Échantillonnage et sondage

Alexis Buckens

20 juin 2016

Objectif

Le but de ce travail est de choisir la meilleure méthode d'échantillonnage parmi les deux méthodes proposées. Dans le premier cas, on choisit un certain nombre de personnes via Facebook. Cette méthode a l'avantage d'être la moins couteuse, mais présente également un gros désavantage : un faible taux de réponses. La deuxième méthode, basée sur les écoles, semble a priori plus coûteuse, mais permet d'avoir un meilleur taux de réponse. Dans les deux cas, le but final est de s'intéresser à l'argent de poche des élèves, et il sera donc intéressant de comparer les propriétés des estimateurs de la variable "Argent" dans les deux cas.

Échantillonnage de la population

Il convient d'abord de sélectionner un échantillon de chacune des bases des données, en respectant la contrainte de coût.

Le premier échantillon doit être sélectionné parmi la base de donnée "OfficialPopu". Cette base de donnée contient une liste d'écoles comprenant une série d'élèves. Comme il s'agit de sélectionner m écoles parmi M, on peut considérer qu'il s'agit d'un échantillonnage en grappes sur les écoles. On peut en effet par ailleurs supposer que l'usage de l'argent de poche sera globalement plus variable entre les élèves d'une école qu'entre les élèves de deux écoles différentes. Il faudra naturellement tenter de vérifier cette hypothèse sur base de l'échantillon. Par ailleurs, étant donné que la contrainte de coût implique de tenir compte aussi bien des villes que des écoles, il peut être tentant de considérer l'échantillonnage comme un échantillonnage à deux degrés, le premier sur les villes, considérées comme des strates, et le deuxième sur les écoles, considérées comme des grappes. Malheureusement, étant donné le grand nombre de villes différentes, il semble difficile de prélever au moins une école dans chaque ville. Il faut donc considérer les écoles comme des grappes et modifier le nombre d'écoles sélectionnées jusqu'à remplir la contrainte de coût. Dans ce cas-ci, 18 écoles ont été sélectionnées, avec une probabilité proportionnelle à leur taille avec un tirage sans remise. La variance de l'estimateur sera donc obtenue par $\frac{M^2}{N^2}(1 - f_{GR})\frac{\sigma^2_{\tau;corr}}{m}$. Il faudra calculer ultérieurement cette valeur sur base de l'échantillon.

Le second échantillon, celui sélectionné à partir de Facebook, est plus simple, il suffit de sélectionner 29000 individus de la population en utilisant un sondage à probabilités égales et sans remise. La variance de l'estimateur sera donc $(1 - f)\frac{\sigma^2_{corr}}{n}$. Dans les deux cas, il faudra tenir compte des non-réponses.

Échantillons

On peut alors s'intéresser aux estimateurs des deux échantillons et à leurs propriétés. L'échantillon "Facebook" étant le plus simple, il parait naturel de commencer par s'intéresser à celui-ci.

On peut commencer par calculer la moyenne de l'argent de poche. Cette moyenne, qu'elle soit calculée à l'aide de la fonction "svymean" tenant compte du sondage ou avec la fonction "mean", est la même. Ce qui n'est pas étonnant, étant donnée la méthode d'échantillonnage utilisée. Cette moyenne est de 22.38. On peut alors calculer l'écart-type de la moyenne. En utilisant la fonction "svymean", celle-ci évalue l'erreur type à 0.062644. En calculant cette valeur explicitement, le résultat est de 0.06264435, ce qui est équivalent en arrondissant. L'intervalle de confiance à 95% pour la moyenne est donc [22.258 ; 22.504]. Finalement, l'incertitude relative est donc de 0.00548.

L'échantillon "Official" est un peu plus complexe. La fonction svymean renvoie une moyenne de 21.18440 et une erreur type de 0.94652. L'erreur semble assez importante et le fait qu'elle soit a priori plus grande que dans le cas d'un simple PESR peut sembler surprenant. Néanmoins, il serait possible que l'hypothèse d'une forte dispersion au sein des écoles soit fausse, et que la variance au sein d'une école soit plus faible qu'entre les différentes écoles. Une manière de vérifier ça est de calculer η^2 en calculant $\frac{\sigma^2_{entre}}{\sigma^2}$ (ou plutôt des estimateurs de la variance corrigée). Cette valeur est a priori de 0.141, mais la valeur très élevée de la moyenne totale semble indiquer une erreur de calcul, en particulier quand celle-ci est comparée avec la valeur fournie dans l'énoncé ou avec la valeur obtenue dans le premier échantillon, qui est plus petite que la valeur de la moyenne "inter" calculée ! Pour vérifier si cette valeur de η^2 est correcte, on peut comparer en utilisant la formule de la variance "inter-grappe" : $\frac{1}{M}\Sigma(\mu_g - \mu)^2$. Si cette valeur n'est naturellement valable que dans le cas de grappes à tailles égales, elle possède l'avantage d'être facile à calculer et de pouvoir nous donner au moins une estimation de η^2. Cette valeur est de 0.12. De nouveau, étant donné que la taille des grappes varie énormément, allant 69 à 527 et on peut donc douter de la validité de cette estimation. Néanmoins, si elle est correcte, la faible valeur de η^2 laisse penser que la variance de l'estimateur devrait être plus petite dans le cas de l'échantillonnage en grappe que dans le cas de l'échantillonne à probabilités égales sans remises. Il est également possible de calculer l'erreur type de l'estimateur explicitement et de calculer cette même valeur dans le cas où il s'agirait d'un PESR. Dans le premier cas, la valeur est de 0.074 et dans le second cas, de 0.134. Il est bien sûr de nouveau possible qu'une erreur de calcul fausse ces valeurs, celles-ci semblent plus probables que les premières si la valeur du rapport de corrélation a été correctement calculée, mais celle-ci ne semble pas correcte si on tient compte de la valeur de la variance pour la population fournie dans l'énoncé. Si l'erreur se trouve dans le calcul de svydesign, l'erreur type est donc de 0.074, si l'erreur se trouve dans mon calcul ultérieur, ce qui est plus probable, l'erreur-type est de 0.94652. La différence entre les deux est assez grande.

En conclusion pour cette première étape, on voit que quelle que soit la valeur de la variance de l'estimateur pour le jeu de données "Official", le jeu de données "Facebook" semble meilleur, étant donné qu'il permet de calculer la moyenne avec une meilleure précision.

Redressement

Cette première partie étant finie, on peut tenter de diminuer la variance de l'estimateur par redressement. Étant donné que l'on possède déjà une série de valeurs sur la population, on peut comparer celles-ci avec les valeurs obtenues dans l'échantillon afin de vérifier qu'elles ne diffèrent pas démesurément.

Dans le cas de Facebook, la moyenne du nombre de séances de cinéma est proche de la moyenne de la population, et l'écart-type de la population ne varie pas énormément non plus. La proportion de fumeurs, si elle n'est pas incroyablement différente est néanmoins 0.55, contre 0.58 pour la population. On peut donc utiliser cette valeur pour essayer de mieux évaluer l'estimateur. La moyenne, après correction serait de 21.4153, ce qui est assez différent de la première valeur trouvée, et se trouve même en dehors de l'intervalle de confiance initial. Cette méthode suppose normalement que la variable "Argent" est proportionnelle à la variable "Fume", et qu'il s'agit donc de deux variables quantitatives, ce qui n'est bien entendu pas le cas. On peut également tenter d'améliorer la précision de l'estimateur de la moyenne par post-stratification. On peut choisir plusieurs variables afin de construire les strates de la population, mais la variable "Pays" semble celle qui permet de réduire au minimum l'imprécision de l'estimateur. Après

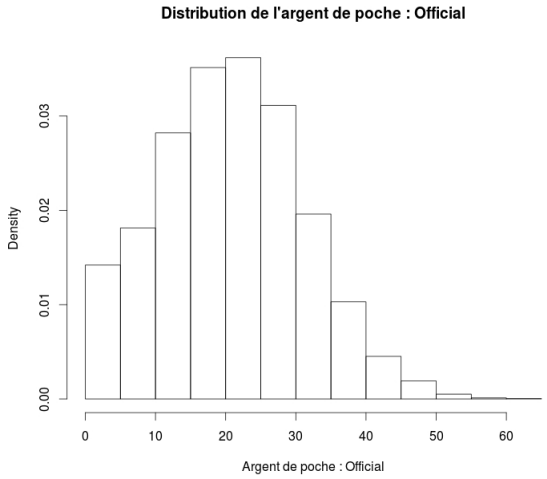
post-stratification, la moyenne est désormais de 22.161 et l'erreur type est de 0.0611. L'incertitude relative est maintenant de 0.00540, ce qui n'est pas considérablement différent de la valeur calculée précédemment, et l'intervalle de confiance à 95% est donc [22.04124 ; 22.28076], ce qui est quand même assez différent.

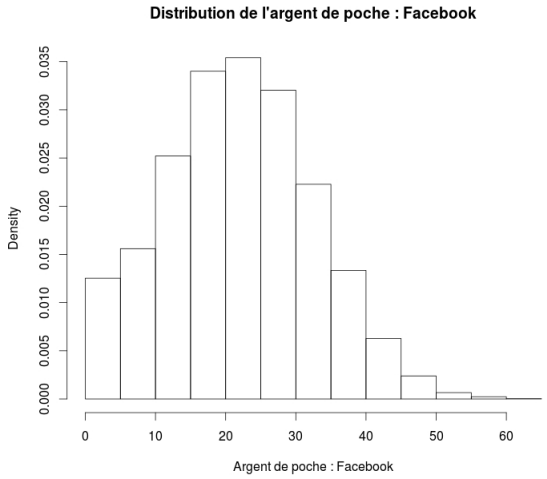
Dans le second cas, comme indiqué précédemment, la valeur calculée de la variance était impossible, on peut donc simplement calculer l'écart-type avec la fonction "sd". On remarque d'ailleurs que si cette valeur de 10.492 avait été utilisée pour calculer η^2, celui-ci aurait été plus grand que un, ce qui aurait été impossible. La valeur de la variance est donc proche de celle fournie dans l'énoncé. La population de fumeur est de 0.598, ce qui n'est pas non plus très éloigné de la valeur fournie dans l'énoncé. La moyenne du nombre de séances de cinéma n'est pas non plus très éloignée. En appliquant la post-stratification par pays à cet échantillon et en supposant qu'il n'y avait pas d'erreur dans l'usage de la fonction "svydesign" pour l'échantillonnage en grappes, on obtient 22.009 comme moyenne, et 0.8623 comme erreur type. Comme ces valeurs se rapprochent de celles obtenues via l'autre échantillon, on peut supposer qu'il n'y avait pas d'erreur dans la fonction et que l'erreur type est simplement importante.

On peut utiliser d'autres méthodes pour le redressement, comme la régression ou le rééchantillonage, mais aucune ne semble réellement utile ici.

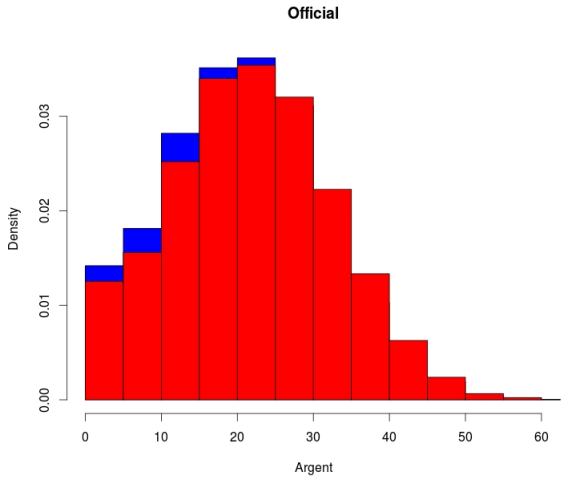
Données manquantes

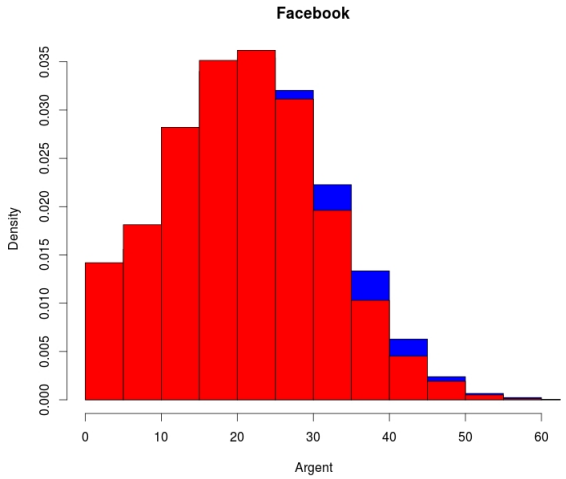
A priori donc, il semblerait que le premier échantillonnage soit le plus efficace. Néanmoins, outre la précision des estimateurs, un autre problème se pose lorsqu'il faut comparer les deux échantillons. En effet, la quantité importante de données manquantes pour le premier échantillon peut dissimuler des informations. Par exemple, on pourrait supposer que les personnes n'ayant pas répondu sur Facebook sont celles qui ont le moins d'argent de poche, ou celles qui ont le plus d'argent de poche. Dans les deux cas, l'estimateur serait faussé, et il n'est pas possible de vérifier ce qu'auraient répondu les individus n'ayant pas répondu. Néanmoins, comme l'on possède également l'information de l'échantillon "Official", on peut comparer la distribution des réponses dans les deux échantillons. Comme le taux de non-réponse dans le second échantillon est bien plus faible, si la distribution est la même dans les deux cas, on peut supposer que l'estimateur n'est pas trop biaisé. Une manière intuitive d'observer les valeurs manquantes pourrait par exemple consister à comparer les histogrammes (en fréquence relative) pour différentes catégorie d'argent de poche. Le premier graphe est celui de Facebook, le second est celui des écoles. À priori les deux semblent assez proches.





Néanmoins si on superpose l'un sur l'autre, on peut voir que les individus ayant moins d'argent de poche sont moins présents dans l'échantillon "Facebook". Dans le premier graphe, la partie en bleu représente la partie plus représentée dans "Official". Dans le second graphe, la partie en bleu représente cette fois ci la partie surreprésentée dans "Facebook".





Le taux de non-réponse dans l'échantillon "Facebook" semble donc susceptible d'affecter la valeur de l'estimateur.

Conclusion

À priori, l'échantillon "Facebook" semble bien meilleur. L'estimation de la moyenne est plus précise, et le coût peut être plus facilement contrôlé qu'en passant par les écoles. Néanmoins, la réalité est plus nuancée. Le taux de non-réponse dans l'échantillon "Facebook" abouti à un estimateur qui semble légèrement biaisé vers le haut. Il faudrait donc, si l'on décide de sélectionner l'échantillon "Facebook", corriger ce biais d'une manière ou d'une autre. Si on préfère éviter ce problème, l'échantillon Official est meilleur. Par ailleurs, même si la variance de celui-ci est moins bonne, elle permet néanmoins d'évaluer l'argent de poche avec une précision qui ne semble somme toute pas si mauvaise.

Annexes

```
#Setup
setwd("/run/media/alexis/6917a5e2-e4da-439f-ada9-c2d93a4db183/alexis/
    stats/sondage/")
rm(list=ls())
load("/run/media/alexis/6917a5e2-e4da-439f-ada9-c2d93a4db183/alexis/
    stats/sondage/FacebookPopu.Rda")
load("/run/media/alexis/6917a5e2-e4da-439f-ada9-c2d93a4db183/alexis/
    stats/sondage/OfficialPopu.Rda")
library(sampling)
library(samplingbook)
library(survey)
library(DBI)
library(dplyr)
dim(FacebookPopu)[1]
set.seed(5)

#Official Cluster?
View(OfficialPopu)
Probasc<-OfficialPopu$NEleve/sum(OfficialPopu$NEleve)
S<-cluster(data = OfficialPopu, clustername = c("NumEcoleUnique"),
    size = 18, method = "srswor", pik=("Probasc"))
Echant6 <- getdata(OfficialPopu, S)
table(Echant6$Ville, Echant6$Pays)
View(Echant6)
dim(Echant6)
save(Echant6, file = "Echantillon_official.Rda")
write.table(Echant6$NumEcoleUnique, "AlexisBuckens_Official.txt", sep
    ="\t", row.names = F, quote = F, col.names = F)
test<-read.table("Echantillon_officiel.txt", sep="\t")


#Facebook PESR
S<-srswor(29000, dim(FacebookPopu)[1])
Echantillon<-FacebookPopu[S!=0,]
View(Echantillon)
dim(Echantillon)
save(Echantillon, file = "Echantillon_facebook.Rda")
write.table(Echantillon$NumEleveUnique, "AlexisBuckens_Facebook.txt",
     sep="\t", row.names = F, quote = F, col.names = F)
length(Echantillon$NumEleveUnique)

## Seconde partie
load("/run/media/alexis/6917a5e2-e4da-439f-ada9-c2d93a4db183/alexis/
    stats/sondage/AlexisBuckens_FacebookSample.Rda")
```

```
load("/run/media/alexis/6917a5e2-e4da-439f-ada9-c2d93a4db183/alexis/
    stats/sondage/AlexisBuckens_OfficialSample.Rda")
### Facebook
## Facebook: Observation et structure
str(FacebookSample)
FacebookSample$Sexe<-as.factor(FacebookSample$Sexe)
FacebookSample$Fume<-as.factor(FacebookSample$Fume)
FacebookSample$Pays<-as.factor(FacebookSample$Pays)
FacebookSample$Ville<-as.factor(FacebookSample$Ville)
FacebookSample$Annee<-as.factor(FacebookSample$Annee)
str(FacebookSample)
View(FacebookSample)
##design
FacebookSample$Prob<-dim(FacebookSample)[1]/dim(FacebookPopu)[1]
Data1 <- svydesign( ids=~1 , fpc=~Prob , data=FacebookSample )
m<-mean(FacebookSample$Argent)
svymean(~Argent, design=Data1, deff=T)
##Calcul de SE
f<-dim(FacebookSample)[1]/dim(FacebookPopu)[1]
sc<-sd(FacebookSample$Argent)^2
n<-dim(FacebookSample)[1]
var<-(1-f)*(sc/n)
se<-sqrt(var)
## Intervalle de confiance et incertitude relative
up<-m+(1.96*se)
lo<-m-(1.96*se)
1.96*se/m



### Official
#officialSample Observation et structure
str(OfficialSample)
OfficialSample$Pays<-as.factor(OfficialSample$Pays)
OfficialSample$Ville<-as.factor(OfficialSample$Ville)
OfficialSample$Ecole<-as.factor(OfficialSample$Ecole)
OfficialSample$Sexe<-as.factor(OfficialSample$Sexe)
OfficialSample$Fume<-as.factor(OfficialSample$Fume)
OfficialSample$Cinema<-as.integer(OfficialSample$Cinema)
str(OfficialSample)
View(OfficialSample)
#officialSample Design survey
tail(OfficialPopu)
OfficialSample$M<-dim(OfficialPopu)[1]
#Prob<-(OfficialSample$NEleve/(sum(OfficialPopu$NEleve)*sum(
    OfficialSample$Prob)))
#N<-sum(OfficialPopu$NEleve)
Datao<-svydesign(ids=~NumEcoleUnique, data=OfficialSample, fpc = ~M)
Datao
```

```
svymean(~Argent, design=Datao, deff=T)
OfficialSample$propor<-(sum(OfficialPopu$NEleve))
Datao<-svydesign(ids=~NumEcoleUnique, data=OfficialSample, fpc = ~
    propor)
Datao
svymean(~Argent, design=Datao, deff=T)
## Calcul de la variance intra et inter ---- NON
Off<-list(Off)
Off1<-filter(OfficialSample, NumEcoleUnique==17)$Argent
Off2<-filter(OfficialSample, NumEcoleUnique==25)$Argent
Off3<-filter(OfficialSample, NumEcoleUnique==27)$Argent
Off4<-filter(OfficialSample, NumEcoleUnique==30)$Argent
Off5<-filter(OfficialSample, NumEcoleUnique==36)$Argent
Off6<-filter(OfficialSample, NumEcoleUnique==46)$Argent
Off7<-filter(OfficialSample, NumEcoleUnique==54)$Argent
Off8<-filter(OfficialSample, NumEcoleUnique==57)$Argent
Off9<-filter(OfficialSample, NumEcoleUnique==65)$Argent
Off10<-filter(OfficialSample, NumEcoleUnique==68)$Argent
Off11<-filter(OfficialSample, NumEcoleUnique==70)$Argent
Off12<-filter(OfficialSample, NumEcoleUnique==89)$Argent
Off13<-filter(OfficialSample, NumEcoleUnique==99)$Argent
Off14<-filter(OfficialSample, NumEcoleUnique==110)$Argent
Off15<-filter(OfficialSample, NumEcoleUnique==111)$Argent
Off16<-filter(OfficialSample, NumEcoleUnique==112)$Argent
Off17<-filter(OfficialSample, NumEcoleUnique==145)$Argent
Off18<-filter(OfficialSample, NumEcoleUnique==156)$Argent
Off<-list(Off1, Off2, Off3, Off4, Off5, Off6, Off7, Off8, Off9, Off10
    , Off11, Off12, Off13, Off14, Off15, Off16, Off17, Off18)
rm(Off1, Off2, Off3, Off4, Off5, Off6, Off7, Off8, Off9, Off10, Off11
    , Off12, Off13, Off14, Off15, Off16, Off17, Off18)

tot<-rep(0,18)
for(i in 1:18){
  tot[i]<-var(Off[[i]])*length(Off[[i]])
  }
total<-sum(tot)/dim(OfficialSample)[1]

moy<-mean(OfficialSample$Argent)
moyt<-rep(0,18)
for(i in 1:18){
  moyt[i]<-(mean(Off[[i]])-moy)^2*(length(Off[[i]])/dim(
      OfficialSample)[1])
}
inter<-sum(moyt)
inter/(total)
#Si on avait GRTE
moyt<-rep(0,18)
for(i in 1:18){
  moyt[i]<-(mean(Off[[i]])-moy)^2
```

```
}
inter<-(sum(moyt)/dim(OfficialPopu)[1])
inter/sd(OfficialSample$Argent)
#GRTE? Non
ng<-rep(0,18)
for(i in 1:18){
  ng[i]<-length(Off[[i]])
}
min(ng)
max(ng)
## Erreur de l'estimateur, recalcule
sc<-sd(OfficialSample$Argent)^2/17
fac<-(dim(OfficialPopu)[1]^2)/(dim(OfficialSample)[1]^2)
facb<-(1-18/dim(OfficialPopu)[1])
var<-sc*fac*facb
se<-sqrt(var)
#intervalle de confiance
up<-m+(1.96*se)
lo<-m-(1.96*se)
1.96*se/m
#Comparaison avec swsor
f<-dim(OfficialSample)[1]/sum(OfficialPopu$NEleve)
sc<-sd(OfficialSample$Argent)^2
n<-dim(OfficialSample)[1]
var<-(1-f)*(sc/n)
sqrt(var)

#Redressement : Facebook
#correction des infos : Fumeurs: 58.46%, Cinema:3.11seances, ecart-
    type de population=10.76
sd(FacebookSample$Argent)
sum(FacebookSample$Fume==1)/(sum(FacebookSample$Fume==1)+sum(
    FacebookSample$Fume==0))
mean(FacebookSample$Cinema)
propo<-sum(FacebookSample$Fume==1)/(sum(FacebookSample$Fume==1)+sum(
    FacebookSample$Fume==0))
m2<-mean(FacebookSample$Argent)*propo/0.5846
m2
mean(filter(FacebookSample, Fume==1)$Argent)
mean(filter(FacebookSample, Fume==0)$Argent)
cor(FacebookSample$Argent, FacebookSample$Fume, method = 'spearman')

#Poststrata - Pays
eff<-c(0,0,0)
eff[1]<-dim(filter(FacebookPopu, Pays==1))[1]
eff[2]<-dim(filter(FacebookPopu, Pays==2))[1]
eff[3]<-dim(filter(FacebookPopu, Pays==3))[1]
eff
Effectifs <- data.frame(Pays=1:3, Freq=eff)
```

```
design2<-postStratify(design= Data1, strata= ~Pays, population =
    Effectifs)
#poststrata-Age
eff<-rep(0,7)
for(i in 1:7){
  eff[i]<-dim(filter(FacebookPopu, Age==(i+12)))[1]
}
eff
Effectifs <- data.frame(Age=13:19, Freq=eff)
design3<-postStratify(design= Data1, strata= ~Age, population =
    Effectifs)
#poststrata-Sexe
eff<-c(0,0)
eff[1]<-dim(filter(FacebookPopu, Sexe==0))[1]
eff[2]<-dim(filter(FacebookPopu, Sexe==1))[1]
eff
Effectifs <- data.frame(Sexe=c(0,1), Freq=eff)
design4<-postStratify(design= Data1, strata= ~Sexe, population =
    Effectifs)
#comparaison
svymean(~Argent, design4)
svymean(~Argent, design3)
svymean(~Argent, design2)
svymean(~Argent, design=Data1)
se<-0.0611
m<-22.161
## Intervalle de confiance et incertitude relative
up<-m+(1.96*se)
lo<-m-(1.96*se)
1.96*se/m

#Redressement : Official
#correction des infos : Fumeurs: 58.46%, Cinema:3.11seances, ecart-
    type de population=10.76
sd(OfficialSample$Argent)
sum(OfficialSample$Fume==1)/(sum(OfficialSample$Fume==1)+sum(
    OfficialSample$Fume==0))
svymean(~Cinema, design=Datao, deff=T)
#Poststrata - Pays
eff<-c(0,0,0)
eff[1]<-dim(filter(FacebookPopu, Pays==1))[1]
eff[2]<-dim(filter(FacebookPopu, Pays==2))[1]
eff[3]<-dim(filter(FacebookPopu, Pays==3))[1]
eff
Effectifs <- data.frame(Pays=1:3, Freq=eff)
designo2<-postStratify(design= Datao, strata= ~Pays, population =
    Effectifs)
svymean(~Argent, designo2)
svymean(~Argent, Datao, Deff=T)
```

```
#reech
designo2.rep <- as.svrepdesign(designo2)
svymean(~Argent, designo2.rep)

#comparaison distributions
hist(OfficialSample$Argent, freq = F, xlab = "Argent de poche : 
    Official", main = "Distribution de l'argent de poche : Official")
hist(FacebookSample$Argent, freq = F, xlab = "Argent de poche : 
    Facebook", main = "Distribution de l'argent de poche : Facebook")
p1 <- hist(OfficialSample$Argent, freq = F)
p2 <- hist(FacebookSample$Argent, freq = F)
plot( p1, col='blue', xlim=c(0,60), freq = F, main="Official", xlab="
    Argent")  # first histogram
plot( p2, col='red', add=T, freq=F)  # second
plot( p2, col='blue', xlim=c(0,60), freq = F, main="Facebook", xlab="
    Argent")  # first histogram
plot( p1, col='red', add=T, freq=F)  # second
```