Exercices de révision
Licence MIASHS 3ème année
MIA0603T - Sondages

01/04/2025

Exercice 17

Une université souhaite réaliser une enquête sur les revenus des étudiants ayant un job étudiant, parmi un total de 10 000 étudiants. Des études ont montré que les revenus varient considérablement selon les types de jobs étudiants, ce qui conduit à une stratification en trois types de jobs. De plus, il a été observé que les revenus des étudiants sont influencés par leur année d'études. Nous allons donc proposer des plans d'échantillonnage pour analyser les revenus en fonction de l'année d'études des étudiants. Voici les informations disponibles :

Type de job	Poids dans l'ensemble des étudiants	Écart-type des revenus (en euros)
1	40%	400
2	35%	200
3	25%	100
Ensemble	100%	300

a. Soit R le revenu moyen des étudiants, et $\hat{R}$ l'estimateur du revenu moyen basé sur un échantillon aléatoire simple à probabilités égales de $n = 100$ étudiants.

   1. L'estimateur $\hat{R}$ étant basé sur un échantillon aléatoire simple, l'écart-type de cet estimateur est donné par :

      $$\sigma_{\hat{R}} = \frac{\sigma}{\sqrt{n}}$$

      où σ est l'écart-type global des revenus dans la population. Ici, $\sigma = 300$ euros (écart-type global), et $n = 100$.

   2. L'écart-type de $\hat{R}$ est

      $$\sigma_{\hat{R}} = \frac{300}{\sqrt{100}} = \frac{300}{10} = 30 \text{ euros}$$

b. On décide de stratifier l'échantillon de 100 étudiants selon les trois types de jobs étudiants.

1. **Répartition "représentative" de l'échantillon :**
   Puisque la stratification se fait selon les types de jobs, la répartition représentative de l'échantillon est basée sur les poids des différents types de jobs. Le nombre d'individus dans chaque strate est donc :

   $$\text{Type 1 : } 100\times 40\% = 40, \quad \text{Type 2 : } 100\times 35\% = 35, \quad \text{Type 3 : } 100\times 25\% = 25$$

   Ainsi, l'échantillon sera composé de 40 étudiants du type 1, 35 étudiants du type 2, et 25 étudiants du type 3.

2. **Formule générale de l'écart-type dans un échantillonnage stratifié :**
   L'écart-type dans un échantillonnage stratifié est donné par la formule :

   $$\text{Ecart-type}(\hat{R}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \cdot \frac{\sigma_h^2}{n_h}}$$

   où : - H est le nombre de strates (ici $H = 3$), - N_h est la taille de la strate h, - N est la taille totale de la population, - σ_h^2 est la variance dans la strate h, - n_h est la taille de l'échantillon dans la strate h.

3. **Formule pour 3 types de jobs étudiants :**
   Dans notre cas, la formule se simplifie ainsi :

   $$\sigma_{\hat{R}} = \sqrt{\left(\frac{N_1}{N}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{N_2}{N}\right)^2 \frac{\sigma_2^2}{n_2} + \left(\frac{N_3}{N}\right)^2 \frac{\sigma_3^2}{n_3}}$$

   où : - $N_1 = 4000, N_2 = 3500, N_3 = 2500$, - $\sigma_1 = 400, \sigma_2 = 200, \sigma_3 = 100$, - $n_1 = 40, n_2 = 35, n_3 = 25$.

4. Par un calcul simple, ona

   $$\begin{aligned}
   \sigma_{\hat{R}} &= \sqrt{\left(\frac{4000}{10000}\right)^2 \cdot \frac{400^2}{40} + \left(\frac{3500}{10000}\right)^2 \cdot \frac{200^2}{35} + \left(\frac{2500}{10000}\right)^2 \cdot \frac{100^2}{25}} \\
   &= \sqrt{(0.4)^2 \cdot \frac{160000}{40} + (0.35)^2 \cdot \frac{40000}{35} + (0.25)^2 \cdot \frac{10000}{25}} \\
   &= \sqrt{0.16 \cdot 4000 + 0.1225 \cdot 1142.857 + 0.0625 \cdot 400} \\
   &= \sqrt{640 + 140.0002 + 25} = \sqrt{805.0002} \approx 28.37 \text{ euros}
   \end{aligned}$$

5. L'écart-type stratifié est plus faible que l'écart-type de l'échantillon simple, qui était de 30 euros. Cela montre l'avantage de la stratification pour réduire l'incertitude de l'estimation du revenu moyen.

c. Quelle serait la répartition optimale de l'échantillon pour minimiser l'écart-type de l'estimateur de $\hat{R}$?

1. Pour minimiser l'écart-type de l'estimateur, la taille de l'échantillon dans chaque strate doit être proportionnelle à la taille de la population et inversement proportionnelle à l'écart-type dans chaque strate. La répartition optimale n_h^* pour chaque strate est donnée par :

$$n_h^* = n \cdot \frac{N_h \cdot \sigma_h}{\sum_{h=1}^{H} N_h \cdot \sigma_h}$$

où N_h est la taille de la strate et σ_h est l'écart-type dans la strate h.

2. Par application du sondage optimal, ona

$$n_1^* = 100 \cdot \frac{4000 \cdot 400}{4000 \cdot 400 + 3500 \cdot 200 + 2500 \cdot 100} \approx 100 \cdot \frac{1600000}{1600000 + 700000 + 250000} \approx 40$$

$$n_2^* = 100 \cdot \frac{3500 \cdot 200}{1600000 + 700000 + 250000} \approx 35$$

$$n_3^* = 100 \cdot \frac{2500 \cdot 100}{1600000 + 700000 + 250000} \approx 25$$

La répartition optimale est donc la même que la répartition "représentative".

3. Comme les tailles des strates et les écarts-types sont proportionnels, la répartition optimale est identique à la répartition représentative, et l'écart-type reste similaire.

Exercice 18 (Modalité A)

Dans le cas d'une proportion pour un sondage aléatoire simple à probabilités égales, nous avons la formule de l'intervalle de confiance suivante :

$$IC_{1-\alpha}(P) = \hat{P} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{(1-f)\hat{P}(1-\hat{P})}{n-1}},$$

où :

- $\hat{P}$ est l'estimateur de la proportion,
- f est le facteur de correction pour la population finie, défini par $f = \frac{n}{N}$,
- n est la taille de l'échantillon,
- N est la taille de la population,
- $z_{1-\alpha/2}$ est le quantile de la loi normale standard pour un niveau de confiance $1 - \alpha$.

Dans l'exercice, nous avons $n = 75$, $N = 1000$, d'où $f = \frac{75}{1000} = 0.075$. Par ailleurs, l'estimateur $\hat{P}$ donne comme valeur sur cet échantillon $\hat{P} = \frac{30}{75} = 0.4$. Au niveau 95%, le quantile $z_{1-\alpha/2}$ est égal à 1.96.

Nous pouvons donc calculer l'intervalle de confiance :

$$IC = 0.4 \pm 1.96 \times \sqrt{\frac{(1 - 0.075) \cdot 0.4 \cdot (1 - 0.4)}{75 - 1}},$$

ce qui donne :

$$IC = 0.4 \pm 1.96 \times \sqrt{\frac{0.925 \cdot 0.4 \cdot 0.6}{74}} \approx 0.4 \pm 1.96 \times \sqrt{\frac{0.222}{74}} \approx 0.4 \pm 1.96 \times 0.0545.$$

Cela donne un intervalle de confiance :

$$IC = [0.293; 0.507],$$

avec une étendue $e = 0.507 - 0.293 = 0.214$.

Taille d'échantillon pour une étendue deux fois plus petite

Pour avoir une étendue deux fois plus petite, cela implique que n vérifie la condition suivante (cf. Section 2.4.2 du cours) :

$$n \geq \frac{\left(\frac{e}{2}\right)^2 + z_{1-\alpha/2}\hat{P}(1-\hat{P})}{\left(\frac{e}{2}\right)^2 + \frac{z_{1-\alpha/2}\hat{P}(1-\hat{P})}{N}}.$$

En remplaçant par les valeurs données ($e = 0.107$, $\hat{P} = 0.4$, $N = 1000$, et $z_{1-\alpha/2} = 1.96$), nous obtenons :

$$n \geq \frac{\left(\frac{0.107}{2}\right)^2 + 1.96 \cdot 0.4 \cdot 0.6}{\left(\frac{0.107}{2}\right)^2 + \frac{1.96 \cdot 0.4 \cdot 0.6}{1000}}.$$

Calculons les différentes parties de cette expression :

$$\left(\frac{0.107}{2}\right)^2 = 0.002857 \quad \text{et} \quad 1.96 \cdot 0.4 \cdot 0.6 = 0.4704.$$

Ainsi, l'expression devient :

$$n \geq \frac{0.002857 + 0.4704}{0.002857 + \frac{0.4704}{1000}} = \frac{0.473257}{0.0033274} \approx 143.$$

Cela donne une taille d'échantillon minimale de $n \geq 143$, ce qui permet d'obtenir un intervalle de confiance $IC = [0.325; 0.475]$, avec une étendue $e = 0.475 - 0.325 = 0.15$, soit deux fois plus petite que l'intervalle précédent.