Master 2

Théorie des sondages

Exercice 3

1. Les données suivantes proviennent d'un échantillon epsem à deux degrés de a = 10 grappes d'une population de N = 3 048 adultes dans A = 78 grappes de taille inégale, où y_α, est le total de la grappe pour l'indice de masse corporelle (IMC), et z_α est le nombre de personnes à qui un médecin a déjà dit qu'elles avaient un taux de cholestérol élevé parmi les adultes x_α sélectionnés dans chaque groupe. La conception est epsem car le taux de sondage au premier degré a été fixe $\frac{a}{A} = \frac{10}{68}$ et le taux au deuxième degré est de $\frac{b}{B} = \frac{10}{45}$. Par conséquent, la taille de l'échantillon est une variable aléatoire avec une taille d'échantillon attendue de 99,61.

α	y_α	z_α	x_α
1	340.25	2	12
2	192.5	3	7
3	100.15	3	4
4	281.43	2	10
5	316.85	3	12
6	344.36	1	13
7	281.78	3	10
8	304.86	5	10
9	243.26	3	8
10	337.86	3	11
Total	2,743.30	28	97

Sur la base de cet échantillon, répondez aux questions suivantes :

a) Calculer l'indice de masse corporelle (IMC) moyen, son erreur standard et son intervalle de confiance à 95 %

Comme les tailles des grappes sont inégales, l'IMC moyen est estimé comme un rapport moyen

$$\bar{y} = r = \frac{\sum_{\alpha=1}^{a} y_\alpha}{\sum_{\alpha=1}^{a} x_\alpha} = \frac{2,743.30}{97} = 28.281$$

En ignorant la correction pour population finie, la variance d'échantillonnage de cette moyenne est estimée comme

$$\operatorname{var}(\bar{y}) \approx \frac{1}{x^2}\left[\operatorname{var}(y) + r^2 \operatorname{var}(x) - 2r \times \operatorname{cov}(y,x)\right]$$

Où

$$\text{var}(y)=\frac{a}{a-1}\left[\sum_{\alpha=1}^{a}y_\alpha^2-\frac{\left(\sum_{\alpha=1}^{a}y_\alpha\right)^2}{a}\right]=\frac{10}{10-1}\left[806,701.31-\frac{2,743.30^2}{10}\right]=60,146.465$$

$$\text{var}(x)=\frac{a}{a-1}\left[\sum_{\alpha=1}^{a}x_\alpha^2-\frac{\left(\sum_{\alpha=1}^{a}x_\alpha\right)^2}{a}\right]=\frac{10}{10-1}\left[1,007-\frac{97^2}{10}\right]=73.444$$

$$\text{cov}(y,x)=\frac{a}{a-1}\left[\sum_{\alpha=1}^{a}y_\alpha x_\alpha-\frac{\left(\sum_{\alpha=1}^{a}y_\alpha\right)\left(\sum_{\alpha=1}^{a}x_\alpha\right)}{a}\right]=\frac{10}{10-1}\left[28,453.22-\frac{2,743.30\times 97}{10}\right]=2,048.011$$

Par conséquent, nous avons

$$\text{var}(\bar{y})\approx\frac{1}{97^2}\left[60,146.465+28.281^2\times 73.444-2\times 28.281\times 2,048.011\right]=0.324$$

Et $se(\bar{y})=\sqrt{0.324}=0.569$

L'intervalle de confiance à 95 % pour cette moyenne est donné par

$$r\pm t_{\left(a-1;1-\frac{\alpha}{2}\right)}\times se(r)=28.281\pm t_{\left(10-1;1-\frac{0.05}{2}\right)}\times 0.569=28.281\pm 2.26\times 0.569=28.281\pm 1.288=(26.944;29.569)$$

b) Estimez la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé et son erreur type.

La proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé est estimée sous la forme d'un rapport moyen

$$p=r=\frac{\sum_{\alpha=1}^{a}z_\alpha}{\sum_{\alpha=1}^{a}x_\alpha}=\frac{28}{97}=0.289$$

En ignorant la correction pour population finie, la variance d'échantillonnage de cette proportion est estimée comme

$$\text{var}(p)\approx\frac{1}{x^2}\left[\text{var}(z)+r^2\,\text{var}(x)-2r\times\text{cov}(z,x)\right]$$

Ou

$$\text{var}(z) = \frac{a}{a-1}\left[\sum_{\alpha=1}^{a} z_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{a} z_\alpha\right)^2}{a}\right] = \frac{10}{10-1}\left[88 - \frac{97^2}{10}\right] = 10.667$$

$$\text{var}(x) = \frac{a}{a-1}\left[\sum_{\alpha=1}^{a} x_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{a} x_\alpha\right)^2}{a}\right] = \frac{10}{10-1}\left[1{,}007 - \frac{97^2}{10}\right] = 73.444$$

$$\text{cov}(z,x) = \frac{a}{a-1}\left[\sum_{\alpha=1}^{a} z_\alpha x_\alpha - \frac{\left(\sum_{\alpha=1}^{a} z_\alpha\right)\left(\sum_{\alpha=1}^{a} x_\alpha\right)}{a}\right] = \frac{10}{10-1}\left[263 - \frac{28\times 97}{10}\right] = -9.556$$

Par conséquent, nous avons

$$\text{var}(p) \approx \frac{1}{97^2}\left[10.889 + 0.289^2 \times 73.444 - 2\times 0.289\times(-9.556)\right] = 0.002370$$

Et $se(p) = \sqrt{0.002370} = 0.0487$

c) L'approximation de la série de Taylor est-elle adéquate pour les erreurs standard calculées en (a) et (b) ?

Le coefficient de variation du dénominateur est

$$cv(x) = \frac{se(x)}{x} = \frac{\sqrt{\text{var}(x)}}{\sum_{\alpha=1}^{a} x_\alpha} = \frac{\sqrt{73.444}}{97} = 0.088$$

Étant donné que cv(x) est inférieur à 0,15, l'approximation en série de Taylor de l'estimation de la variance d'échantillonnage est adéquate.

d) Calculez l'effet de conception et roh pour la proportion en (b).

La variance d'échantillonnage du SAS de cette proportion est donnée par

$$\text{var}_{SRS}(p) = \frac{p(1-p)}{n-1} = \frac{r(1-r)}{x-1} = \frac{0.289(1-0.289)}{97-1} = 0.002139$$

Ainsi, en utilisant la variance d'échantillonnage calculée en (a),

$$deff(p) = \frac{\text{var}(p)}{\text{var}_{SRS}(p)} = \frac{0.002370}{0.002139} = 1.1082$$

$$roh = \frac{deff(p)-1}{\bar{b}-1} = \frac{deff(p)-1}{\frac{x}{a}-1} = \frac{1.1082-1}{\frac{97}{10}-1} = 0.0124$$

2. Le tableau suivant est un résumé des résultats d'un échantillon aléatoire stratifié à allocation égale de taille n = 48 parmi les N = 3 048 adultes (N_h est la taille de la population de la strate, $\bar{y}_h$ est la moyenne de l'échantillon de la strate de l'indice de masse corporelle (IMC), s_h^2 est la variance des éléments de l'échantillon de la strate pour l'IMC et $\bar{z}_h$ est la proportion de l'échantillon de la strate d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé) :

Stratum	N_h	$\bar{y}_h$	s_h^2	$\bar{z}_h$
1	467	25.53	28.663	0.125
2	388	25.07	34.779	0.250
3	479	28.23	21.477	0.375
4	446	28.65	23.348	0.250
5	570	30.04	67.386	0.250
6	698	33.76	36.500	0.375

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

L'IMC moyen est

$$\bar{y}_w = \sum_{h=1}^{H} W_h \bar{y}_h = \frac{467}{3,048} \times 25.53 + \frac{388}{3,048} \times 25.07 + \frac{479}{3,048} \times 28.23 + \frac{446}{3,048} \times 28.65 + \frac{570}{3,048} \times 30.04 + \frac{698}{3,048} \times 33.76 =$$
$$= 29.080$$

La variance d'échantillonnage de cette moyenne est

$$\text{var}(\bar{y}_w) = \sum_{h=1}^{H} W_h^2 \text{var}(\bar{y}_h) = \sum_{h=1}^{H} W_h^2 (1-f_h) \frac{s_h^2}{n_h} = \left(\frac{467}{3,048}\right)^2 \times \left(1-\frac{8}{467}\right)\frac{28.663}{8} + \ldots + \left(\frac{698}{3,048}\right)^2 \times \left(1-\frac{8}{698}\right)\frac{36.500}{8} =$$
$$= 0.805$$

Ensuite, l'erreur standard de l'IMC moyen est

$$se(\bar{y}_w) = \sqrt{\text{var}(\bar{y}_w)} = \sqrt{0.805} = 0.897$$

L'intervalle de confiance à 95 % pour l'IMC moyen est donné par

$$\bar{y}_w \pm t_{\left(1-\frac{\alpha}{2};n-H\right)} se(\bar{y}_w) = \bar{y}_w \pm t_{(0.025;40-5)} se(\bar{y}_w) = 29.080 \pm 2.03 \times 0.897 = 29.080 \pm 1.822 = [27.259; 30.902]$$

b) Estimez la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé ($\bar{z}$) et son erreur type.

La proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé :

$$p_w = \sum_{h=1}^{H} W_h p_h = \frac{467}{3,048} \times 0.125 + \frac{388}{3,048} \times 0.250 + \frac{479}{3,048} \times 0.375 + \frac{446}{3,048} \times 0.250 + \frac{570}{3,048} \times 0.250 + \frac{698}{3,048} \times 0.375 =$$
$$= 0.2791$$

La variance d'échantillonnage de cette proportion est

$$\text{var}(p_w) = \sum_{h=1}^{H} W_h^2 \, \text{var}(\bar{y}_h) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{p_h (1 - p_h)}{n_h - 1} =$$

$$\left(\frac{467}{3,048}\right)^2 \times \left(1 - \frac{8}{467}\right) \frac{0.125(1-0.125)}{8-1} + \ldots + \left(\frac{698}{3,048}\right)^2 \times \left(1 - \frac{8}{698}\right) \frac{0.375(1-0.375)}{8-1} = 0.0048$$

Ensuite, l'erreur type de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé est :

$$se(\bar{y}_w) = \sqrt{\text{var}(\bar{y}_w)} = \sqrt{0.0048} = 0.0694$$

c) Calculez les effets de grappe pour la variance de l'IMC moyen (CONSEIL : vous aurez besoin d'une estimation SAS de s^2) et la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé. Expliquez en une phrase ou deux pourquoi ces effets de grappe estimés sont différents.

Utilisant le fait que

$$S^2 \approx \sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

on trouve une estimation pour S^2 de l'IMC :

$$s^2 \approx \sum_{h=1}^{H} W_h s_h^2 + \sum_{h=1}^{H} W_h \left(\bar{y}_h - \bar{y}_w\right)^2 = 36.571 + 9.306 = 45.877$$

Par conséquent, la variance d'échantillonnage pour l'IMC moyen sous SAS est

$$\text{var}_{SRS}(\bar{y}) = \frac{1-f}{n} s^2 = \frac{1 - \frac{48}{3,048}}{48} 45.877 = 0.941$$

Par conséquent, l'effet de conception pour l'IMC moyen pour cet échantillon stratifié est

$$deff(\bar{y}_w) = \frac{\text{var}(\bar{y}_w)}{\text{var}_{SRS}(\bar{y})} = \frac{0.805}{0.941} = 0.856$$

Une estimation pour le S^2 de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé peut être estimée comme suit :

$$s^2 \approx \sum_{h=1}^{H} W_h \frac{n_n}{n_n - 1} p_h (1 - p_h) + \sum_{h=1}^{H} W_h \left(p_h - p_w\right)^2 = 0.2213 + 0.0076 = 0.2289$$

Par conséquent, la variance d'échantillonnage pour la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé sous SRS est soit

$$\text{var}_{SRS}(p) = \frac{1-f}{n} s^2 = \frac{1-\frac{48}{3,048}}{48} 0.2289 = 0.0047$$

Par conséquent, l'effet de conception pour la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé pour cet échantillon stratifié est soit :

$$deff(p_w) = \frac{\text{var}(p_w)}{\text{var}_{SRS}(p)} = \frac{0.0048}{0.0047} = 1.0273$$

d) Quelle est la répartition proportionnelle de n = 48 entre les strates, l'erreur type de l'IMC moyen sous cette répartition et l'effet de conception de la variance de la moyenne de l'échantillon réparti proportionnellement ? Qu'en est-il de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé ? L'allocation est-elle différente pour chaque variable ? Expliquer.

L'allocation proportionnelle est donnée par n_h=nXW_h. Parconséquent, nous avons que l'allocation proportionnelle dans ce cas est donnée par

$$n_1 = W_1 \times n = \frac{467}{3,048} \times 48 = 7.35 \approx 7$$

$$n_2 = W_2 \times n = \frac{388}{3,048} \times 48 = 6.11 \approx 6$$

$$n_3 = W_3 \times n = \frac{479}{3,048} \times 48 = 7.54 \approx 8$$

$$n_4 = W_4 \times n = \frac{469}{3,048} \times 48 = 7.02 \approx 7$$

$$n_5 = W_5 \times n = \frac{570}{3,048} \times 48 = 8.98 \approx 7$$

$$n_6 = W_6 \times n = \frac{698}{3,048} \times 48 = 10.99 \approx 11$$

Cette répartition est la même pour les deux variables, car la répartition proportionnelle ne nécessite aucune information sur les variables d'enquête, uniquement sur les tailles relatives des strates de la population.

Dans le cadre de cette répartition proportionnelle, la variance d'échantillonnage de l'IMC moyen est réduite à

$$\text{var}\left(\bar{y}_{prop}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{H}W_h s_h^2=$$

$$=\left(\frac{1-\frac{48}{3,048}}{48}\right)\left(\frac{467}{3,048}\times 28.663+\frac{388}{3,048}\times 34.779+\frac{479}{3,048}\times 21.477+\frac{446}{3,048}\times 23.348+\frac{570}{3,048}\times 67.386+\frac{698}{3,048}\times 36.500\right)=$$

$$=0.750$$

Ensuite, l'erreur standard de l'IMC moyen sous cette allocation est

$$se\left(\bar{y}_{prop}\right)=\sqrt{\text{var}\left(\bar{y}_{prop}\right)}=\sqrt{0.750}=0.866$$

En utilisant la variance d'échantillonnage SRS de l'IMC moyen calculé en (c), l'effet de conception pour l'IMC moyen pour l'échantillon stratifié sous répartition proportionnelle est

$$deff\left(\bar{y}_{prop}\right)=\frac{\text{var}\left(\bar{y}_{prop}\right)}{\text{var}_{SRS}\left(\bar{y}\right)}=\frac{0.750}{0.941}=0.797$$

La variance d'échantillonnage de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé dans le cadre d'une allocation proportionnelle est

$$\text{var}\left(p_{prop}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{H}W_h s_h^2=\left(\frac{1-f}{n}\right)\sum_{h=1}^{H}W_h\frac{n_h}{n_h-1}p_h\left(1-p_h\right)=$$

$$\left(\frac{1-\frac{48}{3,048}}{48}\right)\left(\begin{array}{l}\frac{467}{3,048}\times\frac{8}{7}\times 0.125\left(1-0.125\right)+\frac{388}{3,048}\times\frac{8}{7}\times 0.250\left(1-0.250\right)+\\ \frac{479}{3,048}\times\frac{8}{7}\times 0.375\left(1-0.375\right)+\frac{446}{3,048}\times\frac{8}{7}\times 0.250\left(1-0.250\right)+\\ \frac{570}{3,048}\times\frac{8}{7}\times 0.250\left(1-0.250\right)+\frac{698}{3,048}\times\frac{8}{7}\times 0.375\left(1-0.375\right)\end{array}\right)=0.0045$$

Ensuite, l'erreur type de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé dans cette allocation est

$$se\left(p_{prop}\right)=\sqrt{\text{var}\left(p_{prop}\right)}=\sqrt{0.0045}=0.0674$$

En utilisant la variance d'échantillonnage SRS de la proportion d'adultes à qui un médecin a déjà dit qu'ils avaient un taux de cholestérol élevé calculée en (c), l'effet de plan pour cette estimation pour l'échantillon stratifié sous répartition proportionnelle est

$$deff\left(p_{prop}\right)=\frac{\text{var}\left(p_{prop}\right)}{\text{var}_{SRS}\left(\bar{y}\right)}=\frac{0.0045}{0.0047}=0.9669$$

e) Sur la base des informations du tableau, quelle est l'allocation de Neyman de n = 48 dans les strates, l'erreur type de l'IMC moyen sous cette allocation et l'effet de conception de la variance de la moyenne de l'allocation de Neyman ?

L'allocation de Neyman est donnée par

$$n_h = n\frac{W_h s_h}{\sum_{h=1}^{H} W_h s_h}$$

Par conséquent, nous avons que l'allocation de Neyman dans ce cas pour l'IMC moyen est

$$n_1 = n\frac{W_1 s_1}{\sum_{h=1}^{H} W_h s_h} = 6.65 \approx 7$$

$$n_2 = n\frac{W_2 s_2}{\sum_{h=1}^{H} W_h s_h} = 6.08 \approx 6$$

$$n_3 = n\frac{W_3 s_3}{\sum_{h=1}^{H} W_h s_h} = 5.90 \approx 6$$

$$n_4 = n\frac{W_4 s_4}{\sum_{h=1}^{H} W_h s_h} = 5.73 \approx 6$$

$$n_5 = n\frac{W_5 s_5}{\sum_{h=1}^{H} W_h s_h} = 12.44 \approx 12$$

$$n_6 = n\frac{W_6 s_6}{\sum_{h=1}^{H} W_h s_h} = 11.21 \approx 11$$

Sous cette allocation de Neyman, la variance d'échantillonnage de l'IMC moyen est

$$\text{var}\left(\bar{y}_N\right) = \sum_{h=1}^{H} W_h^2 \text{ var}\left(\bar{y}_N\right) = \sum_{h=1}^{H} W_h^2 \left(1 - f_h\right)\frac{s_h^2}{n_h} = 0.720$$

Ensuite, l'erreur standard de l'IMC moyen sous cette allocation est

$$se\left(\bar{y}_N\right) = \sqrt{\text{var}\left(\bar{y}_N\right)} = \sqrt{0.720} = 0.849$$

L'effet de conception pour l'IMC moyen pour l'échantillon stratifié sous allocation de Neyman est

$$deff\left(\bar{y}_N\right) = \frac{\text{var}\left(\bar{y}_N\right)}{\text{var}_{SRS}\left(\bar{y}\right)} = \frac{0.720}{0.941} = 0.766$$

f) En utilisant la notation de Kish pour l'élément de coût dans chaque strate, supposons qu'ils étaient J1 = J2 = J3 = J4 = 100 FCFA et J5 = J6 = 1000 FCFA. Supposons aussi que C – C0 = 30000. Calculez la répartition qui minimise la variance de l'IMC moyen pour ce coût total. Ne dépassez pas le budget de 30000. Estimez l'erreur type et l'effet de grappe de la variance de la moyenne sous cette allocation.

L'allocation qui minimise la variance d'une moyenne pour un coût total donné est l'allocation optimale, donnée par $n_h = k W_h s_h \Big/ \sqrt{J_h}$

Par conséquent, pour obtenir l'allocation optimale pour l'IMC moyen, nous calculons d'abord

$$n_1^* = W_1 s_1 \Big/ \sqrt{J_1} = 0.820$$
$$n_2^* = W_2 s_2 \Big/ \sqrt{J_2} = 0.751$$
$$n_3^* = W_3 s_3 \Big/ \sqrt{J_3} = 0.728$$
$$n_4^* = W_4 s_4 \Big/ \sqrt{J_4} = 0.707$$
$$n_5^* = W_5 s_5 \Big/ \sqrt{J_5} = 0.485$$
$$n_6^* = W_6 s_6 \Big/ \sqrt{J_6} = 0.438$$

Le coût de cette répartition de l'échantillon est

$$J^* = \sum_{h=1}^{H} n_h^* J_h = 12.2359$$

Cependant, nous avons que C-C_0=3000 . Par conséquent, nous pouvons ajuster cette allocation pour ce coût total en utilisant la constante k=$\frac{C-C_0}{J^*}$=$\frac{3000}{122,359}$

Par conséquent, nous avons que l'allocation optimale dans ce cas est

$$n_1 = k \times n_1^* = 24.518 \times 0.820 = 20.112 \approx 20$$
$$n_2 = k \times n_2^* = 24.518 \times 0.751 = 18.406 \approx 18$$
$$n_3 = k \times n_3^* = 24.518 \times 0.728 = 17.856 \approx 18$$
$$n_4 = k \times n_4^* = 24.518 \times 0.707 = 17.335 \approx 17$$
$$n_5 = k \times n_5^* = 24.518 \times 0.485 = 11.902 \approx 12$$
$$n_6 = k \times n_6^* = 24.518 \times 0.438 = 10.727 \approx 11$$

Observez que la taille globale de l'échantillon dans le cadre de cette allocation est

$$n = \sum_{h=1}^{H} n_h = 96$$

et le coût total est de 3030 FCFA Comme cela dépasse le budget total, nous devons ajuster la taille des échantillons pour tenir compte de cela. Ici, nous pouvons supprimer 1 cas des strates 1, 2 et 3, par exemple.

Par conséquent, nous utilisons

n_1=19, n_2=17, n_3=17, n_4=17 , n_5=12, n_6=11 avec

$$n = \sum_{h=1}^{H} n_h = 93$$

Sous la répartition optimale de ce coût total, la variance d'échantillonnage attendue est

$$\text{var}\left(\bar{y}_{opt}\right)=\sum_{h=1}^{H}W_h^2\ \text{var}\left(\bar{y}_{opt}\right)=\sum_{h=1}^{H}W_h^2\left(1-f_h\right)\frac{s_h^2}{n_h}=0.488$$

Ensuite, l'erreur standard de l'IMC moyen sous cette allocation est

$$se\left(\bar{y}_{opt}\right)=\sqrt{\text{var}\left(\bar{y}_{opt}\right)}=\sqrt{0.488}=0.698$$

En utilisant l'estimation pour S^2 calculée en (c), la variance d'échantillonnage pour l'IMC moyen sous SAS de taille n=93

$$\text{var}_{SRS}\left(\bar{y}\right)=\frac{1-f}{n}s^2=\frac{1-\frac{93}{3,048}}{93}45.877=0.478$$

Par conséquent, l'effet de conception pour l'IMC moyen pour l'échantillon stratifié sous allocation optimale pour un coût total de
C-C_0=3000 est

$$deff\left(\bar{y}_{opt}\right)=\frac{\text{var}\left(\bar{y}_{opt}\right)}{\text{var}_{SRS}\left(\bar{y}\right)}=\frac{0.488}{0.478}=1.019$$