Échantillonnage

Chapitre 5 : Choisir la taille de l'échantillon

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 — Le chiffre d'affaires des PME

Une fédération régionale compte $N = 1\,200$ PME. Une enquête de l'an dernier indique un chiffre d'affaires annuel moyen de 4,5 millions de dh, avec un écart-type de 3,6 millions de dh. On veut estimer le chiffre d'affaires moyen par sondage aléatoire simple, au niveau de confiance de 95 % ($z = 1{,}96$).

(a) Quelle taille faut-il pour une marge de $\pm 0{,}5$ million de dh ? Pour $\pm 0{,}25$ million ? Comparer le rapport des deux tailles à celui qu'on obtiendrait dans une population infinie.

(b) Calculer le coefficient de variation. En déduire la taille nécessaire pour une précision relative de 10 %, puis de 5 %.

(c) Faute de moyens, la fédération n'interroge que 180 PME. Quelle marge peut-elle attendre ?

(d) Pourquoi raisonne-t-on volontiers en précision relative pour un montant comme le chiffre d'affaires ?

Exercice 2 — Une table de tailles pour une proportion

Une commune compte $N = 5\,000$ ménages. On veut y estimer des proportions dont on ne sait rien à l'avance, au niveau de confiance de 95 %.

(a) Pour des marges de 5, 4, 3, 2 et 1 points, calculer n_0, puis la taille $n = n_0/(1 + (n_0 - 1)/N)$. Présenter les résultats dans un tableau, avec la taille qu'exigerait une population infinie et le taux de sondage.

(b) Pour une marge de 1 point, quelle part des ménages faut-il interroger ? Que conclure ?

(c) Vérifier qu'avec la taille trouvée pour 3 points, la marge est bien tenue (pour une proportion, $S^2 = \frac{N}{N-1} P(1-P)$).

(d) Refaire le calcul pour 3 points avec la formule approchée $n_0/(1 + n_0/N)$. L'écart a-t-il une importance pratique ?

Exercice 3 — Une enquête pilote

Une association de microcrédit veut estimer l'encours moyen de ses $N = 6\,000$ clientes à ± 300 dh près, à 95 %. Une enquête pilote auprès de $m = 25$ clientes tirées au hasard donne un écart-type $s = 2\,400$ dh.

(a) Quelle taille retenir si l'on prend $S = s$?

(b) Le vrai S peut différer de s. Quelle marge obtiendrait-on avec la taille de (a) si S valait en réalité 2 880 dh (+20 %) ? 1 920 dh (−20 %) ?

(c) Par prudence, on majore s de 20 %. Quelle taille faut-il alors ? De combien augmente-t-elle ?

(d) Si l'encours suit une loi normale, $(m-1)\,s^2/S^2$ suit une loi du khi-deux à $m-1$ degrés de liberté (Statistique inférentielle, chapitre 4). Le quantile d'ordre 5 % de $\chi^2(24)$ vaut 13,848. En déduire une borne supérieure de S au niveau de 95 %, puis la taille correspondante. L'hypothèse de normalité est-elle crédible ici ?

2 Exercice cumulatif (chapitres 5, 4, 3, 2 et 1)

Exercice 4 — Compter les ménages connectés

On reprend la population du cours : $N = 2\,000$ ménages. On veut estimer le **nombre** de ménages ayant accès à internet à ± 100 ménages près, à 95 %, par sondage aléatoire simple sans remise.

(a) Traduire cet objectif en une marge sur la proportion P de ménages connectés. Quelle taille faut-il si l'on ne sait rien de P ? Et si l'on savait que $P \approx 0{,}674$?

(b) On retient $n = 323$. Calculer la probabilité d'inclusion et le poids de chaque ménage tiré.

(c) Le tirage effectué compte 216 ménages connectés parmi les 323. Calculer l'estimation de Horvitz-Thompson du nombre de ménages connectés, comme somme des poids des ménages connectés tirés. Calculer la variance estimée de $\widehat{P}$, l'erreur-type du total estimé et l'intervalle de confiance à 95 % du total. L'objectif de ± 100 ménages est-il tenu ? La vraie valeur est 1 348 : est-elle dans l'intervalle ?

(d) Avec un tirage **avec remise**, de variance σ^2/n où $\sigma^2 = P(1-P)$, combien de tirages faudrait-il pour la même précision, sans rien savoir de P ? Combien de ménages distincts compterait-on alors en moyenne ? Conclure.

3 Applications en sciences économiques

Exercice 5 — Le rendement du blé dans trois régions

Un office agricole veut estimer le rendement moyen du blé tendre (en quintaux par hectare) dans trois régions, avec une précision relative de ± 5 % **dans chaque région**, à 95 %. Les chiffres ci-dessous sont inventés pour l'exercice :

Région	Exploitations N_h	Coefficient de variation du rendement
Gharb	3 200	0,40
Saïss	2 100	0,35
Loukkos	900	0,45

On attend 85 % de réponses ; chaque exploitation visitée coûte 400 dh.

(a) Quel nombre de réponses faut-il obtenir dans chaque région ?

(b) Combien d'exploitations faut-il visiter dans chaque région, et au total ? Quel est le coût de l'enquête ?

(c) Si l'on ne voulait qu'une estimation pour l'ensemble des 6 200 exploitations, avec un coefficient de variation de 0,40, combien faudrait-il en visiter, pour quel coût ? Comparer.

(d) Pourquoi le Loukkos, la plus petite région, demande-t-il le plus grand échantillon ?

Exercice 6 — Le taux d'activité des jeunes

Un observatoire régional de l'emploi veut estimer la part des jeunes de 15 à 24 ans qui sont actifs, à ± 3 points près, à 95 %. Les jeunes sont assez nombreux pour négliger la correction de population finie. Le plan tire des quartiers, puis interroge 17 jeunes par quartier tiré ; avec une corrélation intra-quartier $\rho = 0{,}05$, l'effet de plan vaut $deff = 1 + (17-1) \times 0{,}05 = 1{,}8$ (chapitre 9). On attend 80 % de réponses. L'enquête coûte 150 000 dh de frais fixes et 250 dh par questionnaire (coûts supposés).

(a) Combien de questionnaires faut-il prévoir ? Pour quel coût ?

(b) Le budget n'est que de 600 000 dh. Combien de questionnaires peut-on financer, et combien de réponses attendre ? Quelle est la taille effective $n/deff$ et la marge obtenue ?

(c) On propose d'interroger 9 jeunes par quartier au lieu de 17, en supposant le coût par questionnaire inchangé. Quel est le nouvel effet de plan ? Quelle marge obtient-on avec 600 000 dh ? Combien de questionnaires et quel budget faudrait-il pour tenir ± 3 points ?

(d) Pourquoi le coût par questionnaire risque-t-il en réalité d'augmenter ?

1 Exercices du chapitre

Correction — Exercice 1

(a) $n_0 = (1{,}96 \times 3{,}6/0{,}5)^2 = 199{,}15$ et $n = \dfrac{199{,}15}{1 + 199{,}15/1\,200} = 170{,}80$, soit 171 PME. Pour $\pm 0{,}25$ million, $n_0 = 796{,}59$, exactement quatre fois plus, mais $n = 478{,}77$, soit 479 PME : 2,80 fois plus seulement. Dans une population infinie, diviser la marge par deux multiplierait la taille par quatre ; ici, 479 PME représentent 40 % de la population, et la correction de population finie joue fortement.
(b) $CV = 3{,}6/4{,}5 = 0{,}8$. Pour 10 %, soit $\pm 0{,}45$ million : $n_0 = (1{,}96 \times 0{,}8/0{,}10)^2 = 245{,}86$ et $n = 204{,}05$, soit 205 PME. Pour 5 % : $n_0 = 983{,}45$ et $n = 540{,}49$, soit 541 PME.
(c) $e = 1{,}96 \times 3{,}6 \times \sqrt{1/180 - 1/1\,200} = 0{,}485$ million de dh, soit environ $\pm 485\,000$ dh. Comme $180 > 171$, la marge est un peu meilleure que $\pm 0{,}5$ million, **si** l'écart-type n'a pas changé depuis l'an dernier.
(d) Le coefficient de variation est un nombre sans unité : il ne change pas si tous les chiffres d'affaires augmentent dans la même proportion (inflation, croissance), et il se transpose d'une région ou d'une année à l'autre bien mieux qu'un écart-type exprimé en dirhams.

Correction — Exercice 2

(a) Sans information, on prend $P = 0{,}5$, d'où $n_0 = 1{,}96^2 \times 0{,}25/e^2 = 0{,}9604/e^2$:

Marge	n_0	Population infinie	n	Arrondi	Taux f
± 5 points	384,2	385	356,8	357	7,1 %
± 4 points	600,25	601	536,0	537	10,7 %
± 3 points	1 067,1	1 068	879,6	880	17,6 %
± 2 points	2 401	2 401	1 622,3	1 623	32,5 %
± 1 point	9 604	9 604	3 288,4	3 289	65,8 %

(b) $3\,289/5\,000 = 65{,}8\,\%$: il faudrait interroger les deux tiers de la commune. La correction de population finie divise la taille par 2,9, mais on est presque dans un recensement. Une marge de 1 point se justifie rarement ; si on l'exige, mieux vaut se demander si un recensement ne serait pas plus simple.
(c) $e = 1{,}96\sqrt{\left(\dfrac{1}{880} - \dfrac{1}{5\,000}\right) \times \dfrac{5\,000}{4\,999} \times 0{,}25} = 0{,}02999 \leq 0{,}03$: la marge est tenue.
(d) $1\,067{,}1/(1 + 1\,067{,}1/5\,000) = 879{,}42$, soit encore 880 après arrondi. Les deux formules ne diffèrent que d'une fraction d'unité : l'écart n'a aucune importance pratique.

Correction — Exercice 3

(a) $n_0 = (1{,}96 \times 2\,400/300)^2 = 245{,}86$ et $n = \dfrac{245{,}86}{1 + 245{,}86/6\,000} = 236{,}18$, soit 237 clientes.
(b) Avec $n = 237$: si $S = 2\,880$ dh, $e = 1{,}96 \times 2\,880 \times \sqrt{1/237 - 1/6\,000} = 359{,}4$ dh ; si $S = 1\,920$ dh, $e = 239{,}6$ dh. La marge est proportionnelle à S : une erreur de 20 % sur l'écart-type se reporte entièrement sur la précision obtenue.
(c) $n_0 = (1{,}96 \times 2\,880/300)^2 = 354{,}04$, soit $1{,}2^2 = 1{,}44$ fois plus, et $n = 334{,}31$, soit 335 clientes : 41 % de plus qu'en (a), un peu moins que 44 % grâce à la correction de population finie.
(d) $P(\chi^2(24) \geq 13{,}848) = 0{,}95$. Avec la probabilité 0,95, $24\,s^2/S^2 \geq 13{,}848$, c'est-à-dire

$$S \leq s\sqrt{\frac{24}{13{,}848}} = 2\,400 \times 1{,}3165 = 3\,159{,}5 \text{ dh.}$$

Avec cette borne, $n_0 = (1{,}96 \times 3\,159{,}5/300)^2 = 426{,}1$ et $n = 397{,}8$, soit 398 clientes, 68 % de plus qu'en (a). La normalité est peu crédible : comme les dépenses, les encours ont une longue queue à droite. La borne n'est qu'indicative ; un pilote plus grand ou une enquête antérieure seraient plus sûrs.

2 Exercice cumulatif (chapitres 5, 4, 3, 2 et 1)

Correction — Exercice 4

(**a**) Le total s'estime par $N\widehat{P}$: une marge de 100 ménages sur le total est une marge de $100/2\,000 = 0{,}05$ sur P, soit 5 points. Sans information, $P = 0{,}5$: $n_0 = 384{,}16$ et $n = \dfrac{384{,}16}{1 + 383{,}16/2\,000} = 322{,}4$, soit 323 ménages. Avec $P \approx 0{,}674$, $P(1-P) = 0{,}2197$: $n_0 = 337{,}6$ et $n = 289{,}0$, soit 289 ménages.
(**b**) $\pi_k = 323/2\,000 = 0{,}1615$ pour chaque ménage (chapitre 3), et le poids vaut $d_k = 2\,000/323 = 6{,}192$ (chapitre 1) : chaque ménage tiré représente un peu plus de six ménages.
(**c**) Estimation de Horvitz-Thompson (chapitre 2) : $\widehat{Y}_{HT} = 216 \times 6{,}192 = 1\,337{,}5$ ménages connectés, soit $N\widehat{P}$ avec $\widehat{P} = 216/323 = 0{,}6687$. Variance estimée (chapitre 4) :

$$\widehat{V}(\widehat{P}) = (1-f)\,\frac{\widehat{P}(1-\widehat{P})}{n-1} = 0{,}8385 \times \frac{0{,}22153}{322} = 0{,}0005769,$$

d'où une erreur-type de 0,0240 pour $\widehat{P}$ et de $2\,000 \times 0{,}0240 = 48{,}0$ ménages pour le total. Intervalle à 95 % : $1\,337{,}5 \pm 1{,}96 \times 48{,}0 = 1\,337{,}5 \pm 94{,}2$, soit $[1\,243{,}3\,;\,1\,431{,}6]$. La marge, 94,2 ménages, tient l'objectif de 100 : la vraie proportion étant plus éloignée de 0,5 que la valeur prudente, la précision est un peu meilleure que prévu. La vraie valeur, 1 348, est dans l'intervalle ; l'estimation ne s'en écarte que de 10,5 ménages.
(**d**) Avec remise, il n'y a pas de facteur $1-f$: $n = n_0 = 384{,}16$, soit 385 tirages. On compterait en moyenne $2\,000(1-(1-1/2\,000)^{385}) = 350{,}3$ ménages distincts. Il faut donc 62 tirages de plus que sans remise, et même 27 ménages distincts de plus, pour la même précision : sur une population de 2 000 ménages, le tirage sans remise est nettement plus économique.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) Dans chaque région, $n_0 = (1{,}96 \times CV/0{,}05)^2$, puis $n = n_0/(1 + n_0/N_h)$:

Région	N_h	n_0	n_h	Réponses	Visites ($n_h/0{,}85$)
Gharb	3 200	245,86	228,32	229	268,61 → 269
Saïss	2 100	188,24	172,75	173	203,24 → 204
Loukkos	900	311,17	231,22	232	272,03 → 273
Ensemble	6 200			634	746

(**b**) Il faut visiter $269 + 204 + 273 = 746$ exploitations, pour un coût de $746 \times 400 = 298\,400$ dh.
(**c**) Pour l'ensemble : $n_0 = 245{,}86$, $n = 245{,}86/(1 + 245{,}86/6\,200) = 236{,}5$, soit 237 réponses, et $236{,}5/0{,}85 = 278{,}2$, soit 279 visites pour 111 600 dh. Exiger la précision dans chaque région multiplie la taille et le coût par 2,7 : chaque région coûte presque une enquête entière.
(**d**) La taille dépend surtout de la dispersion, et très peu de la taille de la région. Le Loukkos a le coefficient de variation le plus élevé (0,45), donc le plus grand n_0. Sa petite taille réduit n_0 davantage qu'ailleurs, de 311 à 231 (un quart des exploitations est interrogé), mais pas assez pour compenser.

Correction — Exercice 6

(**a**) $n_0 = 1{,}96^2 \times 0{,}25/0{,}03^2 = 1\,067{,}1$; avec l'effet de plan, $1\,067{,}1 \times 1{,}8 = 1\,920{,}8$ réponses ; avec la non-réponse, $1\,920{,}8/0{,}80 = 2\,401$ questionnaires. Coût : $150\,000 + 250 \times 2\,401 = 750\,250$ dh.
(**b**) $(600\,000 - 150\,000)/250 = 1\,800$ questionnaires, soit environ $1\,800 \times 0{,}80 = 1\,440$ réponses. La taille effective vaut $1\,440/1{,}8 = 800$: ces 1 440 réponses en grappes ne valent que 800 réponses tirées au hasard. Marge : $1{,}96\sqrt{0{,}25/800} = 0{,}0346$, soit $\pm 3{,}5$ points.

(**c**) $deff = 1 + 8 \times 0{,}05 = 1{,}4$. Avec 1 440 réponses, la taille effective vaut $1\,440/1{,}4 = 1\,028{,}6$ et la marge $1{,}96\sqrt{0{,}25/1\,028{,}6} = 0{,}0306$, soit $\pm 3{,}1$ points. Pour ± 3 points, il faut $1\,067{,}1 \times 1{,}4/0{,}80 = 1\,867{,}4$, soit 1 868 questionnaires, pour $150\,000 + 250 \times 1\,868 = 617\,000$ dh : 133 250 dh de moins qu'en (a), mais encore un peu plus que le budget.
(**d**) Avec 9 jeunes par quartier, il faut visiter environ $1\,868/9 \approx 208$ quartiers au lieu de $2\,401/17 \approx 142$: plus de déplacements, plus de quartiers à recenser avant le tirage. Le coût par questionnaire augmente, et le gain de précision peut être en partie absorbé. C'est l'arbitrage entre coût et précision des sondages en grappes (chapitres 9 et 10).