Chapitre 10 : correction des exercices proposés

Exercices théoriques

T.10.1

(a) $\sum_{k\in U} I_k = \sum_{k\in\mathcal{S}} 1 + \underbrace{\sum_{k\in U\setminus\mathcal{S}} 0}_{0} = \sum_{k\in\mathcal{S}} 1 = n.$

(b) La variable indicatrice I_k associée à l'individu k est une variable aléatoire n'ayant que deux valeurs possibles (v.a. dichotomique) : 1 et 0. Les probabilités associées à ces valeurs sont :

$$\begin{cases} P(I_k = 1) = P(k \in \mathcal{S}) = \pi_k \\ P(I_k = 0) = P(k \notin \mathcal{S}) = 1 - \pi_k. \end{cases}$$

En d'autres termes, $I_k \sim \mathcal{B}(1, \pi_k)$. Il s'ensuit que $E(I_k) = \pi_k$.

(c) La probabilité que l'individu k soit sélectionné pour faire partie de l'échantillon coïncide avec la probabilité de sélectionner un échantillon contenant l'individu k :

$$\pi_k = P(k \in \mathcal{S}) = \sum_{s\in\mathbb{S}|k\in s} p(s)$$

où $\mathbb{S}$ désigne l'ensemble de tous les échantillons PESR de taille n qu'il est possible de prélever dans la population (cf. section 10.4) et $p(s)$ est la probabilité de sélectionner l'échantillon particulier s (la *probabilité de sélection* de l'échantillon particulier s).
Or, les échantillons PESR de taille n qu'il est possible de prélever dans la population U de taille N sont au nombre de $\binom{N}{n}$ et ont tous la même probabilité de sélection : il s'ensuit que

$$p(s) = \frac{1}{\binom{N}{n}}, \quad \text{pour tout } s \in \mathbb{S}.$$

Par conséquent,

$$\begin{aligned} \pi_k &= \sum_{s\in\mathbb{S}|k\in s} \frac{1}{\binom{N}{n}} \\ &= \frac{\text{nombre d'échantillons PESR de taille } n \text{ contenant l'individu } k}{\binom{N}{n}}. \end{aligned}$$

Or, le nombre d'échantillons PESR de taille n contenant l'individu k est égal à $\binom{N-1}{n-1}$ puisque, pour obtenir un tel échantillon, il suffit de considérer que l'individu k a été sélectionné, puis sélectionner par sondage PESR les $(n-1)$ individus nécessaires pour compléter l'échantillon dans la population qui ne compte plus que $(N-1)$ unités. Nous avons donc

$$\begin{aligned} \pi_k &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-1-(n-1))!}}{\frac{N!}{n!(N-n)!}} \\ &= \frac{(N-1)!n!(N-n)!}{(n-1)!(N-n)!N!} = \frac{(N-1)!n!}{N!(n-1)!} = \frac{n}{N}. \end{aligned}$$

Ainsi, dans le cas du sondage PESR, tous les individus de la population ont la même probabilité de faire partie de l'échantillon, égale au taux de sondage $f = n/N$.

(d) Nous pouvons écrire

$$\overline{X} = \frac{1}{n}\sum_{k\in\mathcal{S}} x_k = \frac{1}{n}\sum_{k\in U} x_k I_k.$$

Dans cette dernière expression de $\overline{X}$, seules les variables indicatrices I_k ($k \in U$) sont aléatoires (les x_k sont des *valeurs* fixées). Il s'ensuit que

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n}\sum_{k\in U} x_k I_k\right) = \frac{1}{n}\sum_{k\in U} x_k E(I_k) \\ &= \frac{1}{n}\sum_{k\in U} x_k \pi_k = \frac{1}{n}\sum_{k\in U} x_k \frac{n}{N} \\ &= \frac{1}{N}\sum_{k\in U} x_k = \mu. \end{aligned}$$

T.10.2

Soit k un individu de la population U. La probabilité $\pi_k = P(k \in \mathcal{S})$ que l'individu k appartienne à l'échantillon $\mathcal{S}$ obtenu par n tirages *à probabilités égales avec remise* dans U peut être déterminée comme suit :

$$\begin{aligned} \pi_k &= P(k \in \mathcal{S}) \\ &= 1 - P(k \notin \mathcal{S}) \\ &= 1 - P(\text{l'individu } k \text{ n'est sélectionné à aucun des } n \text{ tirages}). \end{aligned}$$

Or, les n tirages se faisant *avec* remise, ils sont réalisés indépendamment les uns des autres ; par ailleurs, à chaque tirage, la probabilité de ne pas sélectionner l'individu k est égale à $\frac{N-1}{N}$. Nous pouvons donc écrire :

$$\begin{aligned} \pi_k &= 1 - \prod_{i=1}^{n} P(\text{l'individu } k \text{ n'est pas sélectionné au } i\text{-ème tirage}) \\ &= 1 - \prod_{i=1}^{n} \left(\frac{N-1}{N}\right) \\ &= 1 - \left(\frac{N-1}{N}\right)^n \\ &= 1 - \left(1 - \frac{1}{N}\right)^n. \end{aligned}$$

Ainsi, dans le cas du sondage PEAR de taille n dans une population de taille N, tous les individus de la population ont la même probabilité de faire partie de l'échantillon, égale à

$$\pi = 1 - \left(1 - \frac{1}{N}\right)^n.$$

T.10.3

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(x_{ih}-\mu)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(x_{ih}-\mu_h+\mu_h-\mu)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left[(x_{ih}-\mu_h)^2+(\mu_h-\mu)^2+2(x_{ih}-\mu_h)(\mu_h-\mu)\right] \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(x_{ih}-\mu_h)^2+\frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(\mu_h-\mu)^2+\frac{2}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(x_{ih}-\mu_h)(\mu_h-\mu) \\
&= \sum_{h=1}^{H}\frac{N_h}{N}\underbrace{\left[\frac{1}{N_h}\sum_{i=1}^{N_h}(x_{ih}-\mu_h)^2\right]}_{\sigma_h^2}+\frac{1}{N}\sum_{h=1}^{H}N_h(\mu_h-\mu)^2+\frac{2}{N}\sum_{h=1}^{H}(\mu_h-\mu)\sum_{i=1}^{N_h}(x_{ih}-\mu_h) \\
&= \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2+\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2+\frac{2}{N}\sum_{h=1}^{H}(\mu_h-\mu)N_h\underbrace{\left[\frac{1}{N_h}\sum_{i=1}^{N_h}(x_{ih}-\mu_h)\right]}_{=\mu_h-\mu_h=0} \\
&= \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2+\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2.
\end{aligned}
$$

Cette égalité nous montre que la variance de $\mathcal{X}$ dans la population se décompose en une somme de deux termes :

- le premier terme $\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2\right)$ est une moyenne pondérée des variances de $\mathcal{X}$ au sein des différentes strates ; il correspond ainsi à une mesure globale de la dispersion de $\mathcal{X}$ à l'intérieur même des strates et porte le nom de *variance intra(-strates)* (σ^2_{dans}) ;
- le second terme $\left(\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2\right)$ est une mesure de la dispersion des moyennes des différentes strates autour de la moyenne globale μ ; il quantifie dans quelle mesure les moyennes de $\mathcal{X}$ dans les différentes strates diffèrent les unes des autres et est appelé *variance inter(-strates)* (σ^2_{entre}).

La variance globale (variance de $\mathcal{X}$ dans l'ensemble de la population) apparaît donc comme la somme de la variance intra-strates et de la variance inter-strates.

T.10.4

Nous recherchons les effectifs n_h qui minimisent

$$
\begin{aligned}
V(\widehat{\mu}_{\text{ST}}) &\approx \sum_{h=1}^{H}\frac{N_h}{N^2}\frac{N_h-n_h}{n_h}\sigma_h^2 \\
&= \sum_{h=1}^{H}\frac{N_h}{N^2}\left(\frac{N_h}{n_h}-1\right)\sigma_h^2
\end{aligned}
$$

sous la contrainte que $\sum_{h=1}^{H} n_h = n$.

La fonction lagrangienne associée à ce problème de minimisation sous contrainte est la fonction

$$\mathcal{L}(n_1,\ldots,n_H;\lambda)=\sum_{\ell=1}^{H}\frac{N_\ell}{N^2}\left(\frac{N_\ell}{n_\ell}-1\right)\sigma_\ell^2+\lambda\left(\sum_{\ell=1}^{H}n_\ell-n\right).$$

Les effectifs n_h recherchés sont alors solutions du système

$$\left\{\begin{array}{l}\frac{\partial}{\partial_{n_h}}\mathcal{L}(n_1,\ldots,n_H;\lambda)=0, \qquad h=1,\ldots,H\\ \frac{\partial}{\partial_\lambda}\mathcal{L}(n_1,\ldots,n_H;\lambda)=0\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{l}\frac{N_h}{N^2}\left(\frac{-N_h}{n_h^2}\right)\sigma_h^2+\lambda=0, \qquad h=1,\ldots,H\\ \sum_{\ell=1}^{H}n_\ell=n\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{l}\frac{N_h^2}{N^2n_h^2}\sigma_h^2=\lambda\\ \sum_{\ell=1}^{H}n_\ell=n\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}n_h^2=\frac{N_h^2\sigma_h^2}{N^2\lambda}\\ \sum_{\ell=1}^{H}n_\ell=n\end{array}\right. \Leftrightarrow \left\{\begin{array}{ll}n_h=\frac{N_h\sigma_h}{N\sqrt{\lambda}}, \quad h=1,\ldots,H & (1)\\ \sum_{\ell=1}^{H}n_\ell=n & (2)\end{array}\right.$$

En introduisant (1) dans (2), on obtient :

$$\sum_{\ell=1}^{H}\frac{N_\ell\sigma_\ell}{N\sqrt{\lambda}}=n\Leftrightarrow\sqrt{\lambda}=\frac{1}{n}\sum_{\ell=1}^{H}\frac{N_\ell}{N}\sigma_\ell.$$

En réinjectant cette expression de $\sqrt{\lambda}$ dans (1), on trouve, pour $h\in\{1,\ldots,H\}$:

$$n_h=\frac{\frac{N_h}{N}\sigma_h}{\frac{1}{n}\sum_{\ell=1}^{H}\frac{N_\ell}{N}\sigma_\ell}=\left(\frac{N_h\sigma_h}{\sum_{\ell=1}^{H}N_\ell\sigma_\ell}\right)n.$$

T.10.5

$$\begin{aligned}
V(\widehat{\mu}_{\mathrm{ST}}) &= \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(\frac{N_h-n_h}{N_h-1}\right)\frac{\sigma_h^2}{n_h}\\
&\approx \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(\frac{N_h-n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} \quad \text{si } N_h \text{ est grand } (h=1,\ldots,H)\\
&= \sum_{h=1}^{H}\frac{N_h}{N^2}\left(\frac{N_h-n_h}{n_h}\right)\sigma_h^2\\
&= \sum_{h=1}^{H}\frac{N_h}{N^2}\left(\frac{N_h}{n_h}-1\right)\sigma_h^2\\
&= \sum_{h=1}^{H}\frac{N_h^2}{N^2n_h}\sigma_h^2-\frac{1}{N}\underbrace{\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2}_{\sigma_{\mathrm{dans}}^2}.
\end{aligned}$$

Or, dans le cas du sondage stratifié optimal, $n_h \approx \frac{\sigma_h}{\overline{\sigma}} f N_h$ (cf. (10.33)). Par conséquent,

$$\begin{aligned} V(\widehat{\mu}_{\mathrm{STO}}) &\approx \sum_{h=1}^{H} \frac{N_h^2}{N^2 \frac{\sigma_h}{\overline{\sigma}} f N_h} \sigma_h^2 - \frac{\sigma_{\mathrm{dans}}^2}{N} \\ &= \overline{\sigma} \sum_{h=1}^{H} \frac{N_h^2 \sigma_h^2}{N^2 \frac{n}{N} N_h \sigma_h} - \frac{\sigma_{\mathrm{dans}}^2}{N} \\ &= \frac{\overline{\sigma}}{n} \underbrace{\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h}_{\overline{\sigma}} - \frac{\sigma_{\mathrm{dans}}^2}{N} \\ &= \frac{\overline{\sigma}^2}{n} - \frac{\sigma_{\mathrm{dans}}^2}{N} \ . \end{aligned}$$

T.10.6

On se convainc aisément que le minimum de $V(\widehat{\mu}_{\mathrm{ST}})$ ne peut être atteint que si la contrainte est saturée, c'est-à-dire si $\sum_{h=1}^{H} n_h C_h = C_0$ (on accepte de dépenser l'ensemble du budget disponible). Nous recherchons donc les effectifs n_h qui minimisent

$$V(\widehat{\mu}_{\mathrm{ST}}) \approx \sum_{h=1}^{H} \frac{N_h}{N^2} \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2$$

sous la contrainte que $\sum_{h=1}^{H} n_h C_h = C_0$.

La fonction lagrangienne associée à ce problème de minimisation sous contrainte est la fonction

$$\mathcal{L}(n_1, \ldots, n_H; \lambda) = \sum_{\ell=1}^{H} \frac{N_\ell}{N^2} \left(\frac{N_\ell}{n_\ell} - 1 \right) \sigma_\ell^2 + \lambda \left(\sum_{\ell=1}^{H} n_\ell C_\ell - C_0 \right).$$

Les effectifs n_h recherchés sont alors solutions du système

$$\left\{ \begin{array}{ll} \frac{\partial}{\partial_{n_h}} \mathcal{L}(n_1, \ldots, n_H; \lambda) = 0, & h = 1, \ldots, H \\ \frac{\partial}{\partial_\lambda} \mathcal{L}(n_1, \ldots, n_H; \lambda) = 0 & \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{ll} \frac{N_h}{N^2} \left(\frac{-N_h}{n_h^2} \right) \sigma_h^2 + \lambda C_h = 0, & h = 1, \ldots, H \\ \sum_{\ell=1}^{H} n_\ell C_\ell = C_0 & \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{l} \frac{N_h^2}{N^2 n_h^2} \sigma_h^2 = \lambda C_h \\ \sum_{\ell=1}^{H} n_\ell C_\ell = C_0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} n_h^2 = \frac{N_h^2 \sigma_h^2}{\lambda N^2 C_h} \\ \sum_{\ell=1}^{H} n_\ell C_\ell = C_0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} n_h = \frac{N_h \sigma_h}{\sqrt{\lambda} N \sqrt{C_h}}, \quad h = 1, \ldots, H & (1) \\ \sum_{\ell=1}^{H} n_\ell C_\ell = C_0 & (2) \end{array} \right.$$

En introduisant (1) dans (2), on obtient :

$$\sum_{\ell=1}^{H} \frac{N_\ell \sigma_\ell}{\sqrt{\lambda} N \sqrt{C_\ell}} C_\ell = C_0 \Leftrightarrow \sqrt{\lambda} = \sum_{\ell=1}^{H} \frac{N_\ell \sigma_\ell \sqrt{C_\ell}}{N C_0} = \frac{1}{N C_0} \sum_{\ell=1}^{H} N_\ell \sigma_\ell \sqrt{C_\ell}.$$

En réinjectant cette expression de $\sqrt{\lambda}$ dans (1), on trouve, pour $h \in \{1, \ldots, H\}$:

$$\begin{aligned} n_h &= \frac{N_h \sigma_h}{\frac{1}{NC_0}\left(\sum_{\ell=1}^{H} N_\ell \sigma_\ell \sqrt{C_\ell}\right) N\sqrt{C_h}} \\ &= \frac{C_0 N_h \sigma_h}{\sqrt{C_h}\sum_{\ell=1}^{H} N_\ell \sigma_\ell \sqrt{C_\ell}} \cdot \end{aligned}$$

T.10.7

(a)

$$\begin{aligned} E(\widehat{\tau}_{\mathrm{HT}}) &= E\left(\sum_{k\in\mathcal{S}} \frac{x_k}{\pi_k}\right) = E\left(\sum_{k\in U} \frac{x_k}{\pi_k} I_k\right) \\ &= \sum_{k\in U} \frac{x_k}{\pi_k} E(I_k) = \sum_{k\in U} \frac{x_k}{\pi_k}\pi_k \\ &= \sum_{k\in U} x_k = \tau. \end{aligned}$$

(b) Puisque $\mu = \frac{1}{N}\sum_{k\in U} x_k = \frac{\tau}{N}$, il est naturel de prendre

$$\widehat{\mu}_{\mathrm{HT}} = \frac{\widehat{\tau}_{\mathrm{HT}}}{N} \cdot$$

Nous avons bien alors

$$E(\widehat{\mu}_{\mathrm{HT}}) = E\left(\frac{\widehat{\tau}_{\mathrm{HT}}}{N}\right) = \frac{E(\widehat{\tau}_{\mathrm{HT}})}{N} = \frac{\tau}{N} = \mu.$$

(c) Dans le cas du sondage PESR, nous avons, pour tout $k \in U$,

$$\pi_k = \frac{n}{N} \cdot$$

Dès lors

$$\begin{aligned} \widehat{\tau}_{\mathrm{HT}} &= \sum_{k\in\mathcal{S}} \frac{x_k}{\pi_k} = \sum_{k\in\mathcal{S}} \frac{x_k}{n/N} \\ &= N\left(\frac{1}{n}\sum_{k\in\mathcal{S}} x_k\right) = N\overline{X} \end{aligned}$$

et

$$\widehat{\mu}_{\mathrm{HT}} = \frac{\widehat{\tau}_{\mathrm{HT}}}{N} = \overline{X}.$$

Les estimateurs de Horvitz-Thompson de τ et μ dans le cas PESR coïncident avec les estimateurs « classiques » de τ et μ utilisés pour ce plan de sondage.

T.10.8

Dans le sondage stratifié, le taux de sondage appliqué dans la strate numéro h est n_h/N_h

$(h = 1, \ldots, H)$. Puisque, dans chaque strate, on prélève un échantillon par tirages PESR, la probabilité d'inclusion d'un individu correspond au taux de sondage appliqué dans la strate à laquelle cet individu appartient :

$$\pi_k = P(k \in \mathcal{S}) = \frac{n_h}{N_h} \quad \text{si } k \in \text{strate numéro } h.$$

Ainsi, les individus d'une strate possèdent tous la même probabilité d'inclusion, mais dès le moment où on applique des taux de sondage différant d'une strate à l'autre, deux individus appartenant à deux strates différentes ont des probabilités d'inclusion différentes (si $k \in$ strate h et $j \in$ strate ℓ : $\pi_k = n_h/N_h$ et $\pi_j = n_\ell/N_\ell$). De manière générale, le sondage stratifié est un sondage à probabilités inégales.

Cependant, dans le cadre d'un sondage stratifié proportionnel, le même taux de sondage est appliqué à toutes les strates : $n_h/N_h = n_\ell/N_\ell = n/N$ pour tout $h \neq \ell \in \{1, \ldots, H\}$. Dans ce cas, tous les individus de la population possèdent la même probabilité d'inclusion : le sondage stratifié proportionnel est donc bien un sondage à probabilités égales.

T.10.9

La moyenne-échantillon est donnée par

$$\overline{X} = \frac{1}{n}\sum_{k\in\mathcal{S}} x_k = \frac{1}{n}\sum_{k\in U} x_k I_k.$$

Par conséquent,

$$\begin{aligned}
V(\overline{X}) &= E\left[(\overline{X}-\mu)^2\right] = E\left[\left(\frac{1}{n}\sum_{k\in U} x_k I_k - \frac{n}{n}\mu\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{k\in U} x_k I_k - \mu\frac{1}{n}\sum_{k\in U} I_k\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{k\in U} (x_k-\mu) I_k\right)^2\right] \\
&= \frac{1}{n^2}E\left[\sum_{k\in U}(x_k-\mu)^2 I_k^2 + \sum_{k\neq\ell\in U}\sum (x_k-\mu)(x_\ell-\mu) I_k I_\ell\right] \\
&= \frac{1}{n^2}\left[\sum_{k\in U}(x_k-\mu)^2 E(I_k^2) + \sum_{k\neq\ell\in U}\sum (x_k-\mu)(x_\ell-\mu) E(I_k I_\ell)\right]. \qquad (1)
\end{aligned}$$

Or, nous avons :

$$\begin{aligned}
&I_k \sim \mathcal{B}(1, \pi_k) \\
&E(I_k^2) = E(I_k) = \pi_k = \frac{n}{N} \quad \text{(PESR)}. \qquad (2)
\end{aligned}$$

Par ailleurs, nous pouvons définir, pour tout $k \neq \ell \in U$:

$$\begin{aligned}
\pi_{k\ell} &= P\left[(k \in \mathcal{S}) \text{ et } (\ell \in \mathcal{S})\right] \\
&= P(k \in \mathcal{S}) P(\ell \in \mathcal{S} | k \in \mathcal{S}) \\
&= \pi_k \pi_{\ell|k},
\end{aligned}$$

où $\pi_k = \frac{n}{N}$ (probabilité d'inclusion d'un individu pour le prélèvement PESR d'un échantillon de taille n dans une population de taille N) et $\pi_{\ell|k} = P(\ell \in \mathcal{S}|k \in \mathcal{S}) = \frac{n-1}{N-1}$ (sachant que l'on a déjà sélectionné l'individu k, il nous reste à sélectionner, par tirages PESR, $(n-1)$ individus dans une population qui ne compte plus que $(N-1)$ unités ; la probabilité $\pi_{\ell|k}$ est la probabilité d'inclusion de l'individu ℓ dans le cadre de ces $(n-1)$ tirages et vaut donc le taux de sondage $\frac{n-1}{N-1}$). Nous avons ainsi

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)} \cdot$$

Si nous définissons la variable aléatoire indicatrice

$$I_{k\ell} = \begin{cases} 1 & \text{si } k \text{ et } \ell \in \mathcal{S} \\ 0 & \text{sinon,} \end{cases}$$

nous avons $I_{k\ell} \sim \mathcal{B}(1, \pi_{k\ell})$. En outre, il est clair que $I_{k\ell} = I_k I_\ell$. Dès lors,

$$E(I_k I_\ell) = E(I_{k\ell}) = \pi_{k\ell} = \frac{n(n-1)}{N(N-1)} \cdot \tag{3}$$

Il découle de (1), (2) et (3) que

$$V(\overline{X}) = \frac{1}{n^2}\left[\frac{n}{N}\sum_{k\in U}(x_k-\mu)^2 + \frac{n(n-1)}{N(N-1)}\sum\sum_{k\neq\ell\in U}(x_k-\mu)(x_\ell-\mu)\right]. \tag{4}$$

Or, $\sum_{k\in U}(x_k-\mu) = 0$ et donc $\left[\sum_{k\in U}(x_k-\mu)\right]^2 = 0$, ce qui implique que :

$$\begin{aligned}
& \left[\sum_{k\in U}(x_k-\mu)\right]\left[\sum_{\ell\in U}(x_\ell-\mu)\right] = 0 \\
\Leftrightarrow \quad & \sum_{k\in U}\sum_{\ell\in U}(x_k-\mu)(x_\ell-\mu) = 0 \\
\Leftrightarrow \quad & \sum\sum_{k\neq\ell\in U}(x_k-\mu)(x_\ell-\mu) + \sum_{k\in U}(x_k-\mu)^2 = 0 \\
\Leftrightarrow \quad & \sum\sum_{k\neq\ell\in U}(x_k-\mu)(x_\ell-\mu) = -\sum_{k\in U}(x_k-\mu)^2.
\end{aligned}$$

Il s'ensuit que

$$\begin{aligned}
V(\overline{X}) &= \frac{1}{n^2}\left[\frac{n}{N}\sum_{k\in U}(x_k-\mu)^2 - \frac{n(n-1)}{N(N-1)}\sum_{k\in U}(x_k-\mu)^2\right] \\
&= \frac{1}{n}\left[\frac{1}{N} - \frac{n-1}{N(N-1)}\right]\sum_{k\in U}(x_k-\mu)^2 \\
&= \frac{1}{n}\cdot\frac{N-1-n+1}{N(N-1)}\sum_{k\in U}(x_k-\mu)^2 \\
&= \frac{N-n}{n(N-1)}\cdot\underbrace{\frac{1}{N}\sum_{k\in U}(x_k-\mu)^2}_{\sigma^2} \\
&= \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\cdot
\end{aligned}$$

T.10.10

(a) Pour tout $k \neq \ell \in U$: $I_k \sim \mathcal{B}(1, \pi_k)$ et $I_{k\ell} \sim \mathcal{B}(1, \pi_{k\ell})$ (voir le corrigé de l'exercice T.10.9). Par conséquent,

$$V(I_k) = \pi_k(1 - \pi_k)$$

et

$$\begin{aligned} \mathrm{Cov}(I_k, I_\ell) &= E(I_k I_\ell) - E(I_k)E(I_\ell) \\ &= E(I_{k\ell}) - E(I_k)E(I_\ell) \\ &= \pi_{k\ell} - \pi_k \pi_\ell. \end{aligned}$$

(b) L'estimateur de Horvitz-Thompson du total τ de la variable d'intérêt $\mathcal{X}$ dans la population est donné par l'expression

$$\widehat{\tau}_{\mathrm{HT}} = \sum_{k \in \mathcal{S}} \frac{x_k}{\pi_k} = \sum_{k \in U} \frac{x_k}{\pi_k} I_k.$$

Dès lors,

$$\begin{aligned} V(\widehat{\tau}_{\mathrm{HT}}) &= V\left(\sum_{k \in U} \frac{x_k}{\pi_k} I_k\right) \\ &= \sum_{k \in U} \frac{x_k^2}{\pi_k^2} V(I_k) + \sum_{k \neq \ell \in U} \sum \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} \mathrm{Cov}(I_k, I_\ell) \\ &= \sum_{k \in U} \frac{x_k^2}{\pi_k^2} \pi_k(1 - \pi_k) + \sum_{k \neq \ell \in U} \sum \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell) \\ &= \sum_{k \in U} \frac{x_k^2}{\pi_k^2} (\pi_k - \pi_k^2) + \sum_{k \neq \ell \in U} \sum \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell) \\ &= \sum_{k \in U} \frac{x_k^2}{\pi_k^2} \left[\sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) - \sum_{\ell \neq k} (\pi_{k\ell} - \pi_k \pi_\ell)\right] + \sum_{k \neq \ell \in U} \sum \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell). \end{aligned}$$

Or,

$$\sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) = \sum_{\ell \in U} \pi_{k\ell} - \pi_k \sum_{\ell \in U} \pi_\ell,$$

où

$$\sum_{\ell \in U} \pi_\ell = \sum_{\ell \in U} E(I_\ell) = E\left(\sum_{\ell \in U} I_\ell\right) = E(n) = n$$

et

$$\begin{aligned} \sum_{\ell \in U} \pi_{k\ell} &= \pi_k + \sum_{\ell \in U, \ell \neq k} \pi_{k\ell} = \pi_k + \sum_{\ell \in U, \ell \neq k} \pi_k \pi_{\ell|k} \\ &= \pi_k + \pi_k \underbrace{\sum_{\ell \in U, \ell \neq k} \pi_{\ell|k}}_{n-1} \\ &= \pi_k(1 + n - 1) = n\pi_k. \end{aligned}$$

Nous avons donc

$$\sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) = n\pi_k - n\pi_k = 0.$$

Par conséquent,

$$\begin{aligned}
V(\widehat{\tau}_{\mathrm{HT}}) &= -\sum_{k\in U} \frac{x_k^2}{\pi_k^2} \left[\sum_{\ell \neq k} (\pi_{k\ell} - \pi_k \pi_\ell) \right] + \sum\sum_{k \neq \ell \in U} \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell) \\
&= -\frac{1}{2} \sum\sum_{k \neq \ell \in U} \frac{x_k^2}{\pi_k^2} (\pi_{k\ell} - \pi_k \pi_\ell) - \frac{1}{2} \sum\sum_{k \neq \ell \in U} \frac{x_k^2}{\pi_k^2} (\pi_{k\ell} - \pi_k \pi_\ell) \\
&\quad + \sum\sum_{k \neq \ell \in U} \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell) \\
&= -\frac{1}{2} \sum\sum_{k \neq \ell \in U} \frac{x_k^2}{\pi_k^2} (\pi_{k\ell} - \pi_k \pi_\ell) - \frac{1}{2} \sum\sum_{k \neq \ell \in U} \frac{x_\ell^2}{\pi_\ell^2} (\pi_{k\ell} - \pi_k \pi_\ell) \\
&\quad + \sum\sum_{k \neq \ell \in U} \frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell),
\end{aligned}$$

cette dernière égalité étant justifiée par le rôle parfaitement symétrique joué par les indices k et ℓ. Nous pouvons donc encore écrire que

$$\begin{aligned}
V(\widehat{\tau}_{\mathrm{HT}}) &= -\frac{1}{2} \sum\sum_{k \neq \ell \in U} \left[\frac{x_k^2}{\pi_k^2} + \frac{x_\ell^2}{\pi_\ell^2} - 2\frac{x_k}{\pi_k} \frac{x_\ell}{\pi_\ell} \right] (\pi_{k\ell} - \pi_k \pi_\ell) \\
&= -\frac{1}{2} \sum\sum_{k \neq \ell \in U} \left(\frac{x_k}{\pi_k} - \frac{x_\ell}{\pi_\ell} \right)^2 (\pi_{k\ell} - \pi_k \pi_\ell).
\end{aligned}$$

(c) Comme nous l'avons montré dans l'exercice T.10.9, dans le cas PESR, $\pi_k = \frac{n}{N}$ et $\pi_{k\ell} = \frac{n(n-1)}{N(N-1)}$ pour tout $k \neq \ell \in U$. Nous obtenons alors

$$\begin{aligned}
V(\widehat{\tau}_{\mathrm{PESR}}) &= -\frac{1}{2} \sum\sum_{k \neq \ell \in U} \left(\frac{x_k}{n/N} - \frac{x_\ell}{n/N} \right) \left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right) \\
&= -\frac{N^2}{2n^2} \sum\sum_{k \neq \ell \in U} (x_k - x_\ell)^2 \left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right)
\end{aligned}$$

et donc

$$\begin{aligned}
V(\overline{X}) &= V\left(\frac{\widehat{\tau}_{\mathrm{PESR}}}{N} \right) = \frac{1}{N^2} V(\widehat{\tau}_{\mathrm{PESR}}) \\
&= -\frac{1}{2n^2} \left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right) \sum\sum_{k \neq \ell \in U} (x_k - x_\ell)^2.
\end{aligned}$$

Or,

$$\begin{aligned}\sum_{k\neq\ell\in U}\sum (x_k - x_\ell)^2 &= \sum_{k\in U}\sum_{\ell\in U}(x_k - x_\ell)^2 \\ &= \sum_{k\in U}\left[\sum_{\ell\in U}(x_k - x_\ell)^2 - (x_k - x_k)^2\right] \\ &= \sum_{k\in U}\sum_{\ell\in U}(x_k - x_\ell)^2 \\ &= 2N^2\sigma^2,\end{aligned}$$

la dernière égalité découlant d'une application de la relation (3.53) pour la variance de $\mathcal{X}$ dans la population U. Par conséquent,

$$\begin{aligned}V(\overline{X}) &= -\frac{1}{2n^2}\left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}\right)2N^2\sigma^2 \\ &= \left(n^2 - \frac{nN(n-1)}{N-1}\right)\frac{\sigma^2}{n^2} \\ &= \left(1 - \frac{N(n-1)}{n(N-1)}\right)\sigma^2 \\ &= \frac{n(N-1) - N(n-1)}{n(N-1)}\ \sigma^2 \\ &= \frac{nN - n - nN + N}{n(N-1)}\ \sigma^2 \\ &= \frac{N-n}{N-1}\ \frac{\sigma^2}{n}\ .\end{aligned}$$

Exercices pratiques

E.10.1

Soit la population $U = \{2, 4, 12\}$ (population constituée des âges des trois enfants d'une famille).

Les probabilités d'inclusion affectées aux trois enfants de la famille valent respectivement 2/10, 3/10 et 5/10.

Le tableau ci-dessous présente la distribution d'échantillonnage de la v.a. X correspondant à l'âge de l'enfant sélectionné.

Valeurs possibles x de X	$p_x = P(X = x)$
2	2/10
4	3/10
12	5/10
	1

En effet,

$$\begin{aligned} P(X=2) &= P(\text{sélectionner le 1er enfant}) = 2/10\ ; \\ P(X=4) &= P(\text{sélectionner le 2e enfant}) = 3/10\ ; \\ P(X=12) &= P(\text{sélectionner le 3e enfant}) = 5/10. \end{aligned}$$

Il s'ensuit que

$$\begin{aligned} E(X) &= \frac{2}{10}\,2 + \frac{3}{10}\,4 + \frac{5}{10}\,12 = 7.6\ ; \\ V(X) &= \frac{2}{10}\,2^2 + \frac{3}{10}\,4^2 + \frac{5}{10}\,12^2 - (7.6)^2 = 19.84. \end{aligned}$$

E.10.2

Dans la population U :

Mme A est âgée de 26 ans ;
M. B est âgé de 33 ans ;
M. C est âgé de 40 ans ;
Mme D est âgée de 21 ans.

L'âge moyen μ dans cette population vaut $\frac{26+33+40+21}{4} = 30$ ans.

(a) et (b) Sondage PEAR de taille $n = 2$

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
(A,A)	1/16	26	26	26
(A,B)	1/16	26	33	29.5
(A,C)	1/16	26	40	33
(A,D)	1/16	26	21	23.5
(B,A)	1/16	33	26	29.5
(B,B)	1/16	33	33	33
(B,C)	1/16	33	40	36.5
(B,D)	1/16	33	21	27
(C,A)	1/16	40	26	33
(C,B)	1/16	40	33	36.5
(C,C)	1/16	40	40	40
(C,D)	1/16	40	21	30.5
(D,A)	1/16	21	26	23.5
(D,B)	1/16	21	33	27
(D,C)	1/16	21	40	30.5
(D,D)	1/16	21	21	21
Total	1			

Notons que, dans le tableau ci-dessus, nous avons défini les échantillons possibles en tenant compte de l'ordre du tirage, de telle sorte que chacun des $4^2 = 16$ échantillons possibles ait la même probabilité de sélection (égale à 1/16) ; x_1 correspond à l'âge de la première personne sélectionnée et x_2 à l'âge de la seconde personne sélectionnée.

La distribution d'échantillonnage de $\overline{X}$ est résumée dans le tableau suivant :

Valeurs possibles $\overline{x}$ de $\overline{X}$	$P(\overline{X} = \overline{x})$
21	1/16
23.5	2/16
26	1/16
27	2/16
29.5	2/16
30.5	2/16
33	3/16
36.5	2/16
40	1/16
Total	1

$$\begin{aligned} E_{\text{PEAR}}(\overline{X}) &= \frac{1}{16}(21) + \frac{2}{16}(23.5) + \ldots + \frac{1}{16}(40) = 30 = \mu \ ; \\ V_{\text{PEAR}}(\overline{X}) &= \frac{1}{16}(21)^2 + \frac{2}{16}(23.5)^2 + \ldots + \frac{1}{16}(40)^2 - (30)^2 = 25.75. \end{aligned}$$

Sondage PESR de taille $n = 2$

Si l'on ne tient pas compte de l'ordre du tirage, le nombre d'échantillons possibles est égal à

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 2} = 6 \ ;$$

chaque échantillon a une même probabilité égale à 1/6 d'être sélectionné. Dans le tableau ci-dessous, x_1 et x_2 désignent respectivement les âges de la première et de la seconde personne constituant l'échantillon s.

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
{A, B}	1/6	26	33	29.5
{A, C}	1/6	26	40	33
{A, D}	1/6	26	21	23.5
{B, C}	1/6	33	40	36.5
{B, D}	1/6	33	21	27
{C, D}	1/6	40	21	30.5
Total	1			

Le tableau ci-dessus nous donne la distribution d'échantillonnage de $\overline{X}$. Nous en déduisons que :

$$\begin{aligned} E_{\text{PESR}}(\overline{X}) &= \frac{1}{6}(29.5) + \frac{1}{6}(33) + \ldots + \frac{1}{6}(30.5) = 30 = \mu \ ; \\ V_{\text{PESR}}(\overline{X}) &= \frac{1}{6}(29.5)^2 + \frac{1}{6}(33)^2 + \ldots + \frac{1}{6}(30.5)^2 - (30)^2 = 17.17. \end{aligned}$$

Sondage STP de taille $n = 2$ (hommes : B et C ; femmes : A et D)

Le nombre d'échantillons possibles est égal à $2 \times 2 = 4$ (sélection d'*un* individu dans la strate des hommes et d'*un* individu dans la strate des femmes) ; chaque échantillon a une probabilité égale à 1/4 d'être sélectionné.

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
$\{B, A\}$	1/4	33	26	29.5
$\{B, D\}$	1/4	33	21	27
$\{C, A\}$	1/4	40	26	33
$\{C, D\}$	1/4	40	21	30.5
Total	1			

Le tableau ci-dessus nous donne la distribution d'échantillonnage de $\overline{X}$. Nous en déduisons que :

$$\begin{aligned} E_{\text{STP}}(\overline{X}) &= \frac{1}{4}(29.5) + \ldots + \frac{1}{4}(30.5) = 30 = \mu \ ; \\ V_{\text{STP}}(\overline{X}) &= \frac{1}{4}(29.5)^2 + \ldots + \frac{1}{4}(30.5)^2 - (30)^2 = 4.625. \end{aligned}$$

(c) Comme attendu d'après les résultats théoriques,

$$V_{\text{STP}}(\overline{X}) < V_{\text{PESR}}(\overline{X}) < V_{\text{PEAR}}(\overline{X}).$$

On vérifie que

- $\mu = 30$, $\mu_{\text{hommes}} = \frac{33+40}{2} = 36.5$, $\mu_{\text{femmes}} = \frac{26+21}{2} = 23.5$,
- $\sigma^2 = \frac{26^2+33^2+40^2+21^2}{4} - 30^2 = 51.5$,
- $\sigma^2_{\text{hommes}} = \frac{33^2+40^2}{2} - (36.5)^2 = 12.25$,
- $\sigma^2_{\text{femmes}} = \frac{26^2+21^2}{2} - (23.5)^2 = 6.25$,
- $\sigma^2_{\text{dans}} = \frac{2}{4}\sigma^2_{\text{hommes}} + \frac{2}{4}\sigma^2_{\text{femmes}} = \frac{1}{2}(12.25) + \frac{1}{2}(6.25) = 9.25$.

La variance intra-strates ne constitue donc que $9.25/51.5 = 0.18 = 18\%$ de la variance des âges dans l'ensemble de la population. La dispersion des âges à l'intérieur même des strates est donc bien plus faible que la dispersion globale des âges dans la population. Ceci explique pourquoi le sondage STP se montre beaucoup plus efficace que le sondage PESR ($V_{\text{STP}}(\overline{X})$ est près de 4 fois plus petite que $V_{\text{PESR}}(\overline{X})$).

E.10.3

Les valeurs de la variable d'intérêt dans la population sont :

$$2 \quad 3 \quad 6 \quad 8 \quad 11.$$

La moyenne μ de cette variable d'intérêt dans la population vaut $\frac{2+3+6+8+11}{5} = 6$.

(a) et (b) Sondage PEAR de taille $n = 2$

Le nombre d'échantillons possibles est égal à $5^2 = 25$.

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
(2,2)	$1/25 = 0.04$	2	2	2
(2,3)	0.04	2	3	2.5
(2,6)	0.04	2	6	4
(2,8)	0.04	2	8	5
(2,11)	0.04	2	11	6.5
(3,2)	0.04	3	2	2.5
(3,3)	0.04	3	3	3
(3,6)	0.04	3	6	4.5
(3,8)	0.04	3	8	5.5
(3,11)	0.04	3	11	7
(6,2)	0.04	6	2	4
(6,3)	0.04	6	3	4.5
(6,6)	0.04	6	6	6
(6,8)	0.04	6	8	7
(6,11)	0.04	6	11	8.5
(8,2)	0.04	8	2	5
(8,3)	0.04	8	3	5.5
(8,6)	0.04	8	6	7
(8,8)	0.04	8	8	8
(8,11)	0.04	8	11	9.5
(11,2)	0.04	11	2	6.5
(11,3)	0.04	11	3	7
(11,6)	0.04	11	6	8.5
(11,8)	0.04	11	8	9.5
(11,11)	0.04	11	11	11
	1			

Notons que, dans le tableau ci-dessus, nous avons défini les échantillons possibles en tenant compte de l'ordre du tirage, de telle sorte que chacun des 25 échantillons possibles ait la même probabilité de sélection (égale à 1/25) ; x_1 correspond à la valeur observée au premier tirage et x_2 à la valeur observée au second tirage.

La distribution d'échantillonnage de $\overline{X}$ est résumée dans le tableau suivant :

Valeurs possibles $\overline{x}$ de $\overline{X}$	$P(\overline{X} = \overline{x})$
2	0.04
2.5	0.08
3	0.04
4	0.08
4.5	0.08
5	0.08
5.5	0.08
6	0.04
6.5	0.08
7	0.16
8	0.04
8.5	0.08
9.5	0.08
11	0.04
Total	1

$$\begin{aligned} E_{\text{PEAR}}(\overline{X}) &= (0.04)2 + (0.08)(2.5) + \ldots + (0.04)11 = 6 = \mu \ ; \\ V_{\text{PEAR}}(\overline{X}) &= (0.04)2^2 + (0.08)(2.5)^2 + \ldots + (0.04)11^2 - 6^2 = 5.4. \end{aligned}$$

Sondage PESR de taille $n = 2$

Si l'on ne tient pas compte de l'ordre du tirage, le nombre d'échantillons possibles est égal à

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 3 \cdot 2} = 10 \ ;$$

chaque échantillon a une même probabilité égale à $1/10 = 0.1$ d'être sélectionné. Dans le tableau ci-dessous, x_1 et x_2 désignent respectivement les valeurs observées dans l'échantillon s.

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
$\{2,3\}$	0.1	2	3	2.5
$\{2,6\}$	0.1	2	6	4
$\{2,8\}$	0.1	2	8	5
$\{2,11\}$	0.1	2	11	6.5
$\{3,6\}$	0.1	3	6	4.5
$\{3,8\}$	0.1	3	8	5.5
$\{3,11\}$	0.1	3	11	7
$\{6,8\}$	0.1	6	8	7
$\{6,11\}$	0.1	6	11	8.5
$\{8,11\}$	0.1	8	11	9.5
Total	1			

Nous déduisons du tableau ci-dessus que :

$$\begin{aligned} E_{\text{PESR}}(\overline{X}) &= (0.1)(2.5) + (0.1)(4) + \ldots + (0.1)(9.5) = 6 = \mu \ ; \\ V_{\text{PESR}}(\overline{X}) &= (0.1)(2.5)^2 + (0.1)(4)^2 + \ldots + (0.1)(9.5)^2 - 6^2 = 4.05. \end{aligned}$$

Sondage stratifié de taille $n = 2$ (pair : 2, 6, 8 ; impair : 3, 11)

Le nombre d'échantillons possibles est égal à $3 \times 2 = 6$ (sélection d'*une* valeur dans la strate des trois nombres pairs et d'*une* valeur dans la strate des deux nombres impairs) ; chaque échantillon a une probabilité égale à 1/6 d'être sélectionné.

$s \in \mathbb{S}$	$p(s)$	x_1	x_2	Valeur observée pour $\overline{X}$ dans s
$\{2,3\}$	1/6	2	3	2.5
$\{2,11\}$	1/6	2	11	6.5
$\{6,3\}$	1/6	6	3	4.5
$\{6,11\}$	1/6	6	11	8.5
$\{8,3\}$	1/6	8	3	5.5
$\{8,11\}$	1/6	8	11	9.5
Total	1			

Le tableau ci-dessus nous donne :

$$\begin{aligned} E_{\text{ST}}(\overline{X}) &= \frac{1}{6}(2.5) + \frac{1}{6}(6.5) + \ldots + \frac{1}{6}(9.5) = 6.1667 \neq \mu \ ; \\ V_{\text{ST}}(\overline{X}) &= \frac{1}{6}(2.5)^2 + \frac{1}{6}(6.5)^2 + \ldots + \frac{1}{6}(9.5)^2 - (6.1667)^2 = 5.56. \end{aligned}$$

On observe que $\overline{X}$ est un estimateur biaisé dans le cas du sondage stratifié (non proportionnel, comme c'est le cas ici [1]). Dans le cas de ce plan de sondage (avec $n = 2$), nous aurions dû estimer la moyenne-population μ par $\widehat{\mu}_{\text{ST}} = \frac{3}{5}x_1 + \frac{2}{5}x_2$. Notons toutefois que les deux strates considérées ne sont pas idéales pour rendre le sondage stratifié plus efficace que le sondage PESR. En effet, on vérifie que $\sigma^2 = 10.8$, $\sigma^2_{\text{pair}} = 6.22$ et $\sigma^2_{\text{impair}} = 16$, ce qui nous donne $\sigma^2_{\text{dans}} = \frac{3}{5}(6.22) + \frac{2}{5}(16) = 10.13$; la variance dans les strates constitue ainsi $10.13/10.8 = 0.94 = 94\%$ de la variance globale de la variable d'intérêt dans la population ; la variance entre les strates représente à peine 6% de cette variance globale.

E.10.4

Nous avons

$$\begin{aligned} N &= 1000 \ ; \\ n &= 100 \ ; \\ f &= 100/1000 = 0.1 \ ; \\ \sigma^2 &= 16. \end{aligned}$$

1. Pour obtenir un échantillon stratifié de taille $n = 2$, nous sommes obligés de tirer une valeur dans chacune des deux strates. Il s'ensuit que le taux de sondage appliqué dans la strate des valeurs paires est égal à 1/3, alors que celui appliqué dans la strate des valeurs impaires est égal à 1/2. Ces deux taux de sondage n'étant pas identiques, le sondage stratifié n'est pas proportionnel.

PEAR

On effectue $n = 100$ tirages à probabilités égales et avec remise dans l'ensemble des 1000 étudiants de la faculté :

$$V_{\text{PEAR}}(\overline{X}) = \frac{\sigma^2}{n} = \frac{16}{100} = 0.160.$$

PESR

On effectue $n = 100$ tirages à probabilités égales et sans remise dans l'ensemble des 1000 étudiants de la faculté :

$$V_{\text{PESR}}(\overline{X}) = \frac{N-n}{N-1}\ \frac{\sigma^2}{n} = \frac{(1000-100)}{(1000-1)}\ \frac{16}{100} = 0.144.$$

STP

On doit appliquer le même taux de sondage de $10\% = 0.1$ dans chacune des 9 strates.

h	N_h	σ_h	n_h	$\frac{N_h}{N}\ \sigma_h^2$
1	80	4.5	8	1.620
2	170	5.0	17	4.250
3	210	4.2	21	3.704
4	290	3.1	29	2.787
5	50	2.4	5	0.288
6	70	2.7	7	0.510
7	90	1.8	9	0.292
8	30	1.5	3	0.068
9	10	1.0	1	0.010
	1000		100	13.529

$$\begin{aligned} V_{\text{STP}}(\widehat{\mu}) = V_{\text{STP}}(\overline{X}) &\approx \frac{1-f}{n}\sum_h \frac{N_h}{N}\ \sigma_h^2 \\ &= \frac{(1-0.1)}{100}(13.529) = 0.122. \end{aligned}$$

STO

L'allocation optimale de Neyman nous indique la taille de l'échantillon à prélever par tirages PESR dans chacune des strates :

$$n_h \approx n_h^* = \left(\frac{N_h\sigma_h}{\sum_\ell N_\ell\sigma_\ell}\right) n.$$

h	N_h	σ_h	$N_h\sigma_h$	n_h^*	n_h	$\frac{N_h}{N}\ \sigma_h^2$
1	80	4.5	360	10.236	10	1.620
2	170	5.0	850	24.168	24	4.250
3	210	4.2	882	25.078	25	3.704
4	290	3.1	899	25.562	26	2.787
5	50	2.4	120	3.412	3	0.288
6	70	2.7	189	5.374	5	0.510
7	90	1.8	162	4.606	5	0.292
8	30	1.5	45	1.279	1	0.068
9	10	1.0	10	0.284	1	0.010
	1000		3517		100	13.529

$$\begin{aligned}
\overline{\sigma} &= \sum_{\ell} \frac{N_\ell}{N}\ \sigma_\ell = \frac{1}{N}\left(\sum_{\ell} N_\ell \sigma_\ell\right) = \frac{3517}{1000} = 3.517\ ; \\
\sigma^2_{\text{dans}} &= \sum_{h} \frac{N_h}{N}\ \sigma_h^2 = 13.529\ ; \\
V_{\text{STO}}(\widehat{\mu}) &= \frac{\overline{\sigma}^2}{n} - \frac{\sigma^2_{\text{dans}}}{N} = \frac{(3.517)^2}{100} - \frac{13.529}{1000} = 0.110.
\end{aligned}$$

Comme attendu,

$$V_{\text{STO}}(\widehat{\mu}) < V_{\text{STP}}(\overline{X}) < V_{\text{PESR}}(\overline{X}) < V_{\text{PEAR}}(\overline{X}).$$

E.10.5

Nous avons

$$\begin{aligned}
N &= 2000\ ; \\
n &= 120\ ; \\
f &= 120/2000 = 0.06\ ; \\
\sigma^2 &= 400.
\end{aligned}$$

PEAR

On effectue $n = 120$ tirages à probabilités égales et avec remise dans l'ensemble des 2000 personnes :

$$V_{\text{PEAR}}(\overline{X}) = \frac{\sigma^2}{n} = \frac{400}{120} = 3.333.$$

PESR

On effectue $n = 120$ tirages à probabilités égales et sans remise dans l'ensemble des 2000 personnes :

$$V_{\text{PESR}}(\overline{X}) = \frac{N-n}{N-1}\ \frac{\sigma^2}{n} = \frac{(2000-120)}{(2000-1)}\ \frac{400}{120} = 3.135.$$

STP

On doit appliquer le même taux de sondage de $6\% = 0.06$ dans chacune des 6 strates.

h	N_h	σ_h	n_h	$\frac{N_h}{N}\ \sigma_h^2$
1	180	1	11	0.09
2	250	3	15	1.125
3	300	2	18	0.6
4	420	9	25	17.01
5	700	10	42	35
6	150	4	9	1.2
	2000		120	55.025

$$\begin{aligned}
V_{\text{STP}}(\widehat{\mu}) = V_{\text{STP}}(\overline{X}) &\approx \frac{1-f}{n} \sum_{h} \frac{N_h}{N}\ \sigma_h^2 \\
&= \frac{(1-0.06)}{120}(55.025) = 0.431.
\end{aligned}$$

STO

L'allocation optimale de Neyman nous indique la taille de l'échantillon à prélever par tirages PESR dans chacune des strates :

$$n_h \approx n_h^* = \left(\frac{N_h \sigma_h}{\sum_\ell N_\ell \sigma_\ell} \right) n.$$

h	N_h	σ_h	$N_h\sigma_h$	n_h^*	n_h	$\frac{N_h}{N}\ \sigma_h^2$
1	180	1	180	1.673	2	0.09
2	250	3	750	6.971	7	1.125
3	300	2	600	5.577	6	0.6
4	420	9	3780	35.136	35	17.01
5	700	10	7000	65.066	65	35
6	150	4	600	5.577	6	1.2
	2000		12910		121	55.025

Quand on compare les tailles n_h des sous-échantillons à sélectionner dans les strates selon le STO et le STP, on voit clairement que, comparativement au STP, le STO a tendance à surreprésenter dans l'échantillon global les strates hétérogènes et à sous-représenter les strates plus homogènes.

$$\begin{aligned}
\overline{\sigma} &= \sum_\ell \frac{N_\ell}{N}\ \sigma_\ell = \frac{1}{N}\left(\sum_\ell N_\ell \sigma_\ell\right) = \frac{12910}{2000} = 6.455\ ; \\
\sigma^2_{\text{dans}} &= \sum_h \frac{N_h}{N}\ \sigma_h^2 = 55.025\ ; \\
V_{\text{STO}}(\widehat{\mu}) &= \frac{\overline{\sigma}^2}{n} - \frac{\sigma^2_{\text{dans}}}{N} = \frac{(6.455)^2}{120} - \frac{55.025}{2000} = 0.320.
\end{aligned}$$

Comme attendu,

$$V_{\text{STO}}(\widehat{\mu}) < V_{\text{STP}}(\overline{X}) < V_{\text{PESR}}(\overline{X}) < V_{\text{PEAR}}(\overline{X}).$$

E.10.6

Nous devons respecter l'allocation suivante :

$$n_h \approx n_h^* = \frac{C_0 N_h \sigma_h}{\sqrt{C_h}\left(\sum_\ell N_\ell \sigma_\ell \sqrt{C_\ell}\right)}, \qquad h = 1, \ldots, H.$$

Par ailleurs,

$$V_{ST}(\widehat{\mu}) = \sum_h \frac{N_h^2}{N^2}\ \frac{N_h - n_h}{N_h - 1}\ \frac{\sigma_h^2}{n_h} \quad \text{(cf. (10.25))}.$$

(a) 1) $C_0 = 120$

h	N_h	σ_h	C_h	$\frac{N_h\sigma_h}{\sqrt{C_h}}$	$N_h\sigma_h\sqrt{C_h}$	n_h^*	n_h	n_hC_h	$\frac{N_h^2}{N^2}\ \frac{N_h-n_h}{N_h-1}\ \frac{\sigma_h^2}{n_h}$
1	380	7.6	1.44	2406.67	3465.6	46.43	46	66.24	0.250
2	200	5.2	2.25	693.33	1560	13.38	13	29.25	0.122
3	120	3.1	4.00	186	744	3.59	4	16	0.053
4	100	1.8	6.25	72	450	1.39	1	6.25	0.051
	800				6219.6		64	117.74	0.475

L'allocation est donnée dans la colonne des n_h ; elle donne lieu à un coût global de 117.74. On obtient avec cette allocation $V_{\mathrm{ST}}(\widehat{\mu}) = 0.475$.

(a) 2) $C_0 = 72$

h	N_h	σ_h	C_h	$\frac{N_h\sigma_h}{\sqrt{C_h}}$	$N_h\sigma_h\sqrt{C_h}$	n_h^*	n_h	n_hC_h	$\frac{N_h^2}{N^2}\ \frac{N_h-n_h}{N_h-1}\ \frac{\sigma_h^2}{n_h}$
1	380	7.6	1.44	2406.67	3465.6	27.86	28	40.32	0.432
2	200	5.2	2.25	693.33	1560	8.03	8	18	0.204
3	120	3.1	4.00	186	744	2.15	2	8	0.107
4	100	1.8	6.25	72	450	0.83	1	6.25	0.051
	800				6219.6		39	72.57	0.794

L'allocation est donnée dans la colonne des n_h ; elle donne lieu à un coût global de 72.57. On obtient avec cette allocation $V_{\mathrm{ST}}(\widehat{\mu}) = 0.794$.

(b) 1) $C_0 = 120$

h	N_h	σ_h	C_h	$\frac{N_h\sigma_h}{\sqrt{C_h}}$	$N_h\sigma_h\sqrt{C_h}$	n_h^*	n_h	n_hC_h	$\frac{N_h^2}{N^2}\ \frac{N_h-n_h}{N_h-1}\ \frac{\sigma_h^2}{n_h}$
1	380	7.6	3	1667.39	5002.16	25.79	26	78	0.468
2	200	5.2	3	600.44	1801.33	9.29	9	27	0.180
3	120	3.1	3	214.77	644.32	3.32	3	9	0.071
4	100	1.8	3	103.92	311.77	1.61	2	6	0.025
	800				6219.6		40	120	0.744

L'allocation est donnée dans la colonne des n_h ; elle donne lieu à un coût global de 120. On peut par ailleurs vérifier qu'elle coïncide avec l'allocation optimale de Neyman correspondant à une taille de l'échantillon fixée à $120/3 = 40$. On obtient, avec cette allocation, $V_{\mathrm{ST}}(\widehat{\mu}) = 0.744$.

(b) 2) $C_0 = 72$

h	N_h	σ_h	C_h	$\frac{N_h\sigma_h}{\sqrt{C_h}}$	$N_h\sigma_h\sqrt{C_h}$	n_h^*	n_h	n_hC_h	$\frac{N_h^2}{N^2}\ \frac{N_h-n_h}{N_h-1}\ \frac{\sigma_h^2}{n_h}$
1	380	7.6	3	1667.39	5002.16	15.47	15	45	0.837
2	200	5.2	3	600.44	1801.33	5.57	6	18	0.275
3	120	3.1	3	214.77	644.32	1.99	2	6	0.107
4	100	1.8	3	103.92	311.77	0.96	1	3	0.051
	800				6219.6		24	72	1.269

L'allocation est donnée dans la colonne des n_h ; elle donne lieu à un coût global de 72. On peut par ailleurs vérifier qu'elle coïncide avec l'allocation optimale de Neyman correspondant à une taille de l'échantillon fixée à $72/3 = 24$. On obtient, avec cette allocation, $V_{\mathrm{ST}}(\widehat{\mu}) = 1.269$.

E.10.7

Population : $U = \{H_1, H_2, H_3, F\}$ de taille $N = 4$.
La proportion π d'hommes dans la population est égale à $3/4 = 0.75$.

PEAR ($n = 2$)

s	$p(s)$	$\widehat{\pi}$	$p(s)\widehat{\pi}$	$p(s)\widehat{\pi}^2$
(H_1, H_1)	1/16	1	0.0625	0.0625
(H_1, H_2)	1/16	1	0.0625	0.0625
(H_1, H_3)	1/16	1	0.0625	0.0625
(H_1, F)	1/16	0.5	0.03125	0.015625
(H_2, H_1)	1/16	1	0.0625	0.0625
(H_2, H_2)	1/16	1	0.0625	0.0625
(H_2, H_3)	1/16	1	0.0625	0.0625
(H_2, F)	1/16	0.5	0.03125	0.015625
(H_3, H_1)	1/16	1	0.0625	0.0625
(H_3, H_2)	1/16	1	0.0625	0.0625
(H_3, H_3)	1/16	1	0.0625	0.0625
(H_3, F)	1/16	0.5	0.03125	0.015625
(F, H_1)	1/16	0.5	0.03125	0.015625
(F, H_2)	1/16	0.5	0.03125	0.015625
(F, H_3)	1/16	0.5	0.03125	0.015625
(F, F)	1/16	0	0	0
	1		0.75	0.65625

On obtient ainsi :

$$\begin{aligned} E_{\text{PEAR}}(\widehat{\pi}) &= 0.75 = \pi \ ; \\ V_{\text{PEAR}}(\widehat{\pi}) &= 0.65625 - (0.75)^2 = 0.09375 = \frac{\pi(1-\pi)}{n} \ . \end{aligned}$$

PESR ($n = 2$)

s	$p(s)$	$\widehat{\pi}$	$p(s)\widehat{\pi}$	$p(s)\widehat{\pi}^2$
$\{H_1, H_2\}$	1/6	1	0.1667	0.1667
$\{H_1, H_3\}$	1/6	1	0.1667	0.1667
$\{H_1, F\}$	1/6	0.5	0.0833	0.0417
$\{H_2, H_3\}$	1/6	1	0.1667	0.1667
$\{H_2, F\}$	1/6	0.5	0.0833	0.0417
$\{H_3, F\}$	1/6	0.5	0.0833	0.0417
	1		0.75	0.625

On obtient ainsi :

$$\begin{aligned} E_{\text{PESR}}(\widehat{\pi}) &= 0.75 = \pi \ ; \\ V_{\text{PESR}}(\widehat{\pi}) &= 0.625 - (0.75)^2 = 0.0625 = \frac{N-n}{N-1} \ \frac{\pi(1-\pi)}{n} \ . \end{aligned}$$

E.10.8

Population : $U = \{S_1, S_2, S_3, P_1, P_2\}$ de taille $N = 5$.
La proportion π d'étudiants de la faculté de philosophie et lettres dans la population est égale à $2/5 = 0.4$.

PEAR ($n = 2$)

s	$p(s)$	$\widehat{\pi}$	$p(s)\widehat{\pi}$	$p(s)\widehat{\pi}^2$
(S_1, S_1)	$1/25 = 0.04$	0	0	0
(S_1, S_2)	0.04	0	0	0
(S_1, S_3)	0.04	0	0	0
(S_1, P_1)	0.04	0.5	0.02	0.01
(S_1, P_2)	0.04	0.5	0.02	0.01
(S_2, S_1)	0.04	0	0	0
(S_2, S_2)	0.04	0	0	0
(S_2, S_3)	0.04	0	0	0
(S_2, P_1)	0.04	0.5	0.02	0.01
(S_2, P_2)	0.04	0.5	0.02	0.01
(S_3, S_1)	0.04	0	0	0
(S_3, S_2)	0.04	0	0	0
(S_3, S_3)	0.04	0	0	0
(S_3, P_1)	0.04	0.5	0.02	0.01
(S_3, P_2)	0.04	0.5	0.02	0.01
(P_1, S_1)	0.04	0.5	0.02	0.01
(P_1, S_2)	0.04	0.5	0.02	0.01
(P_1, S_3)	0.04	0.5	0.02	0.01
(P_1, P_1)	0.04	1	0.04	0.04
(P_1, P_2)	0.04	1	0.04	0.04
(P_2, S_1)	0.04	0.5	0.02	0.01
(P_2, S_2)	0.04	0.5	0.02	0.01
(P_2, S_3)	0.04	0.5	0.02	0.01
(P_2, P_1)	0.04	1	0.04	0.04
(P_2, P_2)	0.04	1	0.04	0.04
	1		0.4	0.28

On obtient ainsi :

$$\begin{aligned} E_{\text{PEAR}}(\widehat{\pi}) &= 0.4 = \pi \ ; \\ V_{\text{PEAR}}(\widehat{\pi}) &= 0.28 - (0.4)^2 = 0.12 = \frac{\pi(1-\pi)}{n} \ . \end{aligned}$$

PESR ($n = 2$)

s	$p(s)$	$\widehat{\pi}$	$p(s)\widehat{\pi}$	$p(s)\widehat{\pi}^2$
$\{S_1, S_2\}$	0.1	0	0	0
$\{S_1, S_3\}$	0.1	0	0	0
$\{S_1, P_1\}$	0.1	0.5	0.05	0.025
$\{S_1, P_2\}$	0.1	0.5	0.05	0.025
$\{S_2, S_3\}$	0.1	0	0	0
$\{S_2, P_1\}$	0.1	0.5	0.05	0.025
$\{S_2, P_2\}$	0.1	0.5	0.05	0.025
$\{S_3, P_1\}$	0.1	0.5	0.05	0.025
$\{S_3, P_2\}$	0.1	0.5	0.05	0.025
$\{P_1, P_2\}$	0.1	1	0.1	0.1
	1		0.4	0.25

On obtient ainsi :

$$\begin{aligned} E_{\text{PESR}}(\widehat{\pi}) &= 0.4 = \pi \ ; \\ V_{\text{PESR}}(\widehat{\pi}) &= 0.25 - (0.4)^2 = 0.09 = \frac{N-n}{N-1} \ \frac{\pi(1-\pi)}{n} \ . \end{aligned}$$

ST ($n = 2$)

s	$p(s)$	$\widehat{\pi}$	$p(s)\widehat{\pi}$	$p(s)\widehat{\pi}^2$
$\{S_1, P_1\}$	1/6	0.5	0.0833	0.04167
$\{S_1, P_2\}$	1/6	0.5	0.0833	0.04167
$\{S_2, P_1\}$	1/6	0.5	0.0833	0.04167
$\{S_2, P_2\}$	1/6	0.5	0.0833	0.04167
$\{S_3, P_1\}$	1/6	0.5	0.0833	0.04167
$\{S_3, P_2\}$	1/6	0.5	0.0833	0.04167
	1		0.5	0.25

Dans le cas stratifié, $E(\widehat{\pi}) = 0.5 \neq \pi$: $\widehat{\pi}$ est biaisé. Il faudrait plutôt utiliser l'estimateur

$$\widehat{\pi}_{\text{ST}} = \frac{3}{5}\widehat{\pi}_S + \frac{2}{5}\widehat{\pi}_P,$$

où $\widehat{\pi}_S$ est la proportion d'étudiants de philo et lettres dans l'échantillon sélectionné en faculté des sciences ($\widehat{\pi}_S = 0$) et $\widehat{\pi}_P$ est la proportion d'étudiants de philo et lettres dans l'échantillon sélectionné dans cette faculté de philo et lettres ($\widehat{\pi}_P = 1$) : $\widehat{\pi}_{\text{ST}} = \frac{3}{5}0 + \frac{2}{5}1 = \frac{2}{5} = \pi$. cet estimateur est de variance nulle.
Nous avons aussi $V(\widehat{\pi}) = 0.25 - (0.5)^2 = 0$.

E.10.9

(a) Population U :

x_k	x_k^2
1	1
2	4
3	9
4	16
5	25
8	64
9	81
10	100
11	121
12	144
65	565

$$\begin{aligned} \mu &= \frac{65}{10} = 6.5 \ ; \\ \sigma^2 &= \frac{565}{10} - (6.5)^2 = 14.25. \end{aligned}$$

(b) 1)

U_1	
x_k	x_k^2
1	1
2	4
3	9
4	16
5	25
15	55

U_2	
x_k	x_k^2
8	64
9	81
10	100
11	121
12	144
50	510

$$\begin{aligned} \mu_1 &= \frac{15}{5} = 3 \ ; \\ \mu_2 &= \frac{50}{5} = 10 \ ; \\ \sigma_1^2 &= \frac{55}{5} - 3^2 = 2 \ ; \\ \sigma_2^2 &= \frac{510}{5} - 10^2 = 2. \end{aligned}$$

Nous avons ainsi :

$$\begin{aligned} \sigma^2_{\text{dans}} &= \frac{5}{10}(2) + \frac{5}{10}(2) = 2 \quad \Rightarrow \quad \frac{\sigma^2_{\text{dans}}}{\sigma^2} = \frac{2}{14.25} = 14\% \ ; \\ \sigma^2_{\text{entre}} &= \frac{5}{10}(3-6.5)^2 + \frac{5}{10}(10-6.5)^2 = 12.25 \quad \Rightarrow \quad \frac{\sigma^2_{\text{entre}}}{\sigma^2} = \frac{12.25}{14.25} = 86\%. \end{aligned}$$

Les deux strates sont bien homogènes (la dispersion y est faible) ; l'essentiel de la dispersion dans la population se fait entre les deux strates. Les deux strates vont dès lors permettre d'avoir un sondage stratifié efficace, c'est-à-dire un sondage stratifié pour lequel la précision des estimateurs sera bien meilleure que pour le sondage PESR.

(b) 2)

U_1	
x_k	x_k^2
2	4
4	16
8	64
10	100
12	144
36	328

U_2	
x_k	x_k^2
1	1
3	9
5	25
9	81
11	121
29	237

$$\begin{aligned}
\mu_1 &= \frac{36}{5} = 7.2 \ ; \\
\mu_2 &= \frac{29}{5} = 5.8 \ ; \\
\sigma_1^2 &= \frac{328}{5} - (7.2)^2 = 13.76 \ ; \\
\sigma_2^2 &= \frac{237}{5} - (5.8)^2 = 13.76.
\end{aligned}$$

Nous avons ainsi :

$$\begin{aligned}
\sigma^2_{\text{dans}} &= \frac{5}{10}(13.76) + \frac{5}{10}(13.76) = 13.76 \quad \Rightarrow \quad \frac{\sigma^2_{\text{dans}}}{\sigma^2} = \frac{13.76}{14.25} = 96.6\% \ ; \\
\sigma^2_{\text{entre}} &= \frac{5}{10}(7.2 - 6.5)^2 + \frac{5}{10}(5.8 - 6.5)^2 = 0.49 \quad \Rightarrow \quad \frac{\sigma^2_{\text{entre}}}{\sigma^2} = \frac{0.49}{14.25} = 3.4\%.
\end{aligned}$$

Les deux strates sont fort hétérogènes ; l'essentiel de la dispersion dans la population se fait à l'intérieur même des deux strates. Le sondage stratifié ne va dès lors pas permettre d'atteindre une meilleure précision que le sondage PESR.

E.10.10

a) Population U

	U_1	U_2	U_3	U_4	U
Taille	4	3	4	3	14
Moyenne	2	6	8	4	
Variance	0.5	0.6667	0.5	0.6667	

$$\mu = \sum_h \frac{N_h}{N} \mu_h = \frac{4}{14}\ 2 + \frac{3}{14}\ 6 + \frac{4}{14}\ 8 + \frac{3}{14}\ 4 = 5$$

$$\begin{aligned}
\sigma^2_{\text{dans}} &= \frac{4}{14}(0.5) + \frac{3}{14}(0.6667) + \frac{4}{14}(0.5) + \frac{3}{14}(0.6667) = 0.5714 \\
\sigma^2_{\text{entre}} &= \frac{4}{14}(2-5)^2 + \frac{3}{14}(6-5)^2 + \frac{4}{14}(8-5)^2 + \frac{3}{14}(4-5)^2 = 5.5714 \\
\sigma^2 &= \sigma^2_{\text{dans}} + \sigma^2_{\text{entre}} = 0.5714 + 5.5714 = 6.1429
\end{aligned}$$

Les quatre strates sont bien homogènes (la dispersion y est faible ; $\sigma^2_{\text{dans}}/\sigma^2 = 9.3\%$) ; l'essentiel de la dispersion dans la population se fait entre les strates ($\sigma^2_{\text{entre}}/\sigma^2 = 90.7\%$). Les

quatre strates vont dès lors permettre d'avoir un sondage stratifié efficace, c'est-à-dire un sondage stratifié pour lequel la précision des estimateurs sera bien meilleure que pour le sondage PESR.

b) Population U^*

	U_1	U_2	U_3	U_4	U
Taille	4	3	4	3	14
Moyenne	4	6	6	4	
Variance	5	2.6667	6.5	6	

$$\mu = \sum_h \frac{N_h}{N}\mu_h = \frac{4}{14}\,4 + \frac{3}{14}\,6 + \frac{4}{14}\,6 + \frac{3}{14}\,4 = 5$$

$$\begin{aligned}
\sigma^2_{\text{dans}} &= \frac{4}{14}(5) + \frac{3}{14}(2.6667) + \frac{4}{14}(6.5) + \frac{3}{14}(6) = 5.1429 \\
\sigma^2_{\text{entre}} &= \frac{4}{14}(4-5)^2 + \frac{3}{14}(6-5)^2 + \frac{4}{14}(6-5)^2 + \frac{3}{14}(4-5)^2 = 1 \\
\sigma^2 &= \sigma^2_{\text{dans}} + \sigma^2_{\text{entre}} = 5.1429 + 1 = 6.1429
\end{aligned}$$

Les deux strates sont fort hétérogènes ; l'essentiel de la dispersion dans la population se fait à l'intérieur même des deux strates ($\sigma^2_{\text{dans}}/\sigma^2 = 83.7\%$ et $\sigma^2_{\text{entre}}/\sigma^2 = 16.3\%$). Le sondage stratifié ne va dès lors pas permettre d'atteindre une meilleure précision que le sondage PESR.

E.10.11

On tire successivement (sans remise) et « au hasard » 10 papiers dans une urne contenant des papiers numérotés de 1 à 666. Les papiers sélectionnés nous indiquent les numéros des pages qui constitueront notre échantillon.

E.10.12

On dresse la liste des chapitres du livre et on les numérote de 1 à 14 (si l'on veut tenir compte du chapitre « En guise de conclusion : la pratique statistique »). On sélectionne alors, par tirages PESR, 3 chapitres parmi ces 14 chapitres du livre.

Dans chacun des trois chapitres sélectionnés au premier degré du sondage, on numérote chacun des paragraphes et on sélectionne ensuite, par tirages PESR, 2 paragraphes parmi l'ensemble des paragraphes du chapitre.

E.10.13

Nous avons $N = 300$ et $n = 10$: le pas du sondage systématique est dès lors égal à $300/10 = 30$.

On tire « au hasard » une page parmi les 30 premières pages du livre : supposons que le hasard nous fasse sélectionner la page 12. Notre échantillon systématique sera alors constitué des pages

12 42 72 102 132 162 192 222 252 282.

E.10.14

$U = \{1, 2, 3, 4\}$: $\tau = 10$ et $\sigma^2 = 1.25$.

(a) PESR ($n = 2$)

s	$p(s)$	x_1	x_2	$\widehat{\tau} = N\overline{x}$	$p(s)\widehat{\tau}$	$p(s)\widehat{\tau}^2$
$\{1,2\}$	1/6	1	2	6	1	6
$\{1,3\}$	1/6	1	3	8	1.3333	10.6667
$\{1,4\}$	1/6	1	4	10	1.6667	16.6667
$\{2,3\}$	1/6	2	3	10	1.6667	16.6667
$\{2,4\}$	1/6	2	4	12	2	24
$\{3,4\}$	1/6	3	4	14	2.3333	32.6667
	1				10	106.6667

On obtient :

$$\begin{aligned} E(\widehat{\tau}) &= 10 \ ; \\ V(\widehat{\tau}) &= 106.6667 - 10^2 = 6.6667. \end{aligned}$$

(b) 1) PISR ($n = 2$)

s	$p(s)$	x_1	π_1	x_2	π_2	$\widehat{\tau} = \frac{x_1}{\pi_1} + \frac{x_2}{\pi_2}$	$p(s)\widehat{\tau}$	$p(s)\widehat{\tau}^2$
$\{1,2\}$	0.050	1	0.2	2	0.2	15	0.75	11.25
$\{1,3\}$	0.075	1	0.2	3	0.8	8.75	0.65625	5.7422
$\{1,4\}$	0.075	1	0.2	4	0.8	10	0.75	7.5
$\{2,3\}$	0.075	2	0.2	3	0.8	13.75	1.03125	14.1797
$\{2,4\}$	0.075	2	0.2	4	0.8	15	1.125	16.875
$\{3,4\}$	0.650	3	0.8	4	0.8	8.75	5.6875	49.765625
	1						10	105.3125

Les probabilités d'inclusion sont obtenues à partir du plan de sondage comme suit :

$$\begin{aligned} P(1 \in \mathcal{S}) &= p(\{1,2\}) + p(\{1,3\}) + p(\{1,4\}) = 0.050 + 0.075 + 0.075 = 0.2 \ ; \\ P(2 \in \mathcal{S}) &= p(\{1,2\}) + p(\{2,3\}) + p(\{2,4\}) = 0.050 + 0.075 + 0.075 = 0.2 \ ; \\ P(3 \in \mathcal{S}) &= p(\{1,3\}) + p(\{2,3\}) + p(\{3,4\}) = 0.075 + 0.075 + 0.650 = 0.8 \ ; \\ P(4 \in \mathcal{S}) &= p(\{1,4\}) + p(\{2,4\}) + p(\{3,4\}) = 0.075 + 0.075 + 0.650 = 0.8. \end{aligned}$$

On vérifie que

$$\sum_{k=1}^{4} P(k \in \mathcal{S}) = 0.2 + 0.2 + 0.8 + 0.8 = 2 = n.$$

Il découle du tableau ci-dessus que :

$$\begin{aligned} E(\widehat{\tau}) &= 10 \ ; \\ V(\widehat{\tau}) &= 105.3125 - 10^2 = 5.3125. \end{aligned}$$

(b) 2) PISR ($n = 2$)

s	$p(s)$	x_1	π_1	x_2	π_2	$\widehat{\tau} = \frac{x_1}{\pi_1} + \frac{x_2}{\pi_2}$	$p(s)\widehat{\tau}$	$p(s)\widehat{\tau}^2$
$\{1,2\}$	0.025	1	0.15	2	0.375	12	0.3	3.6
$\{1,3\}$	0.050	1	0.15	3	0.65	11.2821	0.5641	6.3642
$\{1,4\}$	0.075	1	0.15	4	0.825	11.5151	0.8636	9.9449
$\{2,3\}$	0.100	2	0.375	3	0.65	9.9487	0.9949	9.8977
$\{2,4\}$	0.250	2	0.375	4	0.825	10.1818	2.5454	25.9174
$\{3,4\}$	0.500	3	0.65	4	0.825	9.4639	4.7319	44.7824
	1						10	100.5066

Les probabilités d'inclusion sont obtenues à partir du plan de sondage comme suit :

$$\begin{aligned}
P(1 \in \mathcal{S}) &= p(\{1,2\}) + p(\{1,3\}) + p(\{1,4\}) = 0.025 + 0.050 + 0.075 = 0.15 \ ; \\
P(2 \in \mathcal{S}) &= p(\{1,2\}) + p(\{2,3\}) + p(\{2,4\}) = 0.025 + 0.100 + 0.250 = 0.375 \ ; \\
P(3 \in \mathcal{S}) &= p(\{1,3\}) + p(\{2,3\}) + p(\{3,4\}) = 0.050 + 0.100 + 0.500 = 0.65 \ ; \\
P(4 \in \mathcal{S}) &= p(\{1,4\}) + p(\{2,4\}) + p(\{3,4\}) = 0.075 + 0.250 + 0.500 = 0.825.
\end{aligned}$$

On vérifie que

$$\sum_{k=1}^{4} P(k \in \mathcal{S}) = 0.15 + 0.375 + 0.65 + 0.825 = 2 = n.$$

Il découle du tableau ci-dessus que :

$$\begin{aligned}
E(\widehat{\tau}) &= 10 \ ; \\
V(\widehat{\tau}) &= 100.5066 - 10^2 = 0.5066.
\end{aligned}$$

Le plan de sondage PISR dans lequel les probabilités d'inclusion des éléments de la population sont approximativement proportionnelles aux valeurs de la variable d'intérêt s'avère très efficace : pour ce plan de sondage, l'estimateur de Horvitz-Thompson du total-population τ a une variance très faible.