Échantillonnage

Chapitre 8 : Le sondage systématique

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 — Douze exploitations laitières

Une coopérative laitière compte douze exploitations, listées dans l'ordre de leur adhésion. Leur production journalière, en dizaines de litres, est, dans cet ordre :

$$12,\ 15,\ 9,\ 20,\ 14,\ 18,\ 10,\ 22,\ 16,\ 13,\ 19,\ 24.$$

On veut en interroger quatre.

(a) Calculer $\overline{Y}$, S^2 et la variance de $\overline{y}$ pour un sondage aléatoire simple de 4 exploitations.

(b) Tirage systématique dans l'ordre de la liste : quel est le pas ? Énumérer les échantillons possibles et leurs moyennes, puis calculer $V(\overline{y})$.

(c) Même question après avoir rangé la liste par production croissante.

(d) Le rangement de la question (c) est-il réalisable en pratique ? Par quelle information, connue avant l'enquête, pourrait-on le remplacer ?

Exercice 2 — Le tirage systématique circulaire

Pour garder une taille fixe quand N n'est pas un multiple de n, on peut parcourir la liste comme un cercle : on tire un départ r au hasard parmi $1, \dots, N$, puis on prend les rangs $r, r+k, r+2k, \dots$ en revenant au début de la liste après le rang N. On applique ce tirage à une population de $N = 7$ unités de valeurs

$$3,\ 5,\ 6,\ 8,\ 9,\ 11,\ 14,$$

avec $n = 3$ et un pas $k = 2$.

(a) Énumérer les sept échantillons possibles et leurs moyennes.

(b) Calculer les probabilités d'inclusion π_j. Quelles paires d'unités ne sont jamais tirées ensemble ? Que valent les autres probabilités doubles π_{jl} ?

(c) Vérifier que $E(\overline{y}) = \overline{Y}$. Calculer $V(\overline{y})$ et la comparer à celle d'un sondage aléatoire simple de même taille.

(d) Peut-on estimer $V(\overline{y})$ sans biais à partir de l'échantillon ? Justifier.

Exercice 3 — Une liste ordonnée et les différences successives

Une administration dispose d'une liste de $N = 200$ entreprises rangées par effectif salarié croissant. Elle en tire $n = 10$ par tirage systématique de pas $k = 20$. Leurs chiffres d'affaires, en millions de dh et dans l'ordre de la liste, sont :

$$1{,}2\ ;\ 1{,}6\ ;\ 2{,}0\ ;\ 2{,}3\ ;\ 2{,}9\ ;\ 3{,}1\ ;\ 3{,}8\ ;\ 4{,}4\ ;\ 4{,}6\ ;\ 5{,}5.$$

(a) Calculer $\overline{y}$, s^2, puis la variance estimée de $\overline{y}$ « comme pour un sondage aléatoire simple » et l'erreur-type correspondante.

(b) Calculer $s_d^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (z_{i+1} - z_i)^2$ et $\widehat{V}_d(\overline{y}) = (1-f)\, s_d^2/n$.

(c) Donner les deux intervalles de confiance approchés à 95 % de la moyenne.

(d) Lequel des deux paraît le plus réaliste ? D'où vient l'écart entre les deux estimations ?

2 Exercice cumulatif (chapitres 8, 7, 6, 5 et 4)

Exercice 4 — Tirer systématiquement dans le fichier des ménages

On reprend la population du cours : $N = 2\,000$ ménages, $\overline{Y} = 7\,911{,}44$ dh, $S = 3\,984{,}29$ dh, répartis en quatre provinces de 820, 560, 380 et 240 ménages. Le fichier est rangé par province, puis par milieu, puis par district. On tire $n = 200$ ménages par tirage systématique dans l'ordre du fichier.

(a) Combien de ménages de chaque province contient chacun des échantillons possibles ? Quelle allocation reconnaît-on ? La comparer à l'allocation de Neyman $(90{,}6\,;\,56{,}1\,;\,35{,}3\,;\,18{,}0)$.

(b) Les variances de $\overline{y}$ sont : 71 435,5 (sondage aléatoire simple), 68 139,7 (stratifié proportionnel par province), 67 172,3 (Neyman), 30 287,0 (systématique dans l'ordre du fichier). Calculer pour chaque plan l'effet de plan, la taille effective et la demi-largeur de l'intervalle de confiance à 95 %.

(c) En déduire l'erreur-type de l'estimateur $N\overline{y}$ du total des dépenses, pour le sondage aléatoire simple et pour le tirage systématique.

(d) Sur le même fichier, le rapport de la variance du tirage systématique à celle du sondage aléatoire simple vaut 1,341 avec le pas $k = 4$, 0,508 avec $k = 5$ et 0,785 avec $k = 20$; une stratification proportionnelle par district donnerait 0,841. Que penser du rapport 0,424 obtenu avec $k = 10$?

(e) On veut une marge d'erreur de 300 dh à 95 % sur la dépense moyenne. Quelle taille faut-il pour un sondage aléatoire simple ? Pour un tirage systématique dans l'ordre du fichier, en retenant un effet de plan de 0,841 ? Pourquoi ne pas retenir 0,424 ?

3 Applications en sciences économiques

Exercice 5 — La récolte d'une coopérative oléicole

Une coopérative oléicole regroupe 600 parcelles, listées de l'amont vers l'aval de son canal d'irrigation ; en aval, les parcelles sont plus grandes et produisent davantage. Pour estimer la récolte d'olives de la campagne avant la fin de la cueillette, on pèse la production d'une parcelle sur 30, avec le départ $r = 17$. Les productions des 20 parcelles pesées, en tonnes et dans l'ordre de la liste, sont (chiffres fictifs) :

0,8 ; 1,1 ; 0,9 ; 1,4 ; 1,6 ; 1,5 ; 2,0 ; 1,8 ; 2,3 ; 2,6 ; 2,4 ; 3,1 ; 3,0 ; 3,6 ; 3,9 ; 4,4 ; 4,8 ; 5,5 ; 6,3 ; 7,9.

On donne $\sum (y_j - \overline{y})^2 = 70{,}5695$ et $\sum_i (z_{i+1} - z_i)^2 = 6{,}15$.

(a) Quels sont les rangs des parcelles pesées ? Quelle est la probabilité d'inclusion d'une parcelle ? Estimer la récolte totale.

(b) Estimer la variance de ce total comme pour un sondage aléatoire simple ; en déduire un intervalle de confiance à 95 % et le coefficient de variation.

(c) Même question avec l'estimateur par les différences successives.

(d) Quel intervalle la coopérative devrait-elle annoncer ? Quelle particularité de la liste faudrait-il vérifier avant de lui faire confiance ?

Exercice 6 — Les clients d'un hypermarché

Un hypermarché de Tanger veut connaître le panier moyen de ses clients un samedi. Un enquêteur, posté après les caisses, interroge un client sur 25 ; le premier interrogé est le 18$^{\text{e}}$ client de la journée, rang tiré au hasard entre 1 et 25. À la fermeture, les tickets de caisse indiquent 1 862 passages. Les clients interrogés ont dépensé au total 29 600 dh, avec un écart-type $s = 310$ dh (chiffres fictifs).

(a) Combien de clients ont été interrogés ? Combien l'auraient été avec le départ 5 ? Pourquoi la taille de l'échantillon est-elle aléatoire ?

(b) Quelle est la probabilité d'inclusion d'un client ? Estimer le total des achats de la journée par l'estimateur de Horvitz-Thompson, puis par $N\overline{y}$. Lequel préférer ici ?

(c) Estimer le panier moyen et donner un intervalle de confiance à 95 % en traitant l'échantillon comme un sondage aléatoire simple. En déduire un intervalle pour le total.

(d) Pourquoi le tirage systématique est-il naturel ici ? L'intervalle obtenu est-il plutôt prudent ou plutôt optimiste ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $\overline{Y} = 192/12 = 16$. Écarts : $-4, -1, -7, 4, -2, 2, -6, 6, 0, -3, 3, 8$; somme des carrés 244, d'où $S^2 = 244/11 = 22{,}18$. Sondage aléatoire simple de 4 exploitations : $V(\overline{y}) = (1-4/12)\times 22{,}18/4 = 3{,}697$.
(b) Le pas vaut $k = 12/4 = 3$: trois échantillons possibles.

Départ	Rangs	Productions	$\overline{y}_r$
$r = 1$	1, 4, 7, 10	12, 20, 10, 13	13,75
$r = 2$	2, 5, 8, 11	15, 14, 22, 19	17,5
$r = 3$	3, 6, 9, 12	9, 18, 16, 24	16,75

La moyenne des trois moyennes vaut $48/3 = 16 = \overline{Y}$. Écarts : $-2{,}25$; 1,5 ; 0,75. $V(\overline{y}) = (5{,}0625 + 2{,}25 + 0{,}5625)/3 = 7{,}875/3 = 2{,}625$, soit 0,71 fois la variance du sondage aléatoire simple.
(c) Liste rangée : $9, 10, 12, 13, 14, 15, 16, 18, 19, 20, 22, 24$. Les trois échantillons sont $\{9, 13, 16, 20\}$, $\{10, 14, 18, 22\}$ et $\{12, 15, 19, 24\}$, de moyennes 14,5 ; 16 ; 17,5. $V(\overline{y}) = (2{,}25 + 0 + 2{,}25)/3 = 1{,}5$, soit 0,41 fois celle du sondage aléatoire simple : chaque échantillon prend une exploitation dans chacun des quatre blocs (des plus petites aux plus grandes), c'est une stratification implicite.
(d) Non : la production est précisément ce que l'on veut mesurer. On peut ranger la liste selon une information connue et liée à la production, comme le nombre de vaches laitières déclaré à l'adhésion ou la surface fourragère.

Correction — Exercice 2

(a) Avec $k = 2$, le départ r donne les rangs r, $r+2$, $r+4$ (modulo 7) :

Départ	1	2	3	4	5	6	7
Rangs	1, 3, 5	2, 4, 6	3, 5, 7	4, 6, 1	5, 7, 2	6, 1, 3	7, 2, 4
Valeurs	3, 6, 9	5, 8, 11	6, 9, 14	8, 11, 3	9, 14, 5	11, 3, 6	14, 5, 8
$\overline{y}$	6	8	9,67	7,33	9,33	6,67	9

(b) Chaque unité figure dans 3 échantillons sur 7 : $\pi_j = 3/7 = n/N$. Les sept paires de voisins sur le cercle, $\{1,2\}$, $\{2,3\}$, ..., $\{6,7\}$ et $\{7,1\}$, ne sont jamais tirées ensemble : $\pi_{jl} = 0$. Les paires séparées de deux rangs sur le cercle (comme $\{1,3\}$ ou $\{1,6\}$) apparaissent dans deux échantillons, $\pi_{jl} = 2/7$; celles séparées de trois rangs (comme $\{1,4\}$ ou $\{1,5\}$) dans un seul, $\pi_{jl} = 1/7$. Contrôle : $7\times 0 + 7\times\frac{2}{7} + 7\times\frac{1}{7} = 3 = \binom{3}{2}$ paires par échantillon.
(c) Somme des sept moyennes : 56, d'où $E(\overline{y}) = 8 = \overline{Y}$ (56/7). Écarts : -2, 0, $\frac{5}{3}$, $-\frac{2}{3}$, $\frac{4}{3}$, $-\frac{4}{3}$, 1 ; somme des carrés 106/9, d'où $V(\overline{y}) = 106/63 = 1{,}68$. Pour le sondage aléatoire simple : $S^2 = 84/6 = 14$ et $V = (1 - 3/7)\times 14/3 = 2{,}67$. Le tirage circulaire fait mieux (rapport 0,63) : il sépare les voisins.
(d) Non : l'estimation sans biais de la variance exige $\pi_{jl} > 0$ pour toutes les paires (chapitre 2). Ici, sept paires sur vingt et une ont $\pi_{jl} = 0$.

Correction — Exercice 3

(a) $\overline{y} = 31{,}4/10 = 3{,}14$. Somme des carrés des écarts : 17,924, d'où $s^2 = 17{,}924/9 = 1{,}992$. Avec $f = 10/200 = 0{,}05$: $\widehat{V}(\overline{y}) = 0{,}95\times 1{,}992/10 = 0{,}1892$, erreur-type 0,435.
(b) Différences successives : 0,4 ; 0,4 ; 0,3 ; 0,6 ; 0,2 ; 0,7 ; 0,6 ; 0,2 ; 0,9. Somme des carrés : 2,51, d'où $s_d^2 = 2{,}51/18 = 0{,}1394$ et $\widehat{V}_d(\overline{y}) = 0{,}95\times 0{,}1394/10 = 0{,}01325$, erreur-type 0,115.
(c) Comme un sondage aléatoire simple : $3{,}14 \pm 1{,}96\times 0{,}435$, soit [2,29 ; 3,99] millions de dh. Par les différences successives : $3{,}14 \pm 1{,}96\times 0{,}115$, soit [2,91 ; 3,37]. Avec dix entreprises seulement, l'approximation normale reste grossière.

(**d**) Les valeurs tirées croissent régulièrement : la liste, rangée par effectif, est aussi presque rangée par chiffre d'affaires. s^2 mesure surtout cette tendance, c'est-à-dire les écarts entre blocs de 20 entreprises, que le tirage systématique neutralise puisqu'il prend une entreprise par bloc. Les différences successives ne retiennent que l'écart entre voisins : s_d^2 ne vaut que 7 % de s^2. Le second intervalle est le plus réaliste, à condition que la liste ne cache pas de cycle ; s'il fallait rester prudent, on publierait le premier.

2 Exercice cumulatif (chapitres 8, 7, 6, 5 et 4)

Correction — Exercice 4

(**a**) Les effectifs des provinces sont des multiples de 10 et le fichier les range l'une après l'autre : chaque échantillon prend exactement un ménage sur dix de chaque province, soit 82, 56, 38 et 24 ménages. C'est l'allocation proportionnelle $n_h = nN_h/N$ du chapitre 6. L'allocation de Neyman ($n_h \propto N_h S_h$, chapitre 7) donnerait davantage de ménages à Tanger-Assilah (90,6), où la dépense est la plus dispersée, et moins à Al Hoceïma (18,0).
(**b**) Effet de plan $= V/71\,435{,}5$; taille effective $= 200/\text{deff}$; demi-largeur $= 1{,}96\sqrt{V}$.

Plan ($n = 200$)	$V(\bar{y})$	Effet de plan	Taille effective	Demi-largeur (dh)
Sondage aléatoire simple	71 435,5	1	200	523,9
Stratifié proportionnel	68 139,7	0,954	209,7	511,6
Neyman	67 172,3	0,940	212,7	508,0
Systématique, ordre du fichier	30 287,0	0,424	471,7	341,1

(**c**) L'erreur-type de $N\bar{y}$ vaut N fois celle de $\bar{y}$ (chapitre 4) : $2\,000 \times 267{,}27 \approx 534\,500$ dh pour le sondage aléatoire simple, $2\,000 \times 174{,}03 \approx 348\,100$ dh pour le tirage systématique, sur un total de 15 822 870 dh.
(**d**) Le rangement du fichier garantit à peu près le gain d'une stratification par district (0,841) : c'est ce que l'on obtient en moyenne quand l'ordre des ménages à l'intérieur des districts est quelconque. Le 0,424 ajoute à ce gain un ordre exact favorable ; avec d'autres pas, le même fichier donne des rapports très différents, et même supérieur à 1 pour $k = 4$. Avec seulement k échantillons possibles, la variance exacte dépend beaucoup du hasard de la liste : le 0,424 est une bonne surprise, pas une propriété du plan.
(**e**) Sondage aléatoire simple (chapitre 5) : $n_0 = (1{,}96 \times 3\,984{,}29/300)^2 = 677{,}6$, puis $n = 677{,}6/(1 + 677{,}6/2\,000) = 506{,}1$, soit 507 ménages. Tirage systématique avec l'effet de plan 0,841 : $n_0 = 0{,}841 \times 677{,}6 = 569{,}9$, puis $n = 569{,}9/(1 + 569{,}9/2\,000) = 443{,}5$, soit 444 ménages. Avec 0,424, on trouverait 252 ménages, mais rien ne garantit qu'un autre pas ou un fichier mis à jour reproduise ce rapport : on risquerait de manquer la précision promise.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) Rangs $17, 47, 77, \ldots, 17 + 19 \times 30 = 587$. Chaque parcelle appartient à un seul des 30 échantillons possibles : $\pi_j = 1/30$. Production des parcelles pesées : 60,9 tonnes, d'où une récolte estimée de $30 \times 60{,}9 = 1\,827$ tonnes ($= N\overline{y} = 600 \times 3{,}045$).
(**b**) $s^2 = 70{,}5695/19 = 3{,}714$ et $f = 20/600 = 1/30$. $\widehat{V}(\widehat{Y}) = N^2(1-f)\,s^2/n = 600^2 \times \frac{29}{30} \times 3{,}714/20 = 64\,627$; erreur-type 254,2 tonnes. Intervalle : $1\,827 \pm 1{,}96 \times 254{,}2$, soit $[1\,329\,;2\,325]$ tonnes ; coefficient de variation $254{,}2/1\,827 = 13{,}9\,\%$.
(**c**) $s_d^2 = 6{,}15/(2 \times 19) = 0{,}1618$. $\widehat{V}_d(\widehat{Y}) = 600^2 \times \frac{29}{30} \times 0{,}1618/20 = 2\,816$; erreur-type 53,1 tonnes. Intervalle : $[1\,723\,;1\,931]$ tonnes ; coefficient de variation 2,9 %.
(**d**) La production croît nettement de l'amont vers l'aval : la liste est rangée selon une variable liée à y, et le tirage systématique en profite comme d'une stratification implicite. L'estimateur du sondage aléatoire simple compte cette tendance comme de l'aléa ; celui des différences successives l'écarte. L'intervalle $[1\,723\,;1\,931]$ est donc le plus réaliste. Mais il faut vérifier que la liste n'a pas de cycle : si, par exemple, les parcelles alternaient d'une rive à l'autre du canal, le pas 30, pair, et le départ 17, impair, ne retiendraient que des rangs impairs, donc une seule rive.

Correction — Exercice 6

(**a**) Rangs $18, 43, \ldots, 18 + 73 \times 25 = 1\,843$; le suivant, 1 868, dépasse 1 862 : 74 clients. Avec le départ 5 : rangs $5, \ldots, 5 + 74 \times 25 = 1\,855$, soit 75 clients. Comme $1\,862 = 74 \times 25 + 12$, les départs 1 à 12 donnent 75 clients et les départs 13 à 25 en donnent 74 : le pas est entier mais N n'en est pas un multiple, et N n'est connu qu'à la fermeture.
(**b**) Chaque client est interrogé si et seulement si le départ tombe sur son rang modulo 25 : $\pi_j = 1/25 = 0{,}04$, pour tous. Horvitz-Thompson : $\widehat{Y}_{HT} = 25 \times 29\,600 = 740\,000$ dh. Le même estimateur appliqué au nombre de clients donne $\widehat{N} = 25 \times 74 = 1\,850$, alors que l'on sait que $N = 1\,862$. Le panier moyen est $\overline{y} = 29\,600/74 = 400$ dh et $N\overline{y} = 1\,862 \times 400 = 744\,800$ dh. On préfère ce second estimateur : il corrige l'écart entre $\widehat{N}$ et le vrai N, que l'on connaît (c'est l'idée de l'estimateur par le ratio, chapitre 12).
(**c**) $f = 74/1\,862 = 0{,}0397$; $\widehat{V}(\overline{y}) = (1-0{,}0397) \times 310^2/74 = 1\,247{,}0$; erreur-type 35,3 dh. Intervalle pour le panier moyen : $400 \pm 1{,}96 \times 35{,}3$, soit $[330{,}8\,;469{,}2]$ dh. Pour le total : erreur-type $1\,862 \times 35{,}3 \approx 65\,750$ dh et intervalle $[615\,900\,;873\,700]$ dh.
(**d**) Il n'existe aucune liste des clients avant leur passage, et leur nombre est inconnu à l'avance : l'enquêteur n'a qu'à compter. L'échantillon s'étale sur toute la journée ; or les paniers diffèrent selon l'heure (courses de la semaine le matin, achats d'appoint le soir). C'est une stratification implicite par heure de passage, que la formule du sondage aléatoire simple ignore : l'intervalle est plutôt prudent, sauf si l'affluence suivait un cycle de 25 clients, ce qui est très peu probable.