Échantillonnage

Chapitre 7 : L'allocation optimale et l'effet de plan

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 — L'allocation de Neyman pas à pas

Une population de $N = 1\,000$ unités est partagée en trois strates :

Strate h	1	2	3
N_h	500	300	200
$\overline{Y}_h$	40	60	120
S_h	6	10	30

On tire $n = 100$ unités.

(a) Calculer $\overline{Y}$, puis S^2 par la décomposition $(N-1)\,S^2 = \sum_h (N_h - 1)\,S_h^2 + \sum_h N_h(\overline{Y}_h - \overline{Y})^2$. En déduire la variance de $\overline{y}$ pour un sondage aléatoire simple de 100 unités.

(b) Calculer l'allocation proportionnelle, la variance de $\overline{y}_{st}$ et l'effet de plan.

(c) Calculer l'allocation de Neyman, sa variance par la formule $V_{Ney} = \overline{S}^{\,2}/n - \frac{1}{N}\sum_h W_h S_h^2$, et son effet de plan.

(d) Vérifier que $V_{prop} - V_{Ney} = \frac{1}{n}\sum_h W_h\,(S_h - \overline{S})^2$.

(e) Vérifier que, sous l'allocation de Neyman, le taux de sondage f_h est proportionnel à S_h. Que valent les poids d_k ? Pourquoi l'estimateur n'est-il plus la moyenne simple de l'échantillon ?

Exercice 2 — Le coût linéaire et le problème inverse

Une enquête porte sur deux strates : $N_1 = 3\,000$, $S_1 = 200$ dh, coût unitaire $c_1 = 25$ dh ; $N_2 = 1\,000$, $S_2 = 600$ dh, $c_2 = 100$ dh. Le coût total est $C = c_0 + c_1 n_1 + c_2 n_2$, avec des frais fixes $c_0 = 10\,000$ dh.

(a) Montrer, par la méthode de Lagrange, qu'à budget C donné, la variance de $\overline{y}_{st}$ est minimale pour n_h proportionnel à $N_h S_h/\sqrt{c_h}$.

(b) Le budget total est $C = 46\,000$ dh. Calculer l'allocation optimale et la variance de $\overline{y}_{st}$.

(c) Calculer l'allocation proportionnelle de même coût et sa variance. Quel est le gain de l'allocation optimale ?

(d) Problème inverse : on veut atteindre une variance V_0 au moindre coût. Montrer que la solution a la même forme, avec

$$n_h = \frac{W_h S_h}{\sqrt{c_h}} \times \frac{\sum_k W_k S_k \sqrt{c_k}}{V_0 + \frac{1}{N}\sum_k W_k S_k^2}.$$

Vérifier qu'avec V_0 égale à la variance trouvée en (b), on retrouve l'allocation de (b). Quel budget faut-il pour $V_0 = 60$?

Exercice 3 — Effet de plan, taille effective et formule de Kish

Une enquête sur la dépense des ménages porte sur trois régions de $N_h = 50\,000$, $30\,000$ et $20\,000$ ménages. Pour publier un résultat par région, on tire 400 ménages dans chacune. On admet que les trois régions ont la même dépense moyenne : un sondage aléatoire simple et l'allocation proportionnelle ont alors la même variance. On néglige les corrections de population finie.

(a) Calculer le poids d_k d'un ménage de chaque région, et vérifier que la somme des poids des 1 200 ménages vaut N.

(b) Calculer l'effet de plan de Kish $n\sum_h n_h d_h^2 / \left(\sum_h n_h d_h\right)^2$, la taille effective n_{eff} et $deft$.

(c) On suppose que les trois régions ont aussi le même écart-type. Montrer que l'effet de plan vaut alors $n\sum_h W_h^2/n_h$, et retrouver le résultat de (b).

(d) En réalité, $S_h = 3\,000$, $2\,000$ et $1\,500$ dh. Calculer le rapport $V_{\text{égale}}/V_{prop}$. Pourquoi dépasse-t-il l'effet de plan de Kish ? Quelle serait l'allocation de Neyman ?

(e) On vise une marge de ± 200 dh à 95 %. Les moyennes régionales étant égales, on prend $S^2 \approx \sum_h W_h S_h^2$. Combien de ménages faudrait-il avec un sondage aléatoire simple ? Combien en faut-il avec l'allocation égale, d'après la formule de Kish, puis d'après (d) ?

2 Exercice cumulatif (chapitres 7, 6, 5, 4 et 3)

Exercice 4 — Dimensionner une enquête stratifiée sur la population du cours

On veut estimer la dépense moyenne des 2 000 ménages du cours avec une marge de ± 300 dh à 95 %. On rappelle $S = 3\,984{,}29$ dh et, pour les provinces :

Province	N_h	S_h (dh)	P_h (internet)
Tanger-Assilah	820	4 271,19	0,739
Tétouan	560	3 873,11	0,716
Larache	380	3 593,17	0,589
Al Hoceïma	240	2 899,90	0,488

(a) Quelle taille faut-il pour un sondage aléatoire simple (chapitres 3 et 5) ?

(b) Écrire la variance cible $V = (300/1{,}96)^2$. Quelle taille faut-il pour un sondage stratifié à allocation proportionnelle (chapitre 6) ? On utilisera $n = \sum_h W_h S_h^2 / (V + \frac{1}{N}\sum_h W_h S_h^2)$.

(c) Même question avec l'allocation de Neyman, puis répartir cet échantillon entre les provinces (en entiers) et vérifier que la marge est tenue.

(d) Avec cet échantillon, quelle est l'erreur-type de la proportion de ménages connectés (chapitre 4) ? Comparer à un sondage aléatoire simple de même taille. Conclure.

(e) L'allocation égale exigerait 624 ménages. Que paie-t-on avec ces ménages supplémentaires ?

3 Applications en sciences économiques

Exercice 5 — La production des exploitations agricoles

Une direction provinciale de l'agriculture veut estimer la valeur moyenne de la production des 4 850 exploitations de la province, à partir de 300 exploitations. Les chiffres sont inventés.

Exploitations	N_h	S_h (milliers de dh)
Petites	4 000	20
Moyennes	800	80
Grandes	50	600

(a) Calculer l'allocation proportionnelle.

(b) Calculer l'allocation de Neyman. Quel problème pose-t-elle ? Le résoudre.

(c) Calculer la variance de la moyenne estimée pour les deux allocations. Quelle part de la variance de l'allocation proportionnelle vient des grandes exploitations ?

(d) En déduire l'erreur-type de la production **totale** estimée dans les deux cas.

(e) Quels sont les poids des exploitations tirées sous l'allocation retenue en (b) ? Pourquoi les enquêtes agricoles et les enquêtes auprès des entreprises recensent-elles souvent les plus grandes unités ?

Exercice 6 — Une enquête régionale sur le chômage

Une région compte 400 000 ménages urbains et 200 000 ménages ruraux. On veut estimer la proportion de ménages comptant au moins un chômeur, que l'on suppose voisine de 0,20 en ville et de 0,10 à la campagne. Un questionnaire coûte 100 dh en ville et 400 dh à la campagne (déplacements) ; le budget est de 280 000 dh. Chiffres inventés ; on néglige les corrections de population finie et l'on prend $S_h^2 = P_h(1 - P_h)$.

(a) Calculer la proportion régionale P et les écarts-types S_h.

(b) Calculer l'allocation optimale compte tenu des coûts, et l'erreur-type de $\widehat{P}_{st}$.

(c) Comparer à l'allocation proportionnelle de même coût.

(d) Quelle est la marge d'erreur de la proportion rurale ? Le ministère exige ± 3 points pour le milieu rural. Combien de ménages ruraux faut-il ? Que reste-t-il pour la ville, et que devient l'erreur-type régionale ?

(e) Et si l'on exigeait ± 2 points pour le rural ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $W_h = 0{,}5\,;0{,}3\,;0{,}2$ et $\overline{Y} = 0{,}5{\times}40{+}0{,}3{\times}60{+}0{,}2{\times}120 = 62$. Partie intra : $499{\times}36{+}299{\times}100{+}199{\times}900 = 226\,964$; partie inter : $500 \times 22^2 + 300 \times 2^2 + 200 \times 58^2 = 242\,000 + 1\,200 + 672\,800 = 916\,000$. D'où $(N-1)\,S^2 = 1\,142\,964$ et $S^2 = 1\,144{,}11$. Sondage aléatoire simple : $V(\overline{y}) = (1-0{,}1) \times 1\,144{,}11/100 = 10{,}297$.
(b) $n_h = 100\,W_h$: 50, 30, 20, avec $f_h = 0{,}1$ partout. $\sum_h W_h S_h^2 = 18 + 30 + 180 = 228$, d'où $V_{prop} = \dfrac{0{,}9}{100} \times 228 = 2{,}052$ et $deff = 2{,}052/10{,}297 = 0{,}199$. Les moyennes des strates sont très différentes : la partie inter fait 80 % de la variance totale, et la stratification la supprime.
(c) $N_h S_h = 3\,000\,;3\,000\,;6\,000$, de somme 12 000 : Neyman donne 25, 25, 50. $\overline{S} = 3 + 3 + 6 = 12$, d'où

$$V_{Ney} = \frac{12^2}{100} - \frac{228}{1\,000} = 1{,}44 - 0{,}228 = 1{,}212, \qquad deff = \frac{1{,}212}{10{,}297} = 0{,}118.$$

Contrôle par la formule générale : $0{,}25 \times 0{,}95 \times \frac{36}{25} + 0{,}09 \times \frac{275}{300} \times \frac{100}{25} + 0{,}04 \times 0{,}75 \times \frac{900}{50} = 0{,}342 + 0{,}330 + 0{,}540 = 1{,}212$.
(d) $\sum_h W_h (S_h - \overline{S})^2 = 0{,}5{\times}36{+}0{,}3{\times}4{+}0{,}2{\times}324 = 18{+}1{,}2{+}64{,}8 = 84$, et $84/100 = 0{,}84 = 2{,}052 - 1{,}212$.
(e) $f_h = 25/500 = 0{,}05$; $25/300 = 0{,}083$; $50/200 = 0{,}25$: dans les trois strates, $f_h/S_h = 1/120$. Les poids valent $d_k = 20$, 12 et 4. La strate 3 fournit la moitié de l'échantillon mais ne pèse que 20 % de la population : la moyenne simple des 100 unités vaudrait en espérance $(25{\times}40{+}25{\times}60{+}50{\times}120)/100 = 85$, au lieu de 62. Il faut pondérer : $\overline{y}_{st} = \sum_h W_h \overline{y}_h$.

Correction — Exercice 2

(a) Seul $\sum_h W_h^2 S_h^2 / n_h$ dépend des n_h. On minimise

$$\mathcal{L} = \sum_h \frac{W_h^2 S_h^2}{n_h} + \lambda\Big(c_0 + \sum_h c_h n_h - C\Big), \qquad \frac{\partial \mathcal{L}}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0 \implies n_h = \frac{W_h S_h}{\sqrt{\lambda}\,\sqrt{c_h}}.$$

Comme $W_h = N_h/N$, n_h est proportionnel à $N_h S_h/\sqrt{c_h}$. Chaque terme $1/n_h$ est convexe : ce point est bien un minimum.
(b) $N_h S_h/\sqrt{c_h} = 3\,000 \times 200/5 = 120\,000$ et $1\,000 \times 600/10 = 60\,000$: $n_1 = 2\,n_2$. Le budget variable vaut $46\,000 - 10\,000 = 36\,000 = 25 \times 2n_2 + 100\,n_2 = 150\,n_2$, d'où $n_2 = 240$ et $n_1 = 480$. Avec $W_1 = 0{,}75$ et $W_2 = 0{,}25$:

$$V = 0{,}5625 \times 0{,}84 \times \frac{40\,000}{480} + 0{,}0625 \times 0{,}76 \times \frac{360\,000}{240} = 39{,}375 + 71{,}25 = 110{,}625.$$

(c) Coût moyen d'une unité sous l'allocation proportionnelle : $0{,}75 \times 25 + 0{,}25 \times 100 = 43{,}75$ dh, d'où $n = 36\,000/43{,}75 = 822{,}9$, soit 617,1 et 205,7 unités. $V_{prop} = \dfrac{1 - 822{,}9/4\,000}{822{,}9} \times (0{,}75 \times 40\,000 + 0{,}25 \times 360\,000) = 0{,}000965 \times 120\,000 = 115{,}83$. L'allocation optimale tire 103 unités de moins (720 contre 822,9), mais là où elles servent : la variance baisse de 4,5 % à coût égal.
(d) On minimise $\sum_h c_h n_h$ sous la contrainte $\sum_h W_h^2 S_h^2/n_h = V_0 + K$, avec $K = \frac{1}{N}\sum_h W_h S_h^2$. La condition $c_h - \mu W_h^2 S_h^2/n_h^2 = 0$ donne $n_h = \sqrt{\mu}\,W_h S_h/\sqrt{c_h}$: la même forme. En reportant dans la contrainte, $\sum_h W_h S_h \sqrt{c_h}/\sqrt{\mu} = V_0 + K$, d'où la formule. Ici $\sum_k W_k S_k \sqrt{c_k} = 750 + 1\,500 = 2\,250$, $K = 120\,000/4\,000 = 30$, et $W_h S_h/\sqrt{c_h} = 30$ et 15.

- $V_0 = 110{,}625$: facteur $2\,250/140{,}625 = 16$, d'où $n_1 = 480$ et $n_2 = 240$, l'allocation de (b).
- $V_0 = 60$: facteur $2\,250/90 = 25$, d'où $n_1 = 750$ et $n_2 = 375$, pour $10\,000 + 18\,750 + 37\,500 = 66\,250$ dh.

Le coût variable est proportionnel à $1/(V_0 + K)$: passer de $V = 110{,}6$ à 60 le multiplie par $140{,}625/90 = 1{,}5625$.

Correction — Exercice 3

(a) $d_h = N_h/400 = 125$, 75 et 50. Somme des poids : $400 \times (125 + 75 + 50) = 100\,000 = N$.
(b) $\sum_h n_h d_h^2 = 400 \times (15\,625 + 5\,625 + 2\,500) = 9\,500\,000$, d'où

$$deff_{Kish} = \frac{1\,200 \times 9\,500\,000}{100\,000^2} = 1{,}14, \qquad n_{eff} = \frac{1\,200}{1{,}14} = 1\,052{,}6, \qquad deft = \sqrt{1{,}14} = 1{,}068.$$

Les 1 200 ménages valent 1 053 ménages tirés au hasard simple ; les marges sont 6,8 % plus larges.
(c) Moyennes et écarts-types égaux : $V_{SAS} = S^2/n$ et $V_{\text{égale}} = \sum_h W_h^2 S^2/n_h$, d'où $deff = n\sum_h W_h^2/n_h = 1\,200 \times (0{,}25 + 0{,}09 + 0{,}04)/400 = 3 \times 0{,}38 = 1{,}14$. C'est bien la formule de Kish : $W_h^2/n_h = n_h d_h^2/N^2$ et $\sum_h n_h d_h = N$.
(d) $V_{\text{égale}} = (0{,}25 \times 9\,000\,000 + 0{,}09 \times 4\,000\,000 + 0{,}04 \times 2\,250\,000)/400 = 2\,700\,000/400 = 6\,750$ et $V_{prop} = (0{,}5 \times 9\,000\,000 + 0{,}3 \times 4\,000\,000 + 0{,}2 \times 2\,250\,000)/1\,200 = 6\,150\,000/1\,200 = 5\,125$. Rapport : 1,317, au-delà de 1,14. La région la plus sous-représentée (poids 125) est aussi la plus dispersée : la formule de Kish, qui suppose des dispersions égales, sous-estime la perte. Neyman : $N_h S_h = 150$, 60 et 30 millions, d'où 750, 300 et 150 ménages.
(e) $S^2 \approx 0{,}5 \times 9\,000\,000 + 0{,}3 \times 4\,000\,000 + 0{,}2 \times 2\,250\,000 = 6\,150\,000$, d'où, sans correction de population finie, $n_0 = 1{,}96^2 \times 6\,150\,000/200^2 = 590{,}6$, soit 591 ménages en aléatoire simple. Comme la variance est inversement proportionnelle à n, il faut multiplier cette taille par l'effet de plan. D'après Kish, $590{,}6 \times 1{,}14 = 673{,}3$, soit 674 ménages ; en réalité, $590{,}6 \times 1{,}317 = 777{,}9$, soit 778 ménages. La règle approchée oublierait 104 ménages : un effet de plan se calcule avec les dispersions réelles quand on les connaît.

2 Exercice cumulatif (chapitres 7, 6, 5, 4 et 3)

Correction — Exercice 4

(a) $n_0 = (1{,}96 \times 3\,984{,}29/300)^2 = 677{,}6$ et $n = 677{,}6/(1 + 677{,}6/2\,000) = 506{,}1$, soit 507 ménages.
(b) $V = (300/1{,}96)^2 = 23\,427{,}7$; $\sum_h W_h S_h^2 = 15\,142\,148$ et $\frac{1}{N}\sum_h W_h S_h^2 = 7\,571{,}1$. D'où $n = 15\,142\,148/(23\,427{,}7 + 7\,571{,}1) = 488{,}5$, soit 489 ménages.
(c) $\overline{S} = 3\,866{,}35$ et $\overline{S}^2 = 14\,948\,674$: $n = 14\,948\,674/30\,998{,}8 = 482{,}2$, soit 483 ménages. Parts $N_h S_h / \sum_k N_k S_k$: 45,3 %, 28,0 %, 17,7 %, 9,0 %, d'où 218,77 ; 135,48 ; 85,29 ; 43,47. Arrondis à l'entier le plus proche, ils ne totalisent que 482 : on donne le ménage manquant à Al Hoceïma, la province la moins bien estimée, d'où 219, 135, 85 et 44. Avec ces entiers, $V = \sum_h W_h^2(1 - f_h)S_h^2/n_h = 23\,379{,}2$, erreur-type 152,9 dh, marge $\pm 299{,}7$ dh : l'objectif est tenu.
(d) $S_h^2 = \frac{N_h}{N_h - 1} P_h(1 - P_h) = 0{,}1931$; 0,2037 ; 0,2426 ; 0,2509. Avec 219, 135, 85 et 44 ménages, $V(\widehat{P}_{st}) = 0{,}0003455$, soit une erreur-type de 0,0186 (marge $\pm 3{,}6$ points). Un sondage aléatoire simple de 483 ménages : $(1 - 483/2\,000) \times 0{,}21983/483 = 0{,}0003452$, la même erreur-type ($deff = 1{,}001$). L'allocation optimale pour la dépense n'apporte **rien** pour internet : une allocation n'est optimale que pour une variable.
(e) $624 - 483 = 141$ ménages de plus, soit 29 %. Ils paient la précision **provinciale** : l'allocation égale donne à chaque province, Al Hoceïma comprise, une estimation de qualité comparable (chapitre 7, domaines).

3 Applications en sciences économiques

Correction — Exercice 5

(a) $W_h = 0{,}8247$; 0,1649 ; 0,0103, d'où $n_h = 247{,}4$; 49,5 ; 3,1.
(b) $N_h S_h = 80\,000$; 64 000 ; 30 000, de somme 174 000 : 137,9 ; 110,3 ; 51,7. Il n'y a que 50 grandes exploitations : on les **recense** toutes ($f_h = 1$), et l'on répartit les 250 autres au prorata de 80 000 et 64 000 : 138,9 et 111,1, soit 139 petites et 111 moyennes.

(**c**) Proportionnelle : $V = \sum_h W_h^2(1-f_h)S_h^2/n_h = 1{,}032 + 3{,}301 + 11{,}606 = 15{,}939$, erreur-type 3,99 milliers de dh. Neyman avec strate exhaustive : $1{,}8910 + 1{,}3495 + 0 = 3{,}2405$, erreur-type 1,80. La variance est divisée par près de 5. Sous l'allocation proportionnelle, les 3 grandes exploitations tirées font 72,8 % de la variance.
(**d**) Erreur-type du total = $N\times$ erreur-type de la moyenne : $4\,850 \times \sqrt{15{,}939} = 19\,363$ milliers de dh avec l'allocation proportionnelle, $4\,850 \times \sqrt{3{,}2405} = 8\,731$ avec l'allocation retenue.
(**e**) Poids : $4\,000/138{,}9 = 28{,}8$ pour une petite exploitation, $800/111{,}1 = 7{,}2$ pour une moyenne, 1 pour une grande. Les grandes unités concentrent la dispersion, et souvent une grande part du total : en tirer quelques-unes au hasard ferait dépendre l'estimation de la chance. Les recenser supprime leur erreur d'échantillonnage pour un coût modeste, car elles sont peu nombreuses.

Correction — Exercice 6

(**a**) $W_h = 2/3$ et $1/3$; $P = \frac{2}{3} \times 0{,}20 + \frac{1}{3} \times 0{,}10 = 0{,}167$. $S_h = \sqrt{0{,}2 \times 0{,}8} = 0{,}4$ en ville, $\sqrt{0{,}1 \times 0{,}9} = 0{,}3$ à la campagne.
(**b**) $N_h S_h/\sqrt{c_h} = 400\,000 \times 0{,}4/10 = 16\,000$ et $200\,000 \times 0{,}3/20 = 3\,000$: $n_u = 16k$ et $n_r = 3k$. Budget : $100 \times 16k + 400 \times 3k = 2\,800k = 280\,000$, d'où $k = 100$, $n_u = 1\,600$ et $n_r = 300$.

$$V(\widehat{P}_{st}) = \tfrac{4}{9} \times \frac{0{,}16}{1\,600} + \tfrac{1}{9} \times \frac{0{,}09}{300} = 4{,}444 \cdot 10^{-5} + 3{,}333 \cdot 10^{-5} = 7{,}778 \cdot 10^{-5},$$

soit une erreur-type de 0,0088 et une marge de $\pm 1{,}73$ point.
(**c**) Coût moyen : $\frac{2}{3} \times 100 + \frac{1}{3} \times 400 = 200$ dh, d'où 1 400 ménages (933,3 urbains, 466,7 ruraux). $V = (\frac{2}{3} \times 0{,}16 + \frac{1}{3} \times 0{,}09)/1\,400 = 9{,}762 \cdot 10^{-5}$, erreur-type 0,0099. L'allocation optimale réduit la variance de 20 % au même coût.
(**d**) Marge rurale : $1{,}96\sqrt{0{,}09/300} = 0{,}034$, soit $\pm 3{,}4$ points : l'exigence n'est pas tenue. Il faut $n_r \geq 0{,}09 \times (1{,}96/0{,}03)^2 = 384{,}2$, soit 385 ménages ruraux, pour 154 000 dh. Il reste 126 000 dh, soit 1 260 ménages urbains. Alors $V = \frac{4}{9} \times 0{,}16/1\,260 + \frac{1}{9} \times 0{,}09/385 = 5{,}644 \cdot 10^{-5} + 2{,}597 \cdot 10^{-5} = 8{,}241 \cdot 10^{-5}$, erreur-type 0,0091 (marge $\pm 1{,}78$ point) : la précision régionale ne perd que 3 % en erreur-type. Le plancher rural coûte peu.
(**e**) $n_r \geq 0{,}09 \times (1{,}96/0{,}02)^2 = 864{,}4$, soit 865 ménages ruraux, pour 346 000 dh : plus que tout le budget. Il faut soit l'augmenter, soit assouplir l'exigence, soit réduire le coût rural, par exemple en regroupant les interviews dans des douars tirés au hasard (chapitre 9).