Échantillonnage

Chapitre 14 : La non-réponse et le redressement

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 — Propriétaires et locataires

Une ville compte 5 000 ménages : 3 000 propriétaires et 2 000 locataires. Si l'on tentait d'interroger chacun d'eux sur sa dépense mensuelle, on obtiendrait les taux de réponse et les dépenses moyennes suivants (en dh, chiffres fictifs) :

			Dépense moyenne (dh)	
Statut	Ménages	Taux de réponse τ_c	des répondants	des non-répondants
Propriétaires	3 000	0,9	9 000	9 600
Locataires	2 000	0,6	6 000	6 300

(a) Calculer la dépense moyenne de chaque statut, puis $\overline{Y}$.

(b) Calculer le nombre de répondants, le taux de réponse τ, puis $\overline{Y}_R$ et $\overline{Y}_{NR}$. En déduire le biais de la moyenne des répondants, directement puis par la formule $(1-\tau)\,(\overline{Y}_R - \overline{Y}_{NR})$.

(c) On repondère les répondants par classes de statut. Vers quelle valeur la moyenne repondérée tend-elle sur de grands échantillons ? Calculer le biais résiduel et le décomposer entre les deux classes.

(d) Le biais était positif avant la correction ; il est négatif après. Expliquer ce changement de signe. La correction a-t-elle été utile ? À quelle condition aurait-elle supprimé tout le biais ?

Exercice 2 — Trois revenus manquants

On tire $n = 10$ ménages parmi $N = 500$ par sondage aléatoire simple. Trois d'entre eux acceptent l'entretien mais refusent de déclarer leur revenu mensuel (en milliers de dh) :

Ménage	1	2	3	4	5	6	7	8	9	10
Taille (personnes)	2	3	3	4	4	5	5	6	6	7
Revenu	3,8	5,0	—	6,1	7,0	—	8,2	9,5	—	11,0

On donne, pour les sept revenus déclarés : $\sum y_k = 50{,}6$ et $\sum y_k^2 = 404{,}14$. On suppose d'abord que les revenus manquent au hasard complet.

(a) Calculer $\overline{y}_r$ et s_r^2 des sept revenus déclarés. En les traitant comme un sondage aléatoire simple de 7 ménages parmi 500, calculer l'erreur-type de $\overline{y}_r$ et un intervalle de confiance approché à 95 %.

(b) On impute la moyenne $\overline{y}_r$ aux trois ménages muets. Calculer la moyenne et la variance s_c^2 du fichier complété. Un utilisateur qui ignore l'imputation calcule l'erreur-type $\sqrt{(1-10/500)\,s_c^2/10}$ et l'intervalle correspondant : les donner. Combien vaut le rapport des deux erreurs-types ? D'où vient l'écart ?

(c) On impute plutôt par le ratio sur la taille : $\widehat{R}\,x_k$, où $\widehat{R}$ est le revenu par personne des sept ménages renseignés. Calculer les valeurs imputées et la moyenne du fichier complété. Pourquoi diffère-t-elle de $\overline{y}_r$?

(d) On impute enfin par hot deck, chaque donneur étant tiré avec remise parmi les sept ménages renseignés. Combien de fichiers complétés sont possibles ? Quelle variance ce tirage ajoute-t-il à celle de la moyenne complétée ?

Exercice 3 — Relancer une partie des absents

Une enquête sur les dépenses de santé tire $n = 400$ ménages parmi $N = 8\,000$ par sondage aléatoire simple. 280 ménages répondent, avec une dépense annuelle moyenne de 6 500 dh. On tire au hasard 30 des 120 non-répondants (un sur $k = 4$) ; un enquêteur expérimenté les relance jusqu'à obtenir leur

réponse : ils dépensent en moyenne 8 100 dh. Un questionnaire de premier passage coûte 50 dh, une relance 400 dh. On admet que la variance de l'estimateur obtenu vaut

$$V \approx (1-f)\,\frac{S^2}{n} + W_{NR}\,(k-1)\,\frac{S_{NR}^2}{n},$$

avec $S = 3\,000$ dh, $S_{NR} = 3\,200$ dh (dispersion parmi les non-répondants) et $W_{NR} = 0{,}3$ (part des non-répondants).

(a) Estimer la dépense moyenne. De combien la moyenne des seuls répondants la sous-estime-t-elle ? Pourquoi cette estimation est-elle sans biais ?

(b) Calculer l'erreur-type de l'estimation. La comparer à celle qu'on aurait obtenue si les 400 ménages avaient répondu.

(c) Calculer le coût de l'enquête. Que coûterait la relance des 120 non-répondants ? Quelle erreur-type donnerait-elle ?

(d) On envisage de relancer un non-répondant sur deux. Calculer l'erreur-type et le coût. Quel choix recommander ?

2 Exercice cumulatif (chapitres 14, 13, 12, 11 et 10)

Exercice 4 — Les non-répondants de l'échantillon fil rouge

On reprend l'échantillon fil rouge : 200 ménages tirés par sondage aléatoire simple parmi les $N = 2\,000$ ménages de la population du cours, qui compte 1 438 urbains, 562 ruraux et $X = 7\,859$ personnes. Parmi les ménages tirés, 152 sont urbains et 48 ruraux. Répondent 119 urbains, dont les dépenses mensuelles totalisent 903 640 dh, et 42 ruraux, qui totalisent 302 080 dh. Les 161 répondants comptent 607 personnes et dépensent en moyenne $\overline{y}_r = 7\,488{,}94$ dh.

(a) On traite la réponse comme un second tirage, dont la probabilité est estimée par le taux de réponse du milieu. Calculer, pour un répondant urbain et pour un répondant rural, la probabilité d'être tiré puis de répondre, et le poids correspondant. Quel plan du cours combinait déjà deux probabilités de cette façon ? Vérifier que les poids des 161 répondants somment à N.

(b) Ces probabilités sont inégales : estimer le total des dépenses par l'estimateur de Horvitz-Thompson, puis la dépense moyenne.

(c) Post-stratifier les répondants sur le milieu, avec les effectifs de la population : calculer les poids, l'estimation du total et celle de la moyenne. Quelle information de plus utilise-t-on qu'en (b) ?

(d) Estimer la dépense moyenne par le ratio sur la taille, calculé sur les répondants. Les 39 non-répondants tirés comptent en moyenne 3,95 personnes, contre 3,77 pour les répondants : expliquer pourquoi cet estimateur corrige davantage.

(e) En réalité, les 200 ménages tirés dépensent en moyenne 7 843,05 dh, et $\overline{Y} = 7\,911{,}44$ dh. Décomposer l'écart $\overline{y}_r - \overline{Y}$ en une erreur de tirage et une erreur de non-réponse. Sur la population entière, les biais de non-réponse des méthodes valent $-170{,}6$ dh (classes de milieu), $-130{,}0$ dh (ratio sur la taille) et $-63{,}8$ dh (calage sur le milieu et le nombre de personnes). Que conclure ?

3 Applications en sciences économiques

Exercice 5 — Une enquête de conjoncture auprès des entreprises

Une chambre de commerce régionale interroge, par sondage aléatoire simple dans chaque catégorie, 120 des 2 400 très petites entreprises (TPE), 50 des 500 PME et 40 des 100 grandes entreprises de son ressort. Répondent 60 TPE, 35 PME et 36 grandes entreprises, dont le chiffre d'affaires annuel moyen est respectivement de 0,9, 11 et 140 millions de dh (chiffres fictifs).

(a) Calculer les taux de réponse, les poids de sondage d_k et les poids corrigés de la non-réponse dans chaque catégorie. Vérifier la somme des poids corrigés.

(b) Estimer le chiffre d'affaires total des 3 000 entreprises avec les poids corrigés. Que donnerait le même calcul avec les seuls poids d_k ? Et la moyenne non pondérée des répondants multipliée par 3 000 ?

(c) Sous quelle hypothèse l'estimation corrigée est-elle sans biais ? On soupçonne que les grandes entreprises qui n'ont pas répondu traversent des difficultés : dans quel sens irait le biais ? Pourquoi la chambre a-t-elle intérêt à relancer d'abord les grandes entreprises ?

Exercice 6 — La production céréalière d'une commune

Une commune rurale compte 600 exploitations céréalières, qui couvrent 4 800 hectares d'après le registre agricole. On en tire 12 par sondage aléatoire simple pour estimer la production de la campagne. Trois exploitants ne déclarent pas leur production (en quintaux, chiffres fictifs) :

Surface (ha)	3	4	5	6	7	8	9	10	12	14	3	5
Production	45	58	—	86	100	—	128	150	—	196	40	72

(a) Imputer aux trois exploitations la moyenne des neuf productions déclarées, et estimer la production totale de la commune par $N\,\overline{y}$.

(b) Imputer plutôt le rendement moyen des neuf exploitations renseignées, multiplié par la surface : donner les valeurs imputées et la nouvelle estimation de $N\,\overline{y}$.

(c) Le total des surfaces étant connu, estimer la production totale par l'estimateur par le ratio $X\,\widehat{R}$ calculé sur le fichier complété en (b). Expliquer les écarts entre les trois estimations.

(d) Les rendements des exploitations renseignées vont de 13,3 à 15 quintaux par hectare. Que penser de l'imputation (b) ? Sous quelle hypothèse sur la non-réponse est-elle justifiée ?

1 Exercices du chapitre

Correction — Exercice 1

(**a**) Propriétaires : $0{,}9 \times 9\,000 + 0{,}1 \times 9\,600 = 9\,060$ dh. Locataires : $0{,}6 \times 6\,000 + 0{,}4 \times 6\,300 = 6\,120$ dh. D'où

$$\overline{Y} = \frac{3\,000 \times 9\,060 + 2\,000 \times 6\,120}{5\,000} = \frac{39\,420\,000}{5\,000} = 7\,884 \text{ dh.}$$

(**b**) Répondants : $0{,}9 \times 3\,000 + 0{,}6 \times 2\,000 = 2\,700 + 1\,200 = 3\,900$, soit $\tau = 3\,900/5\,000 = 0{,}78$. Non-répondants : $300 + 800 = 1\,100$.

$$\overline{Y}_R = \frac{2\,700 \times 9\,000 + 1\,200 \times 6\,000}{3\,900} = \frac{31\,500\,000}{3\,900} = 8\,076{,}92 \text{ dh,}$$

$$\overline{Y}_{NR} = \frac{300 \times 9\,600 + 800 \times 6\,300}{1\,100} = \frac{7\,920\,000}{1\,100} = 7\,200 \text{ dh.}$$

Biais : $8\,076{,}92 - 7\,884 = +192{,}92$ dh ($+2{,}45\,\%$). Par la formule : $0{,}22 \times (8\,076{,}92 - 7\,200) = 0{,}22 \times 876{,}92 = 192{,}92$ dh.

(**c**) Sur de grands échantillons, la moyenne repondérée tend vers $\sum_c W_c\, \overline{Y}_{R,c} = 0{,}6 \times 9\,000 + 0{,}4 \times 6\,000 = 7\,800$ dh. Biais résiduel : $7\,800 - 7\,884 = -84$ dh, soit

$$\sum_c W_c\,(1 - \tau_c)\,(\overline{Y}_{R,c} - \overline{Y}_{NR,c}) = 0{,}6 \times 0{,}1 \times (-600) + 0{,}4 \times 0{,}4 \times (-300) = -36 - 48 = -84 \text{ dh.}$$

(**d**) Le biais initial additionne deux effets de sens contraire.

- **La composition** : les propriétaires, qui dépensent plus, répondent davantage ; ils font $2\,700/3\,900 = 69{,}2\,\%$ des répondants contre $60\,\%$ de la population. Cet effet vaut $8\,076{,}92 - 7\,800 = +276{,}92$ dh.
- **Les écarts dans chaque classe** : propriétaires et locataires qui refusent dépensent plus que ceux qui répondent (600 et 300 dh d'écart). Cet effet vaut -84 dh.

Au total $276{,}92 - 84 = 192{,}92$ dh. La repondération supprime l'effet de composition et laisse le second : le biais tombe de 192,92 à 84 dh en valeur absolue ($-56{,}5\,\%$), mais change de signe. Elle l'aurait supprimé entièrement si, dans chaque statut, la réponse ne dépendait pas de la dépense (réponse au hasard conditionnellement au statut : $\overline{Y}_{NR,c} = \overline{Y}_{R,c}$). Leçon : un biais modeste peut cacher deux effets de sens contraire qui se compensent en partie ; corriger l'un découvre l'autre.

Correction — Exercice 2

(**a**) $\overline{y}_r = 50{,}6/7 = 7{,}229$ milliers de dh. Somme des carrés des écarts : $404{,}14 - 50{,}6^2/7 = 404{,}14 - 365{,}766 = 38{,}374$, d'où $s_r^2 = 38{,}374/6 = 6{,}396$. Erreur-type :

$$\sqrt{\Big(1 - \frac{7}{500}\Big)\frac{6{,}396}{7}} = \sqrt{0{,}9009} = 0{,}949, \qquad \text{intervalle } 7{,}229 \pm 1{,}96 \times 0{,}949 = [5{,}37\,;\,9{,}09].$$

Avec sept valeurs, l'approximation normale n'est qu'indicative.

(**b**) Imputer la moyenne ne change pas la moyenne : 7,229. Les trois valeurs imputées ont un écart nul à la moyenne : la somme des carrés reste 38,374, mais on divise par 9 : $s_c^2 = 38{,}374/9 = 4{,}264 = \frac{6}{9}\, s_r^2$. Erreur-type naïve : $\sqrt{0{,}98 \times 4{,}264/10} = \sqrt{0{,}4179} = 0{,}646$, et l'intervalle $7{,}229 \pm 1{,}267 = [5{,}96\,;\,8{,}50]$. Rapport des erreurs-types : $0{,}646/0{,}949 = 0{,}681$; l'intervalle naïf est trop étroit d'un tiers. Deux causes se cumulent : une variance trop petite (facteur $6/9$) et un nombre d'observations trop grand (10 au lieu de 7) : $\frac{6}{9} \times \frac{7}{10} \times \frac{0{,}98}{0{,}986} = 0{,}464 = 0{,}681^2$.

(**c**) $\widehat{R} = 50{,}6/31 = 1{,}632$ millier de dh par personne (les sept ménages renseignés comptent 31 personnes). Valeurs imputées : ménage 3 : $3 \times 1{,}632 = 4{,}90$; ménage 6 : $5 \times 1{,}632 = 8{,}16$; ménage 9 : $6 \times 1{,}632 = 9{,}79$. Somme du fichier complété : $50{,}6 + 14 \times 1{,}632 = 73{,}45$, moyenne 7,345. Elle dépasse $\overline{y}_r$ parce que les trois ménages muets sont un peu plus grands (4,67 personnes en moyenne contre 4,43) et que le revenu croît avec la taille.

(**d**) Chaque valeur manquante peut recevoir l'un des sept revenus : $7^3 = 343$ fichiers complétés. La variance des donneurs, divisée par n_r, vaut $\sigma_r^2 = 38{,}374/7 = 5{,}482$; le tirage des donneurs ajoute $m\,\sigma_r^2/n^2 = 3 \times 5{,}482/10^2 = 0{,}164$ à la variance de la moyenne complétée (un écart-type de 0,41).

Correction — Exercice 3

(**a**) Les 30 relancés représentent les 120 non-répondants :

$$\widehat{\bar{Y}} = \frac{280}{400} \times 6\,500 + \frac{120}{400} \times 8\,100 = 4\,550 + 2\,430 = 6\,980 \text{ dh.}$$

La moyenne des répondants, 6 500 dh, la sous-estime de 480 dh (6,9 %). L'estimation est sans biais : les 30 relancés forment un sondage aléatoire simple parmi les 120 non-répondants de l'échantillon, et tous ont fini par répondre ; leur moyenne estime sans biais celle des 120.
(**b**) $V \approx 0{,}95 \times \dfrac{3\,000^2}{400} + 0{,}3 \times 3 \times \dfrac{3\,200^2}{400} = 21\,375 + 23\,040 = 44\,415$, soit une erreur-type de 210,7 dh. Si les 400 ménages avaient répondu : $V = 21\,375$, erreur-type 146,2 dh. Ne relancer qu'un non-répondant sur quatre double la variance.
(**c**) Coût : $400 \times 50 + 30 \times 400 = 20\,000 + 12\,000 = 32\,000$ dh. Relancer les 120 coûterait $20\,000 + 48\,000 = 68\,000$ dh ; avec $k = 1$, le second terme de la variance s'annule et l'erreur-type revient à 146,2 dh.
(**d**) Avec $k = 2$, on relance 60 ménages : $V \approx 21\,375 + 0{,}3 \times 1 \times \dfrac{3\,200^2}{400} = 21\,375 + 7\,680 = 29\,055$, erreur-type 170,5 dh, pour un coût de $20\,000 + 24\,000 = 44\,000$ dh. Par rapport à $k = 4$, l'erreur-type baisse de 19 % pour un coût accru de 37,5 % ; relancer tout le monde la baisse de 31 % pour un coût plus que doublé. Le choix dépend du budget, mais les trois solutions sont sans biais, contrairement à la moyenne des seuls répondants, qui se trompe de 480 dh quel que soit n.

2 Exercice cumulatif (chapitres 14, 13, 12, 11 et 10)

Correction — Exercice 4

(**a**) Un ménage est tiré avec la probabilité $200/2\,000 = 0{,}1$. Urbain : $0{,}1 \times \dfrac{119}{152} = 0{,}0783$, poids 12,77. Rural : $0{,}1 \times \dfrac{42}{48} = 0{,}0875$, poids 11,43. Comme dans le sondage à deux degrés (chapitre 10), la probabilité finale est un produit : probabilité d'être tiré, puis probabilité de répondre sachant qu'on a été tiré. Somme des poids : $119 \times \dfrac{1\,520}{119} + 42 \times \dfrac{480}{42} = 1\,520 + 480 = 2\,000 = N$.
(**b**) Avec des probabilités inégales, chaque répondant compte pour l'inverse de sa probabilité (chapitres 2 et 11) :

$$\widehat{Y}_{HT} = \frac{1\,520}{119} \times 903\,640 + \frac{480}{42} \times 302\,080 = 11\,542\,292 + 3\,452\,343 = 14\,994\,635 \text{ dh,}$$

soit une dépense moyenne de $14\,994\,635/2\,000 = 7\,497{,}32$ dh : c'est la moyenne repondérée par classes, $0{,}76 \times 7\,593{,}61 + 0{,}24 \times 7\,192{,}38$.
(**c**) Poids : $1\,438/119 = 12{,}08$ (urbain) et $562/42 = 13{,}38$ (rural), de somme 2 000. Total : $\frac{1\,438}{119} \times 903\,640 + \frac{562}{42} \times 302\,080 = 10\,919\,616 + 4\,042\,118 = 14\,961\,734$ dh, moyenne 7 480,87 dh. On utilise les effectifs N_c de la population (chapitre 13) au lieu des effectifs tirés n_c : la post-stratification corrige aussi la composition aléatoire de l'échantillon, qui compte 76 % d'urbains contre 71,9 % dans la population.
(**d**) $\widehat{R} = 1\,205\,720/607 = 1\,986{,}36$ dh par personne (chapitre 12), d'où $X\,\widehat{R} = 7\,859 \times 1\,205\,720/607 = 15\,610\,797$ dh et une moyenne de 7 805,40 dh. Les ménages qui ne répondent pas sont plus grands, et les grands ménages dépensent plus : les répondants comptent trop peu de personnes (3,77 par ménage contre 3,93 dans la population). En ramenant le nombre de personnes à son total connu X, le ratio restitue une partie de la dépense des absents ; les classes de milieu ne le font pas.

(**e**) L'écart $7\,488{,}94 - 7\,911{,}44 = -422{,}49$ dh se décompose en

$$(7\,843{,}05 - 7\,911{,}44) + (7\,488{,}94 - 7\,843{,}05) = -68{,}39 - 354{,}11 :$$

une erreur de tirage de $-68{,}39$ dh (les 200 ménages tirés dépensent un peu moins que la population) et une erreur de non-réponse de $-354{,}11$ dh (les répondants dépensent moins que les ménages tirés). La non-réponse pèse cinq fois plus que le hasard du tirage. Sur cet échantillon, les écarts à $\overline{Y}$ valent $-414{,}1$ dh (classes), $-430{,}6$ dh (post-stratification) et $-106{,}0$ dh (ratio). Le ratio fait ici le mieux, mais sur la population entière le calage a un biais plus faible ($-63{,}8$ dh contre $-130{,}0$) : un seul échantillon ne suffit pas à classer les méthodes. Aucune ne supprime le biais, car la réponse dépend de la dépense elle-même, à milieu et taille donnés.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) Dans chaque catégorie h, le poids corrigé vaut $w_k = d_k\, n_h / r_h$:

Catégorie	Taux de réponse	$d_k = N_h/n_h$	Poids corrigé w_k	Répondants r_h	$r_h\, w_k$
TPE	$60/120 = 0{,}5$	20	40	60	2 400
PME	$35/50 = 0{,}7$	10	14,29	35	500
Grandes	$36/40 = 0{,}9$	2,5	2,78	36	100

Les poids corrigés somment à 3 000 : chaque catégorie retrouve son effectif.
(**b**) Total corrigé : $2\,400 \times 0{,}9 + 500 \times 11 + 100 \times 140 = 2\,160 + 5\,500 + 14\,000 = 21\,660$ millions de dh. Avec les seuls poids d_k : $1\,200 \times 0{,}9 + 350 \times 11 + 90 \times 140 = 1\,080 + 3\,850 + 12\,600 = 17\,530$ millions, soit 19,1 % de moins : les répondants ne représentent alors que 1 640 entreprises. Moyenne non pondérée : $(54 + 385 + 5\,040)/131 = 5\,479/131 = 41{,}82$ millions, et $41{,}82 \times 3\,000 = 125\,473$ millions, 5,8 fois trop : les grandes entreprises font 27,5 % des répondants, mais 3,3 % des entreprises.
(**c**) Il faut que, dans chaque catégorie, les répondantes aient en moyenne le même chiffre d'affaires que les non-répondantes. Si les grandes entreprises muettes sont en difficulté, leur chiffre d'affaires est plus faible que celui des répondantes : le total est **surestimé**. Les grandes entreprises réalisent $14\,000/21\,660 = 64{,}6\,\%$ du total estimé : une erreur sur elles fausse tout le résultat, alors qu'il suffit de relancer quatre entreprises pour les avoir toutes. Relancer en priorité les unités les plus lourdes est la règle des enquêtes auprès des entreprises.

Correction — Exercice 6

(**a**) Les neuf productions déclarées somment à 875 quintaux : moyenne 97,22 quintaux, inchangée par l'imputation. Production totale estimée : $600 \times 97{,}22 = 58\,333$ quintaux.
(**b**) Les neuf exploitations renseignées couvrent 61 hectares : rendement moyen $\widehat{R} = 875/61 = 14{,}34$ quintaux par hectare. Valeurs imputées : $5\,\widehat{R} = 71{,}72$; $8\,\widehat{R} = 114{,}75$; $12\,\widehat{R} = 172{,}13$. Somme du fichier complété : $875 + 358{,}61 = 1\,233{,}61$, moyenne 102,80, d'où $600 \times 102{,}80 = 61\,680$ quintaux.
(**c**) Les 12 exploitations tirées couvrent 86 hectares, et le rendement du fichier complété reste $1\,233{,}61/86 = 14{,}34$: $X\,\widehat{R} = 4\,800 \times 1\,233{,}61/86 = 68\,852$ quintaux.

- De (a) à (b), +3 347 quintaux : les trois exploitations muettes sont plus grandes (8,33 hectares en moyenne contre 6,78) ; leur imputer la production moyenne les sous-estime.
- De (b) à (c), +7 172 quintaux : l'échantillon est tombé sur des exploitations plus petites que la moyenne ($86/12 = 7{,}17$ hectares contre $4\,800/600 = 8$). L'estimateur par le ratio corrige cette erreur de tirage grâce au total connu des surfaces (chapitre 12).

(**d**) Des rendements aussi réguliers montrent que la production est presque proportionnelle à la surface : imputer $\widehat{R}$ fois la surface est crédible. L'imputation suppose que, à surface égale, les exploitants qui ne déclarent pas ont le même rendement que les autres (non-réponse au hasard conditionnellement

à la surface). Si l'on se tait plus volontiers après une mauvaise récolte, la production reste surestimée.