TD 4 de Théorie des Sondages

Ensai
Année 2023-2024

Exercice 13

Rappel : Sous le modèle de travail

$$m : y_k \;=\; \mathbf{x}_k^\top \beta + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma_k^2, \end{array} \right.$$

les résidus de régression calculés sur la population valent

$$E_k = y_k - \mathbf{x}_k^\top B_{MCG} \quad \text{avec} \quad B_{MCG} = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2},$$

et le coefficient de détermination est mesuré par

$$R^2 \;=\; 1 - \frac{\sum_{k \in U} E_k^2}{\sum_{k \in U} (y_k - \mu_y)^2}.$$

Dans une plantation de 500 hectares contenant $N = 1\,000\,000$ d'arbres, nous sélectionnons un échantillon S d'arbres selon un sondage aléatoire simple de taille $n = 500$, et pour chacun de ces arbres le volume y_k est mesuré. Nous obtenons les résultats suivants :

$$\sum_{k \in S} y_k = 181.3 \text{ m}^3 \quad \text{et } \sum_{k \in S} (y_k - \bar{y})^2 = 50.64 \text{ m}^6.$$

1) Donner un estimateur sans biais du total t_y, et calculer sa valeur.
2) Donner un estimateur de variance sans biais associé, et calculer sa valeur sur l'échantillon. En déduire le coefficient de variation estimé.

Nous disposons après enquête de l'information auxiliaire suivante. Parmi l'ensemble des N arbres de la plantation :

- 40 % sont de petite taille (diamètre à la base ≤ 22.5 cm) ; l'échantillon contient $n_1 = 210$ arbres de cette catégorie, avec un volume moyen de 0.12 m^3 et une dispersion de 0.01 m^6,
- 40 % sont de taille moyenne (diamètre à la base compris entre 22.5 et 37.5 cm) ; l'échantillon contient $n_2 = 200$ arbres de cette catégorie, avec un volume moyen de 0.38 m^3 et une dispersion de 0.02 m^6,
- 20 % sont de grande taille (diamètre à la base ≥ 37.5 cm) ; l'échantillon contient $n_3 = 90$ arbres de grande taille ayant un volume moyen de 0.89 m^3 et une dispersion de 0.08 m^6.

3) Ecrire le modèle de travail correspondant au sondage aléatoire simple, et celui correspondant à la post-stratification.
4) Justifier pourquoi, sous ce modèle de travail, le coefficient de détermination se réécrit

$$R^2 \quad = \quad 1 - \frac{\sum_{h=1}^{H} \sum_{k \in U_h} (y_k - \mu_{yh})^2}{\sum_{k \in U} (y_k - \mu_y)^2},$$

avec des notations à préciser.
5) En écrivant le coefficient de détermination comme une fonction de totaux, expliquer pourquoi ce R^2 peut être estimé approximativement sans biais sur l'échantillon par

$$\hat{R}^2_{\pi} \quad = \quad 1 - \frac{\sum_{h=1}^{H} \sum_{k \in S_h} (y_k - \bar{y}_h)^2}{\sum_{k \in S} (y_k - \bar{y})^2}.$$

6) En utilisant les données collectées sur l'échantillon, conclure sur l'utilité de la post-stratification.

7) Donner la formule de l'estimateur post-stratifié du total $\hat{t}_{y,post}$, et calculer sa valeur. Pourquoi obtenons-nous une valeur plus grande que celle de $\hat{t}_{y\pi}$.
8) Donner un estimateur de variance approximativement sans biais pour $\hat{t}_{y,post}$.

Exercice 14

Nous considérons une région agricole contenant $N = 2\ 000$ fermes découpées en 2 strates : 1 400 fermes de moins de 160 hectares (U_1) et 600 fermes de plus de 160 hectares (U_2). Nous notons x_k la surface totale de l'exploitation k et y_k sa surface totale en céréales. Nous cherchons à estimer le total t_y en utilisant un sondage aléatoire simple stratifié de taille $n = 200$.

1) Quelle taille d'échantillon doit-on sélectionner dans chaque strate avec une allocation proportionnelle ?
2) Supposons que la dispersion de y dans la strate U_2 est 2 fois plus grande que la dispersion dans la strate U_1. Quelle taille d'échantillon doit-on sélectionner dans chaque strate avec une allocation optimale pour y ?

Nous utilisons finalement $n_1 = 140$ et $n_2 = 60$. Après enquête, nous obtenons les résultats suivants dans l'échantillon :

Strate	N_h	n_h	$\bar{y}_h$	$\bar{x}_h$	s^2_{yh}	s^2_{xh}	s_{xyh}
1	1 400	140	1 425	6 225	50 160	668 745	154 830
2	600	60	1 300	6 125	89 730	1 677 200	339 100

3) Donner l'estimateur de Horvitz-Thompson du total t_y, et calculer sa valeur sur l'échantillon.
4) Donner le coefficient de variation estimé, et calculer sa valeur. Commenter brièvement la qualité de l'estimation obtenue.
5) En utilisant l'approximation de variance

$$V_p(\hat{t}_{y\pi}) \simeq N^2 \frac{1-f}{n} S^2_{y,intra},$$

montrer que le coefficient de variation de $\hat{t}_{y\pi}$ est inférieur à 1% si

$$n \geq \frac{1}{\frac{1}{N} + \left(\frac{0.01 t_y}{N}\right)^2 \times \frac{1}{S^2_{y,intra}}}.$$

6) En utilisant les données collectées sur l'échantillon, donner une estimation de la taille d'échantillon nécessaire pour que le CV de $\hat{t}_{y\pi}$ soit inférieur à 1%.

Nous supposons maintenant que les sous-totaux de la variable x_k sur les strates sont connus, et valent respectivement $t_{x1} = 9\ 000\ 000$ et $t_{x2} =$

3 500 000. Il est proposé d'utiliser un redressement par le ratio sur t_{xh} dans chaque strate U_h.

7) Le redressement proposé va-t-il conduire à une variance plus faible qu'avec l'estimateur de Horvitz-Thompson ? En utilisant le critère donné dans l'exercice 12, question 9, justifier quantitativement votre réponse.

8) Pour chacune des deux strates, donner l'estimateur de Horvitz-Thompson $\hat{t}_{x_h\pi}$ de t_{xh}, et calculer sa valeur.
9) Ecrire l'estimateur $\hat{t}_{yR}$ obtenu avec le redressement, et calculer sa valeur. Ecrire le modèle de travail sous-jacent.
10) Donner un estimateur de variance approximativement sans biais pour $\hat{t}_{yR}$ (il n'est pas demandé de calculer sa valeur).

11) Proposer un autre estimateur redressé utilisant l'ensemble de l'information auxiliaire, et de variance plus faible. Donner le modèle de travail associé.