Sondages
TP 3 – Les sondages avec R
L3 MIASHS - IDS

Exercice 1. Dans ce premier exercice, nous allons apprendre l'utilisation du package R *survey* disponible sur le CRAN[1]. A partir de données simulées, nous allons estimer plusieurs quantités d'intérêt vues dans les TDs précédents.

```
# Rappelez vous pour installer un package on fait
#install.packages("survey")
# Ensuite, à chaque utilisation on charge le package avec :
library(survey)

## Loading required package:  grid
## Loading required package:  Matrix
## Loading required package:  survival
##
## Attaching package:  'survey'
## The following object is masked from 'package:graphics':
##
##    dotchart
```

1. Tout d'abord nous allons générer les données d'intérêt de notre population grâce à la commande suivante

   ```
   N <- 1000
   index <- 1:N
   y <- rnorm(N)
   population <- data.frame(index, y)
   ```

   Commentez les données stockées dans le dataframe population. Quelle est la moyenne et le total de la variable d'intérêt y dans votre population ?

   ```
   # Le data.frame contient deux colonnes avec respectivement l'index
   # et la valeur de la variable d'intéret. Les paramètres cible sont
   # (Attention: nous utilisons de données générées de manière aléatoire,
   #             les valeurs numériques ne seront pas les mêmes entre deux
   #             executions du code mais qualitativement ils sont comparables)
   y_mean  <- mean(y)
   y_total <- sum(y)
   print(paste0("Moyenne pop: ", y_mean))

   ## [1] "Moyenne pop: 0.0196175389589368"

   print(paste0("Total pop: ", y_total))

   ## [1] "Total pop: 19.6175389589368"
   ```

2. Tirez un échantillon de la population selon un plan SI sans remise de taille $n = 100$. Calculez l'estimateur d'Horvitz-Thompson du total et sa variance. Faites de même pour la moyenne.

   ```
   n <- 100
   id_sample <- sample(N, n)
   echantillon <- population[id_sample, ]

   y_mean_s  <- mean(echantillon$y)
   y_total_s <- y_mean_s * N

   y_varmean_s = var(echantillon$y) * ((1 - (n/N)) / (n))
   y_vartotal_s = y_varmean_s * (N^2)
   ```

3. Nous allons maintenant utiliser les fonctions du packages *survey*. On crée d'abord l'objet contenant les informations du sondage:

[1] https://cran.r-project.org/web/packages/survey/survey.pdf

```
sondage <- svydesign(id=~1,strata=NULL,data=echantillon,fpc=rep(N,n))
```

La variable *echantillon* contient l'échantillon du dataframe *population* tiré en question 2, et n est la taille de l'échantillon. Commentez les paramètres de la fonction *svydesign*. Pour vous aider, consultez l'aide de la fonction en cliquant sur la touche F1 lorsque le curseur est sur la fonction.

```
sondage <- svydesign(id = ~1, strata = NULL,
                     data = echantillon, fpc = rep(N, n))
```

4. On calcule l'estimateur du total avec la fonction suivante

```
total <- svytotal(~y, design=sondage)
```

Commentez le résultat obtenu.

```
total <- svytotal(~y, design = sondage)
print(total)

##    total     SE
## y 90.131 89.921

# La commande affiche l'estimation ponctuelle du total et
# l'écart type estimé.
```

5. On peut calculer simplement un intervalle de confiance de l'estimateur en utilisant:

```
ICtotal <- confint(svytotal(~y, design=sondage))
```

Vérifiez ce résultat.

```
ICtotal <- confint(svytotal(~y, design = sondage))
print(ICtotal)

##       2.5 %   97.5 %
## y -86.11177 266.3731

#Et on vérifie
y_total_s - (1.96 * sqrt(y_vartotal_s))

## [1] -86.11501

y_total_s + (1.96 * sqrt(y_vartotal_s))

## [1] 266.3763
```

6. Faites de même pour la moyenne (fonction *svymean*).

```
### 6 ###
moyenne <- svymean(~y, design = sondage)
ICmoyenne <- confint(moyenne)
y_mean_s - (1.96 * sqrt(y_varmean_s))

## [1] -0.08611501

y_mean_s + (1.96 * sqrt(y_varmean_s))

## [1] 0.2663763
```

7. Comparez les résultats aux valeurs trouvées en question 2.
8. Les paramètres de votre population (question 1) sont-ils bien estimés avec un sondage de taille $n = 100$?

```
# Pour répondre à cette question vérifiez que les intervalles de
# confiance que vousa vez construit couvrent la valeurs des
# paramètres calculés en 1.
```

9. Réalisez un sondage SISR avec $n = 500$. Commentez la différence avec le plan précédent.

```
n <- 500
id_sample <- sample(N,n)
echantillon <- population[id_sample,]
sondage <- svydesign(id = ~1, strata = NULL,
                     data = echantillon)

## Warning in svydesign.default(id = ~1, strata = NULL, data = echantillon): No weights
or probabilities supplied, assuming equal probability

moyenne_2 <- svymean(~y, design = sondage)

confint(moyenne_2)

##         2.5 %    97.5 %
## y -0.03658333 0.1433176

# Malgré le fait que l'échantillon est plus grande, l'intervalle de
# confiance est moins précis du fait qu'on remplace les éléments
# (on perd alors le facteur de correction par population finie).
```

Exercice 2. Le but de cet exercice est d'expérimenter l'utilisation du package *survey* dans le cas d'un sondage stratifié. Nous allons pour cela utiliser les données *Titanic*

1. On commence par charger les données déja intégrées dans R, et par les transformer en dataframe. Cette manipulation nécessite le package *reshape*

```
        data(Titanic)
Titanic <- as.data.frame(Titanic)
library(reshape)

##
## Attaching package: 'reshape'
## The following object is masked from 'package:Matrix':
##
##     expand

Titanic <- untable(Titanic[, c(1, 2, 3, 4)], num = Titanic[, 5])
```

   Commentez les données stockées dans le dataframe.

```
# Le data.frame contient 2201 observations, décrites par 4 variables
# (Classe, Sexe, Age, Survie).
N <- nrow(Titanic)
index <- 1:N
```

2. Évaluez la proportion de survivant sur la population.

```
# La variable d'interet est Survived, on cherche une proportion
survie <- table(Titanic$Survived)[2] / N
```

3. On va considérer un sondage stratifié selon la classe de cabine du Titanic (variable *class*). Réalisez un tirage selon un plan avec allocation proportionnel et $n = 500$. Calculez l'estimateur de la proportion de survivant et sa variance.

```
n <- 500

N_h <- as.vector(table(Titanic$Class))
TT <- N_h/sum(N_h)
```

```r
n_h <- round(TT * n)

compt <- 1
i_sample <- c()
for (j in levels(Titanic$Class)){
  i_sample <- c(i_sample,sample(index[Titanic$Class == j],n_h[compt]))
  compt <-  compt + 1
}
echantillon <- Titanic[i_sample,]

tab <- table(echantillon$Class,echantillon$Survived)
out <- as.matrix(prop.table(tab,margin = 1))
S <- out[,1] * out[,2] * (n_h/(n_h-1))
var_p <- sum(N_h * ((N_h - n_h) / n_h) * S) / (N*N)
```

4. On va maintenant utiliser le package *survey*. On crée l'objet sondage avec strates graçe à la commande

```r
sondage <- svydesign(id=~1,strata=~Class,data=echantillon,fpc=rep(N_h,n_h))
```

La variable *echantillon* contient l'échantillon du dataframe *Titanic* tiré en question 3. *N_h* est un vecteur contenant le nombre d'individus par strates sur la population, *n_h* un vecteur contenant le nombre d'individus par strates sur l'échantillon. Expliquez la différence avec la valeur *fpc* de l'exercice précédent.

```r
sondage <- svydesign(id =~ 1, strata = ~Class,
                     data = echantillon, fpc = rep(N_h, n_h))
```

5. Vérifiez que la fonction *svymean* donne les bonnes valeurs pour l'estimateur de la proportion de survivant et sa variance en utilisant.

```r
svymean(~Survived,sondage)

##                 mean     SE
## SurvivedNo  0.68219 0.0177
## SurvivedYes 0.31781 0.0177
```

6. Calculez l'intervalle de confiance de l'estimateur de la proportion en utilisant la fonction *svyciprop*.

```r
svyciprop(~Survived, sondage)

##                  2.5% 97.5%
## Survived 0.318 0.284  0.35
```

7. Réalisez maintenant un sondage avec allocation optimale. Calculez l'estimateur de la proportion de survivant, sa variance et intervalle de confiance. Commentez la différence avec l'allocation proportionnelle.

```r
tab <- table(Titanic$Class,Titanic$Survived)
out <- as.matrix(prop.table(tab,margin = 1))
var <- out[,1] * out[,2] * (N_h/(N_h-1))

n_h <- round(n * (N * var) / (sum((N * var))))
n_h[4] <- n_h[4] + 1
compt <- 1
i_sample <- c()
for (j in levels(Titanic$Class)){
  i_sample <- c(i_sample,sample(index[Titanic$Class == j],n_h[compt]))
  compt <-  compt + 1
}
echantillon <- Titanic[i_sample,]
```

```r
sondage <- svydesign(id=~1,strata=~Class,data=echantillon,fpc=rep(N_h,n_h))
svymean(~Survived,sondage)
```

```
##                 mean     SE
## SurvivedNo  0.70211 0.0199
## SurvivedYes 0.29789 0.0199
```

```r
svyciprop(~Survived,sondage)
```

```
##                  2.5% 97.5%
## Survived 0.298 0.260  0.34
```