écarts à 7 : $-3{,}6 ; -1{,}8 ; 0{,}4 ; -0{,}4 ; 1{,}8 ; 3{,}6$; carrés $12{,}96+3{,}24+0{,}16+0{,}16+3{,}24+12{,}96 = 32{,}72$, d'où $V = 32{,}72/6 = 5{,}453$. Formule, avec $f_I = 1/2$ et $\overline{M} = 20/4 = 5$: $(1-\frac{1}{2}) \times 545{,}33/(2\times 25) = 5{,}453$.
(d) $E(\overline{y}_R) = 41{,}07/6 = 6{,}845$: biais $-0{,}155$ millier de dh. Écarts à 7 : $-1{,}33 ; -0{,}5 ; 0{,}4 ; -0{,}4 ; 0{,}33 ; 0{,}57$; d'où $EQM = 2{,}785/6 = 0{,}464$. L'erreur quadratique du ratio est 11,7 fois plus petite que la variance de Horvitz-Thompson : on préfère le ratio malgré son biais. Les moyennes des hameaux sont proches (5, 6, 7, 8), alors que leurs tailles vont de 2 à 8 : $\widehat{Y}_{HT}/N$ dépend surtout de la taille des hameaux tirés (3,4 pour les deux plus petits, 10,6 pour les deux plus grands), le ratio corrige cette taille.

Correction — Exercice 2

(a) $\overline{Y} = 72/12 = 6$. Écarts : $-4, -3, -2, -2, -1, 0, 0, 1, 2, 2, 3, 4$; somme des carrés 68 ; $S^2 = 68/11 = 6{,}18$.
(b) Moyennes des grappes : 3, 5, 7, 9 ; $S_b^2 = (9+1+1+9)/3 = 6{,}67$; $deff = 3\times 6{,}67/6{,}18 = 3{,}24$. Les six ménages interrogés ne valent que $6/3{,}24 = 1{,}85$ ménage tiré au hasard : chaque grappe réunit des voisins presque identiques.
(c) Les quatre moyennes valent 6 : $S_b^2 = 0$, $deff = 0$ et $V(\overline{y}) = 0$. Tout échantillon de deux grappes donne exactement $\overline{Y}$: chaque grappe est un « modèle réduit » de la population.
(d) Regroupement 1 : $CM_{\text{inter}} = 3\times 6{,}67 = 20$; les quatre variances internes valent 1, donc $CM_{\text{intra}} = 1$; $\rho = (20-1)/(20+2\times 1) = 19/22 = 0{,}86$. Regroupement 2 : $CM_{\text{inter}} = 0$; variances internes 16, 7, 7 et 4, donc $CM_{\text{intra}} = 8{,}5$; $\rho = -8{,}5/(2\times 8{,}5) = -0{,}5$. C'est la valeur **minimale** $-1/(M-1)$: les ménages d'une même grappe sont aussi différents que possible. On vérifie $1+(M-1)\rho = 1-2\times 0{,}5 = 0 = deff$. Pour le regroupement 1, $1+2\rho = 1+38/22 = 2{,}73$ au lieu de 3,24 : la formule de Kish est approchée, et avec quatre grappes seulement l'écart se voit.
(e) Sur le terrain, les grappes sont des lieux : un quartier, un douar, un immeuble. On ne choisit pas qui y habite, et les voisins se ressemblent (revenu, logement, accès aux services). Les grappes réelles ressemblent donc au regroupement 1 : $\rho > 0$, et $deff > 1$.

Correction — Exercice 3

(a) $\sum Y_i = 1\,890$ milliers de dh et $\sum M_i = 242$. $\widehat{Y}_{HT} = \frac{120}{6}\times 1\,890 = 37\,800$ milliers de dh, soit 37,8 millions ; $\widehat{Y}_{HT}/N = 37\,800\,000/4\,800 = 7\,875$ dh. Ratio : $\overline{y}_R = 1\,890\,000/242 = 7\,810$ dh.
(b) $f_I = 6/120 = 0{,}05$ et $\overline{M} = 4\,800/120 = 40$.

- Horvitz-Thompson : $s_T^2 = 9\,070$ (milliers de dh)^2, d'où $\widehat{V}(\widehat{Y}_{HT}/N) = \dfrac{0{,}95}{6\times 1\,600}\times 9\,070 = 0{,}898$ et un écart-type de 0,947 millier, soit 947 dh.
- Ratio : résidus $\hat{e}_i = Y_i - 7{,}810\,M_i$: $-19{,}9 ; 28{,}6 ; -6{,}8 ; 41{,}7 ; -36{,}5 ; -7{,}1$; $s_e^2 = 875{,}5$, d'où $\widehat{V}(\overline{y}_R) = \dfrac{0{,}95}{9\,600}\times 875{,}5 = 0{,}0866$ et un écart-type de 294 dh.

(**c**) $7\,810 \pm 2{,}571 \times 294{,}3 = 7\,810 \pm 757$, soit [7 053 ; 8 567] dh. On emploie la loi de Student à $n_I - 1 = 5$ degrés de liberté : la variance est estimée sur six grappes, pas sur 242 ménages.
(**d**) Les totaux des districts vont de 190 à 440 milliers, surtout parce que leurs tailles vont de 29 à 51 ménages : la corrélation entre Y_i et M_i vaut 0,99 dans l'échantillon. Les moyennes par ménage, elles, ne vont que de 6 552 à 8 627 dh. Horvitz-Thompson subit la variation des tailles ; le ratio ne garde que celle des moyennes, et sa variance est 10,4 fois plus petite.

2 Exercice cumulatif (chapitres 9, 8, 7, 6 et 5)

Correction — Exercice 4

(**a**) Tailles effectives $200/\mathit{deff}$:

Plan	*deff*	Taille effective
Systématique dans l'ordre du fichier	0,424	471,7
Stratifié, allocation de Neyman	0,940	212,8
Stratifié, allocation proportionnelle	0,954	209,6
Aléatoire simple	1	200
Cinq districts entiers, ratio	8,383	23,9

(**b**) Les deux plans stratifiés éliminent la variance **entre provinces** ; Neyman gagne un peu plus en tirant davantage dans les provinces les plus dispersées. Le systématique, dans un fichier rangé par province et par district, répartit l'échantillon dans tous les districts : c'est une stratification implicite très fine. Les grappes font l'inverse : elles concentrent l'échantillon dans cinq districts dont les ménages se ressemblent ($\rho = 0{,}162$), et dont les tailles varient.
(**c**) $n'_0 = 8{,}383 \times 243{,}9 \approx 2\,045$; $n = 2\,045/(1 + 2\,045/2\,000) = 1\,011{,}1$, soit 1 012 ménages. Avec 40 ménages par district : $1\,011{,}1/40 = 25{,}3$, donc 26 districts entiers (1 040 ménages en moyenne). Avec la formule de variance du ratio, 26 districts donnent une marge de $\pm$486 dh, et 25 districts $\pm$506 dh : il en faut bien 26. L'écart avec les 24 districts du chapitre 5 vient des **tailles inégales** : la formule $1 + (M - 1)\rho = 7{,}32$ suppose des districts de 40 ménages exactement ; l'effet de plan réel du ratio, 8,38, compte en plus la variation des tailles (chaque district pèse M_i^2 dans S_e^2).
(**d**) Une seule origine aléatoire : de nombreuses paires de ménages ont $\pi_{kl} = 0$, et la variance ne peut pas être estimée sans biais à partir de l'échantillon. Son efficacité dépend de l'ordre du fichier, et une périodicité de la liste égale au pas peut la ruiner.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) $\sum Y_i = 572$ t et $\sum M_i = 252$. Avec $N_I/n_I = 10$: $\widehat{Y}_{HT} = 5\,720$ t et $\widehat{N}_{HT} = 2\,520$ exploitations.
(**b**) $\overline{y}_R = 572/252 = 2{,}270$ t par exploitation, soit $\widehat{Y}_{HT}/\widehat{N}_{HT}$. On ne peut pas diviser $\widehat{Y}_{HT}$ par N, puisque N est inconnu : le ratio n'en a pas besoin.
(**c**) $f_I = 8/80 = 0{,}1$. Horvitz-Thompson : $s_T^2 = 1\,040{,}3$, $\widehat{V}(\widehat{Y}_{HT}) = 80^2 \times 0{,}9 \times 1\,040{,}3/8 = 749\,006$, écart-type 865,5 t, coefficient de variation $865{,}5/5\,720 = 15{,}1\,\%$. Ratio : $\overline{m} = 252/8 = 31{,}5$; résidus $\hat{e}_i = Y_i - 2{,}270\,M_i$: 0,52 ; 4,94 ; −9,86 ; −2,44 ; 11,97 ; −5,83 ; 6,10 ; −5,40 ; $s_e^2 = 53{,}03$; $\widehat{V}(\overline{y}_R) = \dfrac{0{,}9}{8 \times 31{,}5^2} \times 53{,}03 = 0{,}00601$, écart-type 0,078 t, coefficient de variation 3,4 %. Le rendement par exploitation est connu bien plus précisément que la production totale.
(**d**) $5\,720 \pm 2{,}365 \times 865{,}5 = 5\,720 \pm 2\,046$, soit [3 674 ; 7 766] t : une précision de $\pm$36 %, loin des $\pm$10 % ($\pm$572 t) demandés. Le total des douars dépend surtout de leur nombre d'exploitations (leur corrélation vaut ici 0,99). Si l'on connaissait le nombre total d'exploitations N, par un recensement agricole, on

estimerait le total par $N\,\overline{y}_R$: pour N voisin de 2 520, son écart-type serait d'environ 195 t, et l'intervalle de $\pm$462 t, soit $\pm$8 %. C'est l'estimateur par le ratio du chapitre 12.

Correction — Exercice 6

(**a**) Souk entier : $deff = 1 + 59 \times 0{,}1 = 6{,}9$. Une allée : $deff = 1 + 14 \times 0{,}1 = 2{,}4$.
(**b**) Souk entier : $1\,200 + 60 \times 40 = 3\,600$ dh par souk, soit 60 dh par commerçant, et $60 \times 6{,}9 = 414$ dh par commerçant effectif. Une allée : $1\,200 + 15 \times 40 = 1\,800$ dh par souk, soit 120 dh par commerçant, et $120 \times 2{,}4 = 288$ dh par commerçant effectif.
(**c**) Souk entier : $72\,000/3\,600 = 20$ souks, 1 200 commerçants, taille effective $1\,200/6{,}9 = 174$. Une allée : $72\,000/1\,800 = 40$ souks, 600 commerçants, taille effective $600/2{,}4 = 250$. Aléatoire simple : 1 240 dh par commerçant, $72\,000/1\,240 = 58$ commerçants.
(**d**) Interroger une seule allée dans deux fois plus de souks donne la meilleure précision : 250 commerçants effectifs, contre 174 pour les souks entiers et 58 pour l'aléatoire simple. Le commerçant interrogé coûte deux fois plus cher, mais il apporte bien plus d'information, car il ressemble moins aux autres. L'exercice annonce le sondage à deux degrés (chapitre 10) : tirer des souks, puis **une partie** des commerçants de chaque souk tiré, et choisir ce nombre selon les coûts et ρ.