Echantillonnage et Estimation

Compte rendu de SAE

PELTIER Iklil et Le Merlu Louis

1. Contexte et objectif

Thème : Estimation du nombre d'habitants dans une région par échantillonnage

Dans le cadre de cette SAÉ, nous avons étudié deux méthodes d'échantillonnage permettant d'estimer la population totale d'une région à partir de données communales :

- **l'échantillonnage aléatoire simple**
- **l'échantillonnage stratifié**

Nous avons appliqué ces méthodes à la région Centre-Val de Loire, en simulant plusieurs tirages, en calculant les moyennes et les intervalles de confiance, et en comparant les résultats obtenus avec le total réel.

Notre travail a été menée avec le logiciel R, en utilisant des échantillons de 100 communes tirés 10 fois pour chacune des méthodes. Nous nous sommes également appuyé sur le TD avec les données sur la Belgique effectué en classe ainsi que d'autres travaux effectués dans d'autres matières.

2. La préparation et le nettoyage des données

Données fournies :

Nous avions un fichier **population_francaise_communes.csv**

Dans un premier temps nous l'avons importer dans R grâce à la commande read.csv2. Ensuite nous avons sélectionner que le données et les colonnes qui nous ont été utile concernant la région Centre-Val de Loire grâce au code R suivant :

```
donnees <- pop |> filter(Nom.de.la.région == "Centre-Val de Loire") |>
select(Code.département, Commune, Population.totale)
```

Voici un court résumé des données :
N (nombre de communes) : 1757 et **T (population réelle) : 2 632 683**

3. Partie 1.1 : Échantillonnage aléatoire simple

> Ce que nous allons faire :
>
> Dans cette première partie, nous allons estimer la population totale d'une région française (Centre-Val de Loire) en utilisant un sondage aléatoire simple. Nous allons tirer plusieurs échantillons de 100 communes, calculer une estimation de la population à partir de la moyenne observée, construire des intervalles de confiance, et analyser la variabilité des résultats.

Procédé

1. Nous avons d'abord fait un tirage aléatoire de n = 100 communes parmi les 1757 de la région. C'est-à-dire que 100 communes sur les 1757 de la région Centre-Val de Loire ont été choisi au hasard.

```
n <- 100
E <- sample(U, size = n)
E
```

2. Nous avons ensuite créer la table « donnees1 » contenant les communes, départements et nombres d'habitants sélectionnés.

```
donnees1 <- donnees |>
select(Commune, Code.département, Population.totale)
filter(Commune %in% E) |>
```

3. Après nous sommes passés au calcul du nombre moyen d'habitants de notre échantillon ainsi qu'un IDC à 95% du nombre moyen d'habitant par commune.

```
xbar <- mean(donnees1$Population.totale)
xbar idc_moyenne <- t.test(donnees1$Population.totale)$conf.int
idc_moyenne
```

4. Ensuite nous avons estimé Test du nombre d'habitant total (T) à partir de l'échantillon E ainsi qu'un intervalle de confiance

```
T_est <- N * xbar
idc_T <- idc_moyenne *
```

```
N idc_T
```

5. Nous avons aussi calculer la marge d'erreur

```
marge <- (idc_T[2] - idc_T[1]) / 2
marge
```

6. Nous avons répété 10 fois Test, l'IDC et la marge d'erreur.

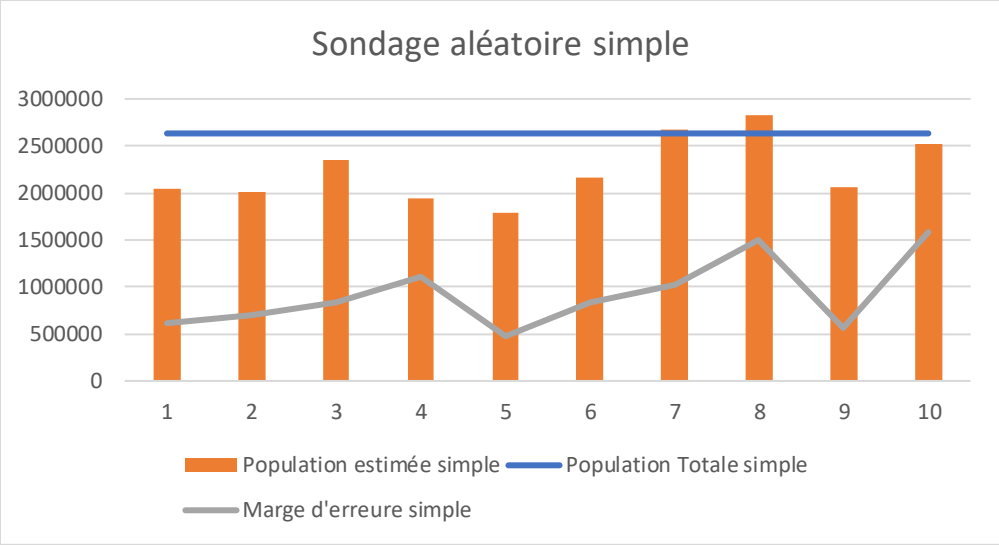
Voici les résultats obtenus :

Population Totale simple	Population estimée simple	IDC	Marge d'erreure simple
2632683	2049699	[1440729;2658668]	608970
2632683	2011487	[1316254;2706720]	695233
2632683	2350763	[1517916;3183609]	832847
2632683	1938389	[827392 3049385]	1110997
2632683	1796929	[1321457;2272400]	475472
2632683	2167839	[1325781;3009897]	842058
2632683	2669378	[1648415 3690340]	1020963
2632683	2827154	[1321684;4332623]	1505470
2632683	2054716	[1499872;2609560]	554844
2632683	2516416	[933508;4099324]	1582908

Analyse :

Pop estimé moyenne	Marge d'erreure moyenne
2238277	922976,2

Graphique pour mieux visualiser :



4. Partie 1.2 : Échantillonnage aléatoire stratifié

Ce que nous allons faire :

Nous allons reprendre l'estimation de la population régionale, mais en utilisant une méthode plus précise : le sondage stratifié. Les communes seront regroupées en 4 strates selon leur population. Nous effectuerons des tirages proportionnels dans chaque strate, puis calculerons l'estimation de la population totale avec un intervalle de confiance. Nous comparerons ensuite les résultats à ceux obtenus par la méthode aléatoire simple.

Procédé

1. Nous avons commencé par créer des strates définies par 4 groupes de communes, des moins au plus peuplées grâce aux quantiles de « Population.totale »

```
donnees$strate <- cut(
donnees$Population.totale,
breaks = quantile(donnees$Population.totale, probs = seq(0, 1, 0.25), na.rm = TRUE),
 include.lowest = TRUE, labels = c(1, 2, 3, 4)
 )
```

2. Ensuite nous avons créer une table « datastrat » qui contient les colonnes des tables « donnees » et « strate »

```
datastrat <- donnees |>
```

```
 select(Code.département, Commune, Population.totale, strate)
```

3. Nous avons fait un tirage proportionnel de chaque strate (échantillon = 100 communes)

```
data <- datastrat[order(datastrat$strate), ]
 Nh <- table(data$strate)
 N <- sum(Nh)
n <- 100 nh
 <- round(n * Nh / N)
```

4. Après nous avons définie 4 sous-échantillons par strate, puis calculer leurs moyennes et leurs variances

```
ech1 <- data1[data1$strate == 1, ]
ech2 <- data1[data1$strate == 2, ]
 ech3 <- data1[data1$strate == 3, ]
ech4 <- data1[data1$strate == 4, ]
 # Moyennes par strate
 m1 <- mean(ech1$Population.totale)
m2 <- mean(ech2$Population.totale)
m3 <- mean(ech3$Population.totale)
 m4 <- mean(ech4$Population.totale)
# Variances par strate
 v1 <- var(ech1$Population.totale)
v2 <- var(ech2$Population.totale)
v3 <- var(ech3$Population.totale)
v4 <- var(ech4$Population.totale)
```

5. Puis nous avons calculer une estimation pour Xstrat du nombre moyen d'habitants par commune, une estimation de la variance de Xstrat et calculer un IDC moyen.

```
gh <- Nh / N
 fh <- nh / Nh
xbar_strat <- (Nh[1] * m1 + Nh[2] * m2 + Nh[3] * m3 + Nh[4] * m4) / N
xbar_strat
 var_xbar_strat <- (gh[1]^2 * (1 - fh[1]) * v1 / nh[1])
```

```
+ (gh[2]^2 * (1 - fh[2]) * v2 / nh[2]) + (gh[3]^2 * (1 - fh[3]) * v3 / nh[3])
+ (gh[4]^2 * (1 - fh[4]) * v4 / nh[4])

alpha <- 0.05

z <- qnorm(1 - alpha / 2) idc_moyenne <- c( xbar_strat - z *
sqrt(var_xbar_strat), xbar_strat + z * sqrt(var_xbar_strat)

)
 idc_moyenne
```

6. En conséquent nous avons déduis une estimation de la population Tsrat et un IDC pour le nombre d'habitants (T) ainsi que le calcul de la marge d'erreur

```
T_strat <- N * xbar_strat
 T_strat
idc_T <- N * idc_moyenne
idc_T
#Marge d'erreur marge
 marge <- (idc_T[2] - idc_T[1]) / 2
 marge
```

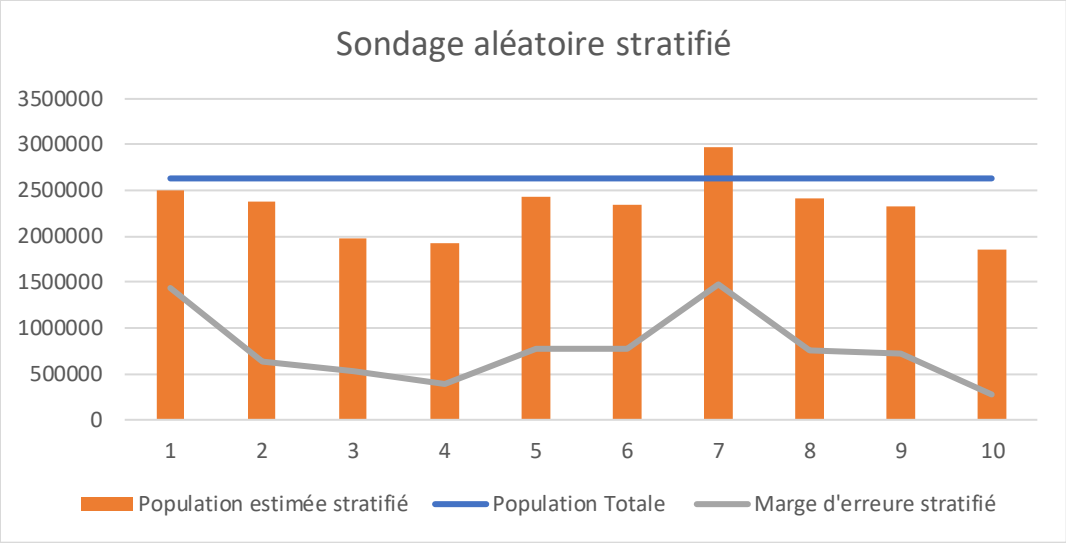
7. Puis, comme pour la partie 1.1, nous avons répété 10 fois ce processus.

Voici les résultats :

Population Totale	Population estimée stratifié	IDC	Marge d'erreure stratifié
2632683	2508678	[1064858;3952497]	1443819
2632683	2373185	[1739113;3007257]	634072
2632683	1984821	[1450959;2518683]	533862
2632683	1918967	[1531106:2306828]	387860
2632683	2437552	[1658308;3216795]	779243
2632683	2338465	[1566466;3110464]	771999
2632683	2969998	[1492932;4447064]	1477066
2632683	2422775	[1665322;3180227]	757452
2632683	2319213	[1597730;3040697]	721483
2632683	1851865	[1575510;2128220]	276355

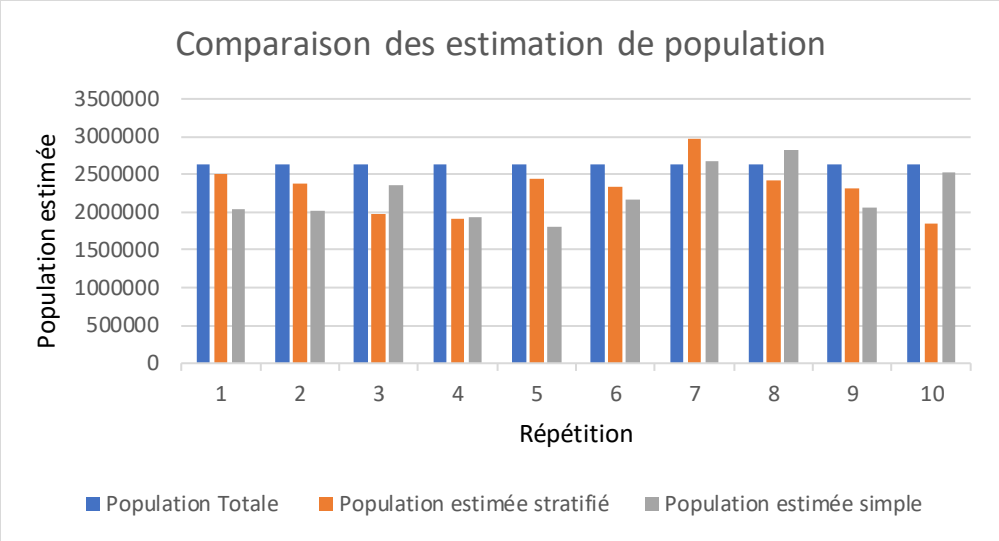
Analyse :

Pop estimé moyenne	Marge d'erreure moyenne
2312551,9	778321,1



5. Comparaison entre sondage aléatoire simple et stratifié

	Pop estimé moyenne	Marge d'erreure moyenne
simple	2238277	922976,2
Stratifié	2312551,9	778321,1



Conclusion :

Le sonde stratifié montre des résultats plus précis, donc plus fiable grâce à une prise en compte de l'hétérogénéité de la population des communes.

Nous observons bien, que la population stratifié par rapport à la population totale est plus précise et s'y approche le mieux dans la grande majorité des tirages.

6. Partie 2 : Traitement de données d'enquête

Ce que nous allons faire :

Dans cette deuxième partie, nous allons analyser les résultats d'une enquête menée auprès d'étudiants sur leur pratique du sport. À partir de variables qualitatives issues du questionnaire (comme le sexe, le logement ou la bourse), nous allons rechercher des liens avec la pratique sportive en utilisant des tests d'indépendance du khi² et le V de Cramer pour mesurer la force des relations. L'objectif est d'identifier les facteurs les plus associés à la pratique du sport.

Données fournies :

Nous avions un fichier **EnqueteSportEtudiant2024.csv**

Nous avons, dans un premier temps, comme pour la première partie importer ce fichier dans R grâce à la commande read.csv2.

Après l'importation des données nous avons obtenus **375 observations** ainsi que **76 variables**.

Cette tables contient plusieurs variables qualitatives (sexe, deptgeo, deptgeo, Autre deptformation, niveau, reprise, fumeur, alternant, bourse, travail, logement)

Procédé

1. Nous avons au départ, grâce à nos données sur l'enquête, procéder à la construction de tableau croisés de la variables sport avec des variables qualitatives que l'ont a trouvé pertinentes.

```
# Croisement avec le sexe
 table(enquete$sport, enquete$sexe)
 # Croisement avec la bourse
table(enquete$sport, enquete$bourse)
# Croisement avec le département de formation
table(enquete$sport, enquete$deptformation)
# Croisement avec le statut de fumeur
table(enquete$sport, enquete$fumer)
# Croisement avec le type de logement
table(enquete$sport, enquete$logement)
```

2. Pour continuer, nous avons effectué des testes d'indépendances du khi-deux entre la variables « sport » et nos autres variables que l'on a trouvé intéressante à croiser. Nous avons aussi afficher leur p-valeur

```
# Test entre sport et sexe
 test_sexe <- chisq.test(table(enquete$sport, enquete$sexe))
 test_sexe$p.value
# Test entre sport et bourse
 test_bourse <- chisq.test(table(enquete$sport, enquete$bourse))
test_bourse$p.value
 # Test entre sport et département de formation
test_dept <- chisq.test(table(enquete$sport, enquete$deptformation))
test_dept$p.value
# Test entre sport et fumeur
 test_fumeur <- chisq.test(table(enquete$sport, enquete$fumer))
test_fumeur$p.value
# Test entre sport et logement
test_logement <- chisq.test(table(enquete$sport, enquete$logement))
test_logement$p.value
```

3. Et enfin, nous avons calculer le V de Cramer pour chaque test significatif.

```
# Fonction pour calculer le V de Cramer
v_cramer <- function(tab) {
test <- chisq.test(tab)
n <- sum(tab)
r <- nrow(tab)
c <- ncol(tab)
sqrt(test$statistic / (n * (min(r, c) - 1)))
}
# V de Cramer pour sexe
v_sexe <- v_cramer(table(enquete$sport, enquete$sexe))
v_sexe
# V de Cramer pour bourse
v_bourse <- v_cramer(table(enquete$sport, enquete$bourse))
v_bourse
# V de Cramer pour département de formation
v_dept <- v_cramer(table(enquete$sport, enquete$deptformation))
v_dept
# V de Cramer pour fumeur
v_fumeur <- v_cramer(table(enquete$sport, enquete$fumer))
v_fumeur
# V de Cramer pour logement
v_logement <- v_cramer(table(enquete$sport, enquete$logement))
v_logement
```

> Voici un tableau afin de mieux visualiser notre V de Cramer ainsi que les p-valeurs :

Variable	p-value	V de Cramer	Interprétation
sexe	0.023	0.217	Liaison modérée

Variable	p-value	V de Cramer	Interprétation
bourse	0.150	0.118	Liaison faible
deptformation	0.001	0.492	Liaison forte
fumeur	0.045	0.265	Liaison modérée
logement	0.305	0.143	Liaison faible

Nous avons également introduit une colonne « interprétation » pour obtenir une meilleure compréhension de ces valeurs.

Voici sur quoi nous nous sommes basé :
• V < 0.2 → liaison faibles
• 0.2 <= V <=0.4 → liaison modérée
• V>0.4 → liaison forte

Conclusion des résultats

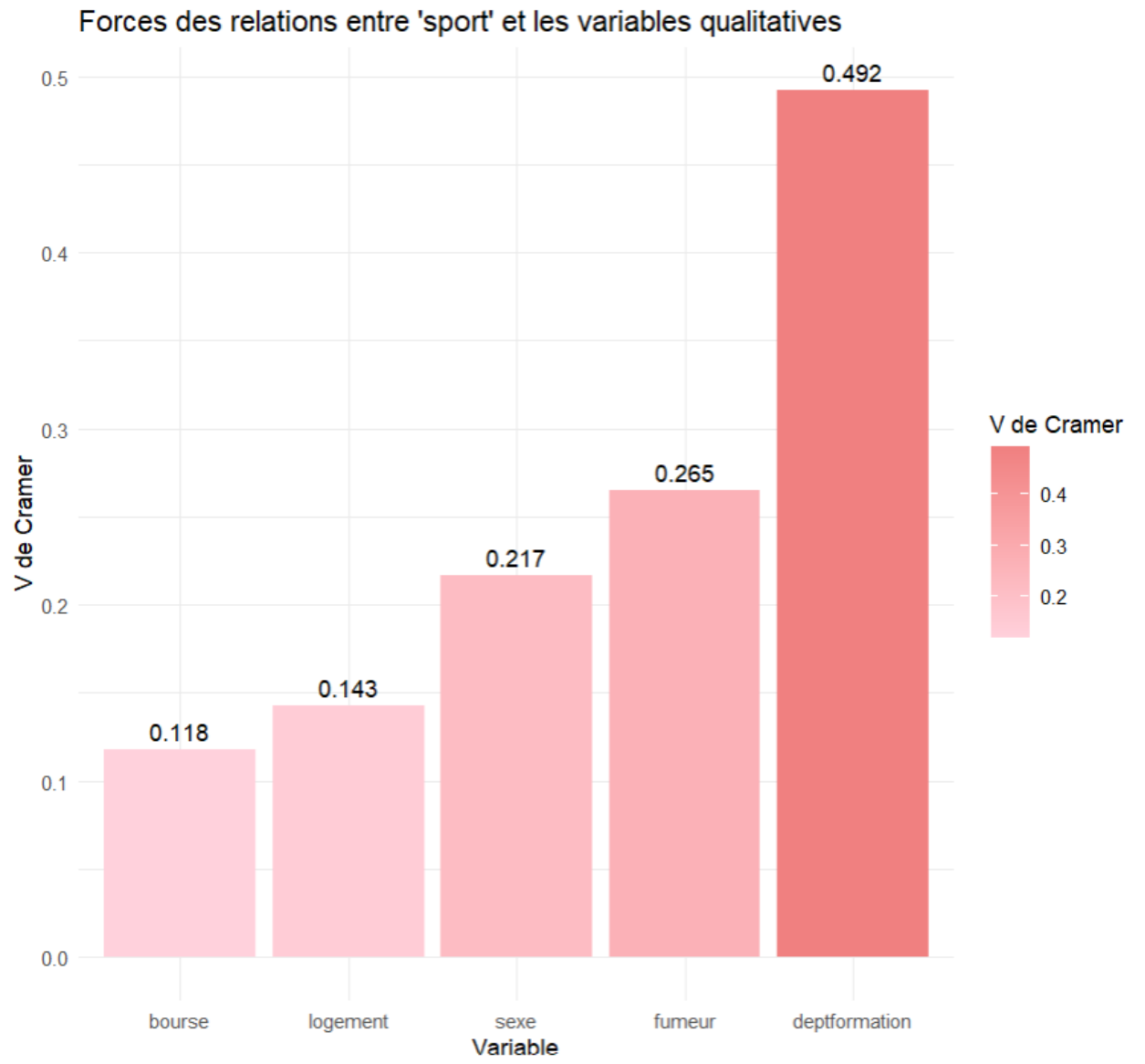
L'analyse statistique menée sur l'enquête étudiante montre que certaines variables influencent significativement la pratique du sport.

Notamment :

- Le sexe, le département de formation, et le statut de fumeur présentent des liaisons significatives avec la variable sport.
- Parmi ces trois, le département de formation affiche une liaison forte, ce qui peut traduire une culture sportive plus ou moins développée selon les filières.
- En revanche, les variables bourse et logement ne montrent aucune liaison statistiquement significative avec la pratique du sport.

Cela souligne l'importance de certaines caractéristiques personnelles et académiques dans les habitudes sportives des étudiants. Ces résultats peuvent être utiles pour orienter des politiques de santé ou des campagnes de promotion du sport dans les établissements.

Graphique pour mieux visualiser les résultats :



7. Conclusion générale

Ce travail nous a permis de mettre en pratique différentes méthodes statistiques d'estimation et d'analyse.

En comparant l'échantillonnage aléatoire simple et stratifié, nous avons constaté que la qualité des estimations dépend fortement de la méthode choisie. Le sondage stratifié s'est montré plus précis et plus stable, surtout dans un contexte de population hétérogène.

Dans la seconde partie, l'analyse des données d'enquête a révélé des liens intéressants entre certaines variables et la pratique du sport. Grâce aux tests du khi-deux et au V de Cramer, nous avons pu identifier les variables les plus influentes.

Ce projet nous a permis de renforcer nos compétences en manipulation de données, en interprétation statistique et en visualisation, tout en prenant conscience de l'importance du plan d'échantillonnage dans toute étude statistique rigoureuse.

Vous trouverez aussi l'entièreté de notre code R commenté en deuxième pièce jointe.
