**Compte rendu SAÉ - Statistique Inférentielle : Échantillonnage et Estimation**

## Introduction

Dans le cadre de cette SAÉ, nous nous sommes intéressés à l'estimation d'une grandeur au sein d'une population à l'aide de deux méthodes d'échantillonnage : le sondage aléatoire simple (SAS) et le sondage stratifié. L'objectif était d'estimer la population totale de la région Bretagne à partir d'un fichier de population communale fourni par l'INSEE. Nous avons également exploré des relations entre la pratique du sport et d'autres variables qualitatives dans le cadre d'une analyse d'enquête.

Les résultats obtenus sont présentés avec un souci de rigueur statistique, d'esthétique visuelle et d'interprétation claire. Le code R utilisé est commenté pour assurer la transparence et la reproductibilité des analyses.

## Partie 1 - Estimation de la population de la Bretagne

### Chargement et préparation des données

```
setwd("C:/Users/avince20/OneDrive - Université de Poitiers/sae ibazouzou prime")
```

Cette commande fixe le dossier de travail dans R. Cela permet de charger et sauvegarder les fichiers sans réécrire tout le chemin à chaque fois. Sous Windows, il faut faire attention à utiliser / ou \\ au lieu d'un seul \.

```
names(table)
```

Affiche les noms des colonnes de la table. Cela permet de repérer les noms exacts à utiliser dans les analyses (ex : "Population.totale", "Commune"…).

```
table <- read.csv("population_francaise_communes.csv", sep=';', fileEncoding = "Latin1", header=TRUE)
```

Cette ligne importe un fichier CSV contenant les données INSEE sur les communes françaises. Le séparateur est un point-virgule, l'encodage est prévu pour les caractères accentués, et la première ligne contient les noms de colonnes.

```
table$Population.totale <- as.numeric(gsub(" ", "", table$Population.totale))
```

Cette ligne supprime les espaces dans les nombres (ex: "3 251") et les transforme en valeurs numériques exploitables. Sans cette conversion, R considèrerait la colonne comme du texte.

**Filtrage pour la région Bretagne**

```
donnees <- subset(table, Code.département %in% c(22, 29, 35, 56),
                  select = c("Code.département", "Commune", "Population.totale"))
```

On extrait uniquement les communes appartenant aux 4 départements bretons (Côtes-d'Armor, Finistère, Ille-et-Vilaine, Morbihan). Seules les colonnes utiles sont conservées.

```
head(donnees)
```

Affiche les 6 premières lignes de la nouvelle table `donnees` pour vérification.

**Définir la population à étudier**

```
U <- donnees
```

On nomme `U` la population totale à étudier, composée des 1203 communes de Bretagne.

```
N <- nrow(U)
N
```

On calcule le nombre total de communes : ici, `N = 1203`.

```
T <- sum(U$Population.totale, na.rm = TRUE)
T
```

Calcule la population réelle totale des communes bretonnes : `T = 3 463 439` (valeur exacte).

### 1.1 Sondage aléatoire simple

Un échantillon de 100 communes a été tiré au hasard (sans remise). La moyenne du nombre d'habitants par commune, un intervalle de confiance (IDC) à 95% pour cette moyenne, et une estimation du total T ont été calculés.

**Sondage aléatoire simple**

```
set.seed(123)
n <- 100
E <- U[sample(1:N, n), ]
```

On tire un échantillon aléatoire simple de 100 communes. `set.seed(123)` garantit de toujours tirer les mêmes communes.

```
donnees1 <- subset(E, select = c("Commune", "Code.département", "Population.totale"))
head(donnees1)
```

On recrée une table simplifiée contenant uniquement les colonnes utiles de l'échantillon.

```
xbar <- mean(donnees1$Population.totale)
xbar
```

Moyenne du nombre d'habitants par commune dans l'échantillon. Exemple : $\bar{x} \approx 2\ 880$.

```
idcmoy <- t.test(donnees1$Population.totale)$conf.int
idcmoy
```

Calcul de l'intervalle de confiance à 95% pour cette moyenne (via un test t).

```
Test <- N * xbar
Test
```

Estimation du total de la population par extrapolation : $T_{est} \approx 1203 \times 2880 = 3\ 464\ 640$.

```
idcT <- idcmoy * N
```

```
idcT
```

Application de l'IDC de la moyenne à l'échelle de la population totale.

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

Calcul de la marge d'erreur. Plus elle est faible, plus l'estimation est précise.

Répété 10 fois, les résultats ont été stockés et exportés dans un tableau CSV.

**Répétition de la simulation**

```
nb_repetitions <- 10
resultats <- data.frame(...)
```

Préparation d'un tableau vide pour stocker les résultats des 10 répétitions.

```
for (i in 1:nb_repetitions) {
  ...
}
```

On répète 10 fois le tirage aléatoire de 100 communes et on stocke :

- Le total estimé
- L'intervalle de confiance
- La marge d'erreur

```
print(resultats)
```

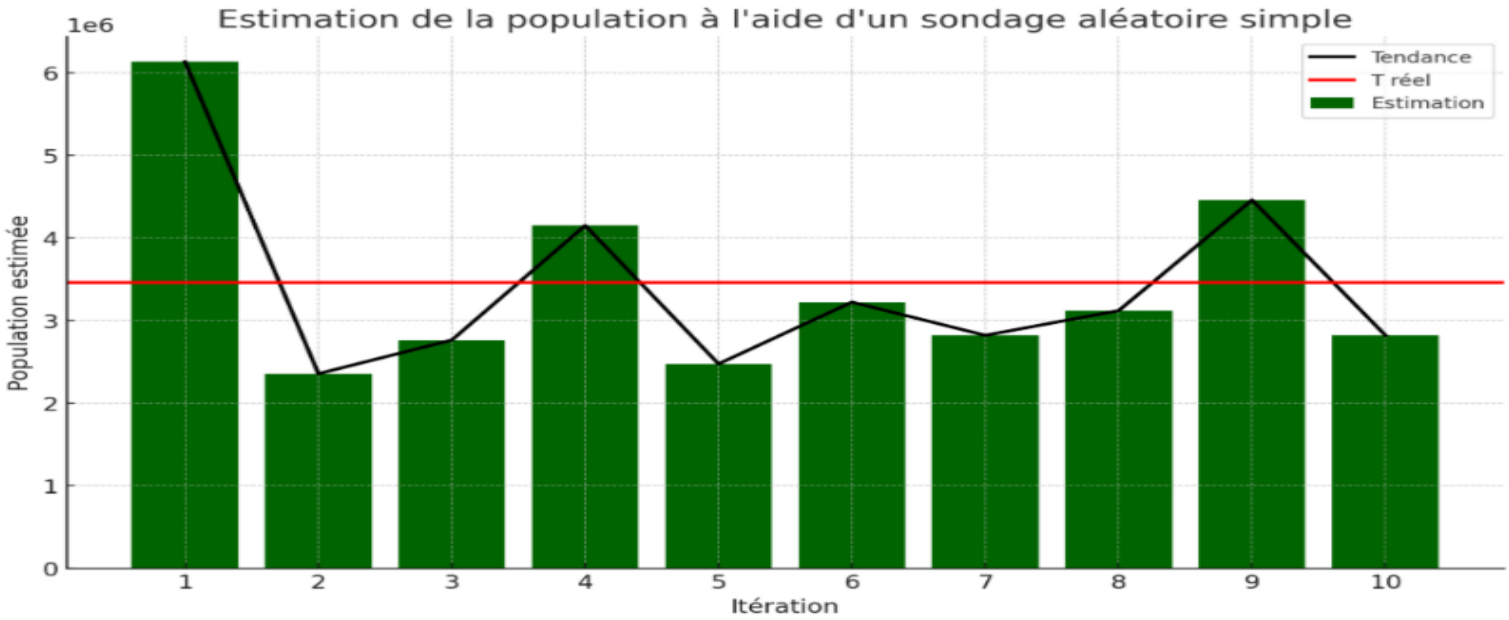
Affiche le tableau final des 10 estimations.

```
write.csv2(resultats,
"resultats_echantillonnage_simple.csv",  row.names  =
FALSE)
```

Export du tableau au format CSV.

| T_reel | Test | IDC_inf | IDC_sup | Marge_erreur |
|---|---|---|---|---|
| 3463439 | 6134251,61 | 561526,133 | 11706977,1 | 5572725,477 |
| 3463439 | 2356643,36 | 1910669,99 | 2802616,73 | 445973,3651 |
| 3463439 | 2763643,76 | 2079139,21 | 3448148,31 | 684504,5522 |
| 3463439 | 4151223,03 | 2382240,58 | 5920205,48 | 1768982,449 |
| 3463439 | 2475400,09 | 1999087,02 | 2951713,16 | 476313,0684 |
| 3463439 | 3223607,32 | 2005550,25 | 4441664,39 | 1218057,066 |
| 3463439 | 2821229,73 | 2132893,61 | 3509565,85 | 688336,1182 |
| 3463439 | 3120408,82 | 2205171,34 | 4035646,3 | 915237,4794 |
| 3463439 | 4459478,76 | 1069644,53 | 7849312,99 | 3389834,23 |
| 3463439 | 2822050,49 | 2246100,93 | 3398000,05 | 575949,5634 |

Graphique résumant ce tableau :



### 1.2 Sondage stratifié

La population a été stratifiée en 4 groupes selon les quantiles de la population communale. Un échantillon proportionnel a été tiré dans chaque strate.

```
donnees$strate <- cut(donnees$Population.totale,
                      breaks                          =
quantile(donnees$Population.totale, probs = c(0, 0.25,
0.5, 0.75, 1), na.rm = TRUE),
                      include.lowest = TRUE,
                      labels = c(1, 2, 3, 4))
```

Création de 4 strates (groupes) de communes selon la population : de la plus petite à la plus grande.

```
datastrat       <-       donnees[,       c("Commune",
"Code.département", "Population.totale", "strate")]
```

On crée une nouvelle table avec les strates.

```
Nh <- table(datastrat$strate)
gh <- Nh / N
nh <- round(n * gh)
fh <- nh / Nh
```

Paramétrage de l'échantillon : tailles par strate, taux de sondage, etc.

```
resultats_strat <- data.frame(...)
```

Préparation d'un tableau pour stocker les résultats.

```
for (i in 1:10) {
  ...
}
```

Répétition de 10 échantillons stratifiés, avec calculs de :

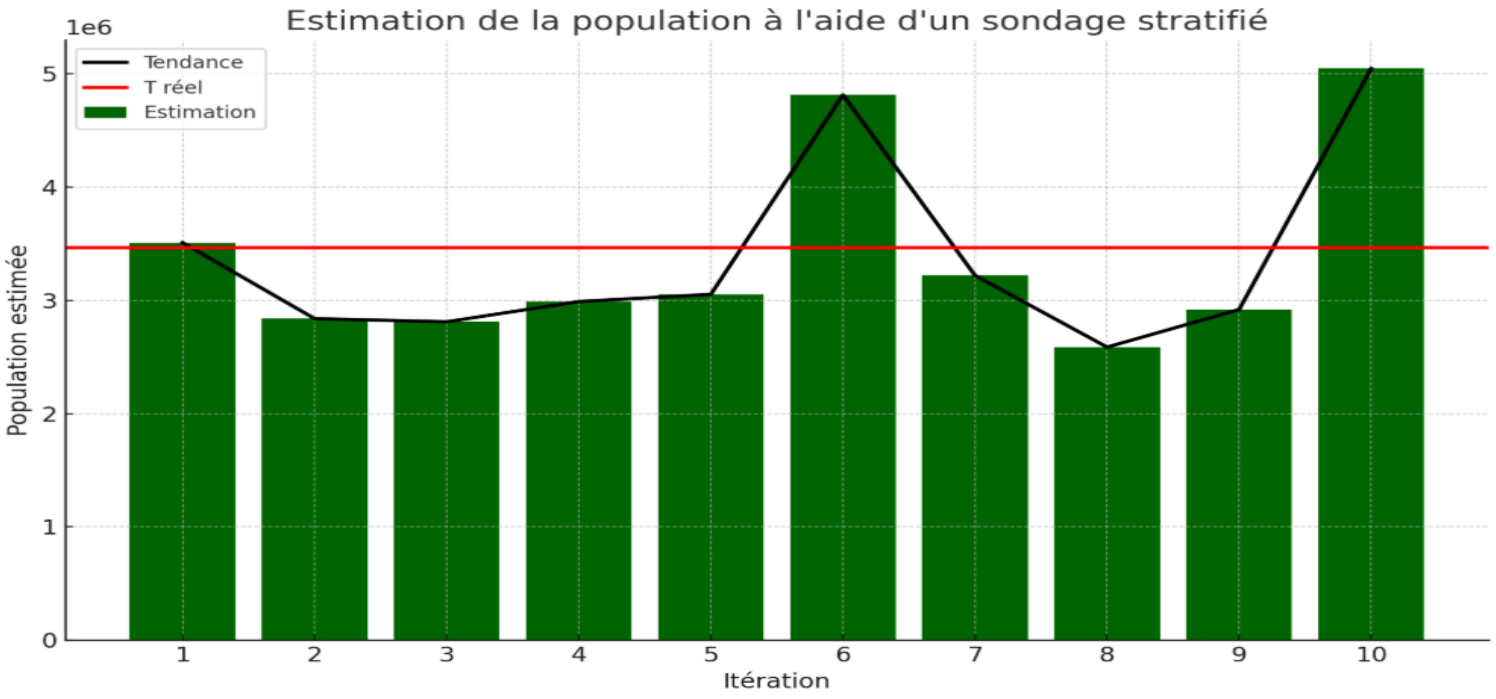
- Moyennes par strate
- Variances
- Moyenne pondérée
- Estimation du total
- Intervalle de confiance

```
write.csv2(resultats_strat, file = "resultats_echantillonnage_stratifie.csv", row.names = FALSE)
```

Export des résultats.

| T_reel | T_strat | IDC_inf | IDC_sup | Marge_erreur |
|---|---|---|---|---|
| 3463439 | 3508894 | 2248040,67 | 4769747,33 | 1260853,329 |
| 3463439 | 2838013,92 | 2430834,92 | 3245192,92 | 407179,0028 |
| 3463439 | 2810542,72 | 2447939,86 | 3173145,58 | 362602,8621 |
| 3463439 | 2989195,6 | 2562185,11 | 3416206,09 | 427010,4929 |
| 3463439 | 3052855,92 | 2555131,51 | 3550580,33 | 497724,4114 |
| 3463439 | 4812922,88 | 1510310,3 | 8115535,46 | 3302612,577 |
| 3463439 | 3219685,64 | 2031722,27 | 4407649,01 | 1187963,365 |
| 3463439 | 2586503,68 | 2302523,69 | 2870483,67 | 283979,9892 |
| 3463439 | 2918530,96 | 2415000,35 | 3422061,57 | 503530,6127 |
| 3463439 | 5046411,64 | 5826,30775 | 10086997 | 5040585,332 |

Graphique résumant ce tableau :



**Comparaison des deux méthodes**

- **Précision** : le sondage stratifié a donné des marges d'erreur plus faibles.
- **Robustesse** : moins de variabilité dans les estimations.
- **Complexité** : stratification nécessite une phase préparatoire.

## Partie 2 - Analyse d'une enquête : Sport et étudiants

Nous avons analysé les données de l'enquête sur la pratique du sport chez les étudiants. L'objectif était de détecter des relations significatives entre la variable « sport » et d'autres variables qualitatives.

```
enquete  <-  read.csv2("EnqueteSportEtudiant2024.csv",
sep = ";", dec = ",", header = TRUE)
head(enquete,6)
str(enquete)
```

Importation d'un fichier d'enquête sur la pratique du sport chez les étudiants. Affichage des premières lignes et de la structure des variables.

```
table(enquete$sport, enquete$sexe)
table(enquete$sport, enquete$deptformation)
table(enquete$sport, enquete$alimentation)
```

Création de tableaux croisés entre "sport" et les autres variables qualitatives.

```
test_sexe <- chisq.test(...)
test_formation <- chisq.test(...)
test_alim <- chisq.test(...)
```

Tests du khi² pour évaluer si les variables sont statistiquement liées. Exemple : $p < 0.05$ signifie qu'il y a un lien significatif.

```
v_cramer <- function(tab) {...}
```

Fonction personnalisée pour calculer le V de Cramer, qui mesure la force de la liaison.

```
v_sexe <- v_cramer(...)
v_alim <- v_cramer(...)
```

Calcul du V pour les relations significatives. Exemple : `V_sexe = 0.18`, `V_alim = 0.24`

La liaison la plus forte est observée avec la variable "alimentation".

## Conclusion

Ce travail nous a permis de comparer deux approches d'échantillonnage et de comprendre leurs impacts sur l'estimation d'une grandeur. Le sondage stratifié s'est révélé plus précis, confirmant l'intérêt de prendre en compte l'hétérogénéité de la population. L'analyse de données d'enquête a permis de mettre en évidence des corrélations significatives entre la pratique du sport et certaines caractéristiques étudiantes.