Échantillonnage

Chapitre 4 : Proportions, totaux et tirage avec remise

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 — Énumérer les estimations d'une proportion

Une petite région compte six coopératives artisanales, dont deux exportent. On en tire trois par sondage aléatoire simple, et l'on note $\widehat{P}$ la proportion de coopératives exportatrices dans l'échantillon.

(a) Calculer P et $S^2 = \frac{N}{N-1}P(1-P)$.

(b) Combien y a-t-il d'échantillons possibles ? Donner la loi du nombre de coopératives exportatrices tirées, puis celle de $\widehat{P}$.

(c) Calculer $E(\widehat{P})$ et $V(\widehat{P})$ par énumération. Retrouver $V(\widehat{P})$ par la formule $\frac{N-n}{N-1}\,\frac{P(1-P)}{n}$.

(d) Calculer, pour chaque valeur possible de $\widehat{P}$, l'estimation $\widehat{V}(\widehat{P}) = (1-f)\,\frac{\widehat{P}(1-\widehat{P})}{n-1}$. Vérifier que $E[\widehat{V}(\widehat{P})] = V(\widehat{P})$.

(e) Un analyste utilise plutôt $\widehat{P}(1-\widehat{P})/n$. Calculer l'espérance de cet estimateur. De combien surestime-t-il la variance, et pourquoi ?

Exercice 2 — Superficie et irrigation

Une commune rurale compte 1 200 exploitations agricoles. Un sondage aléatoire simple de 60 exploitations donne une superficie moyenne de $\overline{y} = 4{,}5$ ha, avec un écart-type $s = 3{,}2$ ha. Parmi les 60 exploitations, 18 sont irriguées ; leur superficie moyenne est de 7,2 ha.

(a) Estimer la superficie agricole totale de la commune, son erreur-type et un intervalle de confiance à 95 %. Calculer le coefficient de variation.

(b) Estimer la proportion et le nombre d'exploitations irriguées, avec leurs erreurs-types et leurs intervalles de confiance à 95 %.

(c) Estimer le nombre d'exploitations non irriguées. Comparer son erreur-type et son coefficient de variation à ceux du nombre d'exploitations irriguées.

(d) Estimer la superficie irriguée totale. Le recensement agricole indique que la commune compte 330 exploitations irriguées : proposer une seconde estimation. Laquelle préférer ?

Exercice 3 — Un tirage avec remise et ses répétitions

Un souk compte $N = 40$ commerces numérotés de 1 à 40. Pour estimer leur chiffre d'affaires hebdomadaire, on effectue 8 tirages **avec remise**. Les numéros tirés et les chiffres d'affaires (milliers de dh) sont :

Tirage	1	2	3	4	5	6	7	8
Numéro du commerce	9	37	25	37	18	4	31	25
Chiffre d'affaires	12	7	15	7	9	20	11	15

(a) Combien de commerces distincts a-t-on observés ? Combien en attendait-on ? Quelle était la probabilité d'observer au moins une répétition ?

(b) Estimer le chiffre d'affaires moyen par $\overline{y}_{AR}$, répétitions comprises, puis estimer sa variance et son erreur-type. En déduire une estimation du chiffre d'affaires total du souk.

(c) Combien de fois la variance de $\overline{y}_{AR}$ est-elle plus grande que celle de la moyenne d'un sondage aléatoire simple de 8 commerces ?

(d) Calculer la moyenne des seuls commerces distincts. Pourquoi diffère-t-elle de $\overline{y}_{AR}$? Pourquoi garde-t-on en général $\overline{y}_{AR}$?

2 Exercice cumulatif (chapitres 4, 3, 2 et 1)

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

On reprend les cinq ménages A, B, C, D, E du cours, de dépenses 4, 6, 7, 9 et 14 milliers de dh ($Y = 40$, $S^2 = 14{,}5$, $\sigma^2 = 11{,}6$). On en tire $n = 3$ par sondage aléatoire simple.

(a) Calculer π_k, π_{kl}, $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ et le poids d_k (chapitres 1 et 2).

(b) L'échantillon tiré est $\{A, C, E\}$. Calculer l'estimation de Horvitz-Thompson du total et vérifier qu'elle vaut $N\overline{y}$. Estimer sa variance.

(c) Calculer la variance exacte de $\widehat{Y} = N\overline{y}$ (chapitre 3).

(d) On s'intéresse à la proportion P de ménages dépensant plus de 8 milliers de dh. Donner $\widehat{P}$ pour chacun des dix échantillons, puis $E(\widehat{P})$ et $V(\widehat{P})$; comparer à la formule du chapitre.

(e) On tire maintenant 3 ménages **avec remise**. Calculer $V(\overline{y}_{AR})$, la comparer à la variance sans remise, et retrouver le rapport $(N-1)/(N-n)$. Combien de ménages distincts attend-on ?

3 Applications en sciences économiques

Exercice 5 — Les éleveurs d'une coopérative laitière

Une coopérative laitière du Gharb compte 1 600 éleveurs adhérents (chiffres inventés). Pour préparer un programme d'équipement, elle tire 160 numéros d'adhérents **avec remise** dans son registre et interroge les éleveurs correspondants. Sur les 160 tirages, 48 correspondent à des éleveurs équipés d'une cuve de réfrigération ; la production moyenne est de 210 litres par jour, avec un écart-type $s = 95$ litres.

(a) Estimer la proportion et le nombre d'éleveurs équipés, avec leurs erreurs-types et leurs intervalles de confiance à 95 %.

(b) Estimer la production laitière totale des adhérents et donner un intervalle de confiance à 95 %.

(c) Combien d'éleveurs distincts la coopérative s'attendait-elle à interroger ?

(d) Avec 160 éleveurs tirés sans remise, de quel facteur les variances estimées auraient-elles été multipliées ? Donner l'erreur-type de la proportion d'éleveurs équipés dans ce cas. La différence justifie-t-elle de renoncer à la simplicité du tirage avec remise ?

Exercice 6 — Taux d'activité et taux de chômage dans une commune

Une enquête communale sur l'emploi (chiffres inventés) tire, par sondage aléatoire simple, 600 personnes parmi les 30 000 habitants de 15 ans et plus d'une commune. Parmi elles, 270 sont actives (en emploi ou au chômage) et 41 sont au chômage.

(a) Estimer le taux d'activité, son erreur-type et un intervalle de confiance à 95 %. Estimer le nombre d'actifs de la commune.

(b) Le taux de chômage est la proportion de chômeurs parmi les actifs. Pourquoi est-ce une proportion de domaine ? Estimer ce taux, son erreur-type et un intervalle de confiance à 95 %.

(c) Estimer le nombre de chômeurs de la commune, avec un intervalle de confiance à 95 % et son coefficient de variation.

(d) Comparer les coefficients de variation du taux d'activité et du taux de chômage. Quelle aurait été l'erreur-type du taux de chômage si l'on avait pu interroger 600 actifs ? Que faudrait-il faire pour publier un taux de chômage précis par quartier ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $P = 2/6 = 1/3$ et $S^2 = \frac{6}{5} \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{15} = 0{,}267$.

(b) $\binom{6}{3} = 20$ échantillons équiprobables. Le nombre X de coopératives exportatrices tirées suit une loi hypergéométrique :

X	0	1	2
Nombre d'échantillons	$\binom{4}{3} = 4$	$\binom{2}{1}\binom{4}{2} = 12$	$\binom{2}{2}\binom{4}{1} = 4$
Probabilité	0,2	0,6	0,2
$\widehat{P} = X/3$	0	1/3	2/3

(c) $E(\widehat{P}) = 0{,}6 \times \frac{1}{3} + 0{,}2 \times \frac{2}{3} = \frac{1}{3} = P$: l'estimateur est sans biais. Les écarts à $\frac{1}{3}$ valent $-\frac{1}{3}$, 0 et $+\frac{1}{3}$, d'où $V(\widehat{P}) = 0{,}2 \times \frac{1}{9} + 0{,}2 \times \frac{1}{9} = \frac{2}{45} = 0{,}0444$. La formule donne $\frac{6-3}{6-1} \times \frac{\frac{1}{3} \times \frac{2}{3}}{3} = \frac{3}{5} \times \frac{2}{27} = \frac{2}{45}$, comme $(1-f)\,S^2/n = 0{,}5 \times \frac{4}{15}/3$.

(d) Ici $f = 0{,}5$ et $n - 1 = 2$. Pour $\widehat{P} = 0$: $\widehat{V} = 0$. Pour $\widehat{P} = \frac{1}{3}$ ou $\frac{2}{3}$: $\widehat{V} = 0{,}5 \times \frac{\frac{1}{3} \times \frac{2}{3}}{2} = \frac{1}{18} = 0{,}0556$. Donc $E(\widehat{V}) = 0{,}8 \times \frac{1}{18} = \frac{2}{45} = V(\widehat{P})$: l'estimateur de la variance est sans biais, bien qu'il se trompe sur chaque échantillon (0 ou 0,0556 au lieu de 0,0444).

(e) $E[\widehat{P}(1-\widehat{P})] = 0{,}8 \times \frac{2}{9} = \frac{8}{45}$, d'où $E[\widehat{P}(1-\widehat{P})/3] = \frac{8}{135} = 0{,}0593$, soit $\frac{4}{3}$ de la vraie variance : une surestimation de 33 %. Deux oublis se combinent : le facteur $1 - f = 0{,}5$ (qui multiplie par 2) et le dénominateur $n - 1$ au lieu de n (qui multiplie par $\frac{2}{3}$). Le second est négligeable pour un grand n ; le premier ne l'est jamais quand le taux de sondage est élevé.

Correction — Exercice 2

(a) $f = 60/1\,200 = 0{,}05$. Erreur-type de la moyenne : $\sqrt{0{,}95 \times 3{,}2^2/60} = 0{,}403$ ha. Superficie totale : $\widehat{Y} = 1\,200 \times 4{,}5 = 5\,400$ ha, erreur-type $1\,200 \times 0{,}403 = 483{,}2$ ha ; intervalle $5\,400 \pm 947$, soit $[4\,453\,;6\,347]$ ha. $CV = 0{,}403/4{,}5 = 8{,}9$ %, le même pour la moyenne et le total.

(b) $\widehat{P} = 18/60 = 0{,}3$; $\widehat{V} = 0{,}95 \times \frac{0{,}3 \times 0{,}7}{59} = 0{,}00338$, erreur-type 0,0581 ; intervalle $0{,}3 \pm 0{,}114$, soit $[0{,}186\,;0{,}414]$. Nombre : $1\,200 \times 0{,}3 = 360$ exploitations, erreur-type $1\,200 \times 0{,}0581 = 69{,}8$; intervalle $[223\,;497]$. $CV = 0{,}0581/0{,}3 = 19{,}4$ %.

(c) $1\,200 \times 0{,}7 = 840$ exploitations non irriguées, avec la **même** erreur-type, 69,8, puisque $N(1-\widehat{P}) = N - N\widehat{P}$. Mais le CV tombe à $69{,}8/840 = 8{,}3$ % : le groupe le plus nombreux est mieux connu en termes relatifs.

(d) Sans connaître le nombre d'exploitations irriguées : $\frac{N}{n}\sum_{k \in S \cap U_d} y_k = 20{\times}18{\times}7{,}2 = 2\,592$ ha, c'est-à-dire $\widehat{N}_d\,\overline{y}_d = 360 \times 7{,}2$. Avec $N_d = 330$ connu : $330 \times 7{,}2 = 2\,376$ ha. On préfère la seconde : l'échantillon a tiré un peu trop d'exploitations irriguées (360 estimées contre 330), et l'estimateur $N_d\,\overline{y}_d$ ne subit plus l'incertitude sur la taille du domaine, seulement celle de la superficie moyenne.

Correction — Exercice 3

(a) Les commerces 37 et 25 ont été tirés deux fois : on a observé 6 commerces distincts. On en attendait $40\left[1 - (39/40)^8\right] = 7{,}33$. La probabilité de n'avoir aucune répétition vaut $\frac{40 \times 39 \times \cdots \times 33}{40^8} = 0{,}473$: on observe au moins une répétition plus d'une fois sur deux (0,527). Avec 8 tirages sur 40 commerces, les répétitions ne sont pas rares.

(b) $\overline{y}_{AR} = 96/8 = 12$ milliers de dh, répétitions comprises. Écarts au carré : 0, 25, 9, 25, 9, 64, 1, 9, de somme 142 ; $s^2 = 142/7 = 20{,}29$. D'où $\widehat{V}(\overline{y}_{AR}) = s^2/n = 2{,}54$ et une erreur-type de 1,59 millier de dh,

sans facteur $1-f$. Total : $40\times 12 = 480$ milliers de dh par semaine, erreur-type $40\times 1{,}59 = 63{,}7$.
(**c**) $\dfrac{V(\bar{y}_{AR})}{V_{SAS}} = \dfrac{N-1}{N-n} = \dfrac{39}{32} = 1{,}22$: la variance est plus grande de 22 %. Le taux de sondage $f = 0{,}2$ est élevé, la perte n'est pas négligeable.
(**d**) Moyenne des six commerces distincts : $74/6 = 12{,}33$. $\bar{y}_{AR}$ compte deux fois les commerces 37 (7) et 25 (15) : chacun y pèse 2/8, contre 1/6 dans la moyenne des commerces distincts. On montre que la moyenne des unités distinctes est, elle aussi, sans biais, et même un peu plus précise ; mais sa variance n'a pas d'expression simple. $\bar{y}_{AR}$ reste l'estimateur d'usage : sa variance, σ^2/n, s'estime par s^2/n sans aucun calcul de probabilités d'inclusion.

2 Exercice cumulatif (chapitres 4, 3, 2 et 1)

Correction — Exercice 4

(**a**) $\pi_k = 3/5 = 0{,}6$; $\pi_{kl} = \dfrac{3\times 2}{5\times 4} = 0{,}3$; $\Delta_{kl} = 0{,}3 - 0{,}36 = -0{,}06$; $d_k = 5/3$.
(**b**) $\widehat{Y}_{HT} = \frac{5}{3}(4+7+14) = \frac{125}{3} = 41{,}67$, et $N\bar{y} = 5\times\frac{25}{3} = 41{,}67$: dans un sondage aléatoire simple, $d_k = N/n$ et l'estimateur de Horvitz-Thompson est $N\bar{y}$. Avec $\bar{y} = 8{,}33$, les écarts valent $-4{,}33$, $-1{,}33$ et $5{,}67$; $s^2 = 52{,}67/2 = 26{,}33$ et $\widehat{V}(\widehat{Y}) = 25\times 0{,}4\times 26{,}33/3 = 87{,}8$, erreur-type estimée 9,37.
(**c**) $V(\widehat{Y}) = N^2(1-f)\,S^2/n = 25\times 0{,}4\times 14{,}5/3 = 48{,}33$, erreur-type 6,95. L'estimation de (b) est presque deux fois trop forte : l'échantillon $\{A, C, E\}$ contient les deux dépenses extrêmes. $\widehat{V}$ n'est juste qu'en moyenne sur les échantillons.
(**d**) D et E dépensent plus de 8 : $P = 0{,}4$.

$\widehat{P}$	0	1/3	2/3
Échantillons	ABC	ABD, ABE, ACD, ACE, BCD, BCE	ADE, BDE, CDE

$E(\widehat{P}) = (6\times\frac{1}{3} + 3\times\frac{2}{3})/10 = 0{,}4$. $V(\widehat{P}) = [0{,}4^2 + 6\times(\frac{1}{3}-0{,}4)^2 + 3\times(\frac{2}{3}-0{,}4)^2]/10 = (0{,}16 + 0{,}0267 + 0{,}2133)/10 = 0{,}04$, et la formule donne $\dfrac{5-3}{5-1}\times\dfrac{0{,}4\times 0{,}6}{3} = 0{,}04$.
(**e**) $V(\bar{y}_{AR}) = \sigma^2/n = 11{,}6/3 = 3{,}87$, contre $(1-0{,}6)\times 14{,}5/3 = 1{,}93$ sans remise : un rapport de $2 = \dfrac{5-1}{5-3}$. On attend $5\left[1-0{,}8^3\right] = 2{,}44$ ménages distincts sur 3 tirages. Avec un taux de sondage de 60 %, le tirage avec remise gaspille beaucoup de tirages.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) $\widehat{P} = 48/160 = 0{,}3$. Avec remise, $s^2 = \frac{n}{n-1}\widehat{P}(1-\widehat{P})$ et $\widehat{V} = s^2/n = \dfrac{0{,}3\times 0{,}7}{159} = 0{,}00132$, sans facteur $1-f$; erreur-type 0,0363 ; intervalle $0{,}3 \pm 0{,}071$, soit $[0{,}229\,;0{,}371]$. Nombre d'éleveurs équipés : $1\,600\times 0{,}3 = 480$, erreur-type 58,1 ; intervalle $[366\,;594]$.
(**b**) Production totale : $1\,600\times 210 = 336\,000$ litres par jour. Erreur-type de la moyenne $95/\sqrt{160} = 7{,}51$ litres, du total 12 017 litres ; intervalle $336\,000 \pm 23\,553$, soit environ $[312\,000\,;360\,000]$ litres par jour. $CV = 3{,}6\,\%$.
(**c**) $1\,600\left[1-(1-1/1\,600)^{160}\right] = 152{,}3$ éleveurs distincts : environ 7,7 tirages retombent sur un éleveur déjà tiré.
(**d**) Sans remise, les variances estimées sont multipliées par $1-f = 0{,}9$ et les erreurs-types par $\sqrt{0{,}9} = 0{,}949$: 0,0345 au lieu de 0,0363 pour la proportion. Le tirage avec remise surestime un peu l'erreur ; il est donc **prudent**. Un écart de 5 % sur une erreur-type ne justifie pas de compliquer le tirage ; il le justifierait si la coopérative interrogeait la moitié de ses adhérents.

Correction — Exercice 6

(a) $f = 600/30\,000 = 0{,}02$. Taux d'activité : $\widehat{P} = 270/600 = 0{,}45$; $\widehat{V} = 0{,}98 \times \dfrac{0{,}45 \times 0{,}55}{599} = 0{,}000405$, erreur-type 0,0201 ; intervalle $0{,}45 \pm 0{,}039$, soit $[0{,}411\,;0{,}489]$. Nombre d'actifs : $30\,000 \times 0{,}45 = 13\,500$, erreur-type 604.

(b) Le taux de chômage se calcule sur la sous-population des actifs, dont l'effectif dans l'échantillon, $n_d = 270$, dépend du tirage et dont l'effectif total N_d est inconnu : c'est une proportion de domaine. $\widehat{P}_d = 41/270 = 0{,}152$. On remplace n_d/N_d par $n_d/\widehat{N}_d = 270/13\,500 = 0{,}02 = f$: $\widehat{V} = 0{,}98 \times \dfrac{0{,}152 \times 0{,}848}{269} = 0{,}000469$, erreur-type 0,0217 ; intervalle $0{,}152 \pm 0{,}042$, soit $[0{,}109\,;0{,}194]$.

(c) Le nombre de chômeurs est le total de l'indicatrice « être au chômage » sur toute la population : $30\,000 \times 41/600 = 2\,050$. Erreur-type : $30\,000 \times \sqrt{0{,}98 \times \dfrac{0{,}0683 \times 0{,}9317}{599}} = 306$; intervalle $2\,050 \pm 600$, soit $[1\,450\,;2\,650]$. $CV = 306/2\,050 = 14{,}9\,\%$.

(d) CV du taux d'activité : $0{,}0201/0{,}45 = 4{,}5\,\%$; du taux de chômage : $0{,}0217/0{,}152 = 14{,}3\,\%$. Le taux de chômage repose sur 270 personnes seulement, et porte sur une proportion plus faible. Avec 600 actifs, son erreur-type aurait été de 0,0143, une fois et demie plus petite. Pour publier un taux par quartier, il faudrait fixer à l'avance un nombre suffisant de personnes dans chaque quartier : c'est le sondage stratifié (chapitres 6 et 7), avec une taille calculée pour la précision voulue (chapitre 5).