TD 2 de Théorie des 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 6

Nous souhaitons comparer la précision de différentes stratégies d'échantillonnage, pour l'estimation du revenu moyen des habitants d'une commune.
Les personnes concernées sont stratifiées selon 3 tranches d'âge : moins de 20 ans, 20 à 49 ans, plus de 50 ans. Le tableau suivant donne, pour chaque strate sa taille N_h, le revenu moyen de la strate μ_{yh} (supposé connu) et la dispersion S^2_{yh} du revenu dans la strate (supposé connu).

h	N_h	μ_{yh}	S^2_{yh}
1	15 000	20	$(6)^2$
2	25 000	35	$(12)^2$
3	20 000	60	$(16)^2$

Nous souhaitons sélectionner un échantillon de taille $n = 1\ 000$ parmi les habitants de la commune.

1) En utilisant l'équation de décomposition de la variance

$$S_y^2 = \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_{yh}^2 + \sum_{h=1}^{H} \frac{N_h}{N - 1} (\mu_{yh} - \mu_y)^2,$$

calculer la dispersion globale S_y^2 du revenu dans l'ensemble de la population. En déduire la variance et le coefficient du revenu moyen estimé, pour un échantillon sélectionné par sondage aléatoire simple sans remise.

```
> N_h <- c(15000,25000,20000)
> mu_yh <- c(20,35,60)
> S_yh <- c(6,12,16)

> N <- ...
> mu_y <- ...

#Calcul de la dispersion intra
> S2_intra <- ...
#Calcul de la dispersion inter
> S2_inter <- ...
#Calcul de la dispersion globale et de la variance SRS
> S2 <- S2_intra+S2_inter
> V_SRS <- ...
> CV <- ...
```

2) La stratification paraît-elle a priori appropriée pour ce problème ? Justifier quantitativement.

3) Quelle allocation par strate obtenons-nous avec un sondage stratifié à allocation proportionnelle ? Donner la variance du revenu moyen estimé, et calculer l'effet de plan.

```
#Allocation proportionnelle
> n_prop <- round(...)
#Variance
> V_PROP <- ...
#Effet de plan
> DEFF_PROP <- ...
```

4) Quelle allocation par strate obtenons-nous avec un sondage stratifié à allocation optimale ? Donner la variance du revenu moyen estimé, et calculer l'effet de plan. Comparer avec le résultat obtenu à la question précédente.

```
#Allocation optimale
> n_opt <- round(...)
#Variance
> V_OPT <- ...
#Effet de plan
> DEFF_OPT <- ...
```

Exercice 7

Nous considérons la population des $N = 67$ principaux aéroports français en 2020. Nous souhaitons sélectionner un échantillon de 30 aéroports dans cette population pour estimer :

- le nombre total t_y de passagers en 2020, avec y_k le nombre de passagers dans l'aéroport k en 2020 (`Pass20`),
- le nombre total t_z de passagers en transit en 2020, avec z_k le nombre de passagers en transit dans l'aéroport k en 2020 (`Trans20`).

Nous utiliserons la variable auxiliaire x_k, donnant le nombre de passagers dans l'aéroport k en 2019 (`Pass19`).
Les données sont contenues dans le fichier `aeroports_complet.xlsx`.

1) Pour la variable `Pass19`, calculer sa valeur moyenne, l'étendue, le rapport interquartile et le coefficient de variation. Commenter les résultats obtenus.

```
> summary <- summary(Pass19)
> range <- ...
> rapport <- ...
> mu_x <- ...
> cv_x <- ...
```

2) Quel est le coefficient de variation de $\hat{t}_{x\pi}$ si l'échantillon est sélectionné selon un SRS ?
3) Quelle taille d'échantillon faudrait-il sélectionner pour que le coefficient de

variation de $\hat{t}_{x\pi}$ soit inférieur à 20 % sous un SRS ? Commenter le résultat obtenu.

```
#Q2 : cas du sondage aléatoire simple
> Npop <- 67
> mu_y <- mean(Pass20)
> S2_y <- var(Pass20)
#Variance sous un SRS de taille n=30
> V_SRS <- ...
> CV_SRS <- ...
#Taille d'échant min pour un CV<=20%
> n_min <- ...
```

Nous découpons la population des $N = 67$ aéroports en trois strates :

- La strate U_1 des $N_1 = 11$ aéroports ayant accueilli plus de 2 000 000 de passagers en 2019,
- La strate U_2 des $N_2 = 30$ aéroports ayant accueilli entre 100 000 et 2 000 000 de passagers en 2019,
- La strate U_3 des $N_3 = 26$ aéroports ayant accueilli moins de 100 000 passagers en 2019.

4) Donner l'allocation optimale pour l'estimation de t_x, pour un sondage aléatoire simple stratifié de taille $n = 30$.
Quel problème se pose dans la strate U_3 avec cette allocation ?
5) Donner la variance de l'estimateur de HT de t_x sous cette allocation, et donner l'effet de plan (design-effect).
Expliquer brièvement la très faible valeur de cet effet de plan.

```
#Q4 : allocation optimale
> N_h <- c(11,30,26)
> S2_xh <- rbind(var(Pass19[Pass19>=2000000]),
                 var(Pass19[100000<=Pass19 & Pass19<2000000]),
                 var(Pass19[Pass19<100000]))
> n_opt <- ...
#Q5 : variance sous l'allocation optimale
> n_opt <- c(...,...,...)
> V_OPT <- ...
> DEFF <- ...
```

Nous retenons finalement l'allocation suivante : nous tirons

— $n_1 = 11$ aéroports dans la strate 1,
— $n_2 = 17$ aéroports dans la strate 2,
— $n_3 = 2$ aéroports dans la strate 3.

6) Utiliser le package `sampling` pour sélectionner un échantillon d'aéroports selon le plan de sondage choisi.
7) Pour l'échantillon sélectionné, donner l'estimateur de HT de t_y et donner un intervalle de confiance à 95 % .
8) Pour l'échantillon sélectionné, donner l'estimateur de HT de t_z et son coefficient de variation estimé.

```
#Q6 : sélection de l'échantillon
#Variable de stratification
> aeroports_complet$U <-
    1+(100000<=Pass19)*(Pass19<2000000)+2*(Pass19<100000)
#Sélection de l'échantillon
> ech=strata(data=...,stratanames="...",size=c(...),method="...")
#Récupération des données
> data<-getdata(aeroports_complet,ech)
> ...
```

Exercice 8 (Dussaix et Grosbras, 1992)

Une entreprise comporte 400 exécutants et 100 cadres. La direction de l'entreprise désire évaluer un indice de satisfaction, assimilable à une variable nu-mérique positive y. Elle décide pour cela de faire réaliser une enquête auprès de 100 personnes employées dans l'entreprise, à l'aide d'un plan de sondage stratifié, avec sondage aléatoire simple dans chaque strate.
Nous supposons que la dispersion de la variable y est la même au sein de chacun des deux groupes. Comment répartir l'échantillon entre les deux strates, selon que l'un des objectifs suivants est visé :

1. obtenir la meilleure précision possible pour la moyenne de l'indice de satisfaction dans l'entreprise,
2. obtenir la même précision pour la moyenne de l'indice de satisfaction dans chacune des deux strates,
3. obtenir la meilleure précision possible sur la différence entre les moyennes de l'indice de satisfaction dans les deux strates.

Exercice 9 (Rousseau, 2009)

Nous voulons sélectionner un échantillon de taille 4 dans une population de 8 entreprises dont la taille, mesurée en termes d'effectif salarié, est connue. L'échantillon est tiré à probabilités proportionnelles à la taille.

Entreprise	1	2	3	4	5	6	7	8
Taille	300	300	150	100	50	50	25	25

1) Donner les probabilités d'inclusion d'ordre 1 des entreprises.
2) Sélectionner l'échantillon selon un tirage systématique.
3) Calculer la matrice des probabilités d'inclusion d'ordre 2. En déduire que la variance de l'estimateur de HT ne peut pas être estimée sans biais.
4) Pour l'échantillon sélectionné, donner l'estimateur $\hat{t}_{x\pi}$ du total de xk, et le comparer au vrai total t_x. Ce résultat était-il prévisible ?

```
> xk <- c(300,300,150,100,50,50,25,25)
#Q1 : probabilités d'inclusion proportionnelles à x_k
> nech <- 4
> pi <- ...
#Q2 : sélection d'un échantillon selon un tirage systématique
> ech <- ...
#Q3 : matrice des probabilités d'inclusion d'ordre 2
> pi_kl <- ...
#Q4 : Estimation du total de x_k
> tx_pihat <- ...
```