Échantillonnage

Chapitre 2 : Estimateur, biais et précision

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 — Horvitz-Thompson sous un plan de Bernoulli

On reprend la mini-population du cours : cinq ménages A, B, C, D, E de dépenses 4, 6, 7, 9 et 14 milliers de dh ($Y = 40$). Chaque ménage est retenu **indépendamment** des autres avec la probabilité $\pi = 0{,}4$ (plan de Bernoulli du chapitre 1).

(a) Donner π_k, π_{kl} pour $k \neq l$, puis Δ_{kk} et Δ_{kl}. Que vaut $E(n)$?

(b) Écrire l'estimateur de Horvitz-Thompson $\widehat{Y}_{HT}$. Montrer que sa variance se réduit à

$$V(\widehat{Y}_{HT}) = \frac{1-\pi}{\pi} \sum_{k \in U} y_k^2,$$

et la calculer.

(c) Comparer au sondage aléatoire simple de taille 2, de même taille moyenne, pour lequel $V(\widehat{Y}_{HT}) = 108{,}75$. Que vaut $\widehat{Y}_{HT}$ si l'échantillon est vide ? s'il contient les cinq ménages ? Expliquer l'écart.

(d) On a tiré l'échantillon $\{B, E\}$. Calculer $\widehat{Y}_{HT}$ et l'estimation sans biais de la variance $\widehat{V} = \sum_{k \in S} (1-\pi_k)\,(y_k/\pi_k)^2$. Pourquoi cette estimation est-elle possible ici ?

Exercice 2 — Un estimateur biaisé peut-il être meilleur ?

Un bureau d'études estime la dépense mensuelle moyenne $\overline{Y}$ des ménages d'une ville par la moyenne $\overline{y}$ d'un sondage aléatoire simple, sans biais et de variance $V(\overline{y}) = 150\,000$ dh^2. Pour « stabiliser » ses résultats, il propose

$$\hat{\theta} = 0{,}8\,\overline{y} + 0{,}2\,a,$$

où $a = 7\,000$ dh est la moyenne trouvée par une enquête ancienne. On suppose que la vraie moyenne est $\overline{Y} = 7\,900$ dh.

(a) Calculer l'espérance, le biais et la variance de $\hat{\theta}$.

(b) Calculer son erreur quadratique moyenne et la comparer à celle de $\overline{y}$. Conclure.

(c) Reprendre la question avec $a = 6\,000$ dh.

(d) Pour quelles valeurs de a l'estimateur $\hat{\theta}$ fait-il mieux que $\overline{y}$?

(e) On quadruple la taille de l'échantillon : $V(\overline{y})$ est divisée par 4. Comparer de nouveau les deux estimateurs avec $a = 7\,000$ dh. Quelle leçon en tirer ?

Exercice 3 — Intervalle de confiance et coefficient de variation

Une enquête auprès des ménages d'une commune qui en compte $N = 6\,000$ estime leur dépense mensuelle totale à $\widehat{Y}_{HT} = 48{,}6$ millions de dh, avec une variance estimée $\widehat{V} = 3{,}24 \times 10^{12}$ dh^2.

(a) Calculer l'erreur-type estimée, l'intervalle de confiance à 95 % du total et le coefficient de variation estimé.

(b) En déduire l'estimation de la dépense moyenne par ménage, son erreur-type et son intervalle de confiance. Le coefficient de variation change-t-il ?

(c) Parmi les phrases suivantes, lesquelles sont correctes ?

- (i) « Le total Y a 95 % de chances d'être dans l'intervalle calculé. »
- (ii) « La méthode employée produit des intervalles qui contiennent Y dans environ 95 % des échantillons. »
- (iii) « 95 % des ménages dépensent entre les deux bornes de l'intervalle de la moyenne. »

(d) Un institut mène 20 enquêtes indépendantes, chacune avec un intervalle à 95 %. Combien d'intervalles manqueront leur paramètre, en moyenne ? Quelle est la probabilité qu'aucun ne le manque ?

(e) Une simulation de 5 000 sondages aléatoires simples de 200 ménages de la population du cours donne une couverture de 0,9424. Construire un intervalle de confiance à 95 % pour la couverture réelle. Contient-il 0,95 ?

2 Exercice cumulatif (chapitres 2 et 1)

Exercice 4 — Cent ménages dans chaque milieu

La population du cours compte 1 438 ménages urbains (dépense moyenne 8 430 dh, écart-type $S_u = 4\,134$ dh) et 562 ménages ruraux (6 585 dh, $S_r = 3\,215$ dh) ; $\overline{Y} = 7\,911{,}44$ dh. On tire 100 ménages par sondage aléatoire simple dans chaque milieu, indépendamment. On admet (chapitre 3) que la moyenne de n ménages tirés au hasard parmi N_h a pour variance $(1 - n/N_h)\, S_h^2/n$.

(a) Rappeler les probabilités d'inclusion et les poids de chaque milieu (chapitre 1).

(b) Montrer que l'estimateur de Horvitz-Thompson de la moyenne s'écrit $W_u\, \overline{y}_u + W_r\, \overline{y}_r$, avec $W_h = N_h/N$.

(c) L'échantillon tiré donne $\overline{y}_u = 8\,766$ dh et $\overline{y}_r = 6\,608$ dh. Calculer l'estimation du total et de la moyenne, et la moyenne brute des 200 ménages.

(d) Calculer la variance de l'estimateur de Horvitz-Thompson de la moyenne, puis le biais, la variance et l'EQM de la moyenne brute.

(e) Sur l'échantillon de la question (c), laquelle des deux estimations tombe le plus près de $\overline{Y}$? Est-ce une raison de préférer la moyenne brute ?

(f) Un sondage aléatoire simple de 200 ménages a une variance de 71 435,5 dh^2. Le plan à 100 ménages par milieu est-il un bon plan pour estimer $\overline{Y}$?

3 Applications en sciences économiques

Exercice 5 — Les exportations des entreprises textiles

Une association professionnelle veut estimer les exportations annuelles totales des 250 entreprises textiles de la région de Tanger (chiffres inventés). Les 10 plus grandes sont toutes interrogées ; elles déclarent 1 850 millions de dh d'exportations. Parmi les 240 autres, on en tire 40 par sondage aléatoire simple : elles exportent au total 312 millions de dh, avec un écart-type $s = 4{,}5$ millions de dh entre entreprises. On admet (chapitre 3) que la variance de $N\overline{y}$ pour un sondage aléatoire simple de n unités parmi N s'estime par $N^2 (1 - n/N)\, s^2/n$.

(a) Donner la probabilité d'inclusion et le poids d'une grande entreprise et d'une autre entreprise.

(b) Calculer l'estimation de Horvitz-Thompson du total des exportations.

(c) Estimer sa variance, son erreur-type, son intervalle de confiance à 95 % et son coefficient de variation. Pourquoi les grandes entreprises n'apportent-elles aucune variance ?

(d) On aurait pu tirer 50 entreprises au hasard parmi les 250, sans traiter à part les grandes. Combien de grandes entreprises l'échantillon contiendrait-il en moyenne ? Quelle est la probabilité qu'il n'en contienne aucune ? Qu'en conclure ?

Exercice 6 — La production céréalière de quatre communes

Quatre communes rurales de la province de Larache cultivent des céréales sur 10, 20, 30 et 40 centaines d'hectares, superficies connues par le cadastre agricole. Leurs productions de la campagne, inconnues de l'enquêteur, sont de 22, 38, 63 et 77 milliers de quintaux (chiffres inventés). Faute de moyens, on ne visite qu'**une** commune, tirée avec une probabilité proportionnelle à sa superficie.

(a) Donner la probabilité d'inclusion de chaque commune, puis l'estimation de Horvitz-Thompson

de la production totale selon la commune tirée. Vérifier que l'estimateur est sans biais.

(b) Calculer sa variance et son erreur-type.

(c) Reprendre les questions avec un tirage à probabilités égales ($\pi_k = 1/4$). Comparer.

(d) Expliquer l'écart à l'aide de la forme de Sen-Yates-Grundy. Dans quel cas le tirage proportionnel à la superficie serait-il dangereux ?

(e) Peut-on estimer la variance à partir de la seule commune visitée ? Pourquoi ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $\pi_k = 0{,}4$ pour chaque ménage. Les tirages étant indépendants, $\pi_{kl} = P(I_k = 1)\,P(I_l = 1) = 0{,}4 \times 0{,}4 = 0{,}16$. D'où $\Delta_{kk} = 0{,}4 \times 0{,}6 = 0{,}24$ et $\Delta_{kl} = 0{,}16 - 0{,}16 = 0$: les indicatrices sont sans corrélation. $E(n) = \sum_k \pi_k = 5 \times 0{,}4 = 2$.
(b) $\widehat{Y}_{HT} = \sum_{k\in S} y_k/0{,}4 = 2{,}5 \sum_{k\in S} y_k$. Dans la double somme $\sum_k \sum_l \Delta_{kl} \dfrac{y_k}{\pi_k}\dfrac{y_l}{\pi_l}$, seuls les termes diagonaux restent :

$$V(\widehat{Y}_{HT}) = \sum_{k\in U} \pi(1-\pi)\,\frac{y_k^2}{\pi^2} = \frac{1-\pi}{\pi}\sum_{k\in U} y_k^2 = \frac{0{,}6}{0{,}4} \times (16 + 36 + 49 + 81 + 196) = 1{,}5 \times 378 = 567.$$

(c) La variance est $567/108{,}75 = 5{,}2$ fois plus forte qu'avec le sondage aléatoire simple. Si l'échantillon est vide (probabilité $0{,}6^5 = 0{,}078$), $\widehat{Y}_{HT} = 0$; s'il contient les cinq ménages (probabilité $0{,}4^5 = 0{,}010$), $\widehat{Y}_{HT} = 2{,}5 \times 40 = 100$. L'estimateur additionne les dépenses des ménages tirés : il suit les variations de la **taille** n de l'échantillon, en plus de celles de sa composition. C'est une raison de plus de préférer les plans de taille fixe (chapitre 1).
(d) $\widehat{Y}_{HT} = 2{,}5 \times (6 + 14) = 50$ milliers de dh, et $\widehat{V} = 0{,}6 \times (15^2 + 35^2) = 0{,}6 \times 1\,450 = 870$. L'estimation est possible parce que toutes les probabilités doubles sont positives ($\pi_{kl} = 0{,}16$) : chaque terme de la variance peut être observé. Ici, en outre, les termes croisés sont nuls ($\Delta_{kl} = 0$).

Correction — Exercice 2

(a) $E(\hat\theta) = 0{,}8 \times 7\,900 + 0{,}2 \times 7\,000 = 7\,720$ dh, d'où un biais $B = 7\,720 - 7\,900 = -180$ dh, c'est-à-dire $0{,}2\,(a - \overline{Y})$. Variance : $V(\hat\theta) = 0{,}8^2 \times 150\,000 = 96\,000$ dh^2 (a est un nombre fixe).
(b) $EQM(\hat\theta) = 96\,000 + 180^2 = 96\,000 + 32\,400 = 128\,400$, contre $EQM(\overline{y}) = V(\overline{y}) = 150\,000$. L'estimateur biaisé fait mieux : il gagne 54 000 en variance et ne perd que 32 400 en biais au carré.
(c) Avec $a = 6\,000$: $E(\hat\theta) = 7\,520$, $B = -380$, $EQM = 96\,000 + 144\,400 = 240\,400 > 150\,000$. Cette fois, le biais l'emporte.
(d) Il faut $0{,}04\,(a - 7\,900)^2 < 150\,000 - 96\,000 = 54\,000$, soit $|a - 7\,900| < \sqrt{1\,350\,000} = 1\,161{,}9$: $6\,738 < a < 9\,062$ dh. L'ennui est que $\overline{Y}$ est inconnu : on ne sait jamais si a est dans cette plage.
(e) $V(\overline{y}) = 37\,500$, et $EQM(\hat\theta) = 0{,}64 \times 37\,500 + 32\,400 = 24\,000 + 32\,400 = 56\,400 > 37\,500$. La variance baisse avec n, **le biais ne baisse pas** : sur un grand échantillon, l'estimateur sans biais l'emporte presque toujours.

Correction — Exercice 3

(a) Erreur-type : $\sqrt{3{,}24 \times 10^{12}} = 1{,}8 \times 10^6$ dh, soit 1,8 million. Intervalle : $48{,}6 \pm 1{,}96 \times 1{,}8 = 48{,}6 \pm 3{,}528$, soit [45,07 ; 52,13] millions de dh. $CV = 1{,}8/48{,}6 = 3{,}7\,\%$.
(b) Tout se divise par $N = 6\,000$: moyenne 8 100 dh, erreur-type 300 dh, intervalle $8\,100 \pm 588 =$ [7 512 ; 8 688] dh. Le coefficient de variation est inchangé : $300/8\,100 = 3{,}7\,\%$.
(c) Seule (ii) est correcte. (i) est fausse : une fois l'échantillon tiré, Y et l'intervalle sont fixes ; l'intervalle contient Y ou non, et les 95 % qualifient la méthode. (iii) est fausse : l'intervalle porte sur la **moyenne**, pas sur les ménages ; les dépenses individuelles sont bien plus dispersées.
(d) En moyenne $20 \times 0{,}05 = 1$ intervalle manque son paramètre. La probabilité qu'aucun ne le manque vaut $0{,}95^{20} = 0{,}358$: près de deux fois sur trois, au moins un intervalle publié est faux, sans qu'on sache lequel.
(e) Erreur-type de la couverture estimée : $\sqrt{0{,}9424 \times 0{,}0576/5\,000} = 0{,}0033$. Intervalle : $0{,}9424 \pm 0{,}0065 =$ [0,9359 ; 0,9489]. Il ne contient pas 0,95 : la couverture réelle est un peu inférieure au niveau annoncé. L'approximation normale est imparfaite pour une dépense asymétrique et $n = 200$.

2 Exercice cumulatif (chapitres 2 et 1)

Correction — Exercice 4

(a) Urbain : $\pi_k = 100/1\,438 = 0{,}0695$, poids 14,38. Rural : $\pi_k = 100/562 = 0{,}178$, poids 5,62.
(b) $\widehat{Y}_{HT} = 14{,}38 \sum_{\text{urbains tirés}} y_k + 5{,}62 \sum_{\text{ruraux tirés}} y_k = \dfrac{1\,438}{100} \times 100\,\bar{y}_u + \dfrac{562}{100} \times 100\,\bar{y}_r = N_u\,\bar{y}_u + N_r\,\bar{y}_r$.
En divisant par N : $\widehat{\bar{Y}}_{HT} = W_u\,\bar{y}_u + W_r\,\bar{y}_r$, avec $W_u = 0{,}719$ et $W_r = 0{,}281$.
(c) Total : $1\,438 \times 8\,766 + 562 \times 6\,608 = 12\,605\,508 + 3\,713\,696 = 16\,319\,204$ dh. Moyenne : $16\,319\,204/2\,000 = 8\,159{,}6$ dh $= 0{,}719 \times 8\,766 + 0{,}281 \times 6\,608$. Moyenne brute : $(8\,766 + 6\,608)/2 = 7\,687$ dh.
(d) $V(\bar{y}_u) = (1 - 100/1\,438) \times 4\,134^2/100 = 159\,015{,}0$ et $V(\bar{y}_r) = (1 - 100/562) \times 3\,215^2/100 = 84\,970{,}4$. Les deux tirages étant indépendants :

$$V(\widehat{\bar{Y}}_{HT}) = 0{,}719^2 \times 159\,015{,}0 + 0{,}281^2 \times 84\,970{,}4 = 82\,204{,}6 + 6\,709{,}3 = 88\,913{,}9,$$

soit une erreur-type de 298,2 dh. Moyenne brute : $E = \frac{1}{2}(8\,430 + 6\,585) = 7\,507{,}5$, biais $7\,507{,}5 - 7\,911{,}44 = -403{,}94$ dh, variance $\frac{1}{4}(159\,015{,}0 + 84\,970{,}4) = 60\,996{,}4$, et $EQM = 60\,996{,}4 + 403{,}94^2 = 60\,996{,}4 + 163\,167{,}5 = 224\,163{,}9$ (racine 473,5 dh).
(e) Sur cet échantillon, l'estimation de Horvitz-Thompson se trompe de +248,2 dh, la moyenne brute de −224,4 dh : la brute tombe plus près. Ce n'est pas une raison de la préférer : un estimateur se juge sur l'ensemble des échantillons possibles. En moyenne, l'erreur quadratique de la brute est 2,5 fois celle de Horvitz-Thompson (224 164 contre 88 914).
(f) Non : 88 913,9 contre 71 435,5, une variance 1,245 fois plus forte que celle du sondage aléatoire simple de même taille. Les ruraux, 28 % de la population, forment la moitié de l'échantillon : leurs poids inégaux coûtent de la précision sans rien rapporter pour la moyenne d'ensemble. Le plan ne se justifie que si l'on veut aussi une moyenne précise des ruraux (chapitre 7).

3 Applications en sciences économiques

Correction — Exercice 5

(a) Grandes entreprises : $\pi_k = 1$, poids 1 (chacune ne représente qu'elle-même). Autres : $\pi_k = 40/240 = 1/6$, poids 6.
(b) $\widehat{Y}_{HT} = 1\,850 + 6 \times 312 = 1\,850 + 1\,872 = 3\,722$ millions de dh.
(c) Seule la partie tirée au hasard est aléatoire :

$$\widehat{V} = 240^2 \times \left(1 - \frac{1}{6}\right) \times \frac{4{,}5^2}{40} = 57\,600 \times \frac{5}{6} \times 0{,}50625 = 24\,300,$$

d'où une erreur-type de 155,9 millions de dh, un intervalle $3\,722 \pm 305{,}5 = [\,3\,416{,}5\,;\,4\,027{,}5\,]$ et $CV = 155{,}9/3\,722 = 4{,}2\,\%$. Les grandes entreprises ont $\pi_k = 1$: $I_k = 1$ dans tous les échantillons, $V(I_k) = 0$. Leurs 1 850 millions, soit 49,7 % de l'estimation, sont connus sans aucune erreur d'échantillonnage.
(d) En moyenne $50 \times 10/250 = 2$ grandes entreprises. La probabilité de n'en tirer aucune vaut $\binom{240}{50}/\binom{250}{50} = 0{,}10$: une fois sur dix, l'estimation ignorerait la moitié des exportations. À l'inverse, un échantillon qui en contient quatre leur donne le poids $250/50 = 5$: elles comptent pour 20 grandes entreprises, le double de la réalité. Interroger toutes les grandes unités supprime cette source de variance. C'est l'idée des strates exhaustives (chapitres 6 et 7) et des probabilités proportionnelles à la taille (chapitre 11).

Correction — Exercice 6

(**a**) Pour un plan d'une seule unité, π_k est la probabilité de tirer k : 0,1 ; 0,2 ; 0,3 ; 0,4. Estimations y_k/π_k : $22/0{,}1 = 220$; $38/0{,}2 = 190$; $63/0{,}3 = 210$; $77/0{,}4 = 192{,}5$ milliers de quintaux. Espérance : $0{,}1 \times 220 + 0{,}2 \times 190 + 0{,}3 \times 210 + 0{,}4 \times 192{,}5 = 22 + 38 + 63 + 77 = 200 = Y$.
(**b**) Écarts à 200 : 20 ; −10 ; 10 ; −7,5. $V = 0{,}1{\times}400 + 0{,}2{\times}100 + 0{,}3{\times}100 + 0{,}4{\times}56{,}25 = 40 + 20 + 30 + 22{,}5 = 112{,}5$; erreur-type 10,6 milliers de quintaux, 5,3 % de Y.
(**c**) À probabilités égales, $\widehat{Y}_{HT} = 4\,y_k$: 88 ; 152 ; 252 ; 308, d'espérance 200. $V = \frac{1}{4}(112^2 + 48^2 + 52^2 + 108^2) = \frac{1}{4} \times 29\,216 = 7\,304$; erreur-type 85,5, soit 42,7 % de Y. La variance est 65 fois plus forte.
(**d**) D'après la forme de Sen-Yates-Grundy, la variance est nulle quand y_k/π_k est constant. Ici y_k/π_k est proportionnel au rendement de la commune : 22 ; 19 ; 21 ; 19,25 quintaux par hectare, presque égaux. À probabilités égales, au contraire, $4\,y_k$ varie comme la taille de la commune. Le tirage proportionnel à la superficie deviendrait dangereux si la production n'était pas proportionnelle à la superficie : une grande commune frappée par la sécheresse, une petite commune irriguée très productive. Le rapport y_k/π_k varierait alors fortement (chapitre 11).
(**e**) Non. Avec une seule commune, deux communes ne sont jamais tirées ensemble : $\pi_{kl} = 0$ pour toutes les paires, et aucun estimateur sans biais de la variance n'existe. Il faut tirer au moins deux unités pour mesurer sa propre précision.