Région
PAYS DE LA LOIRE

BUT Science des données Niort

SAE : Estimation par échantillonnage

Pays de la Loire

Jérémie NDJOYE **Antonin VION**

INTRODUCTION

L'objectif de cette étude est de comprendre comment estimer avec précision le nombre d'habitants d'une population (Pays de la Loire) malgré l'incertitude, en utilisant un intervalle de confiance construit à partir d'un processus d'échantillonnage.

L'étude sera divisée en deux parties. Dans un premier temps, nous utiliserons un sondage aléatoire simple à probabilité égale, où chaque individu a le même poids.

Dans un second temps, nous mettrons en œuvre un sondage par stratification (ou échantillonnage stratifié).

Enfin, nous comparerons les résultats obtenus afin de déterminer la méthode la plus efficace pour l'estimation de la population.

Echantillonnage aléatoire simple

Avant de pouvoir commencer l'analyse des données, il a d'abord fallu les préparer. Pour cela, nous avons commencé par supprimer les premières lignes du fichier Excel mis à notre disposition, puis nous l'avons importé au format CSV dans R.

Une fois les données importées, nous les avons stockées dans un data frame, puis nous avons sélectionné les colonnes pertinentes (Code département, commune, population totale) ainsi que les informations relatives à la région « *Pays de la Loire* ».

Il a également été nécessaire de supprimer les espaces inutiles et de convertir les données concernant la population totale au format numérique, afin de pouvoir les exploiter correctement.

```
# Importation des données
df <- read.csv2("population_francaise_communes.csv")
donnees <- df[df["Nom.de.la.région"]== "Pays de la Loire",c("Code.département","Commune","Population.totale")]
donnees$Population.totale <- as.numeric(gsub(" ","",donnees$Population.totale))


# Affichage des 6 premières lignes
head(donnees)
```

La variable U représente l'ensemble de la population mère, c'est-à-dire toutes les communes de la région Pays de la Loire. La variable T correspond au nombre total d'habitants dans cette population, soit (3922846).

Un échantillon aléatoire simple de 100 communes est tiré à l'aide de la fonction sample (), afin de constituer une base de travail représentative. Les informations relatives aux communes sélectionnées sont ensuite regroupées dans un nouveau data frame, nommé donnees2, qui contient deux variables : le nom de la commune et sa population totale.

```r
# Affichage des 6 premières lignes
head(donnees)


#Chargement de l'ensemble des  communes
U <- donnees$Commune
head(U)

# Nombre total de communes
N <- length(U)
N

# Nombre T total d'habitants en Pays de la Loire
T <- sum(donnees$Population.totale)
T

# Tirage aléatoire simple
n = 100
E <- sample(U,n)
head(E)
```

Différents indicateurs statistiques ont été calculés à partir de l'échantillon. Nous avons notamment déterminé la moyenne de la population des communes tirées, ainsi qu'un intervalle de confiance à 95 % pour estimer le nombre moyen d'habitants par commune dans l'ensemble de la population.

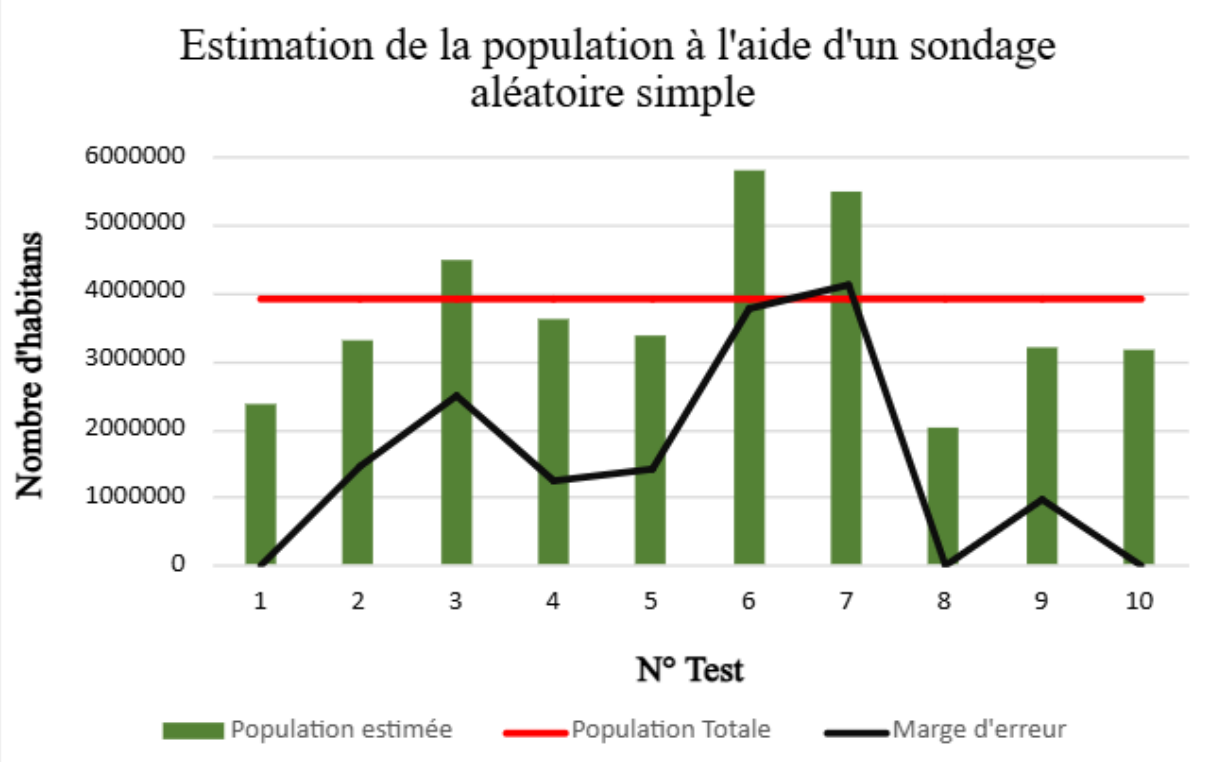
À partir de cette moyenne, nous avons également obtenu une estimation du total de la population (T) pour l'ensemble des communes, accompagnée de son intervalle de confiance à 95 %, ce qui permet d'encadrer l'incertitude de cette estimation.

```r
# Extraction des données correspondant aux 100 communes tirées
donnees1 <- donnees[donnees$Commune %in% E, ]
head(donnees1)  # Affichage des 6 premières lignes

# Création d'un nouveau DataFrame avec uniquement la commune et la population
donnees2 <- subset(donnees1, select = c(Commune, Population.totale))
head(donnees2)

# Calcul de la moyenne de la population sur l'échantillon
xbar <- mean(donnees2$Population.totale)
xbar

# Calcul d'un intervalle de confiance à 95 % pour la moyenne (mu)
idcmoy <- t.test(donnees2$Population.totale)$conf.int
idcmoy

# Estimation de la population totale du Pays de la Loire à partir de l'échantillon
T_test <- N * xbar
T_test

# Intervalle de confiance pour l'estimation de la population totale
idcT <- idcmoy * N
idcT

# Calcul de la marge d'erreur associée à l'estimation
marge = (idcT[2] - idcT[1]) / 2
marge
```

Afin d'évaluer la pertinence et l'efficacité de cette méthode d'échantillonnage, le tirage aléatoire ainsi que les calculs d'estimation de la population totale des Pays de la Loire ont été répétés 10 fois.

Ces 10 répétitions ont permis de constituer un tableau de résultats, récapitulant les différentes estimations obtenues, ainsi qu'un graphique illustrant la variabilité des estimations du total de la population.

	Population Totale	Population estimée	IDC		Marge d'erreur
5	3922846	2349438	1741318	2957557	608119.3
6	3922846	3299162	1843309	4755014	1455852
7	3922846	4492288	1976339	7008237	2515949
8	3922846	3606616	2344108	4869123	1262508
9	3922846	3358809	1933908	4783709	1424901
0	3922846	5800601	2000314	9600889	3800287
1	3922846	5476521	1338013	9615029	4138508
2	3922846	2014137	1603321	2424952	410815.5
3	3922846	3209386	2235745	4183027	973641
4	3922846	3175344	2400980	3949708	774364.3



Bilan des résultats

La méthode d'échantillonnage aléatoire simple s'est révélée peu précise et globalement approximative, comme le montrent les résultats obtenus. Les estimations du total de la population étaient souvent très éloignées de la valeur réelle, atteignant parfois des valeurs maximales ou minimales aberrantes.

Les marges d'erreur associées à ces estimations étaient également particulièrement élevées, soulignant les limites de cette méthode lorsqu'elle est appliquée sans ajustement. Ces écarts s'expliquent en partie par le fait que la taille ou le poids des communes dans la population totale n'a pas été pris en compte. Ainsi, la méthode d'échantillonnage aléatoire simple n'est pas la plus adaptée dans ce contexte, où une forte hétérogénéité des unités statistiques existe.

Echantillonnage aléatoire stratifié

Afin d'améliorer la précision des estimations obtenues précédemment, nous passons désormais à un échantillonnage aléatoire stratifié en prenant en compte cette fois-ci les zones peu peuplées et celles très fortement peuplées. Pour cela, nous utilisons toujours le même jeu de données portant sur les communes des Pays de la Loire.

```
library(sampling)  # Outils d'échantillonnage

# 1. Création des strates à partir des quartiles de population

summary(donnees$Population.totale)  # repère les bornes 525, 1 173 et 2 719

# Découpe en 4 strates :
#   < 525 | 525–1 173 | 1 173–2 719 | > 2 719 habitants
donnees$strate <- cut(donnees$Population.totale,breaks  = c(0, 525, 1173, 2719, Inf),labels  = 1:4)

# Jeu de données réduit aux colonnes utiles
donneesstrat <- donnees[, c("Commune", "Population.totale", "strate")]
head(donneesstrat)
```

Dans un premier temps, la fonction summary() est utilisée pour déterminer les quartiles de la variable *Population.totale*. Ces quartiles servent ensuite à définir quatre strates, correspondant aux quatre classes d'effectifs croissants de population.

Cette stratification permet de regrouper les communes selon leur taille démographique, rendant l'échantillonnage plus représentatif et limitant les écarts extrêmes observés lors de l'échantillonnage aléatoire simple.

```
#2. Tirer, selon un sondage stratifié, un echantillon E de taille n = 100 de communes,
#en prenant des effectifs dans les strates proportionnels aux poids des strates.
data = donneesstrat[order(donneesstrat$strate), ]
head(data)

# effectif des strates
Nh=table(data$strate)
Nh
N=sum(Nh)
N

# Poids des strates
gh=Nh/N
gh

# Tirage d'un echantillon stratifié de taille n=100
n=100
nh=round(c(n*Nh[1]/N, n*Nh[2]/N, n*Nh[3]/N, n*Nh[4]/N))
nh

# taux de sondage dans les strates (on peut assimiler ce tirage sans remise à un tirage avec remise)
fh=nh/Nh
fh
# sondage strat (sans remise dans les strates)
st = strata(data, stratanames = c("strate"), size = nh, method = "srswr")
data1=getdata(data, st)
head(data1)
length((data1$Commune))
```

Une fois les strates définies, nous trions les données par strate afin de faciliter le tirage. Nous calculons ensuite l'effectif total de chaque strate (Nh), ce qui nous permet de connaître la taille totale de la population (N) ainsi que la part relative de chaque strate dans l'ensemble (gh).
Un échantillon de 100 communes est ensuite tiré, en respectant une allocation proportionnelle : le nombre de communes sélectionnées dans chaque strate (nh) est proportionnel à la taille de la strate dans la population.
Le tirage est effectué de manière aléatoire avec remise à l'intérieur de chaque strate (méthode srswr). L'échantillon final est stocké dans l'objet data1 et contient exactement 100 communes, réparties selon les proportions définies.

```r
# Définir les 4 sous échantillons obtenus.
ech1=data1[data1$strate==1, ]
ech1
ech2=data1[data1$strate==2, ]
ech3=data1[data1$strate==3, ]
ech4=data1[data1$strate==4, ]

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

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

Constitution des 4 sous échantillons à partir de l'échantillon des 100 communes tirées et calcul de leurs moyennes et variances. (Explication dans le paragraphe ci-dessous).

```r
#Estimation X barre
#st du nombre d'habitants moyen μ et une estimation de la variance de X barre


# Moyenne des 4 echant reunis
Xbarst= (Nh[1]*m1 + Nh[2]*m2 + Nh[3]*m3 + Nh[4]*m4)/N
# estimation de la variance de Xbarst
varXbarst= ((gh[1])^2)*(1-fh[1])*var1/(nh[1]) + ((gh[2])^2)*(1-fh[2])*var2/(nh[2]) +
  ((gh[3])^2)*(1-fh[3])*var3/(nh[3]) + ((gh[4])^2)*(1-fh[4])*var4/(nh[4])

# idc pour mu à 95%
alpha=0.05
binf = Xbarst - qnorm(1-alpha/2)*sqrt(varXbarst)
bsup = Xbarst + qnorm(1-alpha/2)*sqrt(varXbarst)
idcmoy=c(binf, bsup)

#6.Estimation Tstr du nombre total d'habitants T, ainsi qu'un IDC pour T et sa marge d'erreur.

Tstr= N*Xbarst
Tstr
# Estimation par IDC du total T
binf = idcmoy[1]*N
bsup= idcmoy[2]*N

idcT=c(binf, bsup)
idcT

 # marge d'erreur
marge=(idcT[2]-idcT[1])/2
marge
```

À partir de l'échantillon stratifié obtenu, nous constituons quatre sous-échantillons correspondant à chaque strate. Pour chacun, nous calculons la moyenne et la variance du nombre d'habitants. Ces valeurs permettent ensuite d'estimer la moyenne globale de la population des communes ($\bar{X}st$) en pondérant chaque moyenne par le poids de sa strate dans la population.

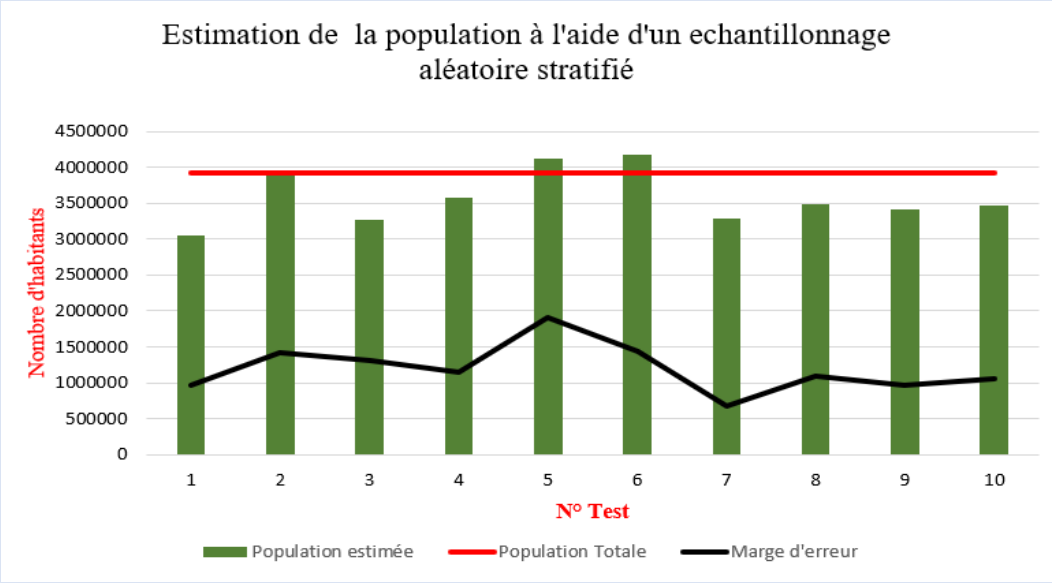
Nous estimons également la variance de cette moyenne, ce qui permet de construire un intervalle de confiance à 95 % pour le nombre moyen d'habitants par commune.

Enfin, en multipliant cette moyenne estimée par le nombre total de communes, nous obtenons une estimation du nombre total d'habitants dans la région. À cela s'ajoute un intervalle de confiance pour cette estimation (idcT) ainsi que la marge d'erreur associée.

Afin d'évaluer la pertinence et l'efficacité de cette méthode d'échantillonnage, le tirage aléatoire ainsi que les calculs d'estimation de la population totale des Pays de la Loire ont été répétés 10 fois.

Ces 10 répétitions ont permis de constituer un tableau de résultats, récapitulant les différentes estimations obtenues, ainsi qu'un graphique illustrant la variabilité des estimations du total de la population.

Population Totale	Population estimée	IDC		Marge d'erreur
3922846	3052902	2093370	4012433	959531,5
3922846	3893707	2480496	5306918	1413211
3922846	3264779	1956787	4572771	1307992
3922846	3579055	2439147	4718963	1139908
3922846	4116118	2199785	6032451	1916333
3922846	4170180	2726343	5614017	1443837
3922846	3293166	2610527	3975805	682639
3922846	3482481	2396701	4568260	1085779,5
3922846	3408648	2435664	4381632	972984
3922846	3475667	2412440	4538893	1063226,5



Les résultats obtenus à l'aide de la méthode d'échantillonnage stratifié sont nettement plus satisfaisants que ceux issus de l'échantillonnage aléatoire simple. En répartissant les communes du Pays de la Loire en quatre strates, définies selon leur population totale (à partir des quartiles), nous avons pu mieux représenter la diversité démographique de la région. Cette stratification a permis de réduire l'effet des communes atypiques (très peu ou très fortement peuplées) sur les estimations, et ainsi d'obtenir des valeurs beaucoup plus proches de la population réelle.

Après avoir comparé les deux méthodes, il apparaît clairement que l'échantillonnage stratifié est préférable lorsqu'on cherche à produire des estimations plus fiables et précises. Il permet de limiter les biais liés à une répartition inégale de la population et améliore la précision des intervalles de confiance.

Cependant, cette méthode pourrait encore être optimisée. Par exemple, augmenter le nombre de strates (passer de 4 à 5 ou 6) permettrait une classification plus fine des communes selon leur taille, ce qui renforcerait la représentativité de chaque groupe. De plus, on pourrait envisager une allocation optimisée (plutôt que proportionnelle) du nombre d'échantillons par strate, en tenant compte non seulement du poids des strates mais aussi de leur variabilité interne.

Au cours de cette partie de la SAÉ, nous avons appris à mettre en œuvre deux méthodes d'échantillonnage : l'échantillonnage aléatoire simple et l'échantillonnage stratifié. Cela nous a permis de mieux comprendre les enjeux liés à la représentativité d'un échantillon dans une étude statistique. Nous avons vu que la méthode aléatoire simple, bien que facile à appliquer, peut produire des résultats très variables et parfois peu fiables. En revanche, l'échantillonnage stratifié, en tenant compte de la structure de la population, permet d'améliorer considérablement la précision des estimations. Cette comparaison nous a permis de développer notre esprit critique face aux choix méthodologiques dans un sondage, et d'acquérir des compétences concrètes en programmation R pour la mise en œuvre de ces techniques.

Traitement de données d'enquête

Pour cette partie, nous allons chercher à identifier des relations significatives entre la variable « sport » et les autres variables qualitatives. Pour cela, nous avons importé notre fichier Excel contenant les réponses de l'enquête dans R afin d'analyser et de calculer les différentes relations.

```
# Chargement des données à partir d'un fichier csv avec séparateur « ; »
donnees <- read.csv2("EnqueteSportEtudiant2024.csv")

# Affiche les 6 premières lignes du jeu de données
head(donnees)
```

Cette table comprend 76 variables, toutes qualitatives, et 375 individus différents. Voici un aperçu des 6 premières lignes de la table avec les 10 premières variables. On peut constater que chaque individu n'a pas systématiquement répondu à toutes les modalités, ce qui limite notre choix de variables qualitatives à corréler avec la variable « sport »

	sexe	deptgeo	deptgeo_Autre	deptformation	niveau	reprise	alternant	bourse	travail	logement
1	Un homme	17		SD	BUT3	Non	Non	Non	Non	Locataire
2	Un homme	17		SD	BUT3	Non	Oui			Locataire
3	Un homme	79		SD	BUT3	Non	Oui		Domicilié.e chez vos (ou un de vos) parents	
4	Un homme	Autre	49	SD	BUT3	Non	Oui			Locataire
5	Un homme	Autre	35	SD	BUT3	Non	Non	Non	Non	Locataire
6	Un homme	Autre	03	SD	BUT3	Non	Oui			Locataire

Nous avons donc retenu pour le croisement les variables suivantes : sexe, département géographique, département de formation, statut de fumeur, état de santé, alimentation et réussite. Ces variables sont complètes et nous semblent intéressantes, car elles sont toutes liées au mode de vie des répondants et peuvent donc influencer leur pratique sportive.

```
> # Table croisée entre le sport et le sexe
> TCD_Sexe = table(donnees$sport, donnees$sexe)
> # Affichage de certains tableaux croisés pour visualiser les données
> TCD_Sexe

     Un homme Une femme
            1         0
Non        48        43
Oui       209        74
```

```
# Test du Chi² pour sport vs sexe
khideux_Sexe = chisq.test(TCD_Sexe)
```

Une fois ces variables choisies, nous avons effectué un test du khi-deux pour chacune d'elles afin de calculer leur p-valeur. Nous avons ensuite identifié les p-valeurs significatives, c'est-à-dire inférieures à 0,01. Enfin, nous avons calculé les coefficients V de Cramer pour mesurer la force des associations. La relation la plus marquée concerne l'alimentation, avec un lien modéré. Cela signifie que les répondants qui font attention à leur alimentation ont plus de chances de pratiquer un sport, et inversement.

	Sexe	Deptgeo	Fumer	Deptform	Sante	alimentation	Reussite
Chi 2 obs	14,742	6,7048	0,8111	18,7777	0,70524	16,661	13,53
p-valeur	0,0006292	0,753	0,6666	0,004557	0,7028	0,000241	0,3535
V de Cramer	0,198274	/	/	0,1582276	/	0,2107832	/

L'analyse des relations entre la variable « sport » et les différentes variables qualitatives révèle que certains aspects du mode de vie des répondants sont effectivement liés à leur pratique sportive. En particulier, l'alimentation apparaît comme un facteur modérément associé à la pratique du sport, suggérant qu'une attention portée à son régime alimentaire peut favoriser une activité physique régulière. D'autres variables, telles que le sexe ou le département de formation, bien que moins fortement corrélées, contribuent également à mieux comprendre les déterminants du sport chez les individus.