TD 3 de Théorie des Sondages

Ensai
Année 2023-2024

Ce sujet contient trois exercices. Certains contiennent une partie TP, à réaliser sous R à l'aide du package `sampling` :

```
#Appel du package sampling
> library(sampling)
> help(package="sampling")
```

Exercice 10

La population `belgianmunicipalities` disponible avec le package `sampling` contient 589 communes. La liste des variables est donnée en annexe.
Nous nous intéressons à la sous-population U des $N = 452$ communes contenant entre 1 000 et 20 000 habitants en 2004, sélectionnée à l'aide du code suivant :

```
> data(belgianmunicipalities)
> loc <- (1000<=belgianmunicipalities$Tot04)
        *(belgianmunicipalities$Tot04<=20000)
> pop <- belgianmunicipalities[loc==1,]
> attach(pop)
```

Nous souhaitons sélectionner un échantillon de taille moyenne $n = 100$, selon un plan de Poisson à probabilités proportionnelles à la variable `Tot04`.

Partie 1

1) S'agit-il d'un plan de taille fixe ?
2) Donner la formule des probabilités d'inclusion proportionnelles à la taille, et les calculer en utilisant le package `sampling`.
3) Quel problème peut-on rencontrer avec cette formule ? Est-ce le cas ici ?
4) Sélectionner un échantillon en utilisant la fonction `UPpoisson`.

```
#Q2 : probabilités d'inclusion prop. à Tot04
> nech <- ...
> Npop <- ...
> pi <- ...
#Q4 : sélection
> set.seed(14121997)
> ech <- ...
> test <- sum(ech)
```

Partie 2

Nous nous intéressons à la variable `TaxableIncome`, notée y_k, donnant le revenu total imposable de la commune k.

5) Rappeler la formule d'estimation de variance pour l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ sous un tirage de Poisson.
6) Calculer l'estimateur de Horvitz-Thompson du revenu total imposable, et donner un intervalle de confiance à 95 %.

```
#Q6 : estimation pour un total
> options("scipen"=-100,digits="4")
> y <- TaxableIncome
#Estimateur de HT
> est_ht_ty=...
> est_ht_ty
#Estimateur de variance
> pikl=pi %*% t(pi) +diag(pi-pi*pi)
> vest_ht_ty=...
#Intervalle de confiance
> Binf <- ...
> Bsup <- ...
```

7) Calculer l'estimateur $\tilde{\mu}_y = \dfrac{\hat{t}_{y\pi}}{N}$ du revenu moyen imposable par commune μ_y, et l'estimateur de variance associé.

```
#Q7 : estimation pour la moyenne
> est_tilde_muy <- ...
> vest_tilde_muy <- ...
```

Partie 3

Nous nous intéressons à l'estimation :

- du revenu moyen imposable par commune μ_y,
- du revenu moyen imposable par individu R.

8) Ecrire chacun de ces paramètres sous la forme d'une fonction de deux totaux, que vous préciserez.
9) Rappeler l'estimateur par substitution pour chacun de ces deux paramètres. Calculer sa valeur sur l'échantillon.

```
#Q9 : estimation de deux ratios
> z <- rep(1,Npop)
> x <- Tot04
#Estimateurs de HT
> est_ht_ty=...
> est_ht_tz=...
> est_ht_tx=...
#Estimateurs par substitution
> est_sub_muy <- ...
> est_sub_R <- ...
```

10) Rappeler la formule d'estimation de variance par linéarisation. La calculer sur l'échantillon :

1. en programmant directement les formules d'estimation,
2. en utilisant la fonction `vartaylor_ratio`.

Vérifier que les deux méthodes donnent le même résultat.

```
#Q10 : estimation de variance
#Programmation directe
#Variables linéarisées
> uk_muy <- ...
> uk_R <- ...
#Estimateurs de variance par linéarisation
> vest_lin1_muy <- varHT(...,pikl[ech==1,ech==1],1)
> vest_lin1_R <- varHT(...,pikl[ech==1,ech==1],1)
#Utilisation de la fonction vartaylor_ratio
> vest_lin2_muy <- vartaylor_ratio(...,...,pikl[ech==1,ech==1])
> vest_lin2_R <- vartaylor_ratio(...,...,pikl[ech==1,ech==1])
```

11) Pour le paramètre μ_y, comparer l'estimation de variance obtenue avec celle obtenue à la question 7. Commenter ce résultat.
12) Donner le coefficient de variation estimé pour les deux paramètres d'intérêt.

```
> options("scipen"=100,digits="4")
> cv_est_sub_muy <- ...
> cv_est_sub_R <- ...
```

Exercice 11

Nous nous intéressons à l'estimation du coefficient de corrélation entre deux variables quantitatives x et y, défini par

$$\rho = \frac{V_{xy}}{\sqrt{V_{xx} \times V_{yy}}} \quad \text{avec} \quad \left\{ \begin{array}{rcl} V_{xy} & = & \sum_{k \in U} (x_k - \mu_x)(y_k - \mu_y), \\ V_{xx} & = & \sum_{k \in U} (x_k - \mu_x)^2, \\ V_{yy} & = & \sum_{k \in U} (y_k - \mu_y)^2. \end{array} \right. \tag{1}$$

Pour un paramètre θ, nous notons $u_k(\theta)$ sa variable linéarisée, et $\hat{u}_k(\theta)$ sa variable linéarisée estimée.

1) En utilisant les règles de calcul de la linéarisation, justifier que :

$$u_k(\rho) = \rho \left[\frac{u_k(V_{xy})}{V_{xy}} - \frac{1}{2} \left\{ \frac{u_k(V_{xx})}{V_{xx}} + \frac{u_k(V_{yy})}{V_{yy}} \right\} \right]. \tag{2}$$

2) En remarquant que

$$V_{xy} = \sum_{k \in U} x_k y_k - \frac{t_x \times t_y}{N}, \tag{3}$$

montrer que

$$\begin{array}{rcl} u_k(V_{xy}) & = & x_k y_k - \{\mu_y x_k + \mu_x y_k - \mu_y \mu_x\} \\ & = & \{x_k - \mu_x\}\{y_k - \mu_y\}. \end{array} \tag{4}$$

3) En déduire (sans calcul, mais en justifiant brièvement) que

$$u_k(V_{xx}) = \{x_k - \mu_x\}^2 \quad \text{et} \quad u_k(V_{yy}) = \{y_k - \mu_y\}^2. \tag{5}$$

4) En déduire que

$$u_k(\rho) = \rho \left[\frac{\{x_k - \mu_x\}\{y_k - \mu_y\}}{V_{xy}} - \frac{1}{2} \frac{\{x_k - \mu_x\}^2}{V_{xx}} - \frac{1}{2} \frac{\{y_k - \mu_y\}^2}{V_{yy}} \right]. \tag{6}$$

5) Donner l'estimateur par substitution de ρ, en spécifiant bien tous les estimateurs impliqués.
6) Donner une expression de la variable linéarisée estimée de ρ.

Exercice 12

Rappel : Soit $\mathbf{x}_k$ un q-vecteur de variables auxiliaires dont le total $t_{\mathbf{x}}$ sur la population est connu. En utilisant le modèle de travail

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

l'estimateur par la régression généralisée (GREG) est donné par

$$\begin{array}{rcl} \hat{t}_{y,greg} & = & \hat{t}_{y\pi} + \hat{\mathbf{b}}_\pi^\top \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \right] \\ \text{avec} & & \hat{\mathbf{b}}_\pi = \left(\sum_{k \in S} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in S} \frac{d_k \mathbf{x}_k y_k}{\sigma_k^2}. \end{array}$$

Partie 1

1) Nous considérons le cas particulier $q = 1$, avec $\mathbf{x}_k = x_k > 0$, et le **modèle ratio**

$$y_k = \beta x_k + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma^2 x_k. \end{array} \right.$$

Montrer que dans ce cas

$$\hat{\mathbf{b}}_\pi = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}, \tag{7}$$

puis que l'estimateur GREG se réécrit sous la forme

$$\hat{t}_{y,greg} = \frac{t_x}{\hat{t}_{x\pi}} \hat{t}_{y\pi} \equiv \hat{t}_{yR}.$$

L'estimateur obtenu est appelé **estimateur par le ratio**.

2) Nous admettons le résultat suivant : sous un sondage aléatoire simple, la variance de $\hat{t}_{yR}$ est plus faible que la variance de l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ si

$$\rho \geq \frac{1}{2} R \sqrt{\frac{S_x^2}{S_y^2}}, \tag{8}$$

avec ρ le coefficient de corrélation linéaire entre x_k et y_k (cf équation 1 de l'exercice 11) et $R = \dfrac{\mu_y}{\mu_x}$.

Interpréter l'équation (8). Sous un sondage aléatoire simple, l'estimateur par le ratio peut-il être préférable à l'estimateur de Horvitz-Thompson si les variables x_k et y_k sont corrélées négativement ?

Partie 2

Nous utilisons la sous-population U des $N = 452$ communes de Belgique contenant entre 1 000 et 20 000 habitants en 2004. Dans cette population, nous sélectionnons un échantillon aléatoire simple de taille $n = 100$ en utilisant le code suivant :

```
#Sélection d'un échant. de n=100 municipalités
> data(belgianmunicipalities)
> loc <- (1000<=belgianmunicipalities$Tot04)
        *(belgianmunicipalities$Tot04<=20000)
> pop <- belgianmunicipalities[loc==1,]

> nech <- 100
> Npop <- nrow(pop)

> set.seed(03111971)
> ech <- srswor(nech,Npop)
> echant <- pop[ech==1,]
> attach(echant)
```

Nous supposons connu le total de la variable $x_k \equiv Tot04_k$ dans la population U, et égal à $t_x = 4\ 391\ 601$. Nous considérons l'estimateur par le ratio sur le total t_x.

3) Pour une variable d'intérêt mesurée sur S, proposer une façon d'estimer si la relation (8) est vraie en utilisant les données de l'échantillon uniquement.

4) Pour les variables d'intérêt quantitatives de l'échantillon, examiner si cette condition est vérifiée. Conclure sur le gain ou la perte en variance de l'estimateur par le ratio pour ces différentes variables.

A Liste des variables de "belgianmunicipalities"

Commune	Municipality name
INS	INS Code INS
Province	Province number
Arrondiss	Administrative division number
Men04	Number of men on July 1, 2004
Women04	Number of women on July 1, 2004
Tot04	Total population on July 1, 2004
Men03	Number of men on July 1, 2003
Women03	Number of women on July 1, 2003
Tot03	Total population on July 1, 2003
Diffmen	Men04 minus Men03
Diffwom	Women04 minus Women03
DiffTOT	Tot04 minus Tot03
TaxableIncome	Total taxable income in euros in 2001
Totaltaxation	Total taxation in euros in 2001
Averageincome	Average of the income-tax return in euros in 2001
Medianincome	median of the income-tax return in euros in 2001.