

# Science des Donnees- S2- Stat. Inferentielle

## SAE- Echantillonnage et Estimation



**GARDERE Théo ERRADDAOUI Yasmine**





```
# =====
# SAE - Échantillonnage et Estimation
# Fait par : Yasmine ERRADAOUI et Théo Gardere
# Objectif : Timer la population d'une région via deux méthodes
#           - Sondage aléatoire simple
#           - Sondage stratifié
# + Analyse de données d'enquête via le test du  $\chi^2$  et V de Cramer
# =====
```

En statistiques, il est rarement possible d'étudier une population entière. C'est pourquoi nous devons avoir recours à l'échantillonnage pour estimer les paramètres d'intérêt. Cette SAE nous permet d'explorer concrètement cette problématique à travers deux situations pratiques. Dans un premier temps, nous nous intéressons à l'estimation de la population totale d'une région française. Cette question peut sembler triviale puisque les données INSEE sont disponibles, mais elle nous permet de comparer l'efficacité de différentes stratégies d'échantillonnage dans un contexte contrôlé. Nous pouvons ainsi mesurer la précision de nos estimations puisque nous connaissons la "vraie" valeur.

La région Grand Est constitue notre terrain d'étude. Cette région présente une grande diversité démographique : des petits villages ruraux aux grandes métropoles comme Strasbourg, Metz ou Nancy. Cette hétérogénéité rend l'exercice d'estimation particulièrement intéressant car elle nous confronte aux défis réels de l'échantillonnage.

Nous comparons deux approches : le sondage aléatoire simple, où chaque commune a la même probabilité d'être sélectionnée, et le sondage stratifié, où nous créons des groupes homogènes de communes selon leur taille avant de procéder à l'échantillonnage. Dans un second temps, nous analysons des données d'enquête sur la pratique sportive chez les étudiants.

L'objectif est d'identifier quelles variables sont significativement liées à la pratique du sport en utilisant le test du  $\chi^2$  (khi-deux) et le coefficient V de Cramer. Ce travail nous permet d'acquérir une expérience pratique de l'échantillonnage et de comprendre l'importance du choix de la méthode selon le contexte d'étude.



## Partie 1 : Estimation du nombre d'habitants de la région Grand Est

Nous commençons par charger les données des communes françaises et nous concentrer sur la région Grand Est. Le code ci-dessous montre notre démarche :

```
# Définition du répertoire de travail contenant les fichiers
setwd("C:/Users/garde/OneDrive - Université de Poitiers/SAÉ stat Yasmine")
# Chargement du fichier csv des communes françaises
table <- read.csv2("population_francaise_communes.csv", sep=";", dec=",", header=TRUE, encoding="UTF-8")
# Sélection des données pour la région Grand Est (Code.région == 44)
donnees <- table[table$Code.région == 44, c("Commune", "Population.totale", "Code.département")]
# Conversion des nombres de population (caractères) en valeurs numériques
donnees$Population.totale <- as.numeric(gsub(" ", "", donnees$Population.totale))
# Suppression des lignes contenant des valeurs manquantes
donnees <- donnees[!is.na(donnees$Population.totale), ]
```

Cette étape de nettoyage est essentielle car les données de population sont initialement stockées sous format texte avec des espaces comme séparateurs de milliers. Après conversion, nous obtenons nos statistiques de base :

```
# Calcul du nombre total de communes dans la région (taille de la population)
N <- nrow(donnees)
# Calcul du total exact des habitants de la région (valeur réelle)
T <- sum(donnees$Population.totale)
```

### Résultats obtenus :

- Nombre de communes N : 5121
- Population totale réelle T : 5668492

#### 1.1 Échantillonnage aléatoire simple

Le principe du sondage aléatoire simple est de sélectionner au hasard 100 communes parmi l'ensemble des communes de la région, chaque commune ayant la même probabilité d'être choisie.

**Méthodologie appliquée :** Nous tirons aléatoirement 100 communes et calculons la moyenne de leur population. Cette moyenne nous permet d'estimer la population totale en la multipliant par le nombre total de communes.



```
# Tirage aléatoire simple de 100 communes de la région
set.seed(123) # Pour garantir la reproductibilité
echantillon <- donnees[sample(1:N, 100), ]
# Calcul de la moyenne de la population des 100 communes
moy <- mean(echantillon$Population.totale)
# Calcul de l'erreur-type et de l'intervalle de confiance à 95 %
et <- sd(echantillon$Population.totale) / sqrt(100)
ic_moy <- c(moy - 1.96*et, moy + 1.96*et)
# Estimation de la population totale à partir de l'échantillon
T_est <- N * moy
```

**Construction de l'intervalle de confiance :** L'intervalle de confiance à 95% nous donne une fourchette dans laquelle nous sommes confiants à 95% que se trouve la vraie valeur. La formule utilisée est :  $IC = [T_{\text{est}} - 1.96 \times N \times \text{erreur\_type} ; T_{\text{est}} + 1.96 \times N \times \text{erreur\_type}]$

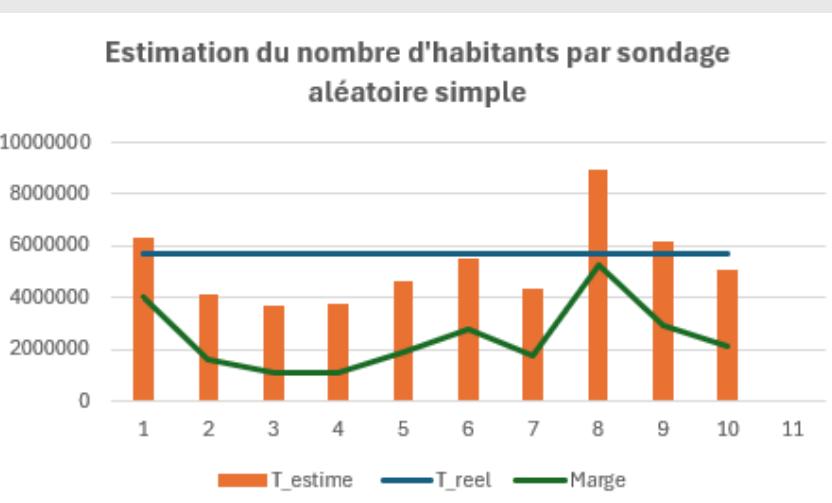
**Analyse de la variabilité :** Pour évaluer la robustesse de notre méthode, nous répétons le processus 10 fois avec différents échantillons. Cela nous permet d'observer la variabilité de nos estimations et de voir si la vraie valeur est bien capturée par nos intervalles de confiance.

```
# Répétition du processus 10 fois pour observer la variabilité
resultats_simple <- data.frame()
for (i in 1:10) {
  ech <- donnees[sample(1:N, 100), ]
  moy_i <- mean(ech$Population.totale)
  et_i <- sd(ech$Population.totale) / sqrt(100)
  T_i <- N * moy_i
  marge_i <- 1.96 * N * et_i
  idc_i <- paste("[", round(T_i - marge_i), ", ", round(T_i + marge_i), "]")
  resultats_simple <- rbind(resultats_simple, data.frame(
    Tirage = i, T_reel = T, T_estime = round(T_i), IDC = idc_i, Marge = round(marge_i)))
}
# Affichage du tableau récapitulatif des 10 tirages
print(resultats_simple)
```



Tirage	T_reel	T_estime	IDC	Marge
1	5668492	6292992	2241783	10344201
2	5668492	4129523	2518439	5740608
3	5668492	3720560	2612902	4828218
4	5668492	3743246	2654716	4831776
5	5668492	4634095	2739085	6529106
6	5668492	5531551	2769484	8293617
7	5668492	4382808	2602688	6162928
8	5668492	8933072	3647342	14218802
9	5668492	6142383	3192314	9092453
10	5668492	5072709	2927652	7217766

Les résultats montrent que le sondage aléatoire simple donne des estimations correctes mais avec une variabilité importante. Cette variabilité s'explique par l'hétérogénéité des communes : un échantillon peut tomber sur beaucoup de petites communes rurales ou au contraire sur quelques grandes villes, ce qui biaise l'estimation.



Ce graphique illustre les résultats de dix estimations du nombre d'habitants obtenues par sondage aléatoire simple. On y observe une forte variabilité des estimations (barres orange) autour de la valeur réelle (ligne bleue), avec des écarts parfois importants. La marge d'erreur (courbe verte) varie également d'un tirage à l'autre, traduisant une incertitude non négligeable. Dans l'ensemble, ce type d'échantillonnage montre ses limites : bien qu'il soit simple à mettre en œuvre, il peut produire des estimations instables, surtout dans une population hétérogène, ce qui justifie l'intérêt d'envisager des méthodes plus structurées comme le sondage stratifié.

## Partie 1.2 : Echantillonnage aléatoire stratifié

Face aux limites du sondage simple, nous testons une approche plus sophistiquée : le sondage stratifié. L'idée est de diviser la population en groupes homogènes (strates) avant d'échantillonner. Construction des strates :

Nous créons 4 strates basées sur les quartiles de la population communale :

- Strate 1 (S1) : Les 25% de communes les moins peuplées
- Strate 2 (S2) : Les communes du 2ème quartile
- Strate 3 (S3) : Les communes du 3ème quartile
- Strate 4 (S4) : Les 25% de communes les plus peuplées

```
# Création de 4 strates selon les quartiles de la population
q <- quantile(donnees$Population.totale, probs = seq(0, 1, 0.25))
donnees$strate <- cut(donnees$Population.totale, breaks = q, include.lowest = TRUE, labels = c("s1", "s2", "s3", "s4"))
```

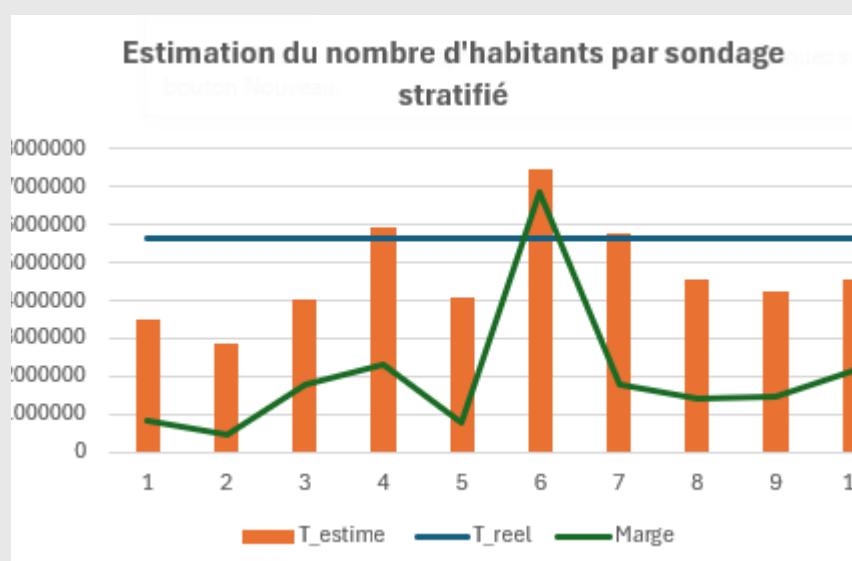


**Échantillonnage proportionnel :** Nous prélevons dans chaque strate un nombre de communes proportionnel à la taille de la strate. Si une strate représente 30% des communes, elle fournira 30% de notre échantillon de 100 communes.

```
# Taille de chaque strate (nombre de communes par strate)
Nh <- table(donnees$strate)
# Définition de la taille totale de l'échantillon (100 communes)
n_total <- 100
# calcul du nombre de communes à tirer dans chaque strate (proportionnellement)
nh <- round(n_total * Nh / N)
```

**Estimation stratifiée :** La moyenne stratifiée est calculée comme une moyenne pondérée des moyennes de chaque strate :  $\bar{X}_{\text{strat}} = \sum(gh \times \text{moyh})$  où gh est le poids de la strate h dans la population totale.

```
# Calcul des moyennes et variances par strate dans l'échantillon
moyennes <- tapply(ech_strat$Population.totale, ech_strat$strate, mean)
variances <- tapply(ech_strat$Population.totale, ech_strat$strate, var)
# Poids de chaque strate dans la population totale
gh <- as.numeric(Nh) / sum(Nh)
# Estimation de la moyenne globale stratifiée
xbar_strat <- sum(gh * moyennes)
```



Le graphique présente les estimations du nombre d'habitants obtenues à partir de dix sondages stratifiés, comparées à la population réelle (ligne bleue horizontale) et accompagnées de la marge d'erreur (courbe verte). On observe que les estimations (barres orange) varient autour de la valeur réelle, avec parfois une surestimation marquée (cas du sondage 6) ou une sous-estimation (cas du sondage 2). La marge d'erreur, plus élevée pour certains sondages comme le 6, met en évidence l'incertitude inhérente à la méthode d'échantillonnage, bien que le sondage stratifié permette globalement une meilleure représentativité en limitant l'écart par rapport à la population réelle. Ce graphique illustre donc concrètement la variabilité des estimations selon les tirages aléatoires, tout en montrant l'intérêt du sondage stratifié pour encadrer cette incertitude.



## Partie 2 : Traitement de données d'enquête

**Contexte et objectifs :** Dans cette seconde partie, nous analysons les résultats d'une enquête menée auprès d'étudiants sur leur pratique sportive. Notre objectif est d'identifier quelles caractéristiques des étudiants sont significativement liées à leur pratique du sport.

**Nous utilisons deux outils statistiques complémentaires :** Le test du  $\chi^2$  (khi-deux) d'indépendance : Ce test nous permet de vérifier s'il existe une relation statistiquement significative entre deux variables qualitatives. L'hypothèse nulle est que les variables sont indépendantes. Et Le coefficient V de Cramer : Cette mesure nous indique l'intensité de la liaison entre deux variables qualitatives. Il varie entre 0 (aucune liaison) et 1 (liaison parfaite).

```
# Fonction de calcul du v de Cramer
v_cramer <- function(chi2, n, r, c) {
  k <- min(r - 1, c - 1) # degrés de liberté
  if (k == 0) return(NA)
  sqrt(chi2 / (n * k))
}
```

Analyse systématique des relations Nous analysons de manière systématique la relation entre la variable "sport" et toutes les autres variables qualitatives de l'enquête :

```
# Récupération de toutes les variables qualitatives sauf "sport"
vars_qualitatives <- names(enquete)[sapply(enquete, function(col) is.factor(col) || is.character(col))]
vars_qualitatives <- setdiff(vars_qualitatives, "sport")

# Initialisation du tableau de résultats
resultats <- data.frame(variable=character(), p_value=numeric(), v_cramer=numeric())
# Boucle sur chaque variable qualitative
for (var in vars_qualitatives) {
  tab <- table(enquete$sport, enquete[[var]]) # tableau croisé entre "sport" et la variable
  if (all(dim(tab) >= 2)) { # on ne garde que les tableaux 2x2 ou plus
    test <- chisq.test(tab) # test du khi² d'indépendance
    v <- v_cramer(test$statistic, sum(tab), nrow(tab), ncol(tab)) # calcul du v de Cramer
    # Ajout des résultats dans la table
    resultats <- rbind(resultats, data.frame(
      Variable = var,
      p_value = round(test$p.value, 4),
      v_cramer = round(v, 4)
    ))
  }
}
```



## Résultats et interprétation

Le tableau ci-dessous présente les 5 variables les plus significativement liées à la pratique du sport. Une p-value inférieure à 0.05 indique une relation statistiquement significative.

Variable	P_value	V_cramer
Fan	0	0.4585
raisonnonsuaps	0	0.6971
Sport avant	0	0.6976
typessportavant_1	0	0.6415
typessportavant_2	0	0.3739

### Variables les plus liées à la pratique sportive :

#### 1.sportavant ( $V = 0.6976$ , $p < 0.001$ )

- Cette variable montre la liaison la plus forte avec la pratique sportive actuelle
- Il existe une très forte corrélation entre avoir pratiqué du sport avant les études et continuer à en faire pendant les études. Cela révèle l'importance des habitudes sportives acquises antérieurement dans le maintien d'une activité physique

#### 2.raisonnonsuaps ( $V = 0.6971$ , $p < 0.001$ )

- Cette variable présente une liaison quasi-équivalente à la précédente
- Les raisons invoquées pour ne pas faire de SUAPS (Service Universitaire des Activités Physiques et Sportives) sont fortement liées à la pratique sportive générale, suggérant que les obstacles au sport universitaire reflètent des obstacles plus larges à la pratique sportive

#### 3.typessportavant\_1 ( $V = 0.6415$ , $p < 0.001$ )

- Le type de sport pratiqué avant les études influence significativement la pratique actuelle
- Cette forte liaison indique que certains types de sports (probablement les sports collectifs ou en club) favorisent davantage la continuité de la pratique sportive que d'autres

#### 4.fan ( $V = 0.4585$ , $p < 0.001$ )

- Le fait d'être fan de sport est modérément lié à la pratique personnelle
- Cette relation montre qu'il existe un lien entre l'intérêt porté au sport en tant que spectateur et la pratique personnelle, mais cette liaison reste plus modérée que les variables précédentes

#### 5.typessportavant\_2 ( $V = 0.3739$ , $p < 0.001$ )

- Un second type de sport antérieur présente également une liaison avec la pratique actuelle
- Cette variable complète l'analyse des sports pratiqués avant les études, confirmant l'impact du passé sportif sur les habitudes actuelles



## **Analyse critique :**

Ces résultats nous éclairent sur les facteurs qui influencent la pratique sportive chez les étudiants. Il est important de noter que ces analyses révèlent des associations statistiques, mais ne permettent pas d'établir des relations de causalité.

Cette SAE nous a permis d'appliquer concrètement des outils fondamentaux de la statistique inférentielle, en particulier autour de l'échantillonnage et de l'analyse de relations entre variables. Elle a montré combien le choix de la méthode d'échantillonnage peut influencer la précision des résultats. Le sondage stratifié, mieux adapté aux caractéristiques de la population, s'est révélé plus pertinent que le sondage aléatoire simple.

L'intégration des intervalles de confiance nous a également rappelé l'importance de prendre en compte l'incertitude inhérente à toute estimation, surtout dans un cadre professionnel où la rigueur des conclusions est essentielle.

L'analyse des données d'enquête nous a permis de mettre en œuvre des tests d'association entre variables qualitatives. Le recours au test du  $\chi^2$ , complété par le V de Cramer, a enrichi notre compréhension en distinguant la simple présence d'un lien de son intensité.

Enfin, plusieurs pistes restent à explorer pour aller plus loin : tester d'autres méthodes de stratification, analyser l'effet de la taille d'échantillon sur la précision, ou encore mobiliser des méthodes d'analyse multivariée. Cette expérience nous a ainsi permis de renforcer notre capacité à mobiliser la statistique pour répondre à des problématiques concrètes et variées.