Tp2 – Sondages

Alain LATOUR

22 janvier 2021

Avant-propos

Vous réalisez aujourd'hui votre seconde séance de travail pratique (TP) du cours de sondages.

Comme pour le premier, vous travaillerez en binôme. Il faut cependant reconstituer des binômes différents de ceux du TP1.

Pour connaître votre « partenaire binomial » voir sur le site :
https ://www-ljk.imag.fr/membres/Alain.Latour/Cours/M1/SOND/germes/germe.html

Le compte rendu est à rendre le 29 janvier 2021

1 Contexte

Rappel : Soit $\mathcal{U} = \{1, 2, \ldots, N\}$ les indices des unités d'une population finie de taille N. On dénote par

$$\pi_k = \Pr(\text{l'unité } k \text{ soit dans l'échantillon})$$

et par

$$\pi_{k\ell} = \Pr(\text{l'unité } k \text{ et l'unité } \ell \text{ soient dans l'échantillon})$$

les probabilités d'inclusion du premier et du second ordre. De plus, on définit les variables aléatoires indicatrices suivantes :

$$Z_k = \begin{cases} 1, & \text{si l'unité } k \text{ est dans l'échantillon ;} \\ 0, & \text{sinon.} \end{cases}$$

Un échantillon est donc une réalisation du vecteur aléatoire $(Z_1, \ldots, Z_N)^\top$, formé des variables indicatrices. La loi de probabilité du vecteur $(Z_1, \ldots, Z_N)^\top$ est le plan de sondage.

Estimateur d'Horvitz-Thompson

Définition Dans le contexte qui nous intéresse, pour estimer τ, le total de la population, on utilise :

$$\hat{\tau} = \sum_{k \in \omega} \frac{y_k}{\pi_k} \tag{1}$$

où ω est l'ensemble des indices des unités appartenant à l'échantillon [1] et π_i la probabilité d'inclusion de l'individu i.

La variance de cet estimateur est :

$$\mathrm{Var}[\hat{\tau}] = \sum_{k=1}^{N-1} \sum_{\ell=k+1}^{N} (\pi_k \pi_\ell - \pi_{k\ell}) \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 . \tag{2}$$

On peut suggérer comme estimateur de la variance de $\hat{\tau}$ l'expression suivante :

$$\widehat{\mathrm{Var}[\hat{\tau}]} = \sum_{k \in \omega} \sum_{\substack{\ell \in \omega \\ \ell > k}} \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}} \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 . \tag{3}$$

Autrement dit, la somme ne se fait que pour les éléments de la population qui se retrouvent dans l'échantillon.

2 Quelques fonctions R utiles

Pour illustrer les éléments détaillés dans cette section, considérons la population utilisée comme exemple en classe, formée de 8 individus

$$\mathcal{U} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8\}$$

et avec comme valeurs de y_i

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

Imaginons que la variable R `omega` contienne les indices des individus d'un échantillon, par exemple, $\omega = \{1,\ 3,\ 5,\ 6\}$. En R, on aurait :

```
> U <- 1:8
> y <- c(1, 2, 3, 4, 4, 7, 7, 7, 8)
> omega <- c(1, 3, 5, 6)
```

Nous disposons d'une fonction nous permettant de savoir si un indice est dans l'échantillon :

```
> un_dedans <- function(x, no) {
+   sum(x == no) >= 1
+ }
> k <- 5
> un_dedans(omega,k)

[1] TRUE

> k <- 4
> un_dedans(omega,k)

[1] FALSE
```

1. $k \in \omega$ signifie que l'individu d'indice k est dans l'échantillon ω ($\subseteq \mathcal{U}$)

Comme on le souhaite, puisque 5 est dans omega, le premier appel à la fonction un_dedans donne TRUE.

On peut aussi vérifier si un couple $(i,\ j)$ est dans l'échantillon.

```
> deux_dedans <- function(x, no1, no2) {
+   as.logical((sum(x == no1) >= 1) * (sum(x == no2) >= 1))
+ }
> i <- 5
> j <- 3
> deux_dedans(omega,i,j)

[1] TRUE

> i <- 4
> j <- 5
> deux_dedans(omega,i,j)

[1] FALSE
```

Ces deux fonctions sont utiles pour calculer les π_k et $\pi_{k\ell}$ comme cela est demandé dans les exercices qui suivent.

3 Exercices

Ces exercices sont tirés de Lohr (2010).

1. Soit $N = 6$ et $n = 3$. Afin d'étudier les distributions des estimateurs, nous supposons connues les valeurs pour toute la population :

$$\begin{array}{lll} y_1 = 98 & y_3 = 154 & y_5 = 190 \\ y_2 = 102 & y_4 = 133 & y_6 = 175 \end{array}$$

   On s'intéresse à $\overline{y}_{\mathcal{U}}$, la moyenne de la population. Deux plans d'échantillonnage sont proposés.

	i_1	i_2	i_3	p[Échant.]
1	1	3	5	0,125
2	1	3	6	0,125
3	1	4	5	0,125
4	1	4	6	0,125
5	2	3	5	0,125
6	2	3	6	0,125
7	2	4	5	0,125
8	2	4	6	0,125

Table 1 – Plan no 1

Plan 1 La table 1 donne les indices des individus de la population composant l'échantillon accompagnés de la probabilité de l'échantillon. Ainsi, pour le premier, les indices des unités choisies sont : 1, 3 et 5. Les valeurs associées Y_{i_j}, $j = 1,\ 2,\ 3$, sont respectivement 98, 154 et 190. La probabilité d'observer cet échantillon est 0,125.

En R, on peut créer une trame de données donnant l'information de la table 1, à l'aide du script 1.

Listing 1 – Création de la trame de données pour le plan de la table 1

```
1  ##
2  ## Une population de 6 unités
3  ##
4  Pop <- c(98, 102, 154, 133, 190, 175)
5  #
6  # Plan 1
7  #
8  i1 <- rep(c(1, 2), each = 4)
9  i2 <- rep((rep(c(3, 4), each = 2)), 2)
10 i3 <- rep(c(5, 6), 4)
11 plan1 <- cbind(i1, i2, i3, p.ech = rep(1 / 8, 8))
12 rownames(plan1) <- 1:8
13 plan1
```

La ligne 13 du script 1 affiche la trame de données du plan 1.

	i_1	i_2	i_3	p[Échant.]
1	1	4	6	0,250
2	2	3	6	0,500
3	1	3	5	0,250

TABLE 2 – Plan no 2

Plan 2 La table 2 donne les indices des individus de la population composant l'échantillon accompagnés de la probabilité de l'échantillon de ce second plan. En R, on peut créer une trame de données donnant l'information de la table 2 en exécutant le script 2.

Listing 2 – Création de la trame de données pour le plan de la table 2

```
1 plan2 <-
2   cbind(
3     i1 = c(1, 2, 1),
4     i2 = c(4, 3, 3),
5     i3 = c(6, 6, 5),
6     p.ech = c(0.25, 0.50, 0.25)
7   )
8 rownames(plan2) <- 1:3
```

(a) Déterminer les probabilités d'inclusion du premier et du second ordre pour ces deux plans de sondages.

(b) Quelle est la valeur de $\overline{y}_{\mathcal{U}}$?

(c) Soit $\overline{y}$ la moyenne des valeurs de l'échantillon. Pour chacun des plans trouver :

i. $\mathrm{E}[\overline{y}]$

ii. $\mathrm{Var}[\overline{y}]$

iii. $\mathrm{Biais}[\overline{y}]$

iv. $\mathrm{EQM}[\overline{y}]$

(d) Lequel des plans est le meilleur ? Pourquoi ?

2. Pour la population utilisée comme exemple en classe, formée de 8 individus

$$\mathcal{U} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8\}$$

et avec comme valeurs de y_i

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

considérer le plan de sondage suivant :

	i_1	i_2	i_3	i_4	p[Échant.]
1	1	3	5	6	0,125
2	2	3	7	8	0,250
3	1	4	6	8	0,125
4	2	4	6	8	0,375
5	4	5	7	8	0,125

TABLE 3 – Plan no 3

Listing 3 – Création de la trame de données pour le plan de la table 3

```
##
## Une population de 8 unités
##
Pop <- c(1, 2, 4, 4, 7, 7, 7, 8)
#
# Plan
#
i1 <- c(1, 2, 1, 2, 4)
i2 <- c(3, 3, 4, 4, 5)
i3 <- c(5, 7, 6, 6, 7)
i4 <- c(6, 8, 8, 8, 8)
p.ech <- c(1 / 8, 1 / 4, 1 / 8, 3 / 8, 1 / 8)
plan <- cbind(i1, i2, i3, i4, p.ech)
rownames(plan) <- 1:5
```

(a) Déterminer la probabilité de sélection π_i, pour chaque unité i.

(b) Quelle est dans ce contexte la distribution de $\hat{t} = 8\overline{y}$?

3. **Estimation d'Horvitz-Thompson** Nous allons utiliser un autre plan de sondage dans lequel il y a 20 échantillons possibles. Pour tous, la probabilité d'être choisi est 1/20.

	i_1	i_2	i_3	i_4	p[Échant.]
1	3	4	5	7	0,050
2	2	4	5	6	0,050
3	3	4	5	8	0,050
4	3	4	7	8	0,050
5	3	4	6	8	0,050
6	3	4	6	7	0,050
7	2	5	7	8	0,050
8	1	6	7	8	0,050
9	1	2	3	6	0,050
10	1	5	6	7	0,050
11	1	2	4	8	0,050
12	4	5	6	8	0,050
13	3	5	6	8	0,050
14	4	5	7	8	0,050
15	1	5	7	8	0,050
16	2	5	6	7	0,050
17	1	4	7	8	0,050
18	3	5	7	8	0,050
19	2	4	6	7	0,050
20	1	2	6	8	0,050

TABLE 4 – Plan no 4

Listing 4 – Création de la trame de données pour le plan de la table 4

```
Pop <- c(1, 2, 4, 4, 7, 7, 7, 8)
N <- length(Pop)
n <- 4
n.ech <- choose(N, n)
ech.all <-
  cbind(choix <-
          t(matrix(combn(1:N, n), n, n.ech)),
        (matrix(Pop[choix], n.ech, n)))
rownames(ech.all) <- seq(1:nrow(ech.all))
colnames(ech.all) <-
  c(paste("i", 1:n, sep = ""), paste("Pop[", 1:n, "]", sep = ""))
set.seed(101)
plan4 <- cbind(ech.all[sample(1:nrow(ech.all),20),][,1:4],pr_ech=1/↘
→20)
rownames(plan4) <- 1:20
plan4
```

(a) Déterminer les probabilités d'inclusion du premier et du second ordre de ce plan de sondage.

(b) Ajouter au tableau du tableau plan4 une colonne donnant l'estimation du total calculé par la formule (1), page 1. Vous définissez ainsi une nouvelle variable

aléatoire.

(c) Évaluer l'espérance mathématique de cette variable aléatoire à l'aide du tableau obtenu. L'estimateur est-il biaisé ?

(d) Évaluer la variance de cette variable aléatoire toujours à l'aide du tableau obtenu.

(e) Calculer la variance avec la formule (2), page 2. Vérifier que vous avez bel et bien le même résultat.

(f) Quelle est la variance estimée de $\hat{\tau}$ si l'échantillon choisi est $\omega = \{1,\ 5,\ 7,\ 8\}$? Voir la formule (3), page 2.

Références

Alalouf, S. 2015, *Introduction aux sondages*, Loze-Dion éditeur inc., ISBN 2923565924, 323 p..

Lohr, S. L. 2010, *Sampling : Design and Analysis*, 2e éd., Advanced Series, Cengage Learning, Pacific Grove, CA, ISBN 0495105279, 608 p..