TD 1 : Se familiariser avec les bases/notations

Exercice 1. Soient une population $\mathcal{U} = \{1, 2, 3\}$ et le plan $p(\cdot)$ suivant

$$p(\{1,2\}) = \frac{1}{2}, \qquad p(\{1,3\}) = \frac{1}{4}, \qquad p(\{2,3\}) = \frac{1}{4}.$$

1. Donnez les probabilités d'inclusion d'ordre un.
2. Pourquoi la somme de ces probabilités d'inclusion vaut nécessairement 2 ?
3. Donnez la matrice de variance–covariance des variables indicatrices $1_{\{k \in S\}}$, $k \in \mathcal{U}$.

Solution 1. 1. On a

$$\pi_1 = \sum_{s \in \mathscr{S} :\, 1 \in s} p(s) = p(\{1,2\}) + p(\{1,3\}) = \frac{3}{4},$$
$$\pi_2 = \sum_{s \in \mathscr{S} :\, 2 \in s} p(s) = p(\{1,2\}) + p(\{2,3\}) = \frac{3}{4},$$
$$\pi_3 = \sum_{s \in \mathscr{S} :\, 3 \in s} p(s) = p(\{1,3\}) + p(\{2,3\}) = \frac{1}{2}.$$

2. Car c'est un plan de taille fixe et la taille est 2 justement.
3. Il s'agit donc de calculer (avec les notations du cours)

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell, & k \neq \ell, \\ \pi_k(1 - \pi_k), & k = \ell. \end{cases}$$

On a

$$\Delta_{11} = \frac{3}{4}\left(1 - \frac{3}{4}\right) = \frac{3}{16}, \quad \Delta_{12} = \frac{1}{2} - \frac{3}{4} \times \frac{3}{4} = -\frac{1}{16}, \quad \Delta_{13} = \frac{1}{4} - \frac{3}{4} \times \frac{1}{2} = -\frac{1}{8},$$
$$\Delta_{22} = \frac{3}{4}\left(1 - \frac{3}{4}\right) = \frac{3}{16}, \quad \Delta_{23} = \frac{1}{4} - \frac{3}{4} \times \frac{1}{2} = -\frac{1}{8}, \quad \Delta_{33} = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4},$$

et donc la matrice de variance–covariance est

$$\Delta = \frac{1}{16}\begin{bmatrix} 3 & -1 & -2 \\ -1 & 3 & -2 \\ -2 & -2 & 4 \end{bmatrix}$$

Exercice 2. Soient une population $\mathcal{U} = \{1, 2, 3\}$ et le plan $p(\cdot)$ suivant

$$p(\{1,2\}) = \frac{1}{2}, \qquad p(\{1,3\}) = \frac{1}{4}, \qquad p(\{2,3\}) = \frac{1}{4}.$$

1. Donnez la distribution de probabilité du π–estimateur de la moyenne μ_y pour un caractère d'intérêt y.
2. En déduire le biais de cet estimateur.

Solution 2. 1. Puisqu'à l'exercice 1 nous avons calculé les probabilités d'inclusions d'ordre 1, nous avons

$$\hat{\mu}_{y,\pi} = \frac{1}{3}\begin{cases} \frac{y_1}{3/4} + \frac{y_2}{3/4}, & s = \{1,2\} \\ \frac{y_1}{3/4} + \frac{y_3}{1/2}, & s = \{1,3\} \\ \frac{y_2}{3/4} + \frac{y_3}{1/2}, & s = \{2,3\} \end{cases} = \begin{cases} \frac{4(y_1+y_2)}{9}, & s = \{1,2\} \\ \frac{4y_1+6y_3}{9}, & s = \{1,3\} \\ \frac{4y_2+6y_3}{9}, & s = \{2,3\} \end{cases}$$

2. Ainsi

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{y,\pi}) &= \frac{1}{2} \times \frac{4(y_1+y_2)}{9} + \frac{1}{4} \times \frac{4y_1+6y_3}{9} + \frac{1}{4} \times \frac{4y_2+6y_3}{9} \\ &= \frac{2y_1+2y_2+y_1+1.5y_3+y_2+1.5y_3}{9} \\ &= \frac{3y_1+3y_2+3y_3}{9} \\ &= \frac{y_1+y_2+y_3}{3} = \mu_y. \end{aligned}$$

Ce qui est logique puisque nous avons vu que le π–estimateur était un estimateur sans biais !

Exercice 3. Soit la matrice de variance–covariance $\Delta = (\Delta_{k\ell})_{k,\ell}$ des indicatrices $1_{\{k\in S\}}$ pour un plan $p(\cdot)$ donné. On sait que

$$\Delta = \frac{6}{25}\begin{bmatrix} 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix}.$$

1. Le plan $p(\cdot)$ est-il de taille fixe ?
2. Satisfait il aux conditions de Sen–Yates–Grundy ?
3. Sachant que $\pi_1 = \pi_2 = \pi_3 > \pi_4 = \pi_5$, calculer les probabilités d'inclusions d'ordre 1.
4. Donnez la matrice des probabilités d'inclusions d'ordre deux.
5. Donnez les probabilités associées à tous les échantillons possibles.

Solution 3. 1. Ce plan n'est pas de taille fixe puisque $\sum_{k\in\mathcal{U}} \Delta_{k\ell} \neq 0$.

2. Ces conditions ne sont pas vérifiées puisque (par exemple) $\Delta_{12} = 1 > 0$.
3. D'après la matrice Δ, on a pour tout $k \in \mathcal{U}$

$$\pi_k(1-\pi_k) = \frac{6}{25} \Longleftrightarrow \pi_k \in \left\{\frac{2}{5}, \frac{3}{5}\right\}.$$

D'après les indications supplémentaires, on a donc

$$\pi_1 = \pi_2 = \pi_3 = \frac{2}{5}, \qquad \pi_4 = \pi_5 = \frac{3}{5}.$$

4. Puisque $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$, $k \neq \ell$ et que nous connaissons déjà les π_k, nous pouvons en déduire que

$$\pi_{12} = \Delta_{12} + \pi_1\pi_2 = \frac{6}{25} + \frac{4}{25} = \frac{2}{5}$$
$$\pi_{14} = \Delta_{14} + \pi_1\pi_4 = -\frac{6}{25} + \frac{6}{25} = 0$$
$$\pi_{45} = \Delta_{45} + \pi_4\pi_5 = \frac{6}{25} + \frac{9}{25} = \frac{3}{5}$$

Avec des calculs similaires, la matrice des probabilités d'inclusion d'ordre deux est

$$\frac{1}{5}\begin{bmatrix} — & 2 & 2 & 0 & 0 \\ 2 & — & 2 & 0 & 0 \\ 2 & 2 & — & 0 & 0 \\ 0 & 0 & 0 & — & 3 \\ 0 & 0 & 0 & 3 & — \end{bmatrix}$$

5. Du fait de la présence de nombreux zéro au sein de la matrice des probabilités d'inclusion d'ordre 2, on sait que les échantillons possibles sont (éventuellement)

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1,2\}, \{1,3\}, \{2,3\}, \{4,5\}, \{1,2,3\}.$$

Du coup on a déjà

$$\Pr(S = \{4,5\}) = \pi_{45} = \frac{3}{5}.$$

D'un côté puisque $\pi_1 = \pi_{12}$, on a

$$\underbrace{\Pr(S = \{1\}) = \Pr(S = \{1,3\})}_{\text{échantillons contenant 1 mais pas 2}} = 0, \qquad \underbrace{\Pr(S = \{1,2\}) + \Pr(S = \{1,2,3\})}_{\text{échantillons contenant 1 et 2}} = \frac{2}{5}.$$

De l'autre puisque $\pi_1 = \pi_{13}$, on a également

$$\Pr(S = \{1\}) = \Pr(S = \{1,2\}) = 0, \qquad \Pr(S = \{1,3\}) + \Pr(S = \{1,2,3\}) = \frac{2}{5}.$$

On en déduit donc que $\Pr(S = \{1,2,3\}) = 2/5$. Le même raisonnement conduit à $\Pr(S = \{2\}) = \Pr(S = \{3\}) = \Pr(S = \{2,3\}) = 0$.
Enfin on a

$$\Pr(S = \{4\}) = \pi_4 - \pi_{45} = \frac{3}{5} - \frac{3}{5} = 0$$
$$\Pr(S = \{5\}) = \pi_5 - \pi_{45} = \frac{3}{5} - \frac{3}{5} = 0.$$

Pour résumer on a donc

$$p(\{4,5\}) = \frac{3}{5}, \qquad p(\{1,2,3\}) = \frac{2}{5},$$

et le plan n'est effectivement pas à taille fixe.

Exercice 4. On considère un plan sans remise effectué sur une population de taille N. On suppose que les probabilités d'inclusions d'ordre 1 et 2 π_k et $\pi_{k\ell}$ sont strictement positives. A partir d'un échantillon aléatoire S, on s'intéresse à l'estimateur suivant

$$\hat{\theta} = \frac{1}{N^2} \sum_{k \in S} \frac{y_k}{\pi_k} + \frac{1}{N^2} \sum_{\substack{k,\ell \in S \\ k \neq \ell}} \frac{y_\ell}{\pi_{k\ell}}.$$

1. Pour quelle fonction d'intérêt cet estimateur est il sans biais ?

Solution 4. On a

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \frac{1}{N^2} \sum_{k \in \mathcal{U}} \frac{y_k}{\pi_k} \mathbb{E}(1_{\{k \in S\}}) + \frac{1}{N^2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} \frac{y_\ell}{\pi_{k\ell}} \mathbb{E}(1_{\{k,\ell \in S\}}) \\
&= \frac{1}{N^2} \sum_{k \in \mathcal{U}} y_k + \frac{1}{N^2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} y_\ell \\
&= \frac{1}{N^2} t_y + \frac{1}{N^2} (N t_y - t_y) \\
&= \frac{1}{N} t_y = \mu_y.
\end{aligned}$$

La fonction d'intérêt est la moyenne !

Exercice 5. 1. Montrez que

$$\frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu_y)^2 = \frac{1}{2N^2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} (y_k - y_\ell)^2.$$

2. Pour un plan sans remise quelconque mais dont les probabilités d'inclusion d'ordre 1 et 2 sont strictement positives, construisez un estimateur sans biais de σ_y^2.

Solution 5. 1. On a

$$\begin{aligned}
\frac{1}{2N^2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} (y_k - y_\ell)^2 &= \frac{1}{2N^2} \sum_{k,\ell \in \mathcal{U}} (y_k - y_\ell)^2 \\
&= \frac{1}{N^2} \sum_{k,\ell \in \mathcal{U}} y_k^2 - \frac{1}{N^2} \sum_{k,\ell \in \mathcal{U}} y_k y_\ell \\
&= \frac{1}{N} \sum_{k \in \mathcal{U}} y_k^2 - \left(\frac{1}{N} \sum_{k \in \mathcal{U}} y_k \right) \left(\frac{1}{N} \sum_{\ell \in \mathcal{U}} y_\ell \right) \\
&= \overline{y^2} - \overline{y}^2 \\
&= \sigma_y^2
\end{aligned}$$

2. On utilise donc le résultat précédent et on utilise le π–estimateur, i.e.,

$$\frac{1}{2N^2} \sum_{\substack{k,\ell \in S \\ k \neq \ell}} \frac{(y_k - y_\ell)^2}{\pi_{k\ell}},$$

qui est un estimateur sans biais dès lors que les $\pi_{k\ell}$ sont tous positifs.

TD 2 : Plans simples

Exercice 6. On souhaite estimer la surface moyenne cultivée dans les fermes d'un canton rural donné. Sur les $N = 2010$ fermes de ce canton, on en tire 100 par sondage aléatoire simple. On mesure y_k la surface cultivée dans la ferme k en hectares, et l'on trouve

$$\sum_{k\in S} y_k = 2907\text{ha}, \qquad \sum_{k\in S} y_k^2 = 154593\text{ha}^2.$$

1. Donnez l'estimateur sans biais classique de la moyenne

$$\mu_y = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k.$$

2. Donnez un intervalle de confiance à 95% pour μ_y.

Solution 6. 1. Dans un plan simple, l'estimateur sans biais classique est

$$\hat{\mu}_y = \frac{1}{n}\sum_{k\in S} y_k = \frac{2907}{100} = 29.07\text{ha}.$$

2. La taille de l'échantillon $n = 100$ étant suffisamment grande, on peut supposer sans trop de risque $\hat{\mu}_y$ suit approximativement une loi normale. L'intervalle de confiance en découle et s'écrit

$$\left[\hat{\mu}_y \pm 1.96\sqrt{\frac{N-n}{N}\frac{\hat{S}_y^2}{n}}\right] = \left[29.07 \pm 1.96\sqrt{\frac{2010-100}{2010}\times\frac{707.94}{100}}\right] = [23.99, 34.15],$$

où nous avons utilisé le fait que

$$\hat{S}_y^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k\in S} y_k^2 - \hat{\mu}_y^2\right) = \frac{100}{99}\left(1545.93 - 29.07^2\right) = 707.94$$

Exercice 7. On s'intéresse à la proportion d'hommes atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. On sait par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type. On se propose de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

Quelle taille d'échantillon faut–il sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance 0.95 soit inférieure à 0.02 pour les plans simples avec et sans remise ?

Solution 7. Le paramètre d'intérêt est donné par

$$p = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k,$$

où les y_k sont des indicatrices codant la présence ou non de la maladie. On estimera ce paramètre par

$$\hat{p} = \frac{1}{n}\sum_{k\in S} y_k,$$

et la variance de cet estimateur est donnée par

$$\mathrm{Var}(\hat{p}) = \begin{cases} \frac{\sigma_y^2}{n}, & \text{avec remise,} \\ \frac{N-n}{N}\frac{S_y^2}{n}, & \text{sans remise,} \end{cases}$$

mais puisque $y_k^2 = y_k$, la variance et la variance corrigée sur la population sont égales à

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k - \left(\frac{1}{N}\sum_{k\in\mathcal{U}} y_k\right)^2 = p - p^2 = p(1-p), \qquad S_y^2 = \frac{N}{N-1}p(1-p).$$

Ainsi on a donc

$$\mathrm{Var}(\hat{p}) = \begin{cases} \frac{p(1-p)}{n}, & \text{avec remise,} \\ \frac{N-n}{N-1}\frac{p(1-p)}{n}, & \text{sans remise.} \end{cases}$$

Si l'on suppose que la taille de l'échantillon est suffisamment grande pour que l'approximation selon la loi normale soit acceptable, on a donc un intervalle de confiance à 95% de la forme

$$\hat{p} \pm 1.96 \times \sqrt{\mathrm{Var}(\hat{p})}.$$

Ainsi on cherche donc la taille de l'échantillon n telle que

$$\begin{aligned} 2 \times 1.96 \times \sqrt{\mathrm{Var}(\hat{p})} \leq 0.02 &\iff \mathrm{Var}(\hat{p}) \leq 196^{-2} \\ &\iff \begin{cases} \frac{p(1-p)}{n} \leq 196^{-2}, & \text{avec remise} \\ \frac{N-n}{N-1}\frac{p(1-p)}{n} \leq 196^{-2}, & \text{avec remise} \end{cases} \\ &\iff \begin{cases} n \geq 196^2 p(1-p) & \text{avec remise} \\ n \geq 196^2 N p(1-p)/\{N-1+196^2 p(1-p)\} & \text{sans remise.} \end{cases} \end{aligned}$$

En prenant $p = 3/10$ et $N = 1500$ on trouve alors que

$$n > \begin{cases} 8067, & \text{avec remise} \\ 1264, & \text{sans remise.} \end{cases}$$

Notons qu'avec remise la taille d'échantillon requise est supérieure à la taille de la population :-(

Exercice 8. Un échantillon de 100 étudiants est constitué au moyen d'un plan aléatoire simple sans remise dans une population de 1000 étudiants. Le résultat obtenu est présenté au sein du Tableau 2.

Table 2: *Nombre de succès/échec selon le sexe pour un échantillon de 100 étudiants pris parmi 1000.*

	Hommes	Femmes	Total
Réussite	$n_{11} = 35$	$n_{12} = 25$	$n_{1\cdot} = 60$
Échec	$n_{21} = 20$	$n_{22} = 20$	$n_{2\cdot} = 40$
Total	$n_{\cdot 1} = 55$	$n_{\cdot 2} = 45$	$n = 100$

— Estimez le taux de réussite des hommes et des femmes.
— Calculez le biais (approché) des taux de réussite.

— Estimez l'erreur quadratique moyenne de ces taux de réussite.

Solution 8. 1. Bon bah là on réfléchit pas plus de 2 secondes et l'on écrit bêtement

$$\hat{R}_F = \frac{25}{45} \approx 55.6\%, \qquad \hat{R}_H = \frac{35}{55} \approx 63.6\%.$$

2. On a vu que la moyenne empirique était un estimateur sans biais—que ce soit sans remise ou avec. Alors pourquoi parlons nous ici de biais approché ? ? ? Le problème vient du fait que l'on pioche (sans remise) 100 étudiants parmi 1000. En conséquence le nombre de filles/garçons présents dans l'échantillon est aléatoire ! ! ! Cela nous introduit un biais...
Rappel de cours (je l'espère inutile) : Le biais approché d'un ratio $R = \mu_y/\mu_x$ est

$$\text{Biais}(R) \approx \frac{1}{\mu_x^2}\frac{N-n}{N}\frac{1}{n}\left(RS_x^2 - S_{xy}\right).$$

Revenons à notre exercice et commençons par les filles galanterie oblige. On a donc

$$\hat{R}_F = \frac{\hat{\mu}_y}{\hat{\mu}_x},$$

où les y sont des variables binaires valant 1 lors de la réussite d'une femme, 0 sinon, et x sont des variables binaires valant 1 s'il s'agit d'une femme, 0 sinon.
On a donc (puisque $x_k^2 = x_k$ et $x_k y_k = y_k$)

$$S_x^2 = \frac{1}{N-1}\left(\sum_{k\in\mathcal{U}} x_k^2 - N\mu_x^2\right) = \frac{1}{N-1}(N\mu_x - N\mu_x^2) = \frac{N}{N-1}\mu_x(1-\mu_x)$$

$$S_{xy} = \frac{1}{N-1}\left(\sum_{k\in\mathcal{U}} x_k y_k - N\mu_x\mu_y\right) = \frac{1}{N-1}(N\mu_y - N\mu_x\mu_y) = \frac{N}{N-1}\mu_y(1-\mu_x).$$

Ainsi le biais approché vaut donc

$$\begin{aligned}
\text{Biais}(\hat{R}_F) &\approx \frac{1}{\mu_x^2}\frac{N-n}{N-1}\frac{1}{n}\left\{R_F\mu_x(1-\mu_x) - \mu_y(1-\mu_x)\right\} \\
&= \frac{1}{\mu_x^2}\frac{N-n}{N-1}\frac{1}{n}\left\{\mu_y(1-\mu_x) - \mu_y(1-\mu_x)\right\}, \qquad \mu_y = R_F\mu_x \\
&= 0.
\end{aligned}$$

On trouvera le même résultat pour les garçons, i.e., un biais approché nul. D'ailleurs c'est nul comme question car pour ce cas particulier le biais approché est toujours égal à 0 ;-)

3. *Cours : La formule de l'erreur quadratique approchée d'un ratio $R = \mu_y/\mu_x$ est*

$$EQM(\hat{R}) \approx \frac{1}{\mu_x^2}\frac{N-n}{N}\frac{1}{n}\left(S_y^2 - 2RS_{xy} + R^2S_x^2\right).$$

Pour notre cas particulier et toujours pour les femmes on a donc

$$\begin{aligned}
\text{EQM}(\hat{R}_F) &\approx \frac{1}{\mu_x^2}\frac{N-n}{N}\frac{1}{n}\left\{\frac{N}{N-1}\mu_y(1-\mu_y) - 2R_F\frac{N}{N-1}\mu_y(1-\mu_x) + R_F^2\frac{N}{N-1}\mu_x(1-\mu_x)\right\} \\
&= \frac{1}{\mu_x^2}\frac{N-n}{N-1}\frac{1}{n}\left\{\mu_y(1-\mu_y) - 2R_F\mu_y(1-\mu_x) + R_F\mu_y(1-\mu_x)\right\} \\
&= \frac{1}{\mu_x^2}\frac{N-n}{N-1}\frac{1}{n}\mu_y\left\{1-\mu_y - R_F(1-\mu_x)\right\} \\
&= \frac{1}{\mu_x^2}\frac{N-n}{N-1}\frac{1}{n}\mu_y(1-R_F).
\end{aligned}$$

On estime cette erreur quadratique moyenne par

$$\begin{aligned}\widehat{\mathrm{EQM}}(\hat{R}_F) &\approx \frac{1}{\hat{\mu}_x^2}\frac{N-n}{N-1}\frac{1}{n}\hat{\mu}_y(1-\hat{R}_F)\\ &= \frac{1}{(45/100)^2}\frac{1000-100}{1000-1}\frac{1}{100}\frac{25}{100}\left(1-\frac{25}{45}\right)\\ &= 4.9\times 10^{-3}.\end{aligned}$$

De même pour les hommes on trouve

$$\widehat{\mathrm{EQM}}(\hat{R}_H) \approx 3.8\times 10^{-3}.$$

TD 3 : Plans à probablités inégales

Exercice 9. Une population est composée de 6 ménages de tailles respectives 2, 4, 3, 9, 1 et 2 (la taille x_k d'un ménage k est le nombre de personnes physiques qu'il comprend). On tire 3 ménages sans remise, avec une probabilité proportionnelle à leur taille.

1. Donnez les probabilités d'inclusion des 6 ménages de la base de sondage—soyez prudent...
2. Réalisez effectivement le tirage par une méthode systématique.
3. A partir de l'échantillon obtenu en 2, donnez une estimation de la taille moyenne $\bar{x}$ des ménages ; le résultat était-il prévisible ?

Solution 9. 1. Pour tout $k \in \mathcal{U}$ on a

$$\pi_k = n\frac{x_k}{\bar{x}} = \frac{3x_k}{21} = \frac{x_k}{7}.$$

On remarque que $\pi_4 = 9/7 > 1$. On corrige le problème en posant $\pi_4 = 1$ et en réajustant les autres probabilités d'inclusions, i.e., pour $k \in \mathcal{U} \setminus \{4\}$

$$\pi_k = 2\frac{x_k}{\bar{x} - x_4} = \frac{2x_k}{12} = \frac{x_k}{6}.$$

On a donc

$$\pi_1 = \frac{1}{3}, \quad \pi_2 = \frac{2}{3}, \quad \pi_3 = \frac{1}{2}, \quad \pi_4 = 1, \quad \pi_5 = \frac{1}{6}, \quad \pi_6 = \frac{1}{3}.$$

Notons que nous avons bien comme attendu $\sum_{k=1}^{6} \pi_k = 3$.

2. Puisque $\pi_4 = 1$, le ménage 4 est forcément pris ; reste donc à tirer deux ménages à l'aide d'une réalisation d'une $U(0,1)$ et d'un pas de tirage de 1.
3. L'échantillon $\mathcal{S}$ peut s'écrire $\mathcal{S} = (k_1, k_2, k_3)$ avec $k_1 = 4$. Ainsi

$$\hat{\bar{x}} = \frac{1}{N}\sum_{k\in\mathcal{S}} \frac{x_k}{\pi_k} = \frac{1}{6}\left(\frac{9}{1} + \frac{x_{k_2}}{x_{k_2}/6} + \frac{x_{k_3}}{x_{k_3}/6}\right) = 3.5.$$

Remarquons que $\hat{\bar{x}} = \bar{x}$. Ce résultat est évident puisque les x_k et π_k sont parfaitement proportionnels et donc un estimateur de variance nulle !

Exercice 10. On a répertorié dans une petite municipalité 6 entreprises dont les chiffres d'affaires (variable x_k) sont respectivement de 40, 10, 8, 1, 0.5 et 0.5 millions d'euros. Dans le but d'estimer l'emploi salarié total, sélectionnez trois entreprises au hasard et sans remise, à probabilités inégales selon le chiffre d'affaires, par la méthode du tirage systématique (en justifiant votre démarche). Pour ce faire, on utilise la réalisation suivante d'une variable aléatoire $U(0,1)$: 0.83021. Que se passe-t-il si on modifie l'ordre du fichier ?

Solution 10. Commençons par dire que le tirage à probabilités inégales semble justifié puisqu'à priori il devrait avoir une relation plus ou moins proportionnelle entre le le chiffre d'affaire et le nombre de salariés.

Commençons nos calculs pour ce tirage systématique. On a $\sum_{k\in\mathcal{U}} x_k = 60$ et puisque

$$\pi_1 = 3\frac{40}{60} = 2 > 1,$$

l'unité 1 est sélectionnée d'office et retirée de la population. De manière analogue, puisque

$$\pi_2 = 2\frac{10}{60-40} = 1,$$

l'unité 2 est également sélectionnée d'office et retirée de la population. Il reste donc à sélectionner une dernière unité parmi celles restantes. On trouve facilement

$$\pi_3 = \frac{8}{10}, \quad \pi_4 = \frac{1}{10}, \quad \pi_5 = \frac{1}{20}, \quad \pi_6 = \frac{1}{20}.$$

Notons que l'on a bien comme attendu $\sum_{k=3}^{6} \pi_k = 1$. Les probabilités d'inclusions cumulées sont

$$\pi_3 = 0.8, \quad \pi_3 + \pi_4 = 0.9, \quad \pi_3 + \pi_4 + \pi_5 = 0.95, \quad \pi_3 + \pi_4 + \pi_5 + \pi_6 = 1.$$

Puisque la réalisation d'une $U(0,1)$ est 0.83021, l'échantillon obtenu est $\{1, 2, 4\}$.

Si l'on modifie l'ordre du fichier en gardant cette réalisation d'une $U(0,1)$, les unités 1 et 2 sont toujours sélectionnées d'office. Si l'unité $x = 8$ est en position 2, 3 ou 4 elle est toujours retenue ; sinon tout est possible...

Morale de l'histoire l'ordre du fichier influe sur l'échantillon sélectionné.

Exercice 11. Soit une population de 5 unités. On veut sélectionner par un tirage systématique à probabilités inégales un échantillon de deux unités avec des probabilités d'inclusion proportionnelles aux valeurs x_i suivantes

$$1, 1, 6, 6, 6.$$

1. Calculez les probabilités d'inclusion d'ordre un.
2. Considérant les deux unités dont la valeur x_i vaut 1, calculez leurs probabilités d'inclusion d'ordre deux pour chacune des permutations possibles du fichier. Conséquence ?

Solution 11.

1. Le total vaut $\sum_{k\in\mathcal{U}} x_k = 20$ et donc

$$\pi_1 = \pi_2 = 2 \times \frac{1}{20} = \frac{1}{10}, \qquad \pi_3 = \pi_4 = \pi_5 = 2 \times \frac{6}{20} = \frac{3}{5}.$$

2. Remarquons qu'il y a $\binom{5}{2} = 10$ permutations possibles du fichiers, i.e., le nombre de manières de placer les valeurs « 1 » parmi 5 cases. Le tableau 3 liste toutes les permutations possibles.

   Ici il s'agit d'être astucieux afin de ne pas faire tous les calculs. En effet pour le tirage systématique seule la position relative des unités importe. Ainsi les permutations du fichier 1, 4, 5, 8 et 10 correspondent à la même situation et les permutations 2,

Table 3: *Toutes les permutations possibles du fichier de l'exercice .*

Permutation	x_1	x_2	x_3	x_4	x_5
1	1	1	6	6	6
2	1	6	1	6	6
3	1	6	6	1	6
4	1	6	6	6	1
5	6	1	1	6	6
6	6	1	6	1	6
7	6	1	6	6	1
8	6	6	1	1	6
9	6	6	1	6	1
10	6	6	6	1	1

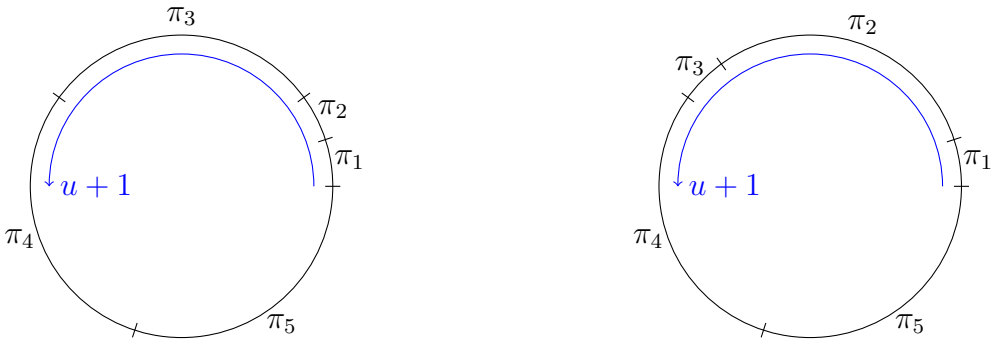

Figure 1: *Représentation graphique des probabilités d'inclusion d'ordre 1 cumulées de l'exercice. Gauche : cas A ; Droite : Cas B.*

3, 6, 7 et 9 à une autre unique situation. Il suffit donc de considérer que ces deux cas particuliers au lieu des 10 possibles, c'est à dire

$$\text{Cas A}: \{1,1,6,6,6\}, \qquad \text{Cas B}: \{1,6,1,6,6\},$$

dont les probabilités d'inclusion d'ordre 1 sont respectivement

$$\text{Cas A}: (1/10, 1/10, 3/5, 3/5, 3/5), \qquad \text{Cas B}: (1/10, 3/5, 1/10, 3/5, 3/5).$$

Graphiquement les probabilités cumulées peuvent se représenter comme sur la Figure 1. On voit notamment qu'il est impossible d'obtenir un échantillon contenant les deux unités « 1 » puisque le pas de tirage vaut 1. Ainsi on a donc une probabilité d'inclusion d'ordre 2 nulle et l'estimation de la variance du π–estimateur est biaisée.

Exercice 12. Soit une population $\mathcal{U}$ composée de 6 unités. On connaît les valeurs prises par un caractère auxiliaire x sur toutes les unités de la population :

$$x_1 = 200, \quad x_2 = 80, \quad x_3 = 50, \quad x_4 = 50, \quad x_5 = 10, \quad x_6 = 10.$$

1. Calculez les probabilités d'inclusion d'ordre un proportionnelles aux x_k pour une taille d'échantillon $n = 4$. Soit 0.48444 une réalisation d'une $U(0,1)$. Sélectionnez un échantillon à probabilités inégales sans remise de taille 4 au moyen d'un tirage systématique, en gardant l'ordre initial du fichier.
2. Donnez la matrice des probabilités d'inclusions d'ordre deux (ordre initial du fichier fixé).

3. On suppose qu'une variable d'intérêt y prend les valeurs suivantes :

$$y_1 = 80, \quad y_2 = 50, \quad y_3 = 30, \quad y_4 = 25, \quad y_5 = 10, \quad y_6 = 5.$$

Constituez un tableau avec, en ligne chaque échantillon s possible, et en colonne les probabilités de tirage $p(s)$, les estimateurs respectifs du total $\hat{Y}(s)$ et de la variance $\widehat{\mathrm{Var}}[\hat{Y}]$. Calculez, sur la base de ce tableau, les espérances $\mathbb{E}[\hat{Y}]$ et $\mathbb{E}[\widehat{\mathrm{Var}}[\hat{Y}]]$. Commentez.

Solution 12. 1. Puisque $\sum_{k\in\mathcal{U}} x_k = 400$ et

$$\pi_1 = 4\frac{200}{400} = 2 > 1,$$

l'unité 1 est sélectionnée d'office et retirée de la population. On a de manière analogue

$$\pi_2 = 3\frac{80}{400-200} = 1.2 > 1,$$

l'unité 2 est sélectionnée d'office et retirée de la population. Enfin on trouve facilement que

$$\pi_3 = \pi_4 = 2\frac{50}{120} = \frac{5}{6}, \qquad \pi_5 = \pi_6 = 2\frac{10}{120} = \frac{1}{6}.$$

On vérifie que l'on a bien $\sum_{k=3}^{6} \pi_k = 2$.
Les probabilités d'inclusion cumulées sont

$$\pi_3 = 5/6, \quad \pi_4 = 10/6, \quad \pi_5 = 11/6, \quad \pi_6 = 2.$$

Figure 2: *Probabilité d'inclusion cumulées de l'exercice.*

Puisque $0.48444 < 5/6$ et $5/6 < 1 + 0.48444 < 10/6$, l'échantillon obtenu est $\{1, 2, 3, 4\}$.

2. La matrice des probabilités d'inclusion d'ordre deux est donnée par

$$\begin{bmatrix} — & 1 & 5/6 & 5/6 & 1/6 & 1/6 \\ 1 & — & 5/6 & 5/6 & 1/6 & 1/6 \\ 5/6 & 5/6 & — & 4/6 & 1/6 & 0 \\ 5/6 & 5/6 & 4/6 & — & 0 & 1/6 \\ 1/6 & 1/6 & 1/6 & 0 & — & 0 \\ 1/6 & 1/6 & 0 & 1/6 & 0 & — \end{bmatrix}$$

Les deux premières lignes et colonnes de cette matrice découlent de la propriété

$$\text{Pour tout } k, \text{ pour tout } \ell : \pi_k = 1 \Longrightarrow \pi_{k\ell} = \pi_\ell.$$

Pour les autres valeurs c'est un peu plus fastidieux puisqu'il faut considérer tous les cas possibles. A ordre fixé, si on note u la valeur tirée au hasard entre 0 et 1, on voit d'après la Figure 2 que
— Si $0 < u \leq 4/6$, alors on tombe dans les intervalles numéro 1 et 2 ;
— Si $4/6 < u \leq 5/6$, alors on tombe dans les intervalles numéro 1 et 3 ;
— Si $5/6 < u \leq 1$, alors on tombe dans les intervalles numéro 2 et 4.

D'où

$$\pi_{34} = \frac{4}{6}, \qquad \pi_{35} = \frac{1}{6}, \qquad \pi_{46} = \frac{1}{6},$$

et les autres probabilités sont nulles.

3. D'après la matrice précédente, il n'y a que 3 échantillons possibles (de taille fixe). Le π–estimateur s'écrit

$$\hat{y} = \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k},$$

et sa variance est estimée par

$$\widehat{\mathrm{Var}}[\hat{y}] = -\frac{1}{2} \sum_{\substack{k,\ell \in \mathcal{S} \\ k \neq \ell}} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2.$$

Notons que le vrai total vaut $y = 200$. Les estimations et variances estimées pour chaque échantillons sont

Table 4: *Probabilités de tirage, estimations et variances estimées pour chaque échantillon possible.*

Échantillon s	$\Pr(S = s)$	$\hat{y}$	$\widehat{\mathrm{Var}}[\hat{y}]$
$\{1, 2, 3, 4\}$	4/6	196	0.75
$\{1, 2, 3, 5\}$	1/6	226	-48
$\{1, 2, 4, 6\}$	1/6	190	0
Total	1	$\mathbb{E}[\hat{y}] = 200$	$\mathbb{E}[\widehat{\mathrm{Var}}[\hat{y}]] = -7.5$

Comme attendu on voit que le π–estimateur est sans biais. En revanche l'estimateur de sa variance est biaisé puisque des probabilités d'inclusion d'ordre deux sont nulles !!!