Échantillonnage

Chapitre 3 : Le sondage aléatoire simple

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 — Tirer en un seul passage : la sélection-rejet

On parcourt une fois la liste des N unités, dans l'ordre. Lorsqu'on arrive à l'unité k et qu'il reste r unités à choisir, on la retient avec la probabilité $\dfrac{r}{N-k+1}$, et l'on passe à la suivante.

(a) On prend $N = 5$ et $n = 2$. Quelle est la probabilité que l'unité 1 soit retenue ? Que l'unité 2 le soit ?

(b) Calculer la probabilité d'obtenir l'échantillon $\{2, 4\}$, en suivant le parcours unité par unité.

(c) Pourquoi la méthode fournit-elle toujours exactement n unités ?

(d) Montrer que, pour N et n quelconques, tout échantillon de n unités a la probabilité $1/\binom{N}{n}$. *Indication : écrire la probabilité d'un échantillon comme un produit de N fractions, puis examiner les dénominateurs, les numérateurs des unités retenues et ceux des unités rejetées.*

(e) Quel avantage pratique cette méthode a-t-elle sur le tri selon une clé aléatoire ?

Exercice 2 — Le facteur de correction dans une petite population

Une zone industrielle compte $N = 1\,200$ entreprises. Une enquête antérieure indique que l'effectif salarié y a un écart-type $S = 45$ salariés. On veut estimer l'effectif moyen par sondage aléatoire simple.

(a) Calculer l'erreur-type de $\bar{y}$ pour $n = 30$, 120, 480 et 1 200, avec le facteur $1 - f$ puis sans lui. Présenter les résultats dans un tableau.

(b) Commenter le cas $n = 1\,200$.

(c) Avec $n = 120$, quelle taille faut-il pour diviser l'erreur-type par deux ? Comparer à la réponse « quatre fois plus, soit 480 », qui oublie le facteur $1 - f$.

Exercice 3 — Dix ménages d'un village

Un village compte $N = 400$ ménages. Un sondage aléatoire simple de $n = 10$ ménages donne les dépenses mensuelles suivantes, en milliers de dh :

5,2	6,8	4,1	9,5	7,3	5,9	12,4	6,1	8,0	4,7

(a) Calculer $\bar{y}$, s^2 et s.

(b) Calculer la variance estimée $\widehat{V}(\bar{y})$, l'erreur-type estimée, et l'intervalle de confiance à 95 % par l'approximation normale.

(c) Estimer la dépense mensuelle totale du village, avec son erreur-type et son intervalle de confiance.

(d) Avec dix ménages seulement, l'approximation normale est fragile. Recalculer l'intervalle de la moyenne avec le quantile de la loi de Student à 9 degrés de liberté, $t_{0,975} = 2{,}262$.

(e) Recalculer $\bar{y}$ et s sans le ménage qui dépense 12,4. Que montre cette comparaison ?

2 Exercice cumulatif (chapitres 3, 2 et 1)

Exercice 4 — La mini-population, trois ménages sur cinq

On reprend la mini-population : cinq ménages A, B, C, D, E dont les dépenses valent 4, 6, 7, 9 et 14 milliers de dh ($Y = 40$, $S^2 = 14{,}5$). On en tire $n = 3$ par sondage aléatoire simple.

(a) Combien d'échantillons sont possibles ? Calculer π_k et π_{kl} par décompte des échantillons, puis par les formules du chapitre 3.

(b) Calculer $\Delta_{kk} = \pi_k(1 - \pi_k)$ et $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ pour $k \neq l$.

(c) Calculer la variance de l'estimateur de Horvitz-Thompson du total, $\widehat{Y}_{HT} = \sum_{k \in S} y_k / \pi_k$, par la formule générale du chapitre 2 :

$$V(\widehat{Y}_{HT}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Comparer au résultat de la formule $N^2 (1 - f)\, S^2 / n$.

(d) Vérifier les deux résultats par l'énumération des échantillons.

(e) Le chapitre 1 a présenté un autre plan de taille 2 : un ménage tiré au hasard parmi A, B, C, et un parmi D, E. Écrire l'estimateur de Horvitz-Thompson du total sous ce plan, puis calculer son espérance et sa variance. Comparer à la variance du total sous un sondage aléatoire simple de taille 2 (108,75) et de taille 3. Que retenir ?

3 Applications en sciences économiques

Exercice 5 — La récolte d'une commune rurale

Une commune rurale de la province de Larache compte 1 500 exploitations céréalières. Pour estimer la récolte de l'année, la direction provinciale de l'agriculture tire 120 exploitations par sondage aléatoire simple. La production moyenne observée est de 38,4 quintaux par exploitation, avec un écart-type de 21,5 quintaux (chiffres fictifs).

(a) Calculer l'erreur-type de la production moyenne et son intervalle de confiance à 95 %.

(b) Estimer la production totale de la commune, avec son intervalle de confiance.

(c) De combien le facteur $1 - f$ réduit-il l'erreur-type ? Calculer le coefficient de variation de l'estimation.

(d) Pour gagner du temps, un agent propose d'interroger plutôt les 120 exploitations les plus proches de la route goudronnée. Pourquoi l'intervalle calculé en (b) n'aurait-il alors plus de sens ?

Exercice 6 — L'audit des factures d'un distributeur

Un distributeur de Tanger a archivé les 24 000 factures papier de l'année, numérotées dans un registre, mais leurs montants n'ont jamais été saisis. Pour estimer le chiffre d'affaires de l'année, un auditeur tire 600 numéros par sondage aléatoire simple et relève les montants : moyenne 412 dh, écart-type 980 dh (chiffres fictifs).

(a) Estimer le montant total des factures, avec son erreur-type.

(b) Donner l'intervalle de confiance à 95 % du total, en millions de dh, et le coefficient de variation de l'estimation.

(c) L'écart-type des montants vaut près de deux fois et demie leur moyenne. Que révèle ce rapport sur la distribution des factures ? Quelles conséquences pour la précision et pour l'approximation normale ?

(d) Le registre indique, pour chaque facture, si le client est un grossiste ou un détaillant. Comment utiliser cette information pour faire mieux avec 600 factures ?

1 Exercices du chapitre

Correction — Exercice 1

(a) À l'unité 1, il reste $r = 2$ unités à choisir parmi 5 : elle est retenue avec la probabilité $2/5$. L'unité 2 est retenue avec la probabilité $1/4$ si l'unité 1 l'a été (il reste alors $r = 1$), et $2/4$ sinon. Par la formule des probabilités totales : $\frac{2}{5} \times \frac{1}{4} + \frac{3}{5} \times \frac{2}{4} = \frac{2}{20} + \frac{6}{20} = \frac{2}{5}$. Les deux unités ont la même probabilité d'inclusion, $n/N = 0{,}4$, comme dans un sondage aléatoire simple.
(b) On suit la liste : unité 1 rejetée $(3/5)$; unité 2 retenue $(r = 2$, 4 unités restantes : $2/4)$; unité 3 rejetée $(r = 1$, 3 restantes : $1 - \frac{1}{3} = \frac{2}{3})$; unité 4 retenue $(r = 1$, 2 restantes : $1/2)$; unité 5 rejetée avec la probabilité 1, puisque $r = 0$. D'où

$$P(\{2,4\}) = \frac{3}{5} \times \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times 1 = \frac{12}{120} = \frac{1}{10} = \frac{1}{\binom{5}{2}}.$$

(c) Lorsque r unités restent à choisir et que la liste n'en contient plus que r, chacune est retenue avec la probabilité $r/r = 1$, puis $(r-1)/(r-1) = 1$, etc. : on ne peut pas en avoir moins de n. Lorsque $r = 0$, la probabilité vaut 0 : on ne peut pas en avoir plus.
(d) La probabilité d'un échantillon s est un produit de N fractions, une par unité de la liste.

- Les dénominateurs valent $N, N-1, \ldots, 1$: leur produit est $N!$.
- Pour une unité retenue, le numérateur est r, le nombre d'unités restant à choisir ; sur les n unités retenues, il prend successivement les valeurs $n, n-1, \ldots, 1$: produit $n!$.
- Pour une unité rejetée, le numérateur est $(N-k+1) - r$, le nombre de rejets restant à faire ; sur les $N-n$ unités rejetées, il prend les valeurs $N-n, \ldots, 1$: produit $(N-n)!$.

Donc $p(s) = \dfrac{n!\,(N-n)!}{N!} = \dfrac{1}{\binom{N}{n}}$ pour **tout** échantillon de taille n : la méthode réalise un sondage aléatoire simple. Dans l'exemple (b) : $\dfrac{2! \times 3!}{5!} = \dfrac{12}{120}$.
(e) Un seul passage dans la liste, dans son ordre, sans tirer ni trier N clés : la méthode convient à une liste lue au fur et à mesure (un registre papier, un très long fichier). Les unités tirées sortent dans l'ordre de la liste, ce qui facilite le travail de terrain.

Correction — Exercice 2

(a) Erreur-type $S\sqrt{(1-f)/n}$ avec $S = 45$ et $f = n/1\,200$; sans correction : $S/\sqrt{n}$.

n	30	120	480	1 200
$1-f$	0,975	0,900	0,600	0
Erreur-type avec $1-f$	8,112	3,897	1,591	0
Erreur-type sans $1-f$	8,216	4,108	2,054	1,299

Pour 30 entreprises, l'écart est négligeable ; pour 480, la correction réduit l'erreur-type de près d'un quart.
(b) Avec $n = N$, on interroge toutes les entreprises : c'est un recensement, $\bar{y} = \overline{Y}$ et l'erreur d'échantillonnage est nulle. La formule sans correction donnerait 1,299 salarié, ce qui n'a pas de sens.
(c) On vise $3{,}897/2 = 1{,}949$, donc une variance divisée par 4. Comme $V(\bar{y}) = S^2\left(\frac{1}{n} - \frac{1}{N}\right)$, il faut

$$\frac{1}{n} - \frac{1}{1\,200} = \frac{1}{4}\left(\frac{1}{120} - \frac{1}{1\,200}\right), \qquad \frac{1}{n} = 0{,}0027083, \qquad n = 369{,}2.$$

Il suffit de 370 entreprises (erreur-type 1,946). Avec 480, l'erreur-type vaudrait 1,591, 18,4 % sous l'objectif : 110 questionnaires de trop. Quand le taux de sondage est notable, oublier le facteur $1-f$ conduit à surdimensionner l'enquête.

Correction — Exercice 3

(**a**) La somme vaut 70, d'où $\overline{y} = 7{,}0$ milliers de dh. Écarts à la moyenne : $-1{,}8$; $-0{,}2$; $-2{,}9$; $2{,}5$; $0{,}3$; $-1{,}1$; $5{,}4$; $-0{,}9$; $1{,}0$; $-2{,}3$. Somme des carrés : 55,5 (ou $\sum y_k^2 - n\overline{y}^2 = 545{,}5 - 490$). D'où $s^2 = 55{,}5/9 = 6{,}167$ et $s = 2{,}483$.
(**b**) $f = 10/400 = 0{,}025$; $\widehat{V}(\overline{y}) = 0{,}975 \times 6{,}167/10 = 0{,}6013$; erreur-type 0,775. Intervalle : $7{,}0 \pm 1{,}96 \times 0{,}775 = 7{,}0 \pm 1{,}52$, soit $[5{,}48\,;8{,}52]$ milliers de dh.
(**c**) $\widehat{Y} = 400 \times 7{,}0 = 2\,800$ milliers de dh, soit 2,8 millions ; erreur-type $400 \times 0{,}7754 = 310{,}2$; intervalle $[2\,192\,;3\,408]$ milliers de dh. Le coefficient de variation, $0{,}775/7{,}0 = 11{,}1\,\%$, est le même pour la moyenne et pour le total.
(**d**) $7{,}0 \pm 2{,}262 \times 0{,}775 = 7{,}0 \pm 1{,}75$, soit $[5{,}25\,;8{,}75]$: un intervalle plus large de 15 % ($2{,}262/1{,}96 = 1{,}154$). La loi de Student tient compte de ce que s est lui-même estimé sur dix valeurs seulement (Statistique inférentielle, chapitre 4).
(**e**) Sans ce ménage : $\overline{y} = 57{,}6/9 = 6{,}4$ et $s = \sqrt{23{,}1/8} = 1{,}699$. Un seul ménage déplace la moyenne de 8,6 % et l'écart-type d'un tiers ($1{,}699/2{,}483 = 0{,}684$). Il ne faut pas pour autant l'écarter : c'est un vrai ménage du village, qui en représente 40. La leçon est qu'avec dix ménages, l'estimation, et surtout l'estimation de sa précision, dépendent beaucoup du hasard qui fait entrer ou non un gros budget dans l'échantillon.

2 Exercice cumulatif (chapitres 3, 2 et 1)

Correction — Exercice 4

(**a**) $\binom{5}{3} = 10$ échantillons. Chaque ménage figure dans $\binom{4}{2} = 6$ d'entre eux : $\pi_k = 6/10 = 0{,}6 = n/N$. Chaque paire figure dans $\binom{3}{1} = 3$ échantillons : $\pi_{kl} = 3/10 = 0{,}3 = \dfrac{3 \times 2}{5 \times 4}$.
(**b**) $\Delta_{kk} = 0{,}6 \times 0{,}4 = 0{,}24$; $\Delta_{kl} = 0{,}3 - 0{,}6^2 = -0{,}06$.
(**c**) Les y_k/π_k valent 6,67 ; 10 ; 11,67 ; 15 ; 23,33. Leur somme vaut 66,67, la somme de leurs carrés 1 050, donc la somme des produits croisés ($k \neq l$) vaut $66{,}67^2 - 1\,050 = 3\,394{,}44$. D'où

$$V(\widehat{Y}_{HT}) = 0{,}24 \times 1\,050 - 0{,}06 \times 3\,394{,}44 = 252 - 203{,}67 = 48{,}33.$$

Et $N^2(1-f)\,S^2/n = 25 \times 0{,}4 \times 14{,}5/3 = 48{,}33$: la formule générale du chapitre 2 et celle du sondage aléatoire simple coïncident.
(**d**) Ici $\widehat{Y}_{HT} = \sum_S y_k/0{,}6 = 5\,\overline{y}$:

s	ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE
$\widehat{Y}_{HT}$	28,33	31,67	40	33,33	41,67	45	36,67	45	48,33	50

La somme vaut 400 : $E(\widehat{Y}_{HT}) = 40 = Y$. Les carrés des écarts à 40 ont pour somme 483,33, d'où $V(\widehat{Y}_{HT}) = 483{,}33/10 = 48{,}33$.
(**e**) Sous ce plan, $\pi_A = \pi_B = \pi_C = 1/3$ et $\pi_D = \pi_E = 1/2$: $\widehat{Y}_{HT} = 3\,y_1 + 2\,y_2$, où y_1 vient du premier groupe. Les six échantillons, de probabilité 1/6, donnent AD : 30 ; AE : 40 ; BD : 36 ; BE : 46 ; CD : 39 ; CE : 49. Espérance $240/6 = 40 = Y$: l'estimateur est sans biais. Variance : $(100 + 0 + 16 + 36 + 1 + 81)/6 = 234/6 = 39$. Contrôle : les deux tirages sont indépendants, donc $V = 9\,V(y_1) + 4\,V(y_2) = 9 \times 1{,}556 + 4 \times 6{,}25 = 14 + 25 = 39$.
Le sondage aléatoire simple a une variance 2,79 fois plus grande avec deux ménages (108,75 contre 39), et encore 1,24 fois plus grande avec trois (48,33 contre 39) : **deux ménages bien répartis valent mieux que trois ménages tirés sans précaution**. Séparer les petits budgets des grands supprime la part de la variance qui vient de l'écart entre les deux groupes : c'est la stratification (chapitre 6). Encore faut-il pondérer : $5\,\overline{y}$ aurait, sous ce plan, une espérance de 42,92 (chapitre 1).

3 Applications en sciences économiques

Correction — Exercice 5

(a) $f = 120/1\,500 = 0{,}08$; erreur-type $\sqrt{0{,}92 \times 21{,}5^2/120} = 1{,}883$ quintal ; intervalle $38{,}4 \pm 1{,}96 \times 1{,}883 = 38{,}4 \pm 3{,}69$, soit $[34{,}71\,;42{,}09]$ quintaux par exploitation.
(b) $\widehat{Y} = 1\,500 \times 38{,}4 = 57\,600$ quintaux ; erreur-type $1\,500 \times 1{,}8825 = 2\,824$; intervalle $[52\,065\,;63\,135]$ quintaux, soit entre 52,1 et 63,1 milliers de quintaux.
(c) $\sqrt{0{,}92} = 0{,}959$: la correction réduit l'erreur-type de 4,1 % (de 1,963 à 1,883). Coefficient de variation : $1{,}883/38{,}4 = 4{,}9$ %, le même pour le total : la récolte est connue à environ ± 10 % près.
(d) L'échantillon ne serait plus aléatoire : les exploitations éloignées de la route auraient une probabilité d'inclusion nulle, et celles qui sont proches sont sans doute mieux équipées et mieux reliées aux marchés. La formule $(1-f)\,s^2/n$ mesure l'erreur due au **hasard du tirage** d'un sondage aléatoire simple ; elle ne dit rien du biais d'un choix raisonné, qui peut être bien plus grand et que rien ne permet de mesurer. L'intervalle n'aurait plus la garantie de 95 % (voir le Literary Digest, introduction).

Correction — Exercice 6

(a) $\widehat{Y} = 24\,000 \times 412 = 9\,888\,000$ dh. Avec $f = 600/24\,000 = 0{,}025$: erreur-type de la moyenne $\sqrt{0{,}975 \times 980^2/600} = 39{,}51$ dh, et du total 24 000 fois plus : 948 122 dh (calcul fait sans arrondir l'erreur-type de la moyenne).
(b) $9\,888\,000 \pm 1{,}96 \times 948\,122$, soit $[8{,}03\,;11{,}75]$ millions de dh. Coefficient de variation : $39{,}51/412 = 9{,}6$ %.
(c) $980/412 = 2{,}38$: la distribution est très asymétrique, faite de nombreuses petites factures et de quelques très grosses. Deux conséquences. La précision est faible : malgré 600 factures, le total n'est connu qu'à ± 19 % près. Et l'approximation normale est moins sûre : un échantillon qui manque les grosses factures a à la fois une moyenne basse et un s faible, donc un intervalle trop bas et trop étroit ; la couverture réelle peut être inférieure à 95 %.
(d) Tirer séparément parmi les factures des grossistes et parmi celles des détaillants : c'est un sondage **stratifié** (chapitre 6). Les factures des grossistes, peu nombreuses mais grosses et très dispersées, doivent être tirées avec un taux de sondage plus élevé, voire toutes vérifiées ; le chapitre 7 calcule la meilleure répartition des 600 factures. Chaque facture tirée est ensuite pondérée par l'inverse de sa probabilité d'inclusion.