Master 2

Théorie des sondages

Exercice 2

Voici les valeurs y_α, le total de la grappe pour l'indice de masse corporelle (IMC) et z_α le nombre de personnes à qui un médecin a déjà dit qu'elles avaient un taux de cholestérol élevé, à partir d'un échantillon composé de a = 10 grappes de taille égale de b = 10 adultes chacune. Supposons que les grappes ont été sélectionnées au hasard et avec remise, et que les adultes ont été sélectionnés avec probabilité égale et sans remise. La fraction d'échantillonnage est f = ab/AB = n/N = 100/3 048

α	y_α	z_α
1	302.91	4
2	305.63	2
3	278.72	5
4	275.19	0
5	276.21	3
6	340.56	6
7	345.15	2
8	270.21	4
9	329.67	5
10	295.41	7
Total	3,019.66	38

Maintenant, répondez aux questions suivantes pour ce sondage en grappe:

a) Calculez $\bar{y}$, l'indice de masse corporelle (IMC) moyen, son erreur standard et son intervalle de confiance à 95 %.

La moyenne est calculée comme suit :

$$\bar{y} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{B} y_{\alpha\beta}}{ab} = \frac{\sum_{\alpha=1}^{a} y_\alpha}{ab} = \frac{302.91+305.63+\ldots+295.41}{10\times 10} = \frac{3,019.66}{100} = 30.197$$

et sa variance d'échantillonnage est donnée par

$$\operatorname{var}(\bar{y}) = (1-f)\frac{s_a^2}{a}$$

Pour s_a^2, nous savons que

$$s_a^2 = \frac{\left[\sum_{\alpha=1}^{a} y_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{a} y_\alpha\right)^2}{a}\right]}{b^2(a-1)} = \frac{\left[918,942.96 - \frac{(3,019.66)^2}{10}\right]}{10^2(10-1)} = \frac{7,108.3}{900} = 7.898$$

Ainsi

$$\text{var}(\bar{y}) = (1-f)\frac{s_a^2}{a} = \left(1 - \frac{100}{3,048}\right)\frac{7.898}{100} = 0.764$$

et l'erreur type de la moyenne de l'échantillon y est donnée par

$$se(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{0.764} = 0.874$$

Ensuite, l'intervalle de confiance à 95 % pour l'IMC est donné par

$$\bar{y} \pm t_{\left(a-1;1-\frac{\alpha}{2}\right)} \times se(\bar{y}) = 30.197 \pm t_{\left(10-1;1-\frac{0.05}{2}\right)} \times 0.874 = 30.197 \pm 2.262 \times 0.764 = 30.197 \pm 1.977 =$$
$$= (28.219; 32.174)$$

b) Estimez l'erreur type de l'IMC moyen auquel vous vous attendriez si l'échantillon était composé de a = 5 grappes avec b=10 chacune

Nous avons déjà une estimation non biaisée de s_a^2 . Nous pouvons l'utiliser pour recalculer la variance sous le nouveau plan :

$$\text{var}(\bar{y}) = (1-f)\frac{s_a^2}{a} = \left(1 - \frac{50}{3,048}\right)\frac{7.898}{5} = 1.554$$

Par conséquent, l'erreur standard est

$$\sqrt{\text{var}(\bar{y})} = \sqrt{1.554} = 1.246$$

c) Maintenant, estimez $\bar{z}$, la proportion de personnes à qui un médecin a dit qu'elles avaient un taux de cholestérol élevé, à partir de l'échantillon du tableau ci-dessus. Calculez son erreur standard, son effet de grappe deff et roh.

$$\bar{z} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{B} z_{\alpha\beta}}{ab} = \frac{\sum_{\alpha=1}^{a} z_\alpha}{ab} = \frac{4+2+\ldots+7}{10\times 10} = \frac{38}{100} = 0.38$$

On peut estimer s_a^2 comme :

$$s_a^2 = \frac{\left[\sum_{\alpha=1}^{a} z_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{a} z_\alpha\right)^2}{a}\right]}{b^2(a-1)} = \frac{\left[184 - \frac{(38)^2}{10}\right]}{10^2(10-1)} = 0.0440$$

Compte tenu de ce résultat, la variance et l'erreur standard sont :

$$\text{var}(\bar{z}) = (1-f)\frac{s_a^2}{a} = \left(1 - \frac{100}{3,048}\right)\frac{0.044}{10} = 0.0043$$

$$\sqrt{\text{var}(\bar{z})} = \sqrt{0.0043} = 0.0652\,.$$

Afin d'obtenir deff et roh, nous avons besoin d'estimations de la variance sous SRS. Nous pouvons obtenir cela avec une proportion puisque nous avons une méthode pour estimer s^2 . Voici l'équation :

$$\text{var}_{SRS}(\bar{z}) = (1-f)\frac{s^2}{n} = (1-f)\frac{p(1-p)}{n-1} = \left(1 - \frac{100}{3,048}\right)\frac{0.38(0.62)}{99} = 0.0023$$

L'étape suivante consiste à estimer l'effet de conception :

$$deff = \frac{\text{var}(\bar{z})}{\text{var}_{SRS}(\bar{z})} = \frac{0.0043}{0.0023} = 1.849$$

Enfin, en utilisant deff, obtenez une estimation de roh :

$$roh = \frac{deff - 1}{b-1} = \frac{0.849}{9} = 0.094$$

d) Estimez l'erreur type de la proportion de personnes à qui un médecin a dit qu'elles avaient un taux de cholestérol élevé à partir d'un échantillon de a = 20 grappes de b = 5 personnes chacune.

La taille globale de l'échantillon est toujours n=axb=20x5=200 et nous avons donc la même variance d'échantillonnage aléatoire simple qu'au point (c), mais l'effet de plan changera lorsque la taille du sous-échantillon passera de b= 10 à b= 5 :

$$deff = 1 + roh(b-1) = 1 + 0.094(5-1) = 1.377$$

Par conséquent, la variance d'échantillonnage estimée sous ce nouveau plan est

$$\text{var}_{b=5}(p) = deff \times \text{var}_{SRS}(p) = 1.377 \times 0.0023 = 0.0032$$

et l'erreur type estimée est

$$se_{b=2}(\bar{z}) = \sqrt{\text{var}_{b=5}(\bar{z})} = \sqrt{0.0032} = 0.0563$$

e) Calculez la taille de grappe optimale si le coût par grappe est c_a = 80000 FCFA et le coût par observation au sein d'une grappe est c_b = 40000 FCFA pour la proportion de personnes à qui un médecin a dit qu'elles avaient un taux de cholestérol élevé.

La taille optimale des grappes est donnée par :

$$b_{opt} = \sqrt{\frac{c_a}{c_b}\frac{1-roh}{roh}} = \sqrt{\frac{800}{400}\frac{1-0.094}{0.0094}} = 4.382 \approx 4$$

f) Quelle serait l'erreur type de la proportion de personnes à qui un médecin a dit qu'elles avaient un taux de cholestérol élevé en utilisant la taille de grappe optimale pour cette même estimation sous la structure des coûts en (e), si le budget total ($C - c_0$) pour une enquête est de 1 5000 000?

Si le budget total est C-C_0= 150 000, alors en utilisant le modèle de coût pour l'échantillonnage à deux degrés, C-C_0 =ac_a+ a(bc_a) et la taille de grappe optimale, il est possible de se permettre

$$a = \frac{C - C_0}{c_a + b_{opt}c_b} = \frac{150,000}{800 + 4\times 400} \approx 62.50 = 63$$

Nous pouvons estimer la variance d'échantillonnage de la proportion de personnes à qui un médecin a dit qu'elles avaient un taux de cholestérol élevé dans ce nouveau plan en utilisant

$$\text{var}_{b=b_{opt}}\left(p\right) = deff \times \text{var}_{SRS}\left(p\right)$$

L'effet de plan de ce nouveau plan d'échantillonnage utilisant la taille de grappe optimale est

$$deff = 1 + roh\left(b_{opt} - 1\right) = 1 + 0.094\left(4 - 1\right) = 1.283$$

La taille globale de l'échantillon est n=axb_{opt}=63x4=252. Par conséquent, la variance d'échantillonnage dsu SAS est donnée par :

$$\text{var}_{SRS}\left(p\right) = \left(1-f\right)\frac{s^2}{n} = \left(1-f\right)\frac{p\left(1-p\right)}{n-1} = \left(1 - \frac{252}{3,048}\right)\frac{0.38\left(1-0.38\right)}{252-1} = 0.000861$$

Par conséquent, la variance d'échantillonnage estimée sous ce nouveau plan est

$$\text{var}_{b=b_{opt}}\left(p\right) = deff \times \text{var}_{SRS}\left(p\right) = 1.283 \times 0.000861 = 0.001105$$

et l'erreur type estimée est

$$se_{b=b_{opt}}\left(\bar{z}\right) = \sqrt{\text{var}_{b=b_{opt}}\left(\bar{z}\right)} = \sqrt{0.001105} = 0.03324$$