SAE – ÉCHANTILLONAGE ET ESTIMATION

Théo PETIT – Gabriel SAGOT

25/05/25

Introduction

Dans le cadre de la SAÉ Échantillonnage et Estimation, nous avons travaillé sur les données des communes françaises, en ciblant plus particulièrement la région Île-de-France. L'objectif principal de ce travail était d'estimer la population totale de cette région à partir d'un échantillon, en recherchant une méthode permettant d'obtenir l'estimation la plus précise possible. Pour ce faire, nous avons utilisé le langage de programmation R et structuré notre démarche en deux approches distinctes :

- Une première utilisant un échantillonnage aléatoire simple (SAS),
- Une seconde reposant sur un échantillonnage stratifié, dans lequel les communes sont regroupées selon leur population.

Cette comparaison nous a permis de mieux comprendre l'intérêt des strates pour réduire la variabilité des estimations, en assurant une meilleure représentativité des données.

Dans une deuxième partie, nous avons analysé les données d'une enquête sur la pratique sportive des étudiants. L'objectif était ici d'identifier les variables qualitatives (comme le tabagisme, l'alimentation ou encore la santé perçue) qui présentent une relation statistiquement significative avec la pratique du sport. Pour cela, nous avons réalisé des tests du khi² d'indépendance, complétés par le calcul du V de Cramer afin d'évaluer la force des relations détectées.

Ce rapport présente les deux volets de notre travail en détaillant notre méthodologie, nos traitements R et nos résultats, accompagnés de commentaires et de représentations graphiques pour en faciliter la lecture et l'analyse.

Estimation de la population de l'Île-de-France

1.1 Estimation par sondage aléatoire simple

Nous avons commencé par filtrer les données de la région Île-de-France et avons extrait les colonnes pertinentes : code département, nom de la commune, et population totale.

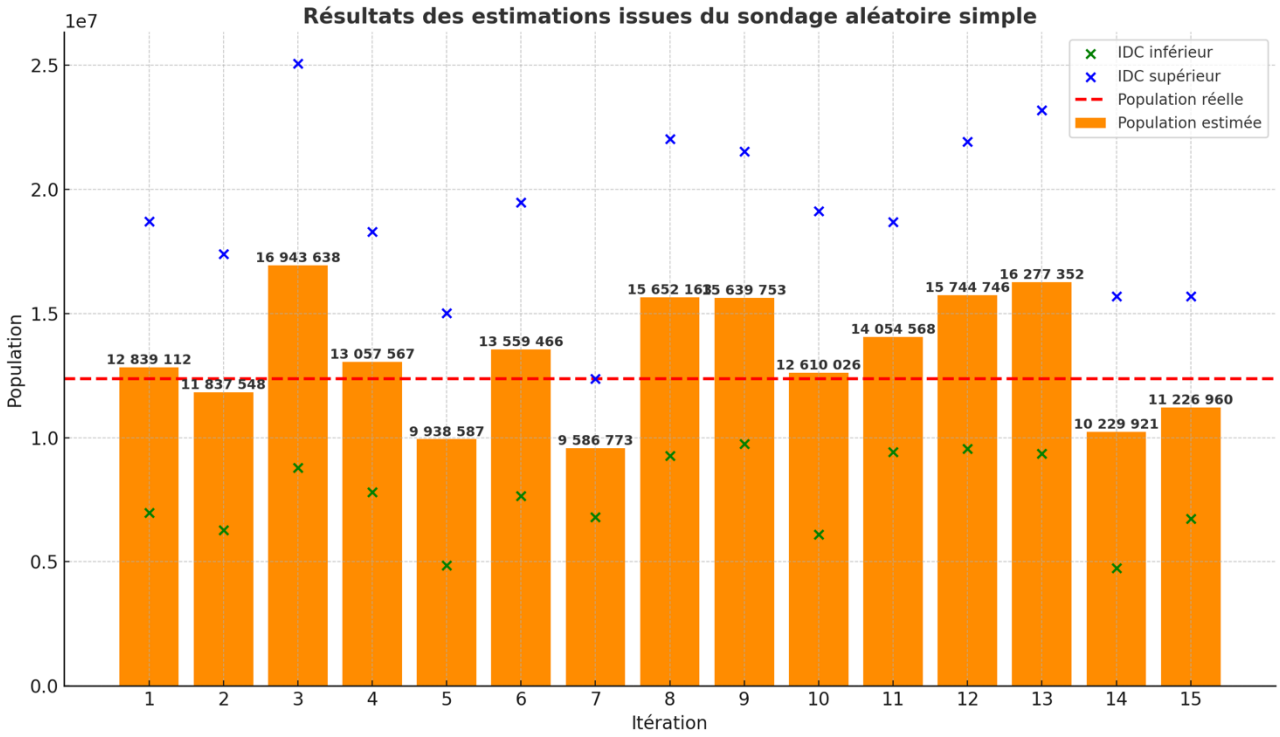
```
28
29  # question 1
30  #-----------#
31  donnees <- table[table$Nom.de.la.région == "Île-de-France", c("Code.département","Commune","Population.totale")]
32  donnees<- unique(donnees)
33  head(donnees)
34
```

```r
# question 4
#-----------#
n=100
E=sample(U, n,replace=FALSE)
head(E)

# question 5
#-----------#
donnees1= donnees[donnees$Commune %in% E, ]
head(donnees1)

# question 6 (moyenne de l'échantillon)
#--------------------------------------#
moy= mean(donnees1$Population.totale)
moy
# idc de mu
idcmoy = t.test(donnees1$Population.totale)$conf.int
idcmoy

# question 7 (Nbre d'habitants total estimé)
#-------------------------------------------#
Test = N*moy
Test
# IDC de T
idcT = idcmoy*N
idcT
#Marge d'erreur
marge=(idcT[2]-idcT[1])/2
marge
```

Un échantillon aléatoire simple de 100 communes a ensuite été tiré à l'aide de la fonction `sample`, et une estimation de la moyenne de population par commune a été obtenue. Le total estimé a été obtenu en multipliant cette moyenne par le nombre total de communes.

Nous avons ensuite répété cette opération 15 fois pour observer la variabilité des estimations.

Le graphique montre une dispersion importante des estimations autour de la population réelle. Les marges d'erreur sont souvent larges, et la population réelle n'est pas toujours incluse dans l'intervalle de confiance. Cela souligne les limites de cette méthode lorsque l'échantillon n'est pas représentatif.



1.2 Estimation par sondage stratifié

Pour améliorer la précision, nous avons réparti les communes en 4 strates selon les quartiles de la population. Chaque strate contient 25 communes (échantillonnage proportionnel).

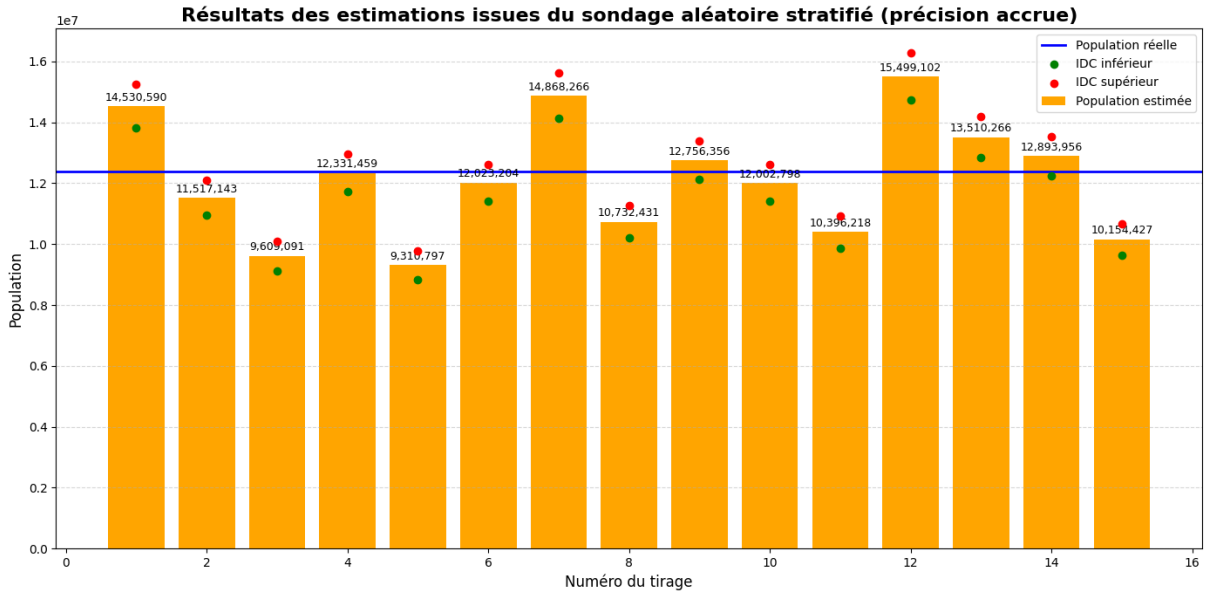
```r

# Question 1 (quartiles)
#----------------------#
summary(donnees$Population.totale)

# Question 2 (strates)
#----------------------#
datastrat=donnees
datastrat$Strate=cut(datastrat$Population.totale, breaks=c(0,569, 1444, 7367, 1000000), labels=c(1,2,3,4))
head(datastrat)
table(datastrat)
```

Ces quatre strates ont considérablement réduit les marges d'erreur, car elles ont permis de constituer un échantillon bien plus représentatif. En effet, au lieu de sélectionner 100 communes dans l'ensemble global, nous en prélevons désormais 25 dans chacune des strates.

```r
# Question 3 : Effectif des strates
#--------------------------------#

# Charger le package
if (!require(sampling)) install.packages("sampling")
library(sampling)

# Trier les données par strate
data <- datastrat[order(datastrat$Strate), ]

# Taille totale de la population
N <- nrow(datastrat)

# Effectifs de chaque strate
Nh <- table(datastrat$Strate)

# Définir nh = 25 pour chaque strate
n <- 100
nh <- rep(25, length(Nh))
names(nh) <- names(Nh)

# Sondage stratifié sans remise
st <- strata(datastrat, stratanames = c("Strate"), size = nh, method = "srswor")

# Obtenir les données échantillonnées
data1 <- getdata(datastrat, st)
data1 <- data1[order(data1$Strate), ]

# Poids des strates (gh)
gh <- Nh / N

# Taux de sondage dans les strates (fh)
fh <- nh / as.numeric(Nh)
```

```r
# Question 4
#-------------#
# Mise en place des 4 échantillons
ech1=data1[data1$Strate==1, ]
ech2=data1[data1$Strate==2, ]
ech3=data1[data1$Strate==3, ]
ech4=data1[data1$Strate==4, ]

# Moyennes des 4 échantillons
m1=mean(ech1$Population.totale)
m2=mean(ech2$Population.totale)
m3=mean(ech3$Population.totale)
m4=mean(ech4$Population.totale)

# Variances des 4 échantillons
var1=var(ech1$Population.totale)
var2=var(ech2$Population.totale)
var3=var(ech3$Population.totale)
var4=var(ech4$Population.totale)
```

Une répétition sur 15 échantillons a été réalisée et stockée dans un tableau similaire au précédent. Les résultats montrent une bien meilleure stabilité des estimations. Les marges d'erreur sont plus faibles, et les estimations sont plus proches de la valeur réelle. (voir graphique ci-dessous).



Le graphique montre une concentration des estimations autour de la population réelle. Les intervalles de confiance sont nettement plus resserrés que dans un sondage simple, et la population réelle est presque toujours incluse dans ces intervalles. Cela met en évidence l'efficacité du sondage stratifié pour améliorer la précision des estimations lorsque les strates sont bien définies.

`Conclusion

Dans cette partie, nous avons mobilisé le logiciel R pour explorer les techniques d'échantillonnage et d'estimation, notamment à travers des simulations de sondage aléatoire simple et stratifié. R nous a permis de manipuler efficacement des données réelles, de réaliser des tirages aléatoires reproductibles, de calculer des estimations et des intervalles de confiance, mais aussi de visualiser clairement les résultats grâce à des graphiques personnalisés.

Nous avons ainsi observé que le sondage aléatoire stratifié permet une meilleure précision que le sondage simple, en réduisant les marges d'erreur, ce qui se traduit visuellement par des intervalles de confiance plus resserrés autour de la population réelle.

Sur le plan technique, cela nous a permis de :

- Renforcer notre maîtrise des fonctions de simulation et de tirage aléatoire dans R (sample(), replicate(), etc.)
- Automatiser des calculs statistiques (moyenne pondérée, erreurs standard, IDC)
- Personnaliser nos visualisations avec ggplot2 ou matplotlib, en respectant une charte graphique claire.

En somme, cette étape nous a permis de mieux comprendre l'impact des méthodes d'échantillonnage sur la qualité des estimations, tout en consolidant nos compétences en programmation statistique avec R.

Enquête Étudiant 2024

2. Traitement de données d'enquête

Résultat du V de Cramer par variable

Variables	V de Cramer	Interprétation
Fumeur	0.309561	Lien modéré
Alimentation	0.213487	Lien faible à modéré
Santé	0.178246	Lien faible à modéré

La variable "Fumeur" présente le V de Cramer le plus élevé (≈ 0.31), ce qui indique qu'elle entretient la liaison la plus forte avec la pratique sportive parmi les trois variables étudiées. Ce résultat suggère que le statut de fumeur influence significativement les habitudes sportives des étudiants : ceux qui pratiquent régulièrement un sport sont plus souvent non-fumeurs.

Les variables "Alimentation" et "Santé" présentent aussi des relations significatives mais moins marquées, avec des V de Cramer respectifs de 0.21 et 0.17. Cela peut s'expliquer par des comportements de santé globalement cohérents : les étudiants qui font attention à leur forme physique déclarent également avoir une alimentation équilibrée ou une meilleure perception de leur santé.

Conclusion

Cette analyse montre que certaines habitudes de vie influencent significativement la pratique sportive des étudiants. Le lien le plus fort est observé avec le tabagisme, suivi par l'alimentation et la perception de la santé. Cela met en évidence des profils d'étudiants plus ou moins sensibles à leur hygiène de vie, notamment dans un contexte urbain comme Paris, où les rythmes de vie peuvent varier fortement selon les habitudes individuelles. Enfin, cette étape a permis d'utiliser R pour automatiser les analyses statistiques et quantifier rigoureusement les relations entre variables qualitatives. L'interprétation du V de Cramer nous a permis de dépasser la simple significativité, en apportant une mesure concrète de l'intensité des liaisons.