Échantillonnage

Chapitre 11 : Les probabilités inégales

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 — Énumérer un tirage proportionnel à la taille

Trois exploitations agricoles ont une surface de 2, 3 et 5 hectares, et ont produit 3, 6 et 11 tonnes de céréales. Pour estimer la production totale Y, on tire deux exploitations **avec remise**, avec des probabilités proportionnelles à la surface, et l'on emploie l'estimateur de Hansen-Hurwitz $\widehat{Y}_{HH} = \frac{1}{2}(z_1 + z_2)$, où $z_j = y_{(j)}/p_{(j)}$.

(a) Calculer les p_i et les $z_i = y_i/p_i$. Vérifier que $E(z) = \sum_i p_i z_i = Y$.

(b) Énumérer les six résultats possibles du tirage (sans tenir compte de l'ordre), leur probabilité et la valeur de $\widehat{Y}_{HH}$. Calculer $E(\widehat{Y}_{HH})$ et $V(\widehat{Y}_{HH})$, puis retrouver cette variance par la formule $V(z)/n$.

(c) Pour chaque résultat, calculer la variance estimée $\widehat{V} = (z_1 - z_2)^2/4$. Vérifier qu'elle est sans biais.

(d) Comparer $V(\widehat{Y}_{HH})$ à la variance de l'estimateur du total pour deux tirages **à probabilités égales** : avec remise (estimateur $\frac{1}{2}(3\,y_{(1)} + 3\,y_{(2)})$), puis sans remise ($N\,\overline{y}$). Expliquer l'écart.

Exercice 2 — Quatre îlots et un tirage systématique

Un quartier compte quatre îlots, rangés dans l'ordre de la liste ci-dessous. On en tire deux par tirage systématique proportionnel au nombre de ménages, et l'on enquête tous leurs ménages.

Îlot	1	2	3	4
Ménages M_i	10	20	30	40
Dépense mensuelle totale Y_i (milliers de dh)	80	200	240	320

(a) Calculer les tailles cumulées, le pas k et les probabilités d'inclusion π_i.

(b) Pour chacun des k départs possibles $r = 1, \dots, k$, quels îlots sont tirés ? En déduire les échantillons possibles, leur probabilité et la valeur de $\widehat{Y}_{HT}$ pour chacun.

(c) Calculer $E(\widehat{Y}_{HT})$ et $V(\widehat{Y}_{HT})$. Retrouver cette variance par la formule de Sen-Yates-Grundy.

(d) Comparer à deux tirages proportionnels à la taille **avec remise** (Hansen-Hurwitz), puis à un sondage aléatoire simple de deux îlots.

(e) Peut-on estimer $V(\widehat{Y}_{HT})$ sans biais à partir de l'échantillon ? Pourquoi ?

Exercice 3 — Un plan autopondéré à deux degrés

Une commune compte $N = 6\,000$ ménages répartis en 150 districts. On tire 6 districts avec remise, avec des probabilités proportionnelles à leur nombre de ménages M_i, puis 10 ménages par sondage aléatoire simple dans chaque district tiré. Les dépenses mensuelles moyennes des six groupes de 10 ménages sont, en dh :

$$6\,850 \qquad 8\,120 \qquad 7\,430 \qquad 9\,260 \qquad 7\,880 \qquad 6\,540$$

(a) Montrer que chaque ménage de la commune a la même chance $n_I \frac{M_i}{N} \times \frac{m}{M_i}$ ($n_I = 6, m = 10$) d'être enquêté (en ignorant les répétitions). Quel est son poids ?

(b) Estimer la dépense moyenne par ménage et la dépense totale de la commune.

(c) Estimer la variance de la moyenne estimée, son erreur-type, et donner un intervalle de confiance à 95 % (loi de Student à 5 degrés de liberté : $t = 2{,}571$).

(d) Pourquoi calcule-t-on la variance sur les 6 moyennes de district, et non sur les 60 dépenses des ménages enquêtés ?

(e) Un district compte 35 ménages. Quelle est sa probabilité d'être tiré à un tirage donné ? D'être tiré au moins une fois ?

2 Exercice cumulatif (chapitres 11, 10, 9, 8 et 7)

Exercice 4 — Dix districts de la population du cours

La population du cours compte $N = 2\,000$ ménages en $N_I = 50$ districts de 13 à 71 ménages. On donne : la variance des totaux de district $S_T^2 = 20\,446\,316\,150$; $\sum_i \frac{M_i}{N}(\overline{Y}_i - \overline{Y})^2 = 2\,844\,644$; $\sum_i M_i^2/N = 44{,}528$; $S^2 = 15\,874\,548$. L'enquête doit porter sur 10 districts.

(a) (*Chapitre 9.*) Les 10 districts sont tirés par sondage aléatoire simple et entièrement enquêtés. Calculer la variance et l'erreur-type de l'estimateur de Horvitz-Thompson de la dépense moyenne, $\widehat{Y}_{HT}/N$. Combien de ménages enquête-t-on en moyenne ?

(b) (*Chapitre 11.*) Les 10 districts sont tirés avec remise, proportionnellement à leur taille, et entièrement enquêtés. Calculer la variance et l'erreur-type de l'estimateur de Hansen-Hurwitz de la moyenne, et le nombre moyen de ménages enquêtés.

(c) (*Chapitre 7.*) Calculer l'effet de plan de chacun des deux plans, en le rapportant à un sondage aléatoire simple de même nombre moyen de ménages, et la taille effective correspondante. Commenter.

(d) (*Chapitre 10.*) Le chapitre 10 tirait 10 districts par sondage aléatoire simple, puis 20 ménages dans chacun : écart-type 1 031,9 dh, *deff* = 14,9, dont un premier degré de $4{,}089 \cdot 10^{12}$ sur la variance du total. Relier ce premier degré au résultat de (a). On tire maintenant les 10 districts selon leur taille, avec remise, puis 20 ménages dans chacun (les trois districts de moins de 20 ménages sont enquêtés en entier) : le second degré vaut 34 664 sur la variance de la moyenne. Calculer l'écart-type et l'effet de plan par rapport au sondage aléatoire simple de 200 ménages ($V_{SAS} = 71\,435{,}5$).

(e) (*Chapitre 8.*) On tire plutôt les 10 districts par tirage systématique proportionnel à la taille, dans l'ordre du fichier, avec un pas de 200. Les 200 départs donnent 45 échantillons distincts, et 780 des 1 225 paires de districts ne sont jamais tirées ensemble. Peut-on estimer la variance sans biais ? Sur les 200 départs, la variance exacte de la moyenne vaut 324 138, et la variance calculée comme avec remise, s^2/n_I, vaut en moyenne 280 056. Commenter.

3 Applications en sciences économiques

Exercice 5 — La production d'olives d'une province

Une province compte 6 000 exploitations oléicoles réparties dans des douars dont on connaît, par le dernier recensement agricole, le nombre d'exploitations M_i. Pour estimer la production totale d'olives, on tire 8 douars avec remise, proportionnellement à M_i, et l'on mesure la production Y_i de toutes leurs exploitations. Les chiffres sont fictifs.

Douar tiré	1	2	3	4	5	6	7	8
Exploitations M_i	40	30	60	50	20	70	40	30
Production Y_i (tonnes)	96	66	162	105	50	182	92	57

(a) Combien de fois le douar de 70 exploitations avait-il plus de chances d'être tiré que celui de 20 ?

(b) Montrer que $z_j = Y_i/p_i = 6\,000 \times Y_i/M_i$. Estimer la production totale de la province.

(c) Estimer l'erreur-type et le coefficient de variation de cette estimation, et donner un intervalle de confiance à 95 % (Student à 7 degrés de liberté : $t = 2{,}365$).

(d) Un collègue propose $6\,000 \times \sum Y_i / \sum M_i$. Que vaut cette estimation ? Pourquoi est-elle mal adaptée à ce plan ?

(e) L'an prochain, on visitera seulement 5 exploitations par douar tiré. Quelle sera la chance d'être enquêtée de chaque exploitation (en ignorant les répétitions) ? Comment s'estimera la production totale ?

Exercice 6 — Une mesure de taille périmée

Une zone industrielle emploie $X = 10\,000$ salariés d'après le registre des entreprises. Pour estimer la valeur ajoutée annuelle totale de la zone, on tire 4 entreprises avec remise, proportionnellement à leur nombre de salariés *inscrit au registre*. Les chiffres sont fictifs.

Entreprise tirée	A	B	C	D
Salariés au registre	120	80	250	10
Valeur ajoutée (millions de dh)	6,0	4,4	12,0	3,0

(a) Calculer les p_i, les z_i, et l'estimation de Hansen-Hurwitz de la valeur ajoutée totale.

(b) Estimer son erreur-type et son coefficient de variation.

(c) Quelle entreprise fait l'essentiel de l'estimation ? Calculer la valeur ajoutée par salarié de chacune, et expliquer.

(d) L'enquêteur constate que l'entreprise D emploie en réalité 60 salariés : elle s'est agrandie depuis son inscription. L'estimation de (a) est-elle biaisée ? Que deviendrait-elle avec une mesure de taille à jour (on négligera le changement de X) ? Quelles précautions prendre ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $X = 10$ hectares : $p_1 = 0{,}2$, $p_2 = 0{,}3$, $p_3 = 0{,}5$. D'où $z_1 = 3/0{,}2 = 15$, $z_2 = 6/0{,}3 = 20$, $z_3 = 11/0{,}5 = 22$, et $E(z) = 0{,}2 \times 15 + 0{,}3 \times 20 + 0{,}5 \times 22 = 3 + 6 + 11 = 20 = Y$.
(b) Un résultat $\{i, j\}$ avec $i \neq j$ a la probabilité $2\,p_i p_j$ (deux ordres possibles), un résultat $\{i, i\}$ la probabilité p_i^2 :

Résultat	$\{1,1\}$	$\{1,2\}$	$\{1,3\}$	$\{2,2\}$	$\{2,3\}$	$\{3,3\}$
Probabilité	0,04	0,12	0,20	0,09	0,30	0,25
$\widehat{Y}_{HH}$	15	17,5	18,5	20	21	22
$\widehat{V} = (z_1 - z_2)^2/4$	0	6,25	12,25	0	1	0

$E(\widehat{Y}_{HH}) = 0{,}6 + 2{,}1 + 3{,}7 + 1{,}8 + 6{,}3 + 5{,}5 = 20 = Y$: l'estimateur est sans biais. $E(\widehat{Y}_{HH}^2) = 9 + 36{,}75 + 68{,}45 + 36 + 132{,}3 + 121 = 403{,}5$, donc $V(\widehat{Y}_{HH}) = 403{,}5 - 20^2 = 3{,}5$. Par la formule : $V(z) = 0{,}2 \times 5^2 + 0{,}3 \times 0^2 + 0{,}5 \times 2^2 = 5 + 0 + 2 = 7$, et $V(z)/2 = 3{,}5$.
(c) $E(\widehat{V}) = 0{,}12 \times 6{,}25 + 0{,}20 \times 12{,}25 + 0{,}30 \times 1 = 0{,}75 + 2{,}45 + 0{,}3 = 3{,}5 = V(\widehat{Y}_{HH})$: la variance estimée est sans biais. Elle vaut 0 quand la même exploitation est tirée deux fois.
(d) À probabilités égales avec remise, $z_i = 3\,y_i = 9, 18, 33$: $V(z) = \frac{1}{3}(11^2 + 2^2 + 13^2) = 98$ et $V = 98/2 = 49$. Sans remise : $S^2 = 32{,}67/2 = 16{,}33$ et $V(N\overline{y}) = 3^2 \times (1 - \frac{2}{3}) \times 16{,}33/2 = 24{,}5$. Le tirage proportionnel à la surface est 7 fois plus précis que le sondage sans remise à probabilités égales, et 14 fois plus que le tirage avec remise à probabilités égales : les rendements y_i/x_i (1,5 ; 2 ; 2,2 tonnes par hectare) varient peu, alors que les productions varient beaucoup.

Correction — Exercice 2

(a) Cumuls : 10, 30, 60, 100. Pas $k = 100/2 = 50$. $\pi_i = 2\,M_i/100$: 0,2 ; 0,4 ; 0,6 ; 0,8 ; tous inférieurs à 1 car $M_i \leq 50$.
(b) Les points sont r et $r + 50$. Pour $r = 1$ à 10 : îlots 1 et 3 ($r + 50 \in]30\,;60]$). Pour $r = 11$ à 30 : îlots 2 et 4. Pour $r = 31$ à 50 : îlots 3 et 4. Les valeurs Y_i/π_i sont 400, 500, 400, 400 :

Échantillon	$\{1,3\}$	$\{2,4\}$	$\{3,4\}$
Probabilité	$10/50 = 0{,}2$	$20/50 = 0{,}4$	$20/50 = 0{,}4$
$\widehat{Y}_{HT}$	800	900	800

On vérifie $\pi_3 = 0{,}2 + 0{,}4 = 0{,}6$ et $\pi_4 = 0{,}4 + 0{,}4 = 0{,}8$.
(c) $E(\widehat{Y}_{HT}) = 0{,}2 \times 800 + 0{,}4 \times 900 + 0{,}4 \times 800 = 840 = Y$. $V(\widehat{Y}_{HT}) = 0{,}2 \times 40^2 + 0{,}4 \times 60^2 + 0{,}4 \times 40^2 = 320 + 1\,440 + 640 = 2\,400$. Sen-Yates-Grundy : $\pi_{13} = 0{,}2$, $\pi_{24} = \pi_{34} = 0{,}4$, les autres π_{kl} sont nuls. Seules les paires où $Y_k/\pi_k \neq Y_l/\pi_l$ comptent, celles qui contiennent l'îlot 2 :

$$V = (0{,}08 - 0) \times 100^2 + (0{,}24 - 0) \times 100^2 + (0{,}32 - 0{,}4) \times 100^2 = 800 + 2\,400 - 800 = 2\,400.$$

(d) Avec remise : $p_i = 0{,}1\,;0{,}2\,;0{,}3\,;0{,}4$; $z_i = 800, 1\,000, 800, 800$; $V(z) = 0{,}1 \times 40^2 + 0{,}2 \times 160^2 + 0{,}3 \times 40^2 + 0{,}4 \times 40^2 = 6\,400$, d'où $V(\widehat{Y}_{HH}) = 3\,200$. Sondage aléatoire simple : $S^2 = 30\,000/3 = 10\,000$ et $V = 4^2 \times 0{,}5 \times 10\,000/2 = 40\,000$. Le tirage selon la taille divise la variance par près de 17 ; sans remise, il gagne encore un quart sur le tirage avec remise.
(e) Non : les îlots 1 et 2, 1 et 4, 2 et 3 ne sont jamais tirés ensemble ($\pi_{kl} = 0$), et l'estimateur de Sen-Yates-Grundy divise par π_{kl}. On calcule en pratique la variance comme si le tirage était avec remise.

Correction — Exercice 3

(**a**) Un district est tiré en moyenne $6\,M_i/6\,000$ fois ; un ménage d'un district tiré est retenu avec la probabilité $10/M_i$. Le produit vaut $\frac{6\times10}{6\,000} = 0{,}01$ pour **tous** les ménages : le plan est autopondéré, et chaque ménage enquêté a le poids 100.
(**b**) Moyenne simple des six moyennes : $46\,080/6 = 7\,680$ dh. Total : $6\,000 \times 7\,680 = 46\,080\,000$ dh, soit 46,08 millions de dh par mois.
(**c**) Écarts à 7 680 : -830 ; 440 ; -250 ; 1 580 ; 200 ; $-1\,140$. Somme des carrés 4 781 000, donc $s_b^2 = 4\,781\,000/5 = 956\,200$ et $\widehat{V} = 956\,200/6 = 159\,366{,}7$; erreur-type 399,2 dh. Intervalle : $7\,680 \pm 2{,}571 \times 399{,}2 = 7\,680 \pm 1\,026{,}4$, soit [6 653,6 ; 8 706,4] dh. Pour le total : [39,92 ; 52,24] millions de dh.
(**d**) Les six moyennes sont indépendantes et de même loi : leur variance contient à la fois l'écart entre districts et l'aléa du choix des 10 ménages. Les 60 ménages ne sont pas indépendants : ceux d'un même district se ressemblent. Les traiter comme un sondage aléatoire simple de 60 ménages sous-estimerait la variance.
(**e**) À chaque tirage, $p = 35/6\,000 = 0{,}00583$. Probabilité d'être tiré au moins une fois en six tirages : $1 - (1 - 0{,}00583)^6 = 0{,}0345$, à peine moins que $6\,p = 0{,}035$.

2 Exercice cumulatif (chapitres 11, 10, 9, 8 et 7)

Correction — Exercice 4

(**a**) $V(\widehat{Y}_{HT}/N) = \dfrac{N_I^2\,(1 - n_I/N_I)\,S_T^2}{n_I\,N^2} = \dfrac{2\,500 \times 0{,}8 \times 20\,446\,316\,150}{10 \times 4\,000\,000} = 1\,022\,316$; erreur-type 1 011,1 dh. On enquête en moyenne $10 \times 40 = 400$ ménages.
(**b**) $V = 2\,844\,644/10 = 284\,464$; erreur-type 533,4 dh. Chaque tirage apporte en moyenne $\sum_i M_i^2/N = 44{,}5$ ménages : 445,3 ménages au total. La variance est divisée par 3,6 : seuls comptent les écarts entre *moyennes* de district, et non plus les écarts de taille.
(**c**) Sondage aléatoire simple de 400 ménages : $V = (1 - 0{,}2) \times 15\,874\,548/400 = 31\,749{,}1$, d'où $deff = 1\,022\,316/31\,749{,}1 = 32{,}2$ et $n_{eff} = 400/32{,}2 = 12{,}4$ ménages. De 445,3 ménages : $V = (1 - 0{,}2226) \times 15\,874\,548/445{,}3 = 27\,713{,}4$, d'où $deff = 10{,}26$ et $n_{eff} = 43{,}4$. Les grappes entières restent très coûteuses en précision, mais le tirage selon la taille divise l'effet de plan par trois. (Avec 5 districts, le chapitre 9 trouvait déjà $deff = 32{,}2$: le facteur $1 - f$ est le même pour les districts et pour les ménages.)
(**d**) Le premier degré du chapitre 10 est la variance du tirage des 10 districts, celle de (a) exprimée sur le total : $2\,000^2 \times 1\,022\,316 = 4{,}089 \cdot 10^{12}$. Selon la taille, le premier degré devient celui de (b) : $V = 284\,464 + 34\,664 = 319\,128$, erreur-type 564,9 dh, $deff = 319\,128/71\,435{,}5 = 4{,}47$ contre 14,9. Les 200 ménages valent 45 ménages tirés au hasard au lieu de 13. Le premier degré fait encore 89 % de la variance.
(**e**) Non : 780 paires sur 1 225, soit 64 %, ont $\pi_{kl} = 0$, et l'estimateur sans biais de la variance divise par π_{kl}. Ici, l'ordre du fichier ne fait pas gagner : la variance exacte, 324 138 (écart-type 569,3 dh), dépasse même celle du tirage avec remise, 284 464. Et la variance calculée comme avec remise la **sous-estime** de 14 % en moyenne. Le chapitre 8 l'a montré : la précision d'un tirage systématique dépend de l'ordre de la liste, et l'approximation n'est prudente que si cet ordre regroupe des unités semblables.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) $p_i = M_i/6\,000 : 70/6\,000 = 0{,}01167$ contre $20/6\,000 = 0{,}00333$, soit 3,5 fois plus de chances.
(**b**) $z_j = Y_i/(M_i/6\,000) = 6\,000 \times Y_i/M_i$. Les productions par exploitation Y_i/M_i valent 2,4 ; 2,2 ; 2,7 ; 2,1 ; 2,5 ; 2,6 ; 2,3 ; 1,9 tonnes, de moyenne $18{,}7/8 = 2{,}3375$. Production totale estimée : $6\,000 \times 2{,}3375 = 14\,025$ tonnes.
(**c**) Écarts à 2,3375 : 0,0625 ; $-0{,}1375$; 0,3625 ; $-0{,}2375$; 0,1625 ; 0,2625 ; $-0{,}0375$; $-0{,}4375$. Somme

des carrés 0,49875, $s^2 = 0{,}49875/7 = 0{,}07125$, et $\widehat{V}(\widehat{Y}_{HH}) = 6\,000^2 \times 0{,}07125/8$; erreur-type $6\,000 \times \sqrt{0{,}0089063} = 6\,000 \times 0{,}0944 = 566{,}2$ tonnes, soit un coefficient de variation de 4,0 %. Intervalle : $14\,025 \pm 2{,}365 \times 566{,}2 = 14\,025 \pm 1\,339$, soit [12 686 ; 15 364] tonnes.
(d) $\sum Y_i / \sum M_i = 810/340 = 2{,}3824$, d'où 14 294 tonnes. Le plan tire déjà plus souvent les grands douars ; les pondérer une seconde fois par M_i leur donne un poids proportionnel à M_i^2. Cette moyenne pondérée est biaisée sous ce plan ; la moyenne simple des Y_i/M_i est l'estimateur sans biais.
(e) $8 \times \frac{M_i}{6\,000} \times \frac{5}{M_i} = \frac{40}{6\,000} = \frac{1}{150}$ pour chaque exploitation : le plan est autopondéré, de poids 150. La production totale s'estimera par 150 fois la production des 40 exploitations enquêtées, soit 6 000 fois leur production moyenne.

Correction — Exercice 6

(a) $p_A = 0{,}012$, $p_B = 0{,}008$, $p_C = 0{,}025$, $p_D = 0{,}001$. $z_A = 6/0{,}012 = 500$, $z_B = 550$, $z_C = 480$, $z_D = 3/0{,}001 = 3\,000$. Estimation : $(500 + 550 + 480 + 3\,000)/4 = 1\,132{,}5$ millions de dh.
(b) Écarts : −632,5 ; −582,5 ; −652,5 ; 1 867,5 ; somme des carrés 4 652 675 ; $s_z^2 = 1\,550\,891{,}7$; $\widehat{V} = 387\,722{,}9$; erreur-type 622,7 millions de dh, soit un coefficient de variation de 55 % : l'estimation est presque inutilisable.
(c) L'entreprise D fournit 66 % de la somme des z_j. Valeur ajoutée par salarié inscrit : 50 000, 55 000, 48 000 dh pour A, B, C, mais 300 000 dh pour D. Tirée avec une petite probabilité, D compte, à elle seule, pour $1/(4\,p_D) = 250$ fois sa valeur, contre 10 à 31 fois pour les autres.
(d) L'estimation n'est pas biaisée : $p_D = 0{,}001$ était bien sa probabilité d'être tirée, et l'estimateur de Hansen-Hurwitz est sans biais quels que soient les p_i. Mais sa variance est énorme, parce que y_i/x_i n'est plus à peu près constant. Avec 60 salariés, $p_D = 0{,}006$, $z_D = 500$, et l'estimation deviendrait $(500 + 550 + 480 + 500)/4 = 507{,}5$ millions de dh, moins de la moitié. Précautions : actualiser la mesure de taille avant le tirage (registre à jour, ou dénombrement des unités), et repérer les rapports y_i/x_i anormaux, qui signalent une taille périmée.