Rapport SAE : Estimation de la population d'une région par sondage

Partie 1 : Estimation du nombre d'habitants d'une région de France

---

Partie 1.1 : Échantillonnage aléatoire simple

1. **Filtrage des données**
   Les données ont été filtrées pour ne conserver que les communes de la région Île-de-France, avec les colonnes : Code département, Commune, Population totale.

```r
# Charger les données à partir d'un fichier CSV
donnees = read.csv2("population_francaise_communes.csv")

# Filtrer les données pour ne garder que les communes de la région Île-de-France
donnees <- donnees[donnees$Nom.de.la.région == "Île-de-France", ]

# Sélectionner seulement les colonnes nécessaires : Code du département, Commune, Population totale
donnees <- donnees[, c("Code.département", "Commune", "Population.totale")]
head(donnees)
```

2. **Population U**
   L'ensemble de la population U est composé des communes d'Île-de-France.
   Nombre total de communes N : 1281

```r
# Variable U qui contient les noms des communes
U = donnees$Commune

# Nombre total de communes dans U
N = length(U)
```

3. **Nombre total exact d'habitants**
   T = 12278210 habitants (somme de la variable Population.totale)

```r
# Calcul du nombre total d'habitants en Île-de-France
T = sum(donnees$Population.totale)
```

4. **Tirage aléatoire simple d'un échantillon de taille n = 100**
   Un échantillon E de 100 communes a été tiré au hasard sans remise.

```r
# Tirage aléatoire simple pour obtenir un échantillon de taille n=100
n = 100
E = sample(U, n)
```

5. **Données de l'échantillon**
   Table contenant les communes tirées avec leur département et population totale.

```r
# Extraire les données correspondant à l'échantillon sélectionné
donnees1 = donnees[donnees$Commune %in% E, ]
head(donnees1)
```

6. **Moyenne et intervalle de confiance (IDC) à 95% pour la moyenne**
   Moyenne de l'échantillon : $\bar{x}$ = 9203,61 habitants
   IDC95% pour la moyenne : [7364,71 ; 11042,51]

```
# Calcul de la moyenne de la population totale dans l'échantillon
xbar = mean(donnees1$Population.totale)

# Calcul de l'intervalle de confiance pour la moyenne de la population
idcmoy = t.test(donnees1$Population.totale)$conf.int
```

7. **Estimation de T et IDC pour T**
   T_est = N * $\bar{x}$ = 1281 * 9203,61 = 11790622,8
   IDC pour T : [9433080,2 ; 14148165,5]
   Marge d'erreur : (14148165,5 - 9433080,2)/2 = 2357542,65

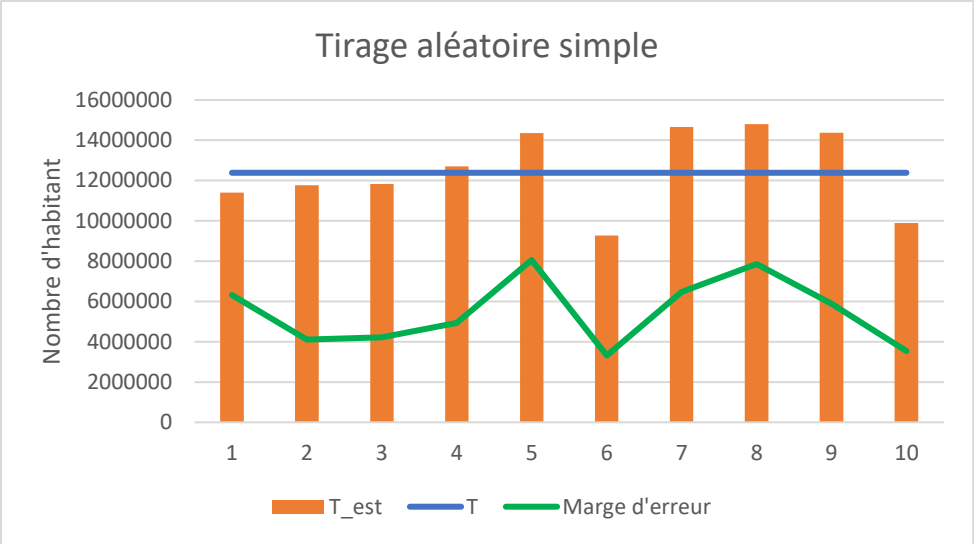
```
# Estimation du total d'habitants dans toute la région à partir de la moyenne de l'échantillon
T_est = N * xbar
T_est

# Estimation de l'intervalle de confiance pour le total d'habitants
idcT = idcmoy * N
idcT

# Calcul de la marge d'erreur pour l'intervalle de confiance du total
marge = (idcT[2] - idcT[1]) / 2
```

8. **Reproduction de l'échantillonnage 10 fois**
   Un tableau Excel a été produit contenant 5 colonnes :
   Population totale réelle | Population estimée | IDC pour T | Marge d'erreur
   Un graphique en barres a été généré pour représenter les estimations et les IC.

	T	T_est	IDC(T)		Marge d'erreur
1	12384734	11393849	5085216	17702483	6308633,5
2	12384734	11766848	7664772	15868923	4102075,5
3	12384734	11825239	7611235	16039243	4214004
4	12384734	12702410	7771315	17633505	4931095
5	12384734	14353988	6298439	22409537	8055549
6	12384734	9275485	5960872	12590098	3314613
7	12384734	14663074	8193460	21132688	6469614
8	12384734	14800976	6939098	22662854	7861878
9	12384734	14366314	8486291	20246338	5880023,5
10	12384734	9883568	6345195	13421941	3538373



9. Conclusion échantillonnage aléatoire
Les résultats montrent une forte variabilité selon les tirages. La précision est moyenne et les marges d'erreur parfois importantes. L'échantillonnage simple ne garantit pas toujours une bonne représentativité.

Partie 1.2 : Échantillonnage aléatoire stratifié

1. Création des strates
4 strates définies par les quantiles de la variable Population.totale :

- Strate 1 : [0 ; 569.5]
- Strate 2 : [569.5 ; 1444]
- Strate 3 : [1444 ; 7367]
- Strate 4 : [7367 ; 231186]

```
# Définir les strates en fonction des déciles de la population
donnees$strate = cut(donnees$Population.totale, breaks=c(0,569.5, 1444, 7367, 231186), labels=c(1,2,3,4))
```

2. Table datastrat
Table contenant les colonnes : Code département, Commune, Population totale, Strate

```
donneesstrat = donnees[, c("Code.département", "Commune", "Population.totale", "strate")]
head(donneesstrat)
```

3. Tirage stratifié (taille n=100, effectifs proportionnels)
Tirage effectué sans remise dans chaque strate, avec taille proportionnelle aux effectifs des strates dans la population.

```r
# Trier les données par strate
data = donneesstrat[order(donneesstrat$strate), ]
head(data)

# Calculer l'effectif de chaque strate
Nh = table(data$strate)
N = sum(Nh)

# Calcul des poids des strates (proportions de chaque strate dans la population totale)
gh = Nh / N

# Tirage aléatoire stratifié pour un échantillon 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

# Calcul du taux de sondage dans chaque strate
fh = nh / Nh
fh

# Effectuer un tirage aléatoire simple sans remise dans chaque strate
st = strata(data, stratanames = c("strate"), size = nh, method = "srswr")

# Extraire les données correspondant à l'échantillon stratifié
data1 = getdata(data, st)
head(data1)
```

4. Définition des sous-échantillons et statistiques

Calcul des moyennes et variances dans chaque strate :

- Moyennes : m1, m2, m3, m4
- Variances : var1, var2, var3, var4

```r
# Diviser l'échantillon stratifié par strate
ech1 = data1[data1$strate == 1, ]
ech2 = data1[data1$strate == 2, ]
ech3 = data1[data1$strate == 3, ]
ech4 = data1[data1$strate == 4, ]

# Calcul de la moyenne de la population totale dans chaque sous-échantillon
m1 = mean(ech1$Population.totale)
m2 = mean(ech2$Population.totale)
m3 = mean(ech3$Population.totale)
m4 = mean(ech4$Population.totale)

# Calcul de la variance de la population totale dans chaque sous-échantillon
var1 = var(ech1$Population.totale)
var2 = var(ech2$Population.totale)
var3 = var(ech3$Population.totale)
var4 = var(ech4$Population.totale)
```

5. Moyenne stratifiée et intervalle de confiance

$\bar{X}$strat = moyenne pondérée des strates

Var($\bar{X}$strat) estimée selon la formule classique

IDC pour μ : $\bar{X}$strat ± 1,96 * √(Var($\bar{X}$strat))

```r
# Calcul de la moyenne des quatre sous-échantillons
Xbarst = (Nh[1] * m1 + Nh[2] * m2 + Nh[3] * m3 + Nh[4] * m4) / N

# Estimation de la variance de la moyenne stratifiée
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]

# Calcul de l'intervalle de confiance pour la moyenne stratifiée 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 du total Tstrat et de son IDC

Tstrat = N * $\bar{X}$strat

IDC pour Tstrat = IDC moyenne * N

Marge d'erreur = largeur IDC / 2

```r
# Estimation du total d'habitants pour toute la région par la moyenne stratifiée
Tstr = N * Xbarst
Tstr

# Estimation de l'intervalle de confiance pour le total d'habitants
binf = idcmoy[1] * N
bsup = idcmoy[2] * N
idcT = c(binf, bsup)
idcT

# Calcul de la marge d'erreur pour l'intervalle de confiance du total
marge = (idcT[2] - idcT[1]) / 2
```

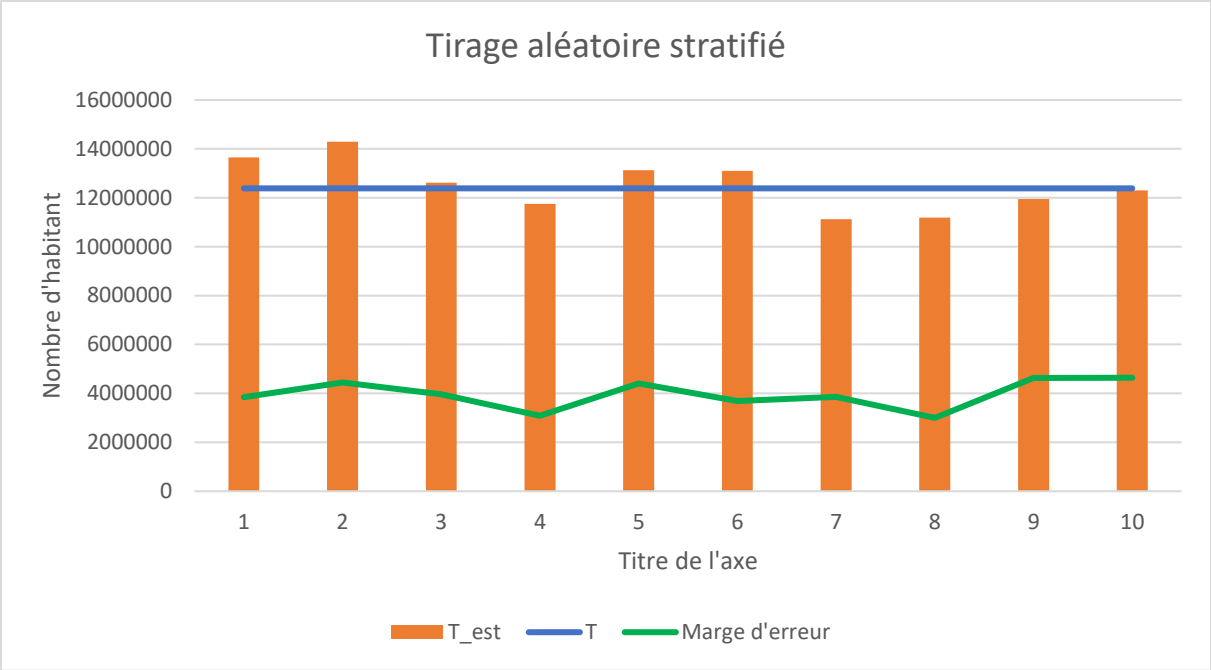
7. Reproduction 10 fois

Comme pour la 1.1, une répétition du tirage stratifié a été faite 10 fois avec calculs des indicateurs dans un tableau Excel :

Population totale | Population estimée | IDC | Marge d'erreur

Un graphique a été généré.

	T	T_est	IDC(T)		Marge d'erreur
1	12384734	13643948	9800354	17487543	3843594,5
2	12384734	14286565	9832306	18740825	4454259,5
3	12384734	12613383	8648052	16578713	3965330,5
4	12384734	11751674	8661383	14841964	3090290,5
5	12384734	13122245	8713760	17530730	4408485
6	12384734	13105888	9421967	16789809	3683921
7	12384734	11119013	7256548	14981478	3862465
8	12384734	11186748	8185563	14187934	3001185,5
9	12384734	11944882	7315829	16573936	4629053,5
10	12384734	12305616	7663929	16947303	4641687



8. Améliorations proposées
Pour augmenter la précision :

- Définir plus de strates (par exemple 5 ou 6)
- Créer des strates par département ou type de commune
- Utiliser d'autres variables pour stratifier (densité, urbanisation)

9. Conclusion sur les deux méthodes
Le sondage stratifié est nettement plus précis que l'aléatoire simple. Les marges d'erreur sont réduites, les intervalles plus resserrés. Il permet une meilleure représentation de la population, surtout en cas d'hétérogénéité marquée.

10. Conclusion générale
Cette SAE m'a permis de comprendre :

- Comment estimer une grandeur d'une population avec des outils statistiques
- L'importance des intervalles de confiance pour traduire l'incertitude
- La supériorité du sondage stratifié dans beaucoup de cas
- L'intérêt de la rigueur dans la mise en place de l'échantillonnage
  Je retiens que les statistiques inférentielles sont indispensables pour prendre des décisions à partir de données partielles.

Partie 2 :

Traitement de données d'enquête Dans cette partie, on reprend les données d'enquête sur les ´étudiants et la pratique du sport. Ces données ont ´été traitées dans la SAE "Tableaux de données et analyse exploratoire" du semestre 1. Le but est de trouver des relations significatives entre la variable "sport" et plusieurs autres variables qualitatives de votre choix.

1- Importer la table EnqueteSportEtudiant2024.csv dans R.

```r
#1- Importer la table EnqueteSportEtudiant2024.csv dans R
df = read.csv2("C:/Users/msabi/Downloads/EnqueteSportEtudiant2024.csv")
```

2- Afficher les 6 premières lignes de cette table. Que contient-elle ? individus ? variables ? types de variables ?

```r
# Affichage des 6 premières lignes
head(df)
str(df)
```

3- Construire et afficher les tableaux croisés de la variable ”sport” avec les autres variables qualitatives que vous pensez intéressantes.

```r
# Exemple 1 : sport x sexe
TCD_Sexe = table(df$sport, df$sexe)

# Exemple 2 : sport x niveau
TCD_Niveau = table(df$sport, df$niveau)

# Exemple 3 : sport x bourse
TCD_Bourse = table(df$sport, df$bourse)

# Exemple 4 : sport x alternant
TCD_Alternant = table(df$sport, df$alternant)

# Exemple 5 : sport x fumer
TCD_Fumeur = table(df$sport, df$fumer)

# Exemple 6 : sport x logement
TCD_Logement = table(df$sport, df$logement)
```

```r
# Exemple 7 : sport x alimentation
TCD_alimentation = table(df$sport, df$alimentation)
```

4- Effectuer un test d’indépendance du khi-deux entre la variable “sport” et toutes les autres variables qualitatives choisies. Afficher les p-valeurs et en déduire les relations significatives.

```r
# Test du khi2
khideux_Sexe= chisq.test(TCD_Sexe)
khideux_Sexe #p-value = 0.0006292 --> relation significative

khideux_Alternant= chisq.test(TCD_Alternant)
khideux_Alternant #p-value = 0,14 --> preuve modérée contre l'hypothèse nulle, peut-être que c'est dû au hasard

khideux_Bourse= chisq.test(TCD_Bourse)
khideux_Bourse #p-value = 0.0497 --> relation significative

khideux_Niveau= chisq.test(TCD_Niveau)
khideux_Niveau #p-value = 0.1238 --> c'est moyen

khideux_Logement= chisq.test(TCD_Logement)
khideux_Logement #p-value = 0.3084 --> preuve modérée pour l'hypothèse nulle

khideux_fumer= chisq.test(TCD_Fumeur)
khideux_fumer #p-value = 0.6666 --> très élevé, il n'y a quasiment pas de lien
```

```r

khideux_alimentation= chisq.test(TCD_alimentation)
khideux_alimentation #p-value = 0.000241 --> relation significative
```

5- Pour chaque test significatif, calculer le V de Cramer. Construire un tableau qui donne pour chaque test significatif le V de Cramer, en soulignant la liaison la plus forte. Commenter vos résultats et faire une conclusion générale.

```r
# V de Cramer
n<-dim(df)[1]
p <- nrow(TCD_Sexe)
q <- ncol(TCD_Sexe)
m <- min(p-1, q-1)
V_Sexe =sqrt(khideux_Sexe$statistic/(n*m))
V_Sexe #lui il est bien

n<-dim(df)[1]
p <- nrow(TCD_Bourse)
q <- ncol(TCD_Bourse)
m <- min(p-1, q-1)
V_Alternant =sqrt(khideux_Bourse$statistic/(n*m))
V_Alternant #nul
```

```r
n<-dim(df)[1]
p <- nrow(TCD_Alternant)
q <- ncol(TCD_Alternant)
m <- min(p-1, q-1)
V_Alternant =sqrt(khideux_Alternant$statistic/(n*m))
V_Alternant #nul

n<-dim(df)[1]
p <- nrow(TCD_Niveau)
q <- ncol(TCD_Niveau)
m <- min(p-1, q-1)
V_niveau =sqrt(khideux_Niveau$statistic/(n*m))
V_niveau #nul

n<-dim(df)[1]
p <- nrow(TCD_Logement)
q <- ncol(TCD_Logement)
m <- min(p-1, q-1)
V_logement =sqrt(khideux_Logement$statistic/(n*m))
V_logement #nul
```

```r
n<-dim(df)[1]
p <- nrow(TCD_alimentation)
q <- ncol(TCD_alimentation)
m <- min(p-1, q-1)
V_alimentation =sqrt(khideux_alimentation$statistic/(n*m))
V_alimentation #lui il est bien et mieux

```

Conclusion

Parmi toutes les comparaisons, ce sont surtout deux liens qui ressortent : celui entre la pratique du sport et le sexe des étudiants, avec un V de Cramer de 0,198, et celui entre le sport et l'alimentation avec un V de Cramer de 0,21. Cela signifie qu'il existe une légère différence dans la pratique sportive selon qu'on soit un homme ou une femme, et que les habitudes alimentaires semblent également jouer un rôle. En revanche, d'autres facteurs comme le fait d'être boursier, en alternance, le niveau d'études ou le type de logement n'ont pas vraiment d'impact sur la pratique du sport.