Chapitre 3 – Échantillonnage aléatoire stratifié

18. Les valeurs de la variable de réponse d'une population sont: $u_1 = 2$, $u_2 = 3$, $u_3 = 4$, $u_4 = 5$, $u_5 = 7$, $u_6 = 9$.

    (a) Déterminer la moyenne μ et la variance σ^2 de la population.

    (b) Calculer la moyenne et la variance de $\overline{y}$ pour un échantillon aléatoire simple de taille 4 de cette population.

    (c) Supposons que la population soit divisée en deux strates: la strate 1 contient $u_1 = 2$, $u_2 = 3$, $u_3 = 4$, tandis que la strate 2 contient $u_4 = 5$, $u_5 = 7$, $u_6 = 9$. Déterminer la moyenne et la variance empirique dans chacune des deux strates.

    (d) Énumérer tous les échantillons de taille 4 qui peuvent être sélectionnés en choisissant des échantillons aléatoires simples de 2 unités dans chacune des strates. Pour chaque échantillon global de taille 4, donner la probabilité qu'il soit sélectionné.

    (e) Pour chaque échantillon obtenu en (d), calculer la moyenne stratifiée $\overline{y}_{\text{STR}}$ de l'échantillon.

    (f) Vérifier que $E(\overline{y}_{\text{STR}}) = \mu$ et

$$V(\overline{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right).$$

    (g) Pour la population considérée, comparer $\overline{y}_{\text{EAS}}$ et $\overline{y}_{\text{STR}}$ comme estimateurs de μ en termes de biais d'échantillonnage et de variabilité. Pourquoi la variance de $\overline{y}_{\text{EAS}}$ est-elle plus élevée que celle de $\overline{y}_{\text{STR}}$?

Solution:

(a) C'est simple: nous avons un population de taille $N = 6$ et

$$\mu = (2 + 3 + 4 + 5 + 7 + 9)/6 = 30/6 = 5$$
$$\sigma^2 = (2^2 + 3^2 + 4^2 + 5^2 + 7^2 + 9^2)/6 - 5^2 = 17/3 = 5.6667.$$

(b) Pour un EAS de taille $n = 4$,

$$E(\overline{y}_{\text{EAS}}) = \mu = 5 \text{ et } V(\overline{y}_{\text{EAS}}) = \frac{\sigma^2}{n} \left(\frac{N - n}{N - 1} \right) = \frac{17/3}{4} \left(\frac{6 - 4}{6 - 1} \right) = 17/30 = 0.56667.$$

(c) Puisque $N_1 = N_2 = 3$, nous obtenons

$$\mu_1 = (2 + 3 + 4)/3 = 3, \quad \sigma_1^2 = (2^2 + 3^2 + 4^2)/3 - 3^2 = 2/3$$
$$\mu_2 = (5 + 7 + 9)/3 = 7, \quad \sigma_2^2 = (5^2 + 7^2 + 9^2)/3 - 7^2 = 8/3.$$

(d) Il y a $\binom{3}{2} \cdot \binom{3}{2} = 9$ différents échantillons de taille $(2, 2)$:

Échantillon	$\overline{y}_{\text{STR}}$
$(2, 3, 5, 7)$	4.25
$(2, 4, 5, 7)$	4.50
$(3, 4, 5, 7)$	4.75
$(2, 3, 5, 9)$	4.75
$(2, 4, 5, 9)$	5.00
$(3, 4, 5, 9)$	5.25
$(2, 3, 7, 9)$	5.25
$(2, 4, 7, 9)$	5.50
$(3, 4, 7, 9)$	5.75

(e) Voir réponse précédente.

(f) On calcule les quantités recherchées directement:

$$E(\overline{y}_{\text{STR}}) = (4.25 + 4.5 + 4.75 + 4.75 + 5 + 5.25 + 5.25 + 5.5 + 5.75)/9 = 5$$
$$V(\overline{y}_{\text{STR}}) = (4.25^2 + 4.5^2 + 4.75^2 + 4.75^2 + 5^2 + 5.25^2 + 5.25^2 + 5.5^2 + 5.75^2)/9 - 5^2 = 5/24 = 0.2083;$$

cette dernière quantité s'obtient aussi directement à l'aide de la formule:

$$V(\overline{y}_{\text{STR}}) = \frac{1}{N^2}\sum_{i=1}^{2} N_i^2 \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right) = \frac{1}{6^2}\left[3^2 \cdot \frac{2/3}{2}\left(\frac{3-2}{3-1}\right) + 3^2 \cdot \frac{8/3}{2}\left(\frac{3-2}{3-1}\right)\right] = 5/24.$$

(g) Ni l'un, ni l'autre des estimateurs n'admet de biais d'échantillonnage, mais l'estimateur de la moyenne de l'échantillon stratifié est beaucoup plus précis selon la mesure de la variation totale. Cela est dû au fait que l'ensemble des moyennes d'échantillons STR possibles est un sous-ensemble propre de l'ensemble des moyennes d'échantillons possibles selon un EAS. Par conséquent, $V(\overline{y}) = V(\overline{y}_{\text{st}}) + K^2$ pour un certain K. ■

19. Pour une population divisée en M strates distinctes, le coût total d'obtention d'un échantillon STR de taille n (contenant n_i unités dans la $i$ème strate, $i = 1, \ldots, M$) est donné par

$$C = c_0 + \sum_{i=1}^{M} c_i n_i^{3/4}.$$

Si l'on souhaite utiliser $\overline{y}_{\text{STR}}$ afin de donner un estimé de la moyenne de la population μ, déterminer les poids d'échantillonnage qui minimiseront $V(\overline{y}_{\text{STR}})$ en respectant la contrainte de coût total ci-dessus.

Solution: Nous utilisons la méthode des multiplicateurs de Lagrange afin de minimiser la fonction

$$f(n_1, \ldots, n_M) = V(\overline{y}_{\text{st}}) = \frac{1}{N^2} \sum_{i=1}^{M} N_i^2 \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right),$$

soumise à la contrainte

$$g(n_1, \ldots, n_M) = c_0 + \sum_{i=1}^{M} c_i n_i^{3/4} - C = 0.$$

En différentiant, on obtient

$$\nabla f(n_1, \ldots, n_M) = -\frac{1}{N^2} \left(\frac{N_1^3 \sigma_1^2}{n_1^2 (N_1 - 1)}, \cdots, \frac{N_M^3 \sigma_M^2}{n_M^2 (N_M - 1)} \right)$$
$$\nabla g(n_1, \ldots, n_M) = \frac{3}{4} \left(\frac{c_1}{n_1^{1/4}}, \cdots, \frac{c_M}{n_M^{1/4}} \right)$$

On résoud $\nabla f = \lambda \nabla g$ en fonction de $(n_1, \ldots, n_M)$ et l'on obtient:

$$(n_1, \ldots, n_M) = -\left(\frac{4}{3N^2 \lambda} \right)^{4/7} \left(\left(\frac{N_1^3 \sigma_1^2}{c_1 (N_1 - 1)} \right)^{4/7}, \cdots, \left(\frac{N_M^3 \sigma_M^2}{c_M (N_M - 1)} \right)^{4/7} \right).$$

Puisque $n = n_1 + \cdots + n_M$, le schéma général d'allocation optimale est donné par

$$w_i = \frac{n_i}{n} = \frac{\left(\frac{N_i^3 \sigma_i^2}{c_i (N_i - 1)} \right)^{4/7}}{\sum_{k=1}^{M} \left(\frac{N_k^3 \sigma_k^2}{c_k (N_k - 1)} \right)^{4/7}}, \quad i = 1, \ldots, M.$$

Si nous supposons en outre que $N_i - 1 \approx N_i$, $i = 1, \ldots, M$, les poids d'échantillonnage sont approximativement les suivants

$$w_i \approx \frac{\left(\frac{N_i^2 \sigma_i^2}{c_i} \right)^{4/7}}{\sum_{k=1}^{M} \left(\frac{N_k^2 \sigma_k^2}{c_k} \right)^{4/7}}, \quad i = 1, \ldots, M,$$

ce qui n'est pas bien différent du cas où nous ne faisons pas l'hypothèse simplificatrice. ■

20. Une chercheuse souhaite donner un estimé du revenu moyen des employés d'une grande entreprise de Montréal. Les employés sont répertoriés par ancienneté (en général, le salaire augmente avec l'ancienneté). Discuter des mérites relatifs de l'EAS et de l'échantillonnage STR dans ce cas. Laquelle de ces approches devrait-on préconiser? À quoi ressemblerait le plan d'échantillonnage ?

Solution: En général, un STR présente plusieurs avantages par rapport à un EAS :

(a) un STR offre une plus grande précision (variance plus faible) qu'un EAS de même taille;

(b) en raison de cette plus grande précision, la taille requise par un STR afin d'obtenir la même précision qu'un EAS est en général plus petite – un STR est ainsi moins dispendieux, en géeral;

(c) un STR peut fournir une protection contre les échantillons "non représentatifs".

Le principal inconvénient du STR est qu'il peut nécessiter un effort administratif plus important que le EAS.

Cependant, dans ce cas, nous disposons déjà d'une base de sondage stratifiée : les dossiers répertoriant les employés par ancienneté. Comme les salaires augmentent avec l'ancienneté (en général), il semble adéquat de stratifier les employés en fonction de leur ancienneté. Mais pour tirer parti du STR, nous devons disposer de strates dans lesquelles la variance est relativement faible. Nous devons donc supposer que les employés de différents secteurs (par exemple, ceux de la recherche et les employés de bureau réguliers) qui ont la même ancienneté ont des salaires comparables. Ensuite, nous devons sélectionner les strates : étant donné que les employés les plus récents ont "plus de marge" pour évoluer (promotions, augmentation de salaire, etc.) et que les employés plus âgés pourraient atteindre un "plafond salarial", on pourrait s'y prendre avec les strates suivantes:

- moins d'un an d'ancienneté;
- entre 1 et 3 ans d'ancienneté;
- entre 3 et 7 ans d'ancienneté;
- entre 7 et 12 ans d'ancienneté;
- plus de 12 ans d'ancienneté,

avec une allocation proportionnelle, puisque le coût et les variances sont égaux entre les strates. On peut aussi utiliser des strates et des allocations différentes (en justifiant les choix, bien entendu). ■

21. Dans l'utilisation de l'estimateur STR $\overline{y}_{\text{STR}}$ en tant qu'estimateur de $\overline{Y}$, il peut s'avérer avantageux de trouver une répartition et une taille d'échantillon qui minimise la variance $V(\overline{y}_{\text{STR}})$, pour un coût fixe C. Autrement dit, le coût C autorisé pour l'enquête est fixe, et nous cherchons la meilleure répartition des ressources qui permet de maximiser l'information au sujet de $\overline{Y}$. la répartition optimale dans ce cas demeure toujours

$$n_i = n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{\sum N_j\sigma_j/\sqrt{c_j}}\right).$$

(a) Montrer que le meilleur choix pour n est

$$n = \frac{(C-c_0)\sum N_k\sigma_k/\sqrt{c_k}}{\sum N_k\sigma_k\sqrt{c_k}},$$

où c_0 représente les frais généraux fixes du sondage. (Noter que $C = c_0 + \sum c_k n_k$.)

(b) Si $V(\overline{y}_{\text{STR}}) = V$ est fixe, montrez que le choix approprié de n est

$$n = \frac{\left(\sum \frac{N_k\sigma_k/\sqrt{c_k}}{N}\right)\left(\sum \frac{N_k\sigma_k\sqrt{c_k}}{N}\right)}{V + \sum \frac{N_k\sigma_k^2}{N^2}}.$$

(c) Une entreprise souhaite obtenir des renseignements sur l'efficacité d'une machine commerciale qu'elle produit. Elle demande aux répondants d'évaluer l'équipement sur une échelle numérique. Le coût par entretien et les variances approximatives des évaluations et du nombre d'éléments pour trois strates sont donnés par (c_1 = \$9, $\sigma_1^2 = 2.25$, $N_1 = 112$), (c_2 = \$25, $\sigma_2^2 = 3.24$, $N_2 = 68$) et (c_3 = \$36, $\sigma_3^2 = 3.24$, $N_3 = 39$). L'entreprise veut donner un estimé de la note moyenne tout en respectant la condition $V(\overline{y}_{\text{STR}}) = 0.1$. Déterminer la taille d'échantillon n qui permet d'atteindre cette borne, et trouver la répartition appropriée.

Solution:

(a) On le constate par la simple manipulation suivante :

$$C - c_0 = \sum n_k c_k = \sum \frac{n(N_k\sigma_k/\sqrt{c_k})}{M} c_k = \frac{n}{M}\sum N_k\sigma_k\sqrt{c_k},$$

où

$$M = \sum \frac{N_k\sigma_k}{\sqrt{c_k}},$$

d'où

$$n = \frac{M(C-c_0)}{\sum n_k\sigma_k\sqrt{c_k}} = \frac{(C-c_0)\sum N_k\sigma_k/\sqrt{c_k}}{\sum N_k\sigma_k\sqrt{c_k}}.$$

(b) Soit a_i tel que $n_i = a_i n$. Puisque

$$V(\overline{y}_{\text{st}}) = \frac{1}{N^2}\sum N_k^2\left(\frac{N_k - a_k n}{N_k}\right)\left(\frac{\sigma_k^2}{a_k n}\right) = V,$$

on obtient (après quelques simplifications)

$$V = \frac{1}{N^2}\sum\left(N_k - \frac{n}{M}(N_k\sigma_k/\sqrt{c_k})\right)\left(\frac{M}{n}\sigma_k\sqrt{c_k}\right)$$

où M est comme à la partie (a). Ainsi ,

$$VN^2 = \frac{1}{n}\sum N_k M\sigma_k\sqrt{c_k} - \sum N_k\sigma_k^2$$

de sorte que

$$\begin{aligned} n &= \frac{\sum N_k M\sigma_k\sqrt{c_k}}{VN^2+\sum N_k\sigma_k^2} = \frac{M\sum N_k\sigma_k\sqrt{c_k}}{VN^2+\sum N_k\sigma_k^2} = \frac{\left(\sum \frac{N_k\sigma_k}{\sqrt{c_k}}\right)\left(\sum N_k\sigma_k\sqrt{c_k}\right)}{VN^2+\sum N_k\sigma_k^2} \\ &= \frac{1/N^2}{1/N^2}\cdot\frac{\left(\sum \frac{N_k\sigma_k}{\sqrt{c_k}}\right)\left(\sum N_k\sigma_k\sqrt{c_k}\right)}{VN^2+\sum N_k\sigma_k^2} = \frac{\left(\sum \frac{N_k\sigma_k/\sqrt{c_k}}{N}\right)\left(\sum \frac{N_k\sigma_k\sqrt{c_k}}{N}\right)}{V+\sum \frac{N_k\sigma_k^2}{N^2}}. \end{aligned}$$

(c) En utilisant les données disponibles, on obtient

$$\begin{aligned} n &= \frac{\left(\frac{N_1\sigma_1}{N\sqrt{c_1}}+\frac{N_2\sigma_2}{N\sqrt{c_2}}+\frac{N_3\sigma_3}{N\sqrt{c_3}}\right)\left(\frac{N_1\sigma_1\sqrt{c_1}}{N}+\frac{N_2\sigma_2\sqrt{c_2}}{N}+\frac{N_3\sigma_3\sqrt{c_3}}{N}\right)}{V+\left(\frac{N_1\sigma_1^2}{N^2}+\frac{N_2\sigma_2^2}{N^2}+\frac{N_3\sigma_3^2}{N^2}\right)} \\ &= \frac{\left(\frac{112\sigma_1}{219\sqrt{9}}+\frac{68\sigma_2}{219\sqrt{25}}+\frac{39\sigma_3}{219\sqrt{36}}\right)\left(\frac{112\sigma_1\sqrt{9}}{219}+\frac{68\sigma_2\sqrt{25}}{219}+\frac{39\sigma_3\sqrt{36}}{219}\right)}{0.1+\left(\frac{112\sigma_1^2}{219^2}+\frac{68\sigma_2^2}{219^2}+\frac{39\sigma_3^2}{219^2}\right)} = 26.266. \end{aligned}$$

En calculant $n_k = \frac{n}{M}N_k\sigma_k$, où $M = 92.18$, on obtient l'allocation $(16, 7, 3)$. ■

22. Pour donner un estimé du nombre total, τ, de sièges du parti social-démocrate dans tous les conseils municipaux d'un pays, la population a été stratifiée en quatre strates en utilisant le nombre total de sièges dans chaque conseil. On retrouve des renseignements sur ces strates dans le tableau suivant.

# sièges	N_i	$\sum_k Y_{k,i}$ (pop)	$\sum_k Y_{k,i}^2$ (pop)	$\sum_k y_{k,i}$ (éch)	$\sum_k y_{k,i}^2$ (éch)
$31-40$	44	756	13784	89	1647
$41-50$	168	3383	72223	441	9735
$51-70$	56	1545	44529	250	8294
71+	16	617	24137	102	5294

(a) Distribuer un échantillon total de taille $n = 40$ dans les 4 strates en utilisant la répartition proportionnelle.

(b) Donner un estimé du total des sièges socio-démocratiques à l'aide d'un échantillon STR de taille $n = 40$ selon cette répartition. Construire un intervalle de confiance pour le total à environ 95%.

(c) Donner un estimé du nombre total de sièges socio-démocratiques si un EAS avait été utilisé à la place afin de sélectionner un échantillon de taille $n = 40$. Construire un intervalle de confiance pour le total à environ 95%.

(d) Laquelle des deux méthodes utilisées en (b) et (c) est la plus efficace? Pourquoi?

Solution:

(a) Sous l'allocation proportionnelle, $n_1 = n\frac{N_i}{N}$. Ici, $n = 40$ et $N = 284$; avec les données, la répartition appropriée est $(n_1, n_2, n_3, n_4) = (6, 24, 8, 2)$.

(b) Dans ce cas, le total τ est approché à l'aide de

$$\hat{\tau} = \sum_{i=1}^{4} N_i \overline{y}_{\mathrm{st,i}} = \sum_{i=1}^{4} \frac{N_i}{n_i} \sum_k y_{k,i} = 6305,67$$

et la variance de τ par

$$\hat{V}(\tau) = \sum_{i=1}^{4} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right),$$

où l'écart type de l'échantillon dans chaque strate est

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_k y_{i,k}^2 - \frac{1}{n_i}\left(\sum_k y_{k,i}\right)^2\right).$$

En utilisant les données fournies dans le tableau, on obtient

$$(s_1^2, s_2^2, s_3^2, s_4^2) = (65.37, 70.94, 68.70, 92.00),$$

de sorte que $\hat{V}(\tau) = 123141.4..$ Ainsi,

$$\hat{\tau} \pm 2\sqrt{\hat{V}(\overline{y}_{\mathrm{st}})} = 6305.67 \pm 701.83.$$

est un I.C. à environ 95

(c) Dans ce cas, le total τ est estimé par

$$N\overline{y} = \frac{284}{40}\sum_{i=1}^{40} y_i = \frac{284}{40}\left(\sum_{i=1}^{4}\sum_{k} y_{k,i}\right) = \frac{284 \cdot 882}{40} = 6262.2$$

et la variance de τ est approchée par

$$\hat{V}(\tau) = N^2\left(\frac{N-n}{N}\right)\left(\frac{s^2}{n}\right),$$

où $s^2 = \frac{1}{39}(24970 - 40 \cdot 22.05^2) = 141.59$. Ainsi,

$$\hat{V}(\tau) = 284^2\left(\frac{244}{284}\right)\left(\frac{141.59}{40}\right) = 245290.52$$

et on obtient un I.C. à environ 95% pour τ à l'aide de

$$N\overline{y}_{\text{st}} \pm 2\sqrt{\hat{V}(\overline{y}_{\text{st}})} = 6262.2 \pm 990.54$$

.

(d) L'estimateur calculé avec un échantillonnage stratifié est plus efficace que l'estimateur calculé avec un échantillonnage aléatoire simple car sa variance est la plus petite des deux; ce n'est pas surprenant puisque les strates sont bien différentes les unes des autres. ■

23. Les salariés d'une grande entreprise sont stratifiés en deux classes: les cadres et les employé.e.s de bureau, la première de taille $N_1 = 121$ et la seconde de taille $N_2 = 589$. On cherche à évaluer l'attitude à l'égard de la politique de congé de maladie en prélevant des échantillons aléatoires indépendants de taille $N_1 = n_2 = 35$ dans chacune des classes. On sépare de plus les réponses selon le genre des répondants. Dans le tableau des résultats, a =nombre d'individus qui aiment la politique; b =nombre d'individus qui n'aiment pas la politique, et c =nombre d'individus qui n'ont pas d'opinion.

	Cadres $N_1 = 121$	**Bureau** $N_2 = 589$	**Total** $N = 710$
Hommes	$a = 3$ $b = 15$ $c = 3$	$a = 10$ $b = 2$ $c = 1$	34
Femmes	$a = 6$ $b = 6$ $c = 2$	$a = 15$ $b = 7$ $c = 0$	36
Total	$n_1 = 35$	$n_2 = 35$	$n = 70$

Donner un estimé et une variance approximative de cet estimé pour les paramètres suivants:

(a) Proportion des cadres en faveur de cette politique.

(b) Proportion des employé.e.s en faveur de cette politique.

(c) Nombre total d'employées qui ne supportent pas la politique.

(d) Différence entre la proportion de cadres masculins et la proportion de cadres féminins en faveur de la politique.

(e) Différence entre la proportion des cadres en faveur de la politique et les cadres qui ne supportent pas la politique.

Solution:

(a) Nous observons que $\hat{p} = \frac{3+6}{35} \approx 0.26$, et que

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n_1 - 1}\left(\frac{N_1 - n_1}{N_1}\right) = \frac{0.26(0.74)}{34}\left(\frac{86}{121}\right) = 0.004.$$

(b) Nous observons que $\hat{p}_{\text{st}} = \frac{1}{N}[N_1 p_1 + N_2 p_2] = \frac{1}{710}[121(0.26) + 589(0.71)] = 0.637$, et que

$$\begin{aligned}\hat{V}(\hat{p}_{\text{st}}) &= \frac{1}{N_2}\sum_{i=1}^{2} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{p_i(1-p_i)}{n_i - 1} \\ &= \frac{1}{(710)^2}\left[(121)^2\left(\frac{86}{121}\right)\frac{0.26(0.74)}{34} + (589)^2\left(\frac{554}{589}\right)\frac{0.71(0.29)}{34}\right] = 0.004.\end{aligned}$$

(c) Nous définissons une nouvelle variables sur les cadres selon

$$z_{1,i} = \begin{cases} 1 & \text{le/la } i\text{ème cadre n'aime pas la politique} \\ 0 & \text{autrement} \end{cases}$$

et sur les employé.e.s de bureau selon

$$z_{2,i} = \begin{cases} 1 & \text{le/la } i\text{ème employé.e de bureau n'aime pas la politique} \\ 0 & \text{autrement} \end{cases}$$

Nous obtenons alors:

Cadre	Bureau
$N_1 = 121$	$N_2 = 589$
$n_1 = 35$	$n_2 = 35$
$\sum z_{1,k} = 21$	$\sum y_{2,k} = 9$
$\sum z_{1,k}^2 = 21$	$\sum y_{2,k}^2 = 9$

Avec ces données, nous obtenons $n = n_1 + n_2 = 70$, $s_1^2 = \frac{34}{35}(21/35(1 - 21/35)) = 0.233$, $s_2^2 = \frac{34}{35}(9/35(1 - 9/35)) = 0.186$, $A_1 = \frac{N_1}{N} = \frac{121}{710}$ et $A_2 = \frac{N_2}{N} = \frac{589}{710}$, d'où

$$\hat{\tau} = N\overline{z} = N\sum_i A_i \overline{y}_i = 710\left[\frac{121}{710}(21/35) + \frac{580}{710}(9/35)\right] = 221.7,$$

et la variance approximative est

$$\begin{aligned}\hat{V}(\hat{\tau}) = N^2\hat{V}(\overline{z}) &= \left[N_1^2\frac{s_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + N_2^2\frac{s_2^2}{n_2}\left(1 - \frac{n_2}{N_2}\right)\right] \\ &= \left[(121)^2\frac{0.233}{35}\left(1 - \frac{35}{121}\right) + (589)^2\frac{0.186}{35}\left(1 - \frac{35}{589}\right)\right] = 1803.358.\end{aligned}$$

(d) Selon un communiqué de presse de l'Organisation internationale du travail émis en 1997, les femmes représenteraient 42% des cadres canadiens en 1997 (ils ne donnent pas l'intervalle de confiance, donc nous ne devrions probablement pas leur faire confiance à ce point, mais en tout cas). Nous allons donc supposer que $0.42(121) \approx 51$ des cadres de l'entreprise sont des femmes et donc que $121 - 51 = 70$ des cadres sont des hommes (vos chiffres pourraient être différents), d'où

$$\hat{p}_{\mathrm{F}} - \hat{p}_{\mathrm{H}} = \frac{6}{14} - \frac{3}{21} = \frac{2}{7} = 0.286$$

et

$$V(\hat{p}_{\mathrm{H}}) = \frac{(3/21)(1 - 3/21)}{21}\left(\frac{70 - 21}{69}\right) = 0.0041,\ V(\hat{p}_{\mathrm{F}}) = \frac{(6/14)(1 - 6/14)}{14}\left(\frac{51 - 14}{50}\right) = 0.0127,$$

d'où

$$V(\hat{p}_{\mathrm{F}} - \hat{p}_{\mathrm{H}}) = V(\hat{p}_{\mathrm{F}}) + V(\hat{p}_{\mathrm{H}}) = 0.0129 + 0.0041 = 0.0171,$$

puisque les proportions sont indépendantes.

(e) Ces quantités sont corrélées (négativement), mais nous supposons ici qu'elles ne le sont pas. La différence entre les proportions est ainsi

$$\hat{p}_1 - \hat{p}_2 = \frac{15 + 6}{35} - \frac{6 + 3}{35} = 0.343,$$

et

$$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p_1}) + V(\hat{p_2}) = \left(\frac{121 - 35}{120(35)}\right)((21/35)(1 - 21/35) + (9/35)(1 - 9/35)) = 0.0088,$$

ce qui complète l'exercice. ■

24. (a) Prélever un échantillon aléatoire de 20 tailles d'hommes à partir d'une distribution binomiale (la taille correspond au nombre de succès de n expériences de Bernouilli indépendentes avec chance de succès p) avec paramètres $n = 142$ et $p = 0.5$, et un échantillon aléatoire distinct de 20 tailles de femmes à partir d'une distribution binomiale avec paramètres $n = 130$ et $p = 0.5$. À partir de ces données, donner un estimé de la taille moyenne des adultes et calculez la marge d'erreur sur l'estimation.

    (b) Prélever un EAS de 40 tailles d'adultes à partir d'une distribution binomiale avec paramètres $n = 135$ et $p = 0.5$. À partir de ces données, donner un estimé de la taille moyenne de tous les adultes et donner une marge d'erreur sur l'estimation.

    (c) Comparer les résultats de (a) et (b). Discuter des cas où la stratification semble produire des gains en précision des estimations.

Solution: On suppose que $N_{\mathrm{H}} \approx N_{\mathrm{F}} \approx 0.5N$ et que

$$\frac{n_{\mathrm{H}}}{N_{\mathrm{H}}} \approx \frac{n_{\mathrm{F}}}{N_{\mathrm{F}}} \approx \frac{n}{N} \approx 0.$$

(a) On se sert du programme suivant afin d'obtenir la moyenne et la variance de chacuns des échantillons:

```
> set.seed(0) # replicabilite
> n = 20
> x.h = rbinom(n,142,0.5) # ech. hommes
> x.f = rbinom(n,130,0.5) # ech. femmes
> (x.h.moy = mean(x.h))   # moy. ech. hommes
[1] 69.85
> (x.h.s2 = var(x.h))     # var. ech. hommes
[1] 29.18684
> (x.f.moy = mean(x.f))   # moy. ech. femmess
[1] 66.3
> (x.f.s2 = var(x.f))     # var. ech. femmes
[1] 26.95789
> (x.st = 0.5*x.h.moy + 0.5*x.f.moy) # moy. str.
[1] 68.07
> (V=(0.5)^2/n*(x.h.s2+x.f.s2)) # var. str.
[1] 0.7018092
> (B=2*sqrt(V))
[1] 1.675481
```

(b) On s'y prend de la même manière:

```
> set.seed(5) # replicabilite
> n = 40
> x = rbinom(n,135,0.5)
> (x.moy = mean(x))
[1] 68.375
> (x.s2 = var(x))
[1] 39.88141
> (V=x.s2/n)
[1] 0.9970353
> (B=2*sqrt(V))
[1] 1.997033
```

(c) Dans ces exemples, la marge d'erreur pour l'échantillon STR est légèrement inférieure à celle de l'échantillon EAS. Est-ce un accident? Le code suivant répète la procédure à 400 reprises pour voir ce qui en découle.

```
> set.seed(6) # replicabilite
> M = 400
> str.meilleure = c()
> for(j in 1:M){
    # STR
    n = 20
    x.h = rbinom(n,142,0.5)
    x.f = rbinom(n,130,0.5)
    x.h.moy = mean(x.h)
    x.f.moy = mean(x.f)
    x.h.s2 = var(x.h)
    x.f.s2 = var(x.f)
    V=(0.5)^2/n*(x.h.s2+x.f.s2)
    B=2*sqrt(V)

    # EAS
    n = 40
    x = rbinom(n,135,0.5)
    x.moy = mean(x)
    x.s2 = var(x)
    V=x.s2/n
    B1=2*sqrt(V)

    # comparaison entre STR et SRS
    str.meilleure[j] = B<B1
> }
> summary(str.meilleure)
   Mode   FALSE    TRUE
logical     210     190
```

L'approche STR semble plus ou moins équivalente à l'EAS avec les paramètres du problème.

En général, lorsque la distribution est multimodale, ce qui n'est pas le cas en (a), la stratification centrée sur les différents modes produit de meilleures estimations, car la variabilité est réduite. Ou encore, si la distribution de la population a une "grande" variance et est une "somme" de distributions distinctes avec de "petites" variances, la stratification est préférable. ■

25. Une école souhaite donner un estimé du score moyen de ses élèves de sixième année à un examen de compréhension de l'écrit. Les élèves de l'école sont regroupés en trois filières: les élèves plus rapides étant regroupés dans la filière I et les élèves plus lents dans la filière III. L'école décide de stratifier par rapport aux filières. La sixième année compte 50 élèves dans la voie I, 90 dans la voie II et 60 dans la voie III. Un échantillon stratifié de 50 élèves est réparti proportionnellement dans les filières (on obtient $n_{\mathrm{I}} = 14$, $n_{\mathrm{II}} = 20$ et $n_{\mathrm{III}} = 16$, respectivement). Les résultats de l'échantillon sont présentés ci-dessous:

Filière i	$\overline{y}_i$	s_i^2
I	79.71	105.14
II	64.75	158.20
III	37.44	186.13

(a) En considérant l'enquête ci-dessus comme une étude pilote, trouver la taille de l'échantillon nécessaire pour donner un estimé du score moyen avec une marge d'erreur sur l'estimation $B = 4$. Utiliser la répartition proportionnelle.

(b) Répéter la partie (a) en utilisant la répartition de Neyman. Comparer les résultats.

Solution: Nous avons $N_I = 50$, $N_{II} = 90$, $N_{III} = 60$ et $N = 200$. On utilise $\sigma_i^2 \approx s_i^2$.

(a) Dans un scénario de répartition proportionnelle, la taille de l'échantillon est

$$n = \frac{N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2}{NB^2/4 + \frac{1}{200}(N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2)} = \frac{50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2}{200(4^2/4) + \frac{1}{200}(50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2)}$$
$$= \frac{50(105.14) + 90(158.20) + 60(186.13)}{200(4) + \frac{1}{200}(50(105.14) + 90(158.20) + 60(186.13))} = 32.16443 \approx 33.$$

La répartition proportionnelle serait alors

$$(n_I, n_{II}, n_{III}) = \frac{n}{N}(N_I, N_{II}, N_{III}) = (8.04, 14.47, 9.65) \approx (8, 14, 10);$$

mais cela ne nous donne que 32 unités, alors qu'il n'en faut au moins 33. Le meilleur candidat est donc $(8, 15, 10)$.

(b) Dans un scénario de répartition de Neyman, la taille de l'échantillon est

$$n = \frac{(N_I\sigma_I + N_{II}\sigma_{II} + N_{III}\sigma_{III})^2}{N^2B^2/4 + (N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2)} = \frac{(50\sigma_I + 90\sigma_{II} + 60\sigma_{III})^2}{200^2(4^2/4) + (50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2)}$$
$$= \frac{(50\sqrt{105.14} + 90\sqrt{158.20} + 60\sqrt{186.13})^2}{200^2(4) + (50(105.14) + 90(158.20) + 60(186.13))} = 31.82409 \approx 32.$$

Pour ceux que cela intéresserait, la répartition de Neyman serait alors

$$(n_I, n_{II}, n_{III}) = \frac{n}{N_I\sigma_I + N_{II}\sigma_{II} + N_{III}\sigma_{III}}(N_I\sigma_I, N_{II}\sigma_{II}, N_{III}\sigma_{III}) \approx (6.62, 14.62, 10.57).$$

Si on arrondi, on obtient $(7, 15, 11)$, ce qui donne 33 unités. À toutes fins pratiques, il n'y a pas vraiment de différence entre la répartition de Neyman et la répartition proportionnelle étant données les variances de strates et la marge d'erreur recherchée, si ce n'est qu'avec l'allocation de Neyman, on donne un peu plus d'importance à la strate qui a une variance plus élevée (compatible avec la définition de cette répartition). ■

26. Une entreprise souhaite obtenir des renseignements sur l'efficacité d'une imprimante commerciale. Un certain nombre de chefs de division seront interrogés par téléphone et il leur sera demandé d'évaluer l'équipement sur une échelle numérique. Les divisions sont situées en Amérique du Nord, en Europe et en Asie. Par conséquent, un échantillonnage stratifié est utilisé. Les coûts sont plus élevés pour les entretiens avec les chefs de division situés en dehors de l'Amérique du Nord. Le tableau suivant indique les coûts par entretien, les variances approximatives des évaluations et la taille des strates.

Strate	N_i	σ_i^2	c_i
Amérique du Nord	127	2.31	\$9
Europe	58	3.33	\$25
Asie	79	3.21	\$36

(a) L'entreprise souhaite donner un estimé de la cote moyenne en préservant $V(\overline{y}_{\text{STR}}) = 0.1$. Obtenir la taille d'échantillon stratifié n qui permet d'atteindre cette marge et trouvez la répartition appropriée.

(b) Un budget de 800\$ est disponible, duquel 125\$ doivent être réservés pour les frais généraux fixes. Déterminer la taille de l'échantillon et la taille optimale des échantillons dans chaque strate.

(c) Répéter (a) et (b) en utilisant un logiciel.

Solution:

(a) Selon la répartition optimale, nous avons

$$n = \frac{\left(\sum_{i=1}^{3} N_i\sigma_i\sqrt{c_i}\right)\left(\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}\right)}{N^2V(\overline{y}_{\text{st}}) + \sum_{i=1}^{3} N_i\sigma_i^2}$$

$$= \frac{\left(127\sqrt{2.31}\sqrt{9} + 58\sqrt{3.33}\sqrt{25} + 79\sqrt{3.21}\sqrt{36}\right)\left(\frac{127\sqrt{2.31}}{\sqrt{9}} + \frac{58\sqrt{3.33}}{\sqrt{25}} + \frac{79\sqrt{3.21}}{\sqrt{36}}\right)}{(264)^2(0.1) + 127(2.31) + 58(3.33) + 79(3.21)} = 27.7;$$

nous devons ainsi choisir (au moins) $n = 28$ chefs de division.
Les poids d'échantillonnage sont donnés par

$$w_i = \frac{n_i}{n} = \frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}} = \frac{1}{142.548} \cdot \frac{N_i\sigma_i}{\sqrt{c_i}}, \quad i = 1, 2, 3.$$

Ainsi,

$$w_1 = \frac{1}{109.1} \cdot \frac{127\sqrt{2.31}}{\sqrt{9}} = 0.590, \; w_2 = \frac{1}{109.1} \cdot \frac{58\sqrt{3.33}}{\sqrt{25}} = 0.194, \; w_3 = \frac{1}{109.1} \cdot \frac{79\sqrt{3.21}}{\sqrt{36}} = 0.216.$$

En gros, nous devrions échantillonner $0.589n = 16.49$ chefs de division en Amérique du Nord, $0.1940n = 5.43$ chefs de division en Europe et $0.216n = 5.98$ chefs de division en Asie (cela ne donne que 27... on en rajoute un en Europe, mettons, puisque la variance dans cette strate est plus élevée): $(n_1, n_2, n_3) = (16, 6, 6)$. On pourrait aussi utiliser $(17, 5, 6)$.

(b) Nous avons $C = 800$ and $c_0 = 125$. La taille totale de l'échantillon minimisant $V(\overline{y}_{\text{st}})$ est donnée par

$$n = (C - c_0)\frac{\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} N_i\sigma_i\sqrt{c_i}} = \frac{(800 - 125)\left(\frac{127\sqrt{2.31}}{\sqrt{9}} + \frac{58\sqrt{3.33}}{\sqrt{25}} + \frac{79\sqrt{3.21}}{\sqrt{36}}\right)}{127\sqrt{2.31}\sqrt{9} + 58\sqrt{3.33}\sqrt{25} + 79\sqrt{3.21}\sqrt{36}} = 37.62;$$

de sorte que nous choisissons 38 chefs de division dans ce schéma. Les poids d'échantillonnage sont les mêmes que ceux en (a), d'où $(n_1, n_2, n_3) \approx (22, 8, 8)$.

(c) Pour la troisième question, on laisse tomber. ■

27. Un gouvernement municipal souhaite agrandir les installations d'une garderie pour enfants à besoins spéciaux. Cette extension augmentera le coût d'inscription d'un enfant dans la garderie. Un sondage sera mené afin de donner un estimé de la proportion de familles ayant des enfants à mobilité réduite qui utiliseront les nouvelles installations. Les familles sont divisées entre celles qui utilisent les installations existantes et celles qui ne les utilisent pas. Certaines familles vivent dans la municipalité, d'autres dans les banlieues et les zones rurales environnantes. On utilise donc un plan d'échantillonnage STR avec les strates suivantes: (1) utilisateurs actuels provenant de la municipalité, (2) utilisateurs actuels provenant des régions environnantes, (3) non-utilisateurs actuels provenant de la municipalité, et (4) non-utilisateurs actuels provenant des régions environnantes. Le coût d'obtention d'une observation pour un utilisateur actuel est de \$4; il est de \$8 pour un non-utilisateur actuel. Selon les dossiers de la municipalité, les populations sont $N_1 = 97$, $N_2 = 43$, $N_3 = 45$ et $N_4 = 68$.

(a) Déterminer la taille de l'échantillon et la répartition requise afin de donner un estimé de la proportion de la population avec une marge d'erreur sur l'estimation de $B = 0.05$.

(b) Supposons que l'enquête soit menée et qu'elle donne les proportions suivantes: $\hat{p}_1 = 0.87$, $\hat{p}_2 = 0.93$, $\hat{p}_3 = 0.60$ et $\hat{p}_4 = 0.53$. Estimez la proportion dans la population et placer une borne sur l'erreur d'estimation. La limite souhaitée en (a) a-t-elle été atteinte?

(c) Supposons qu'un budget de 475\$ soit disponible, mais que 75\$ doivent être réservés pour les frais généraux fixes. Déterminer la taille de l'échantillon STR et la taille optimale de l'échantillon dans chaque strate en utilisant les informations de l'énoncé du problème comme valeurs plausibles pour les proportions des strates (et non celles de la partie (b)).

Solution: On résume la situation comme suit :

Strate i	N_i	c_i	$\hat{p}_{\mathrm{i}}$
1	97	4	0.87
2	43	4	0.93
3	45	8	0.60
4	68	8	0.53

(a) Nous ne connaissons pas les proportions exactes p_{i}, alors nous utilisons $p_{\mathrm{i}} = 0.5$ afin de déterminer les poids d'échantillonnage pour la répartition optimale:

$$w_i = \frac{n_i}{n} \approx \frac{\dfrac{N_i(0.5)}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}} = 0.0091 \cdot \frac{N_i}{\sqrt{c_i}}, \quad i = 1, 2, 3, 4.$$

Nous obtenons alors

$$w_1 = 0.009 \cdot \frac{97}{2} \approx 0.44,\ w_2 = 0.009 \cdot \frac{43}{2} \approx 0.20,\ w_3 = 0.009 \cdot \frac{45}{\sqrt{8}} \approx 0.14.\ w_4 = 0.009 \cdot \frac{68}{\sqrt{8}} \approx 0.22,$$

d'où

$$n = \frac{\left(\displaystyle\sum_{i=1}^{4} N_i(0.5)\sqrt{c_i}\right)\left(\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}\right)}{N^2 D + \displaystyle\sum_{i=1}^{4} N_i(0.25)}, \quad \text{avec} \quad D = \frac{B^2}{4} = 0.000625.$$

Ainsi, $n = 159.622 \approx 160$. La répartition $n_i = w_i n$, $i = 1, 2, 3, 4$ devient:

$$(n_1, n_2, n_3, n_4) = (70.577, 31.287, 23.152, 34.895) \approx (71, 31, 23, 35).$$

Le coût associé au sondage est donc $4(71 + 31) + 8(23 + 35) = 872\$$ (sans compter les frais généraux).

(b) L'estimateur de la proportion prend la valeur

$$\hat{p}_{\mathrm{st}} = \frac{1}{N} \sum_{i=1}^{4} N_{\mathrm{i}} \hat{p}_{\mathrm{i}} = \frac{1}{253} \left(97(0.87) + 43(0.93) + 45(0.60) + 68(0.53)\right) = 0.741.$$

La marge d'erreur sur l'estimation est ainsi

$$B = 2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{st}}\right)} = \frac{2}{N} \sqrt{\sum_{i=1}^{4} N_{\mathrm{i}}^2 \frac{\hat{p}_{\mathrm{i}} \hat{q}_{\mathrm{i}}}{n_{\mathrm{i}} - 1} \left(1 - \frac{n_{\mathrm{i}}}{N_{\mathrm{i}}}\right)} = 0.045;$$

la marge souhaitée a été atteinte.

(c) Puisque nous ne connaissons pas les proportions exactes p_{i}, nous utilisons $p_{\mathrm{i}} = 0.5$ afin de déterminer la taille de l'échantillon :

$$n = (C - c_0) \frac{\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} N_i(0.5)\sqrt{c_i}} = (C - c_0) \frac{\displaystyle\sum_{i=1}^{4} \frac{N_i}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} N_i \sqrt{c_i}} = 400 \cdot \frac{109.951}{599.612} \approx 73.$$

Les poids d'échantillonnage de la répartition optimale sont toujours valides: on utilise ainsi $n_i = w_i(73)$, $i = 1, 2, 3, 4$:

$$(n_1, n_2, n_3, n_4) \approx (32, 14, 11, 16),$$

et le coût total du sondage est alors $75 + 4(32 + 14) + 8(11 + 16) = 475$. ■

28. Un forestier souhaite donner un estimé du nombre total d'acres agricoles plantés d'arbres dans sa province. Comme la superficie des arbres varie considérablement en fonction de la taille de l'exploitation en question, il décide de procéder à une stratification en fonction de la taille des exploitations. Les 263 fermes de la province sont placées dans l'une des quatre catégories en fonction de leur taille. Un échantillon aléatoire stratifié de 40 exploitations, sélectionné en utilisant la répartition proportionnelle, donne les résultats indiqués dans le tableau ci-dessous.

Strate	N_i	n_i	$\overline{y}_\text{i}$	s_i
<200 acres	96	14	63.36	32.74
200 à <400 acres	82	12	183.0	95.2
400 à <600 acres	55	9	340.6	129.6
600+ acres	30	5	472.0	269.0

(a) Donner un estimé de la superficie totale (en acres) d'arbres dans les exploitations de la province, et donner une marge d'erreur sur l'estimation.

(b) Supposons que l'on souhaite obtenir une marge d'erreur sur l'estimation de 5000 acres. En considérant ce qui précède comme une enquête préliminaire, trouver la taille de l'échantillon nécessaire pour atteindre cette borne si on utilise la répartition de Neyman.

Solution:

(a) L'estimateur du total prend la valeur

$$\tau_\text{st.} = N\overline{y}_\text{st.} = \sum_{i=1}^{4} N_\text{i}\overline{y}_\text{i} = 96(63.36) + 82(183.0) + 55(340.6) + 30(472.0) = 53981.56,$$

et la marge d'erreur sur l'estimation approche

$$\begin{aligned} B &= 2\sqrt{\hat{V}\left(\tau_\text{st.}\right)} = 2\sqrt{\sum_{i=1}^{4} N_i^2 \cdot \frac{s_\text{i}^2}{n_\text{i}}\left(1-\frac{n_\text{i}}{N_\text{i}}\right)} \\ &= 2\sqrt{96^2 \cdot \frac{32.74^2}{14}\left(1-\frac{14}{96}\right) + 82^2 \cdot \frac{95.2^2}{12}\left(1-\frac{12}{82}\right) + 55^2 \cdot \frac{129.6^2}{9}\left(1-\frac{9}{55}\right) + 30^2 \cdot \frac{269^2}{5}\left(1-\frac{5}{30}\right)} \\ &= 9058.391. \end{aligned}$$

(b) Avec la répartion de Neyman, la taille d'échantillon minimale devrait être

$$n = \frac{\left(\sum_{i=1}^{4} N_i\sigma_i\right)^2}{\frac{B^2}{4} + \sum_{i=1}^{4} N_i\sigma_i^2} \approx \frac{\left(\sum_{i=1}^{4} N_i s_i\right)^2}{\frac{B^2}{4} + \sum_{i=1}^{4} N_i s_i^2} = 67.08 \approx 68$$

La répartition de Neyman donne

$$n_i = w_i n \approx \left(\frac{N_i s_i}{\sum N_j s_j}\right) \cdot 68 \approx (8, 20, 19, 21),$$

ce qui donne bien une taille de $n = 68$. ■