Compte-Rendu TP2

Thomas Laurent

25 Avril 2017

Import des données et création de la variable indicatrice correspondant aux individus dont l'activité est "Agriculture, sylviculture et pêche"

```
# Import des données
agri = read.csv(file = "fpempl01.csv", header = TRUE, sep = ";")

# Codage de la variable agri et ajout d'un identifiant

agri = agri %>% mutate(agri = ifelse(NAFG16 == "EA", 1, 0), identifiant = seq(1,
    17687, 1))
```

On vérifie que le nombre d'individus dans la population est égal à 17687.

```
# Nombre d'individus dans la population
dim(agri)[1]
```

```
## [1] 17687
```

On calcule ensuite la proportion d'agriculteurs dans la population totale ainsi que la variance correspondante.

```
# Proportion d'agriculteurs dans la population cible
proportion = sum(agri$agri)/dim(agri)[1]
cat(paste0("Proportion d'agriculteurs dans la population cible=", round(proportion *
    100, digits = 2)), "%")
```

```
## Proportion d'agriculteurs dans la population cible=1.69 %
```

```
# Variance de la proportion d'agriculteurs dans la population cible
variance = 17687/17686 * proportion * (1 - proportion)
cat(paste0("\n", "Variance=", variance))
```

```
##
## Variance=0.0166202297645078
```

On observe que dans la population cible, le taux d'agriculteurs est de 1.69%.

Selection du plan PEAR

On choisit arbitrairement de fixer une marge d'erreur relative de 20% avec un niveau de confiance de 95%. On détermine la taille de l'échantillon en utilisant la formule suivante:

$$n \geq \frac{z_{97.5\%}^2(1-p)}{\epsilon^2 p}$$

avec ϵ étant égal à 20%. On peut appliquer la formule précédente en se basant sur la taux d'agriculteurs dans la population d'étude.

```r
# Calcul de la taille de l'échantillon
taille = qnorm(0.975) * (1 - proportion)/(0.2^2 * proportion)
cat(paste0("\n", "Taille de l'échantillon ", round(taille, digits = 0)))
```

```
##
## Taille de l'échantillon 2849
```

On fixe donc la taille de l'échantillon à 2849.

```r
# Selection de l'échantillon selon un plan PEAR
set.seed(123)
echantillon = srswr(taille, 17687)

# Creation du vecteur de poids
poids = 17687/taille * echantillon
```

Estimation de la proportion et de la variance

On estime la proportion d'agriculteurs et la variance correspondante en utilisant l'échantillon sélectionné.

```r
#Estimation de la proportion
PEAR=svydesign(id=~identifiant, weights=poids[which(poids!=0)],

               data=agri[which(echantillon!=0),])
mean=svymean(~agri,PEAR)
cat(paste0("Proportion estimée:",round(mean[1]*100,digits = 3),"%"))
```

```
## Proportion estimée:1.79%
```

```r
#Variance de l'estimation
variance=SE(mean)^2
cat(paste0("\n","Variance estimée:",variance))
```

```
##
## Variance estimée:7.3804534479277e-06
```

```r
#Intervalle de confiance a 95%
cat(paste0("\n","Intervalle de confiance: "))
```

```
##
## Intervalle de confiance:
```

```r
svyciprop(~agri,PEAR)
```

```
##               2.5% 97.5%
## agri 0.0179 0.0133  0.02
```

La proportion d'agriculteur estimée est de 1.79% ce qui est proche de la proportion d'agriculteur dans l'ensemble de la population (1.69%). La différence entre l'estimation de la proportion et la proportion dans la population est inférieure à 10%. De plus, on obtient un intervalle de confiance pour la proportion d'agriculteurs : [1.3%;2%].

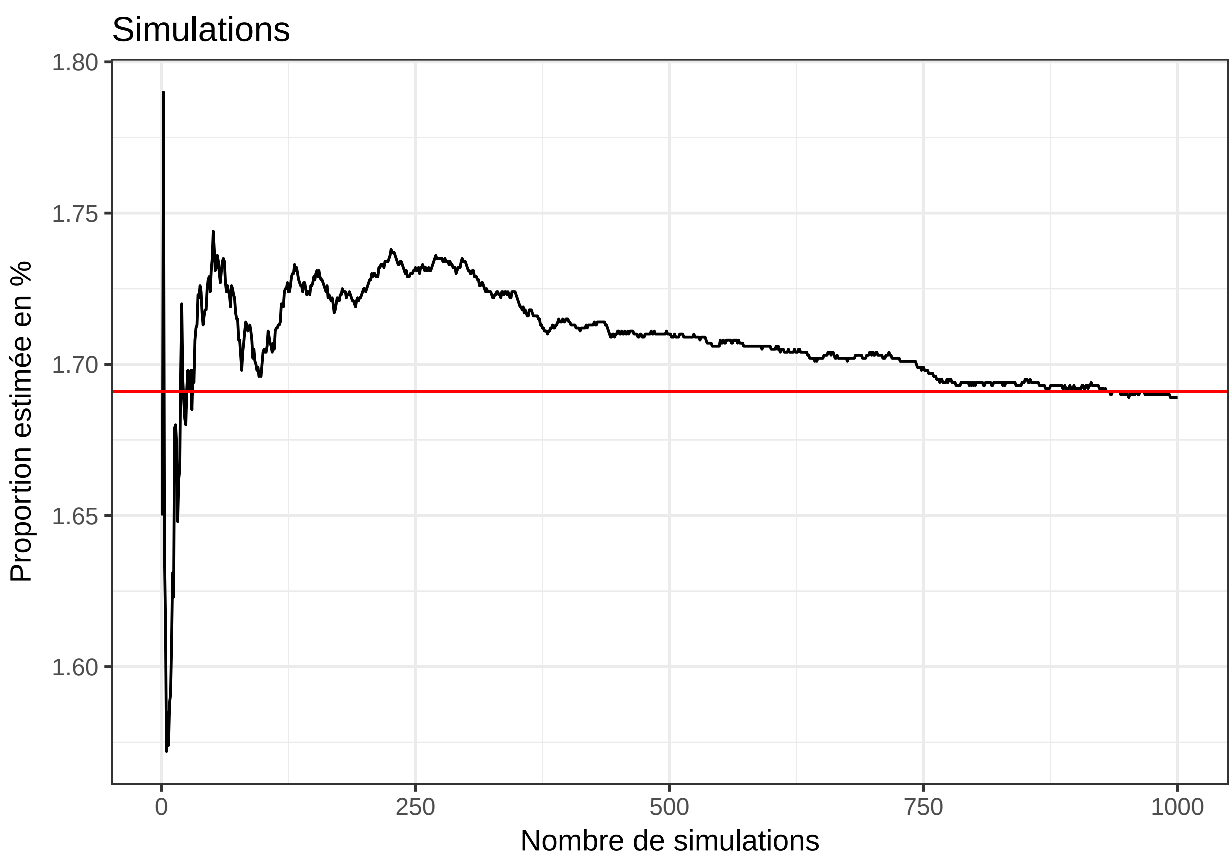
Simulations avec le plan PEAR

Proportion

Pour évaluer si l'estimateur de la proportion d'agriculteur de la population est sans biais, on procède à 1000 simulations de selections d'échantillons et on s'intéresse à la convergence de l'estimateur.

```
# tirage de 1000 echantillons PEAR et de 1000 estimations
nb.simul <- matrix(1:1000, 1000, 1)
y.esti <- matrix(1:1000, 1000, 1)
for (i in 1:1000) {
    echantillon <- srswr(taille, 17687)
    poids_pear = (17687/taille) * echantillon
    ech.pear <- svydesign(id = ~identifiant, weights = poids_pear[which(poids_pear !=
        0)], data = agri[which(echantillon != 0), ])
    estimation = svymean(~agri, ech.pear)
    y.esti[i] <- estimation[1]
}
```

```
# Representation graphique de la variation de l'estimation en fonction du
# nombre de simulations
simul = as.data.frame(cbind(nb.simul, y.esti))
ggplot(data = simul, aes(x = nb.simul, y = round(cumsum(y.esti)/nb.simul * 100,
    digits = 3))) + geom_line() + labs(title = "Simulations") + ylab("Proportion estimée en %") +
    geom_hline(yintercept = round(proportion * 100, digits = 3), color = "red") +
    xlab("Nombre de simulations") + theme_bw()
```



Pour un grand nombre de simulations, la moyenne des estimations tend vers la valeur de la proportion d'agriculteurs de la population. On visualise bien que l'estimateur de la proportion est un estimateur sans biais.

On s'intéresse ensuite au coefficient de variation.

```
# Estimation du coefficient de variation par simulation

var_esti = var(y.esti)
mean_esti = mean(y.esti)
CV_empir = sqrt(var_esti)/mean_esti
cat(paste0("\n", "Coefficient de variation empirique: ", round(CV_empir * 100,
    digits = 2), "%"))
```

```
##
## Coefficient de variation empirique: 14.27%
```

```
# Coefficient de variation theorique
CV = sqrt((1 - proportion)/(taille * proportion))
cat(paste0("\n", "Coefficient de variation (theorique):", round(CV * 100, digits = 2),
    "%"))
```

```
##
## Coefficient de variation (theorique):14.29%
```

On remarque que le coefficient de variation empirique et le coefficient de variation théorique présentent des valeurs proches.