TD de Sondages

Ensai
Année 2023-2024

Certains exercices contiennent une partie TP, à réaliser sous R à l'aide du package `sampling` :

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

Exercice 1

Soit la population $U = \{1, 2, 3, 4\}$. Le plan de sondage $p(\cdot)$ est défini par :

$$\begin{array}{lllllllll} p(\{1,2\}) & = & 0.2 & p(\{1,4\}) & = & 0.1 & p(\{3,4\}) & = & 0.2 \\ p(\{1,2,3\}) & = & 0.3 & p(\{2,3,4\}) & = & 0.1 & p(\{1,2,4\}) & = & 0.1 \end{array}$$

1) Calculer les probabilités d'inclusion d'ordre 1, et en déduire la taille moyenne d'échantillon sélectionné.
2) Calculer les probabilités d'inclusion d'ordre 2, et en déduire la matrice Δ de variance-covariance. Le plan de sondage est-il de taille fixe?

Les valeurs prises par une variable d'intérêt y sont supposées connues sur l'ensemble de la population, et sont données dans le tableau suivant :

k	1	2	3	4
y_k	1	4	5	2

3) Nous supposons que l'échantillon $s = \{1, 2\}$ est sélectionné. Calculer l'estimateur de Horvitz-Thompson du total, et donner son estimateur de variance. Cet estimateur de variance est-il sans biais?

Partie TP

Compléter le code suivant pour mettre en oeuvre sous R le calcul de l'estimateur de HT et de l'estimateur de variance associé pour l'échantillon sélectionné:

```
> y_ech <- ...
> pi_ech <- ...
> pikl_ech <- matrix(data=...,nrow=...)
> est_ht <- HTestimator(...,...)
> evar_ht <- varHT(...,...,method="...")
```

Exercice 2

Nous considérons une population de $N = 6$ entreprises dont le nombre d'employés est connu :

Unité	1	2	3	4	5	6
x	200	80	50	50	10	10

Nous voulons sélectionner un échantillon de taille $n = 2$, à probabilités proportionnelles au nombre d'employés.

1) Calculer les probabilités d'inclusion $\pi_1, \ldots, \pi_6$ correspondantes.

Il est proposé de sélectionner l'échantillon de la façon suivante :

- L'unité 1 est sélectionnée d'office.
- Une unité est tirée au hasard parmi les 5 restantes, avec les probabilités $\pi_2, \ldots, \pi_6$.

2) Cette méthode de tirage respecte t-elle les probabilités d'inclusion voulues?
3) Calculer les probabilités d'inclusion d'ordre 2. Est-il possible d'estimer sans biais la variance de l'estimateur de Horvitz-Thompson?

Il est maintenant proposé de sélectionner l'échantillon de la façon suivante :

- Pour chaque unité $k \in U$, une variable aléatoire u_k est générée indépendamment selon une loi de Bernoulli de paramètre π_k.

- Les unités k pour lesquelles $u_k = 1$ sont sélectionnées.

4) Cette méthode de tirage respecte t-elle les probabilités d'inclusion voulues?
5) Calculer les probabilités d'inclusion d'ordre 2. Est-il possible d'estimer sans biais la variance de l'estimateur de Horvitz-Thompson?

Exercice 3 (adapté de Ardilly et Tillé, 2003)

Nous nous intéressons à la proportion p de personnes atteintes par une maladie professionnelle, dans une entreprise comptant $N = 1\ 500$ employés. Nous souhaitons sélectionner un échantillon par sondage aléatoire simple (sans remise) afin d'estimer cette proportion.

1) Définir la population, l'unité statistique, le plan de sondage, la variable d'intérêt et le paramètre d'intérêt.

2) Comment s'interprète la contrainte suivante sur la précision de l'estimation : *la longueur de l'intervalle de confiance pour le paramètre p est inférieure à* $2\ \beta$.

3) En l'absence d'information auxiliaire, calculer la taille n d'échantillon nécessaire pour que la longueur de l'intervalle de confiance associé à p (avec un niveau de 95%) soit inférieure à 0.02.

Nous supposons maintenant connue l'information auxiliaire suivante : dans les entreprises du même type, la proportion de personnes atteintes de cette maladie est habituellement de 30% .

4) Utiliser cette information pour recalculer la taille n d'échantillon nécessaire. Commenter les résultats obtenus.

5) Calculer la taille n d'échantillon nécessaire pour que le pourcentage de personnes malades soit estimé avec un CV inférieur à 1% . Expliquer le lien entre le CV obtenu et l'intervalle de confiance.

Exercice 4

Nous souhaitons estimer le nombre total de $M\&Ms$ et le nombre total de $M\&Ms$ bleus dans un carton contenant 50 sachets de $M\&Ms$. Nous sélectionnons pour celà 4 sachets par sondage aléatoire simple sans remise. Le tableau suivant donne pour chaque sachet de l'échantillon sa composition en $M\&Ms$ selon la couleur (R=Rouge, J=Jaune, O=Orange, B=Bleu, V=Vert, M=Marron).

1	RJJJOOBBBBVMMMMMM
2	RRRJJJOBBVVVMMM
3	RRRRJJJJOVVVVVMM
4	RRJJJOOOBBBVVVMMMM

1) Donner un estimateur du nombre de total de $M\&Ms$ bleus dans le carton, et un intervalle de confiance à 95% pour ce total.
2) Donner un estimateur du nombre de total de $M\&Ms$ dans le carton, et un intervalle de confiance à 95% pour ce total.
3) Donner un estimateur de la proportion de $M\&Ms$ bleus dans le carton, et calculer sa valeur.

Partie TP

Compléter le code R suivant pour calculer les estimateurs demandés dans les questions précédentes, ainsi que les intervalles de confiance :

```
#Taille d'échantillon et taille de population
> nech=...
> Npop=...
#Probabilités d'inclusion pour un SRS
> pi_ech=...
#Probabilités d'inclusion d'ordre 2 pour un SRS
> pikl_ech=matrix(data=rep(nech*(nech-1)/(Npop*(Npop-1)),
                nech*nech),nech)
> pikl_ech
  for (i in 1:4){
    pikl_ech[i,i]=nech/Npop
  }

#Variable y_k : nb de M&Ms bleus dans le sachet k
```

```
> y_ech=c(4,2,0,3)
#Variable x_k : nb de M&Ms dans le sachet k
> x_ech=c(16,15,16,18)

#Q1: estimation du nb total de M&Ms bleus et intervalle de
    confiance
est_ht <- HTestimator(...,...)
Binf <- est_ht-1.96*sqrt(varHT(...,...,method="..."))
Bsup <- est_ht+1.96*sqrt(varHT(...,...,method="..."))

#Q2: estimation du nb total de M&Ms et intervalle de confiance

#Q3: estimation de la proportion de M&ms bleus
> R_ht <- HTestimator(...,...)/HTestimator(...,...)
```

Exercice 5

Nous considérons une population U de taille N, dans laquelle un échantillon S est tiré selon un sondage aléatoire simple sans remise (SRS) de taille n. Nous relevons sur cet échantillon les valeurs prises par deux variables y_k et z_k.

1) Nous nous intéressons à l'estimation du paramètre $\theta = \mu_y - \mu_z$. Proposer un estimateur basé sur l'échantillon S et calculer sa variance.

Nous sélectionnons maintenant dans U un échantillon S_1 selon un SRS de taille n, puis de façon indépendante un échantillon S_2 selon un SRS de taille n. Nous relevons sur S_1 les valeurs prises par la variable z_k, et sur S_2 les valeurs prises par la variable y_k.

2) Proposer un estimateur de θ basé sur les échantillons S_1 et S_2 et calculer sa variance. Quand cet estimateur est-il préférable au premier?

3) Nous nous intéressons maintenant à l'estimation du paramètre $M = \frac{\mu_y + \mu_z}{2}$. Proposer :

- un estimateur de M basé sur l'échantillon S,
- un estimateur de M basé sur les échantillons S_1 et S_2,

et comparer l'efficacité de ces deux estimateurs.

Partie TP

Nous considérons une population de 10 boulangeries. Nous nous intéressons au prix de la baguette de pain. Ces prix (en euros) sont donnés dans le tableau suivant pour les mois de juin 2009 et juillet 2009.

k	1	2	3	4	5	6	7	8	9	10
Prix en juin	0.85	0.82	0.80	0.83	0.90	0.87	0.84	0.85	0.88	0.83
Prix en juillet	0.85	0.84	0.84	0.85	0.90	0.86	0.85	0.87	0.88	0.86

Compléter le programme R suivant pour obtenir une comparaison de l'efficacité des deux stratégies :

- pour l'estimation de l'évolution du prix entre juin et juillet 2009,
- pour l'estimation du prix moyen de la baguette sur la période juin-juillet 2009

```
> y_ech=c(0.85,0.82,0.80,0.83,0.90,0.87,0.84,0.85,0.88,0.83)
> z_ech=c(0.85,0.84,0.84,0.85,0.90,0.86,0.85,0.87,0.88,0.86)

> sy2 <- var(y_ech)
> sz2 <- var(z_ech)
> syz <- cov(y_ech,z_ech)

#Comparaison pour l'estimation de l'évolution du prix
...

#Comparaison pour l'estimation du prix moyen
...
```