Échantillonnage

Chapitre 15 : Mise en œuvre sur R et Python

Exercices corrigés

Jaouad Madkour
Faculté d'Économie et de Gestion de Tanger
Échantillonnage

Table des matières

Partie 1 — Énoncés 1

Partie 2 — Corrections 5

1 Exercices du chapitre

Dans toute la feuille, `d` est le fichier `menages.csv` de la population du cours ($N = 2\,000$ ménages, 50 districts), chargé sous R avec les paquets `sampling` et `survey`, ou sous Python avec `numpy` et `pandas`. Les sorties reproduites sont celles des programmes exécutés.

Exercice 1 — Déclarer quatre plans

On a tiré sous R quatre échantillons de la population du cours :

- `a` : un sondage aléatoire simple de 300 ménages ;
- `b` : 60 ménages ruraux et 140 urbains, tirés au hasard dans chaque milieu (562 ruraux et 1 438 urbains dans la population) ;
- `cc` : 8 districts entiers tirés au hasard parmi les 50, soit 279 ménages ;
- `dd` : 20 districts tirés au hasard, puis 10 ménages tirés au hasard dans chacun ; la colonne `Mi` donne la taille M_i du district.

(a) Pour chaque échantillon, écrire l'appel à `svydesign`, en créant les colonnes nécessaires à l'argument `fpc`.

(b) Calculer les poids des ménages tirés dans chaque plan. Pour `dd`, donner la probabilité d'inclusion d'un ménage d'un district de M_i ménages, puis les poids extrêmes, sachant que les districts tirés comptent de 18 à 68 ménages.

(c) Combien de degrés de liberté `degf` donnera-t-il pour chaque plan ?

(d) `svymean(~depense, ...)` donne les estimations et erreurs-types suivantes :

```
         depense
sas     8035.77 212.78
strates 8054.07 269.50
grappes 8220.04 834.31
deux    7644.95 366.10
```

Pourquoi le plan `cc`, avec 279 ménages, est-il le moins précis ? Calculer le rapport de sa variance à celle du plan `a`. Pourquoi le plan à deux degrés fait-il bien mieux avec 200 ménages seulement ?

Exercice 2 — Lire la sortie d'un sondage par grappes

On tire cinq districts avec `sampling` ; le générateur de R ne donne pas les districts du chapitre 9 :

```
set.seed(2026)
cl <- sampling::cluster(d, "district", size = 5,
                        method = "srswor")
g <- getdata(d, cl); g$NI <- 50
aggregate(depense ~ district, g,
          function(v) c(M = length(v), Y = sum(v)))
pg2 <- svydesign(ids = ~district, fpc = ~NI, data = g)
svytotal(~depense, pg2)
svymean(~depense, pg2, deff = TRUE)
confint(svymean(~depense, pg2), df = degf(pg2))
```

```
  district depense.M depense.Y
1       27        44    425150
2       29        27    183900
3       33        20    146480
4       38        18    169930
5       45        47    320520
           total      SE
depense 12459800 2535492
          mean     SE   DEff
depense 7987.1  690.5 5.8942
            2.5 %   97.5 %
depense 6069.917 9904.186
```

(a) Combien de ménages a-t-on enquêtés ? Quels sont leur probabilité d'inclusion et leur poids ?
(b) Retrouver, à partir du tableau des districts, le total estimé et la moyenne estimée. Pourquoi la moyenne n'est-elle pas le total divisé par $N = 2\,000$?
(c) Retrouver l'erreur-type du total, $\sqrt{N_I^2 (1 - f_I)\, s_T^2 / n_I}$, sachant que l'écart-type des cinq totaux vaut $s_T = 119\,524{,}2$ dh.
(d) Retrouver l'intervalle de confiance. Quel intervalle aurait donné `confint` sans l'argument `df` ? Lequel retenir ?
(e) Interpréter `DEff`. Comparer aux cinq districts du chapitre 9 (moyenne 8 320,93 dh, erreur-type 529,49).

Exercice 3 — Post-stratifier par province, en Python

Sur l'échantillon fil rouge `e` (200 ménages, sondage aléatoire simple), Python donne l'effectif et la dépense moyenne de chaque province :

```
                size     mean
province
Al Hoceïma        28  5115.00
Larache           29  7524.48
Tanger-Assilah    83  8723.37
Tétouan           60  8052.33
```

Les provinces comptent 820, 560, 380 et 240 ménages dans la population.
(a) Écrire en Python, avec `pandas`, le calcul de l'estimation post-stratifiée $\sum_h W_h \bar{y}_h$ et des poids N_h / n_h des ménages.
(b) Faire le calcul à la main. Comparer à la moyenne simple $\bar{y} = 7\,843{,}05$ dh et à $\overline{Y} = 7\,911{,}44$ dh.
(c) Quelle province a le plus petit poids ? Pourquoi ?
(d) Écrire l'appel R correspondant (`postStratify`). Il donne une erreur-type de 246,56 dh ; la formule du sondage stratifié appliquée avec les n_h observés en donne 248,64. Commenter.

2 Exercice cumulatif (chapitres 15, 14, 13, 12 et 11)

Exercice 4 — Redresser les répondants du fil rouge sous R

Sur les 200 ménages du fil rouge, 161 ont répondu (colonne `repond`). On redresse leurs réponses de plusieurs façons (`pm` contient les effectifs connus des milieux, 562 ruraux et 1 438 urbains) :

```
er <- e[e$repond == 1, ]                # les 161 répondants
pr0 <- svydesign(ids = ~1, fpc = ~N, data = er)
svymean(~depense, pr0)
table(er$milieu)
cn <- data.frame(milieu = c("rural", "urbain"),
                 Freq = c(480, 1520))  # 10 x les tirés
svymean(~depense, postStratify(pr0, ~milieu, cn))
svymean(~depense, postStratify(pr0, ~milieu, pm))
pcal <- calibrate(pr0, ~milieu + taille,
                  c(`(Intercept)` = 2000, milieuurbain = 1438,
                    taille = 7859))
svymean(~depense, pcal)
rr <- predict(svyratio(~depense, ~taille, pr0), total = 7859)
c(rr$total, rr$se) / 2000

          mean     SE
depense 7488.9 266.97
```

```
 rural urbain
    42    119
          mean     SE
depense 7497.3 267.26
          mean     SE
depense 7480.9 266.58
          mean    SE
depense 7684.9 194.3
[1] 7805.3983  223.7413
```

(a) L'échantillon tiré comptait 48 ruraux et 152 urbains. Calculer les taux de réponse par milieu. Quel est l'écart de la moyenne des répondants à $\overline{Y}$? Le comparer au biais attendu de la moyenne des répondants, −219,4 dh (chapitre 14).

(b) Montrer que la deuxième estimation est la repondération par classes de réponse du chapitre 14 : quels poids reçoivent les répondants ?

(c) La troisième cale les répondants sur les effectifs de la population. En quoi diffère-t-elle de la deuxième ? Pourquoi les deux corrections sont-elles si faibles ?

(d) La quatrième est un calage linéaire (chapitre 13), la cinquième l'estimateur par le ratio (chapitre 12). Calculer leur écart à $\overline{Y}$. Les ménages tirés qui n'ont pas répondu comptent en moyenne 3,95 personnes, les répondants 3,77 : expliquer pourquoi la taille corrige davantage que le milieu. Ces corrections suppriment-elles le biais ?

(e) Le tirage systématique proportionnel à la taille du chapitre 11 (pas 400, départ 341) se déclare ainsi :

```
M <- as.numeric(table(d$district))     # tailles des districts
tires <- findInterval(341 + 400 * (0:4) - 1, cumsum(M)) + 1
h <- d[d$district %in% tires, ]
h$pi <- 5 * M[h$district] / 2000       # pi_i = 5 M_i / N
ph <- svydesign(ids = ~district, probs = ~pi, data = h)
tt <- svytotal(~depense, ph)
c(coef(tt), SE(tt)) / 2000
```

```
  depense
7475.0028  925.6163
```

Quelle variance `survey` calcule-t-il sans argument `fpc` ? Retrouver l'estimation à partir des moyennes des cinq districts tirés (10 953,29 ; 5 828,14 ; 6 146,35 ; 7 688,91 ; 6 758,33 dh).

3 Applications en sciences économiques

Exercice 5 — Une enquête auprès des très petites entreprises

Une association de commerçants de Tanger veut estimer le chiffre d'affaires mensuel moyen de ses 1 000 très petites entreprises : 600 commerces, 300 ateliers d'artisanat et 100 entreprises de services (chiffres inventés). Elle en tire 6, 4 et 4 au hasard dans chaque secteur, et obtient (en milliers de dh) :

Commerce	42 ; 55 ; 38 ; 61 ; 47 ; 53
Artisanat	28 ; 35 ; 31 ; 22
Services	85 ; 120 ; 96 ; 104

Sous R :

```
pt <- svydesign(ids = ~1, strata = ~secteur, fpc = ~Nh, data = tpe)
svymean(~ca, pt)
svytotal(~ca, pt)
degf(pt)
confint(svymean(~ca, pt), df = degf(pt))
mean(tpe$ca)
```

```
     mean    SE
ca 48.425 2.361
   total   SE
ca 48425 2361
[1] 11
      2.5 %   97.5 %
ca 43.22849 53.62151
[1] 58.35714
```

(a) Écrire le `data.frame` `tpe` qui a servi à ces calculs.

(b) Calculer les poids. Pourquoi la moyenne simple, 58,36, est-elle fausse ?

(c) Retrouver l'estimation et son erreur-type à la main (moyennes et variances par secteur à calculer).

(d) D'où viennent les 11 degrés de liberté ? Retrouver l'intervalle, avec $t_{0,975;11} = 2{,}201$.

(e) Lire le chiffre d'affaires mensuel total estimé des 1 000 entreprises, et son erreur-type. Comment se déduisent-ils de la moyenne ?

(f) Écrire le même calcul en Python, avec `pandas`.

Exercice 6 — Caler une petite enquête sur l'emploi

Dans un quartier de 600 adultes (290 hommes et 310 femmes ; 260 jeunes et 340 moins jeunes, chiffres inventés), on interroge 12 adultes tirés au hasard ; l'indicatrice vaut 1 s'ils ont un emploi :

Groupe	Hommes jeunes	Hommes âgés	Femmes jeunes	Femmes âgées
Personnes tirées	4	4	3	1
Dont en emploi	3	3	1	0

(a) Calculer la proportion de personnes en emploi non pondérée, et les poids de départ d_k.

(b) Faire à la main un tour d'ajustement proportionnel itératif : sexe, puis âge. Donner les quatre poids obtenus et la proportion pondérée.

(c) Écrire en Python la boucle de l'ajustement itératif (`pandas`), qui répète trente tours.

(d) Sous R, `rake` donne, après convergence :

```
           mean     SE
emploi 0.45895 0.1242
  F âgé F jeune   H âgé H jeune
136.390  57.870  50.902  21.598
```

Vérifier que ces poids reproduisent les deux marges. Comparer à la post-stratification sur le seul sexe, qui donne 0,49167.

(e) Une seule femme âgée a été interrogée. Que pèse-t-elle ? Quel risque cela fait-il courir à l'estimation ?

1 Exercices du chapitre

Correction — Exercice 1

(**a**) Les quatre déclarations, avec les tirages qui les ont produites (`sampling::` évite les fonctions homonymes de `survival`) :

```
set.seed(2026)
a <- d[sample(2000, 300), ]; a$N <- 2000
pa <- svydesign(ids = ~1, fpc = ~N, data = a)
dmil <- d[order(d$milieu), ]           # rural, puis urbain
set.seed(2026)
b <- getdata(dmil, sampling::strata(dmil, "milieu",
             size = c(60, 140), method = "srswor"))
b$Nh <- ifelse(b$milieu == "rural", 562, 1438)
pb <- svydesign(ids = ~1, strata = ~milieu, fpc = ~Nh, data = b)
set.seed(2026)
cc <- getdata(d, sampling::cluster(d, "district", size = 8,
              method = "srswor"))
cc$NI <- 50
pcc <- svydesign(ids = ~district, fpc = ~NI, data = cc)
set.seed(2026)
ms <- sampling::mstage(d, stage = list("cluster", ""),
        varnames = list("district", "id"),
        size = list(20, rep(10, 20)),
        method = list("srswor", "srswor"))
dd <- getdata(d, ms)[[2]]
dd$NI <- 50
dd$Mi <- as.numeric(table(d$district)[as.character(dd$district)])
pdd <- svydesign(ids = ~district + id, fpc = ~NI + Mi, data = dd)
```

Les tailles du tirage stratifié suivent l'ordre d'apparition des strates : on trie d'abord le fichier, d'où 60 ruraux puis 140 urbains. À deux degrés, `ids` nomme les unités des deux degrés et `fpc` les deux tailles de population : $N_I = 50$ districts, puis M_i ménages dans le district i.
(**b**) **et** (**c**) R le confirme :

```
plans <- list(sas = pa, strates = pb, grappes = pcc, deux = pdd)
sapply(plans, degf)
sapply(plans, function(x) round(range(weights(x)), 4))
```

```
    sas strates grappes    deux
    299     198       7      19
        sas strates grappes deux
[1,] 6.6667  9.3667    6.25  4.5
[2,] 6.6667 10.2714    6.25 17.0
```

Sondage aléatoire simple : poids $2\,000/300 = 6{,}67$, $n-1 = 299$ degrés de liberté. Strates : $562/60 = 9{,}37$ pour les ruraux, $1\,438/140 = 10{,}27$ pour les urbains ; $n - H = 198$. Grappes : $50/8 = 6{,}25$ pour tous ; $n_I - 1 = 7$. Deux degrés : $\pi_k = \frac{20}{50} \times \frac{10}{M_i} = \frac{4}{M_i}$, poids $M_i/4$, de $18/4 = 4{,}5$ à $68/4 = 17$; $n_I - 1 = 19$: seuls les districts comptent.
(**d**) Les 279 ménages de `cc` viennent de huit districts : la variance repose sur huit totaux, et les voisins se ressemblent ($\rho = 0{,}162$, chapitre 9). Le rapport des variances vaut $834{,}31^2/212{,}78^2 = 15{,}4$, pour un nombre de ménages presque égal. Le plan à deux degrés visite 20 districts et n'y interroge que 10 ménages : l'effet de grappe tombe vers $1 + 9\rho \approx 2{,}5$. Son erreur-type, 366,10 dh, reste supérieure à celle d'un aléatoire simple de 200 ménages (267,3 dh), mais deux fois et demie inférieure à celle des huit districts entiers.

Correction — Exercice 2

(a) $44+27+20+18+47=156$ ménages, au lieu des 200 attendus en moyenne. Chaque ménage d'un district tiré a la probabilité $\pi_k = 5/50 = 0{,}1$ et le poids 10.
(b) $\sum Y_i = 425\,150+183\,900+146\,480+169\,930+320\,520 = 1\,245\,980$ dh. Total de Horvitz-Thompson : $10\times 1\,245\,980 = 12\,459\,800$ dh. La moyenne est le ratio $1\,245\,980/156 = 7\,987{,}05$ dh : `svymean` divise par le nombre estimé de ménages, $\widehat{N} = 10\times 156 = 1\,560$, et non par N. Le total divisé par 2 000 donnerait 6 229,9 dh : il pâtit des districts tirés, petits (chapitre 9).
(c) $\sqrt{50^2\times 0{,}9\times 119\,524{,}2^2/5} = 2\,535\,492$ dh : c'est la formule du chapitre 9, appliquée aux cinq totaux.
(d) `degf` vaut $n_I - 1 = 4$ et $t_{0{,}975;4} = 2{,}776$: $7\,987{,}05 \pm 2{,}776\times 690{,}5 = 7\,987{,}05 \pm 1\,917{,}1$, soit $[6\,069{,}9\,;9\,904{,}2]$. Sans `df`, `confint` prendrait 1,96 : $\pm 1\,353{,}4$, soit $[6\,633{,}7\,;9\,340{,}4]$, trop étroit pour une variance estimée sur cinq grappes. On retient l'intervalle de Student, qui contient $\overline{Y} = 7\,911{,}44$.
(e) La variance estimée est 5,9 fois celle d'un sondage aléatoire simple de 156 ménages : ils en valent $156/5{,}89 = 26{,}5$. Les cinq districts du chapitre 9 donnaient 8 320,93 dh (529,49) : 334 dh d'écart entre deux tirages, de l'ordre de l'erreur-type. Avec cinq grappes, le résultat dépend beaucoup des districts tirés.

Correction — Exercice 3

(a) Un calcul possible (`Np` contient les effectifs des provinces) :

```
Np = pd.Series(Np)                          # 820, 560, 380, 240
gp = e.groupby("province").depense
print(gp.agg(["size", "mean"]).round(2))
post_p = (Np / 2000 * gp.mean()).sum()
w_p = e.province.map(Np / gp.size())        # poids N_h / n_h
print(round(post_p, 2), round(w_p.sum(), 6))

7874.69 2000.0
```

`pandas` aligne `Np` et les moyennes sur le nom des provinces : l'ordre des lignes n'importe pas.
(b) $0{,}12\times 5\,115{,}00+0{,}19\times 7\,524{,}48+0{,}41\times 8\,723{,}37+0{,}28\times 8\,052{,}33 \approx 613{,}80+1\,429{,}65+3\,576{,}58+2\,254{,}65$, soit 7 874,69 dh. C'est 31,6 dh de plus que $\overline{y}$, et 36,8 dh de moins que $\overline{Y}$: l'erreur est presque divisée par deux ($-68{,}4$ dh pour $\overline{y}$).
(c) Poids : $240/28 = 8{,}57$ (Al Hoceïma), $560/60 = 9{,}33$ (Tétouan), $820/83 = 9{,}88$ (Tanger-Assilah), $380/29 = 13{,}10$ (Larache). Al Hoceïma a le plus petit poids : elle est sur-représentée (28 ménages tirés pour 24 attendus), Larache sous-représentée (29 pour 38). La post-stratification rend à chaque province sa part ; c'est surtout le moindre poids d'Al Hoceïma, la province qui dépense le moins, qui relève l'estimation.
(d) En R :

```
pp <- data.frame(province = c("Tanger-Assilah", "Tétouan",
                              "Larache", "Al Hoceïma"),
                 Freq = c(820, 560, 380, 240))
pps2 <- postStratify(p, ~province, pp)
svymean(~depense, pps2)

          mean     SE
depense 7874.7 246.56
```

Même estimation. Les deux erreurs-types diffèrent de 0,8 % : `survey` linéarise l'estimateur, la formule du stratifié traite les n_h observés comme fixés. Ce sont deux approximations de la même variance, comme au chapitre 13 pour le milieu (252,43 et 253,1 dh).

2 Exercice cumulatif (chapitres 15, 14, 13, 12 et 11)

Correction — Exercice 4

(**a**) Taux de réponse : $42/48 = 0{,}875$ chez les ruraux, $119/152 = 0{,}783$ chez les urbains. La moyenne des répondants, 7 488,9 dh, s'écarte de $\overline{Y}$ de $-422{,}5$ dh : près du double du biais attendu, $-219{,}4$ dh. L'échantillon cumule une erreur de tirage ($\overline{y} = 7\,843{,}05$, soit $-68{,}4$ dh) et des non-répondants qui dépensent beaucoup (9 304,87 dh en moyenne, chapitre 14).
(**b**) Les effectifs 480 et 1 520 sont ceux des ménages **tirés** multipliés par 10. La post-stratification donne aux répondants les poids $480/42 = 11{,}43$ (ruraux) et $1\,520/119 = 12{,}77$ (urbains), soit $d_k\, n_c/r_c = 10 \times 48/42$ et $10 \times 152/119$: la repondération par classes du chapitre 14, qui donnait 7 497,3 dh.
(**c**) Elle cale sur les effectifs de la **population**, 562 et 1 438 : poids 13,38 et 12,08. Elle corrige à la fois la non-réponse et la sous-représentation des ruraux dans l'échantillon tiré (24 % au lieu de 28,1 %) : c'est le calage du chapitre 13 appliqué aux répondants. Les deux corrections sont faibles ($+8$ et -8 dh) parce que le milieu explique peu la dépense, et que, dans chaque milieu, les non-répondants dépensent plus que les répondants (non-réponse non ignorable).
(**d**) Calage sur le milieu et le nombre de personnes : $7\,684{,}9 - 7\,911{,}4 = -226{,}5$ dh ; ratio sur la taille : $7\,805{,}4 - 7\,911{,}4 = -106{,}0$ dh. Les non-répondants sont de plus grands ménages : les répondants, trop petits en moyenne, dépensent trop peu, et caler sur le nombre de personnes $X = 7\,859$ relève l'estimation. Mais une part du biais demeure : la réponse dépend de la dépense elle-même (chapitre 14). La taille réduit aussi l'erreur-type (194,3 et 223,7 dh contre 266,97).
(**e**) Sans `fpc`, `survey` traite les districts comme tirés **avec remise** : c'est la variance prudente du chapitre 11. Avec $\pi_i = 5M_i/2\,000$, chaque district contribue $Y_i/\pi_i = 400\,\overline{Y}_i$ au total, d'où une moyenne égale à celle des cinq moyennes de districts :

$$\frac{10\,953{,}29 + 5\,828{,}14 + 6\,146{,}35 + 7\,688{,}91 + 6\,758{,}33}{5} = \frac{37\,375{,}02}{5} = 7\,475{,}00 \text{ dh},$$

et une erreur-type $s/\sqrt{5} = 925{,}6$ dh, s étant l'écart-type de ces cinq moyennes : les valeurs du chapitre 11. `degf(ph)` vaut 4.

3 Applications en sciences économiques

Correction — Exercice 5

(**a**) Une colonne pour la strate, une pour sa taille N_h, une pour la variable :

```
tpe <- data.frame(
  secteur = rep(c("commerce", "artisanat", "services"), c(6, 4, 4)),
  Nh = rep(c(600, 300, 100), c(6, 4, 4)),
  ca = c(42, 55, 38, 61, 47, 53, 28, 35, 31, 22, 85, 120, 96, 104))
```

(**b**) Poids N_h/n_h : $600/6 = 100$ (commerce), $300/4 = 75$ (artisanat), $100/4 = 25$ (services). Les services, 10 % des entreprises, font $4/14 = 28{,}6$ % de l'échantillon, et ont le plus gros chiffre d'affaires : la moyenne simple, 58,36, les surpondère.
(**c**) Moyennes : $296/6 = 49{,}333$; $116/4 = 29$; $405/4 = 101{,}25$. Variances (en $n_h - 1$) : 73,867 ; 30 ; 216,917.

$$\overline{y}_{st} = 0{,}6 \times 49{,}333 + 0{,}3 \times 29 + 0{,}1 \times 101{,}25 = 29{,}6 + 8{,}7 + 10{,}125 = 48{,}425$$

$$\widehat{V} = 0{,}36 \times 0{,}99 \times \frac{73{,}867}{6} + 0{,}09 \times \frac{296}{300} \times \frac{30}{4} + 0{,}01 \times 0{,}96 \times \frac{216{,}917}{4} = 4{,}3877 + 0{,}6660 + 0{,}5206 = 5{,}5743$$

d'où l'erreur-type 2,361 millier de dh : la sortie de R.
(**d**) Une moyenne est estimée par strate : $14 - 3 = 11$ degrés de liberté. $48{,}425 \pm 2{,}201 \times 2{,}361 = 48{,}425 \pm 5{,}197$, soit [43,23 ; 53,62] milliers de dh.

(**e**) 48 425 milliers de dh, soit 48,4 millions de dh par mois, d'erreur-type 2 361 milliers de dh : la moyenne et son erreur-type multipliées par $N = 1\,000$, car la somme des poids vaut exactement N.
(**f**) En Python :

```
tpe = pd.DataFrame({
    "secteur": ["commerce"] * 6 + ["artisanat"] * 4 + ["services"] * 4,
    "ca": [42, 55, 38, 61, 47, 53, 28, 35, 31, 22, 85, 120, 96, 104]})
Nh5 = pd.Series({"commerce": 600, "artisanat": 300, "services": 100})
g5s = tpe.groupby("secteur").ca
nh5, W5 = g5s.size(), Nh5 / Nh5.sum()
moy5 = (W5 * g5s.mean()).sum()
V5 = (W5**2 * (1 - nh5 / Nh5) * g5s.var() / nh5).sum()
print(round(moy5, 3), round(np.sqrt(V5), 3), t.ppf(0.975, 11).round(3))

48.425 2.361 2.201
```

Correction — Exercice 6

(**a**) $7/12 = 0{,}583$. Poids de départ : $d_k = 600/12 = 50$.
(**b**) Sexe : les hommes représentent $8 \times 50 = 400$ adultes pour 290, les femmes 200 pour 310 ; facteurs 0,725 et 1,55, poids 36,25 et 77,5. Âge : les jeunes représentent alors $4 \times 36{,}25 + 3 \times 77{,}5 = 377{,}5$ adultes pour 260, les autres 222,5 pour 340 ; facteurs 0,6887 et 1,5281. Poids après un tour : hommes jeunes 24,97, hommes âgés 55,39, femmes jeunes 53,38, femme âgée 118,43 ; somme 600. Proportion : $(3 \times 24{,}97 + 3 \times 55{,}39 + 53{,}38)/600 = 0{,}4908$. Mais les hommes représentent maintenant 321,4 adultes : il faut recommencer.
(**c**) La boucle, et les poids distincts aux tours 1, 2, 3 et 30, rangés par ordre croissant :

```
emp = pd.DataFrame({"sexe": ["H"] * 8 + ["F"] * 4,
    "age": ["jeune"] * 4 + ["âgé"] * 4 + ["jeune"] * 3 + ["âgé"],
    "emploi": [1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0]})
cibles = {"sexe": pd.Series({"H": 290, "F": 310}),
          "age": pd.Series({"jeune": 260, "âgé": 340})}
w6 = np.full(12, 50.0)                    # 600 / 12
for it in range(1, 31):
    for v, c in cibles.items():
        tot = pd.Series(w6).groupby(emp[v].values).sum()
        w6 = w6 * emp[v].map(c / tot).values
    if it <= 3 or it == 30:
        print(it, np.unique(w6.round(2)),
              round(np.average(emp.emploi, weights=w6), 4))

1 [ 24.97  53.38  55.39 118.43] 0.4908
2 [ 21.83  51.23  57.56 135.09] 0.4612
3 [ 21.61  50.93  57.85 136.3 ] 0.4591
30 [ 21.6   50.9   57.87 136.39] 0.4589
```

La convergence est rapide : après trois tours, les poids sont presque fixés.
(**d**) Hommes : $4 \times 21{,}598 + 4 \times 50{,}902 = 290$; femmes : $3 \times 57{,}870 + 136{,}390 = 310$; jeunes : $4 \times 21{,}598 + 3 \times 57{,}870 = 260$; âgés : $4 \times 50{,}902 + 136{,}390 = 340$. Les deux marges sont reproduites, et la proportion vaut 0,459, comme en Python. La post-stratification sur le sexe seul corrige la sous-représentation des femmes (4 sur 12, au lieu de 52 %), moins souvent en emploi dans l'échantillon : 0,492. Caler aussi sur l'âge relève le poids des personnes âgées (5 sur 12 au lieu de 57 %), dont la femme âgée sans emploi : l'estimation baisse encore.
(**e**) Elle représente 136,4 adultes, 23 % du quartier. Si elle avait eu un emploi, l'estimation serait passée à $0{,}459 + 136{,}39/600 = 0{,}686$: une seule réponse déplace le résultat de 23 points. Des poids aussi inégaux gonflent la variance (chapitre 7) ; on regroupe alors des cases ou l'on borne les poids (chapitre 13).