TD4–Chapitre 4 - Sondages stratifiées

Exercice 15. (**Les animaux du cirque**) Un directeur de cirque possède 100 éléphants classés en deux catégories : "mâles" et "femelles". Le directeur veut estimer le poids total de son troupeau pour gérer le transport. Il a la possibilité de ne faire peser seulement 10 éléphants de son troupeau, pour une raison de coût et de temps. Cependant, en 1998, ce même directeur a pu faire peser tous les éléphants de son troupeau, et il a obtenu les résultats suivants (en tonnes) :

	Effectif	Moyenne	Variance (corrigée)
Mâles	60	6	4.00
Femelles	40	4	2.25

1. Calculer la variance dans la population de la variable "poids de l'éléphant" en 1998.
2. Si, en 1998, le directeur avait procédé à un sondage aléatoire simple sans remise de 10 éléphants, quelle aurait été la variance de l'estimateur du poids total du troupeau ?
3. Si le directeur avait procédé à un sondage stratifié, avec sondage aléatoire simple dans chaque strate, avec allocation proportionnelle de 10 éléphants, quelle aurait été la variance de l'estimateur du poids total du troupeau ?
4. Si le directeur avait procédé à un sondage stratifié optimal, avec sondage aléatoire simple dans chaque strate, de 10 éléphants, quels auraient été les effectifs de l'échantillon dans les strates, et quelle aurait été la variance de l'estimateur du poids total du troupeau ?

Corrigé :

Nous avons $N = 100, n = 10$. Rappelons les données ci-dessous.

	Effectif	Moyenne	Variance (corrigée)
Mâles	60	6	4.00
Femelles	40	4	2.25

1. Pour calculer la variance de la variable "poids" en 1998 dans la population, on va utiliser la formule de l'analyse de la variance. On a ici deux strates $J = 2$. Cela nous donne

$$\sigma_Y^2 = \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 + \frac{N_1}{N}\left(\bar{Y}_1 - \bar{Y}\right)^2 + \frac{N_2}{N}\left(\bar{Y}_2 - \bar{Y}\right)^2$$

où N_1, N_2 sont les tailles respectives des strates "Mâles", "Femelles", σ_1^2 et σ_2^2 les variances non corrigées sur chaque strate, $\bar{Y}_1, \bar{Y}_2$, et $\bar{Y}$ les moyennes respectives de la variable sur chaque strate et sur la population. On a

$$\bar{Y} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2 = \frac{60}{100}\cdot 6 + \frac{40}{100}\cdot 4 = 3.6 + 1.6 = 5.2$$

De là

$$\sigma_Y^2 = \frac{60}{100}\frac{59\times 4}{60} + \frac{40}{100}\frac{39\times 2.25}{40} + \frac{60}{100}(6-5.2)^2 + \frac{40}{100}(4-5.2)^2 = 4.1975$$

2. Pour un sondage aléatoire simple sans remise de $n = 10$ éléphants, la variance de l'estimateur du poids total du troupeau est :

$$\begin{aligned}
\mathrm{Var}\left(\hat{t}_Y\right) &= 100^2\left(1 - \frac{10}{100}\right)\frac{\tilde{\sigma}_Y^2}{10} \\
&= 100^2(1 - 10/100) \times \frac{1}{10}\frac{100\times\sigma_Y^2}{99} \\
&= 100^2(1 - 10/100) \times \frac{10\times 4.1975}{99} \\
&= 3815.9
\end{aligned}$$

3. Si le directeur procède à un sondage stratifié, avec sondage aléatoire simple dans chaque strate, avec allocation proportionnelle de 10 éléphants, alors le taux de sondage est $f = 10/100 = 0.1$ et $n_1 = 0.1 \times 60 = 6$ et $n_2 = 0.1 \times 40 = 4$. D'après le cours, la variance de l'estimateur $\hat{t}_Y$ du total est

$$\begin{aligned} \operatorname{Var}\left(\hat{t}_Y\right) &= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left(N_1\tilde{\sigma}_1^2 + N_2\tilde{\sigma}_2^2\right) \\ &= 10(1-0.1)(60 \times 4 + 40 \times 2.25) = 2970 \end{aligned}$$

4. Pour un sondage stratifié optimal, on a donc

$$n_j = nN_j\tilde{\sigma}_j / \sum_{\ell=1}^{J} N_\ell\tilde{\sigma}_\ell,$$

d'où

$$n_1 = 10\frac{60 \times 2}{60 \times 2 + 40 \times 1.5} \simeq 7 \quad \text{et} \quad n_2 = 10\frac{40 \times 1.5}{60 \times 2 + 40 \times 1.5} \simeq 3$$

$$\operatorname{Var}\left(\hat{t}_Y\right) = N_1\left(N_1 - n_1\right)\frac{\tilde{\sigma}_1^2}{n_1} + N_2\left(N_2 - n_2\right)\frac{\tilde{\sigma}_2^2}{n_2} = 60(60-7)4/7 + 40(40-3)2.25/3 \simeq 2927.14$$

Exercice 16. (**Contrôle de réception**) Un contrôle de réception de deux fournisseurs est effectué sur le poids de comprimés. L'année dernière un recensement complet a été effectué sur l'ensemble des unités des deux lots reçus. Les données sont rassemblées dans le tableau ci-dessous

	Effectif	Moyenne	Variance (corrigée)
Lot 1	200	60.21	43.85
Lot 2	150	54.14	43.81

1. Calculer la variance de la variable "poids d'un comprimé" sur l'ensemble des unités réceptionnées l'année dernière.
2. Si, l'année dernière, le directeur avait procédé à un sondage aléatoire simple sans remise de 10 unités, quelle aurait été la variance de l'estimateur du poids moyen des comprimés ?
3. Si le directeur avait procédé à un sondage stratifié, avec sondage aléatoire simple dans chaque strate, avec allocation proportionnelle de 10 unités, quelle aurait été la variance de l'estimateur du poids moyen ?
4. Si le directeur avait procédé à un sondage stratifié optimal, avec sondage aléatoire simple dans chaque strate, de 10 unités, quels auraient été les effectifs de l'échantillon dans les strates, et quelle aurait été la variance de l'estimateur du poids moyen du troupeau ?
5. Établir les intervalles de confiance (IC) de la moyenne dans les cas précédents.

Corrigé :

1. Pour calculer la variance dans la population, on utilise la formule de l'analyse de la variance. On a ici deux strates $J = 2$. Cela nous donne

$$\sigma_Y^2 = \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 + \frac{N_1}{N}\left(\bar{Y}_1 - \bar{Y}\right)^2 + \frac{N_2}{N}\left(\bar{Y}_2 - \bar{Y}\right)^2$$

où $N_1 = 200, N_2 = 150$ sont les tailles respectives des strates "Lot1", "Lot2", $\sigma_1^2 = 199\tilde{\sigma}_1/200$ et $\sigma_2^2 = 149\tilde{\sigma}_2/150$ les variances non corrigées sur chaque strate, $\bar{Y}_1, \bar{Y}_2$, et $\bar{Y}$ les moyennes respectives de la variable sur chaque strate et sur la population. On a

$$\bar{Y} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2 = \frac{200}{350} \cdot 60.21 + \frac{150}{350} \cdot 54.14 = 57.60$$

De là

$$\sigma_Y^2 = \frac{200}{350}\frac{199 \times 43.85}{200} + \frac{150}{350}\frac{149 \times 43.81}{150} + \frac{200}{350}(60.21 - 57.60)^2 + \frac{150}{350}(54.14 - 57.60)^2 = 52.48$$

2. Pour un sondage aléatoire simple sans remise de $n = 10$ unités, la variance de l'estimateur du total de Y est

$$\begin{aligned}\operatorname{Var}\left(\hat{t}_Y\right) &= 350^2\left(1 - \frac{10}{350}\right)\frac{\tilde{\sigma}_Y^2}{10} \\ &= 350^2(1 - 10/350) \times \frac{1}{10}\frac{350 \times \sigma_Y^2}{349} \\ &= 350^2(1 - 10/350) \times \frac{10 \times 52.48}{349} \\ &= 6939025\end{aligned}$$

3. Si l'on procède à un sondage stratifié, avec sondage aléatoire simple dans chaque strate, avec allocation proportionnelle de 10 unités, alors le taux de sondage est $f = 10/350 \simeq 0.0286$ et $n_1 = 200 \times 10/350 = 5.71 \simeq 6$ et $n_2 = 150 \times 10/350 = 4.29 \simeq 4$. D'après le cours, la variance de l'estimateur $\hat{t}_Y$ du total est

$$\begin{aligned}\operatorname{Var}\left(\hat{t}_Y\right) &= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left(N_1\tilde{\sigma}_1^2 + N_2\tilde{\sigma}_2^2\right) \\ &= 523094\end{aligned}$$

4. Pour un sondage stratifié optimal, on a donc

$$n_j = nN_j\tilde{\sigma}_j / \sum_{\ell=1}^{J} N_\ell \tilde{\sigma}_\ell,$$

d'où

$$n_1 = 10\frac{200 \times 43.85}{200 \times 43.85 + 150 \times 43.81} \simeq 6 \quad \text{et} \quad n_2 = 10\frac{150 \times 43.81}{200 \times 43.85 + 150 \times 43.81} \simeq 4$$

$$\operatorname{Var}\left(\hat{t}_Y\right) = N_1\left(N_1 - n_1\right)\frac{\tilde{\sigma}_1^2}{n_1} + N_2\left(N_2 - n_2\right)\frac{\tilde{\sigma}_2^2}{n_2} = 200(200 - 6)43.85/6 + 150(150 - 4)43.81/4 \simeq 523094.$$