Master 2

Théorie des sondages

Exercice Solution

1. (a) L'échantillon se compose des 40 éléments suivants :

CASEID	CHLEV	BMI
157	0	26.54
177	0	26.62
310	1	31.62
569	1	34.69
575	1	24.81
745	0	26.32
803	0	21.63
857	0	23
908	0	20.37
922	0	32.37
937	0	22.61
950	1	31.97
986	0	30.26
998	0	25.08
1090	0	27.45
1184	0	26.96
1225	1	29.16
1233	0	22.24
1246	1	25.85
1479	1	28.9
1531	1	36.35
1543	0	18.78
1691	0	31.95
1790	0	38.94
1835	0	30.47
1886	1	43.55
1911	1	32.29
1987	0	32.47
2222	0	42.89
2314	0	21.77
2337	0	25.12
2340	0	22.33
2363	0	22.75
2535	0	23.3
2672	0	29.52
2812	0	28.13
2852	0	23.3
2884	1	21.79
2932	0	20.81
3047	0	21.41

(b) L'IMC moyen de l'échantillon est :

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{29.16 + 32.29 + \cdots + 26.96}{40} = 27.659$$

L'IMC moyen de la population est

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{N} = \frac{18.83 + 19.74 + \cdots + 31.41}{3{,}048} = \frac{85{,}838.79}{3{,}048} = 28.162$$

L'estimation de l'échantillon de l'IMC moyen n'est pas exactement égale à la moyenne de la population. On s'attend à ce que la moyenne d'un échantillon donné diffère de la vraie moyenne de la population en raison de l'erreur inhérente au fait de ne pas observer toutes les unités de la population, connue sous le nom d'erreur d'échantillonnage.

(c) La variance élémentaire de l'IMC est

$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{(\sum_{i=1}^{n} y_i^2 - \frac{y^2}{n})}{n-1} = \frac{1397{,}96}{40-1} = 35{,}85$

La variance d'échantillonnage de l'IMC moyen est

$\text{var}(\bar{y}) = (1\text{-f})\frac{s_y^2}{n} = (1 - \frac{40}{3048})\frac{35{,}85}{40} = 0{,}8844$

et l'erreur standard de l'IMC moyen est

$\text{se}(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{0{,}8844} = 0{,}9404$

(d) L'intervalle de confiance à 95 % pour l'IMC est donné par

$$\bar{y} \pm t_{\left(n-1;1-\frac{\alpha}{2}\right)} \times se(\bar{y}) = 27.659 \pm t_{\left(40-1;1-\frac{0.05}{2}\right)} \times se(\bar{y}) = 28.736 \pm 2.02 \times 0.9404 = (25.757, 29.561)$$

e) La proportion de l'échantillon de personnes à qui on a déjà dit qu'elles avaient un taux de cholestérol élevé est

$$p = \frac{z}{n} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{1+1+\ldots+0}{40} = \frac{11}{40} = 0.275$$

L'erreur type de cette proportion est

$$se(p) = \sqrt{\text{var}(p)} = \sqrt{(1-f)\frac{p(1-p)}{n-1}} = \sqrt{\left(1 - \frac{40}{3{,}048}\right)\frac{0.275(1-0.275)}{40-1}} = \sqrt{0.00505} = 0.07103$$

Et son intervalle de confiance à 90% est donné par

$$p \pm t_{\left(n-1;1-\frac{\alpha}{2}\right)} \times se(p) = 0.275 \pm t_{\left(40-1;1-\frac{0.10}{2}\right)} \times 0.07103 = 0.275 \pm 1.68 \times 0.07103 = 0.275 \pm 0.1197 =$$
$$= (0.155; 0.395)$$

f) Lorsque la taille de l'échantillon passe de n = 40 à n = 100, s_y^2 ne change pas et, à partir de (c), nous avons déjà une estimation sans biais de s_y^2 . Donc, si n=100

$\text{se}(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{(1 - \text{f})\frac{s_y^2}{n}} = \sqrt{(1 - \frac{100}{3048})\frac{35{,}85}{100}} = \sqrt{0{,}3467} = 0{,}5888$

(g) La précision souhaitée est en termes de coefficient de variation pour la moyenne de l'IMC. Réécrivons-le en fonction de la variance d'échantillonnage souhaitée, c'est-à-dire

$$CV(\bar{y}) = 0.025 \rightarrow \frac{se(\bar{y})}{27.659} = 0.025 \rightarrow\ se(\bar{y}) = 0.691 \rightarrow Var(\bar{y}) = 0.691^2 = 0.478$$

Par conséquent, en utilisant s_y^2 calculé en (c) comme estimation de la variance de la population d'éléments, S_y^2 , la taille d'échantillon requise pour cette précision cible est donnée par

$$n_{BMI} = \frac{s_y^2}{Var(\bar{y}) + \frac{s_y^2}{N}} = \frac{35.85}{0.478 + \frac{35.85}{3048}} = 73.17\ \cong 74$$

De manière équivalente, pour la proportion de personnes à qui on a déjà dit qu'elles avaient un taux de cholestérol élevé,

$$CV(p) = 0.025 \rightarrow \frac{se(p)}{p} = 0.025 \rightarrow \frac{se(p)}{0.275} = 0.025 \rightarrow se(p) = 0.006875 \rightarrow Var(p) = 0.006875^2 = 0.000047$$

Maintenant, en utilisant le fait que pour les proportions, la variance de l'élément peut être approchée par

$s^2 = \frac{n}{n-1} p(1-p) \approx p(1-p)$, la taille d'échantillon requise pour cette précision cible pour cette proportion est

$$n_{HYPEV} = \frac{p(1-p)}{Var(\bar{y}) + \frac{p(1-p)}{N}} = \frac{0.275(1-0.275)}{0.000047 + \frac{0.275(1-0.275)}{3,048}} = 1769.43 \cong 1,770$$

Étant donné qu'il est nécessaire que les deux estimations aient un coefficient de variation d'au plus 0,025, la taille de l'échantillon doit être n = 1 770 (c'est-à-dire la plus grande taille d'échantillon)

(h) Calculez un intervalle de confiance à 99 % pour le nombre total de personnes à qui un médecin a déjà dit qu'elles avaient un taux de cholestérol élevé

Le nombre total de personnes qui se sont déjà fait dire par un médecin qu'elles avaient un taux de cholestérol élevé peut être estimé à

$$\hat{Z} = Np = 3,048 \times 0.275 = 838.2$$

et son erreur standard est

$$se(\hat{Z}) = \sqrt{var(\hat{Z})} = \sqrt{var(Np)} = \sqrt{N^2 var(p)} = N\sqrt{var(p)} = Nse(p) = 3,048 \times 0.07103 = 216.469$$

Par conséquent, l'intervalle de confiance à 99 % pour ce total est donné par

$$\hat{Z} \pm t_{\left(n-1;1-\frac{\alpha}{2}\right)} se(\hat{Z}) = 838.2 \pm t_{\left(40-1;1-\frac{0.01}{2}\right)} \times 216.469 = 838.2 \pm 2.708 \times 586.252$$
$$= (251.948; 1,424.452)$$