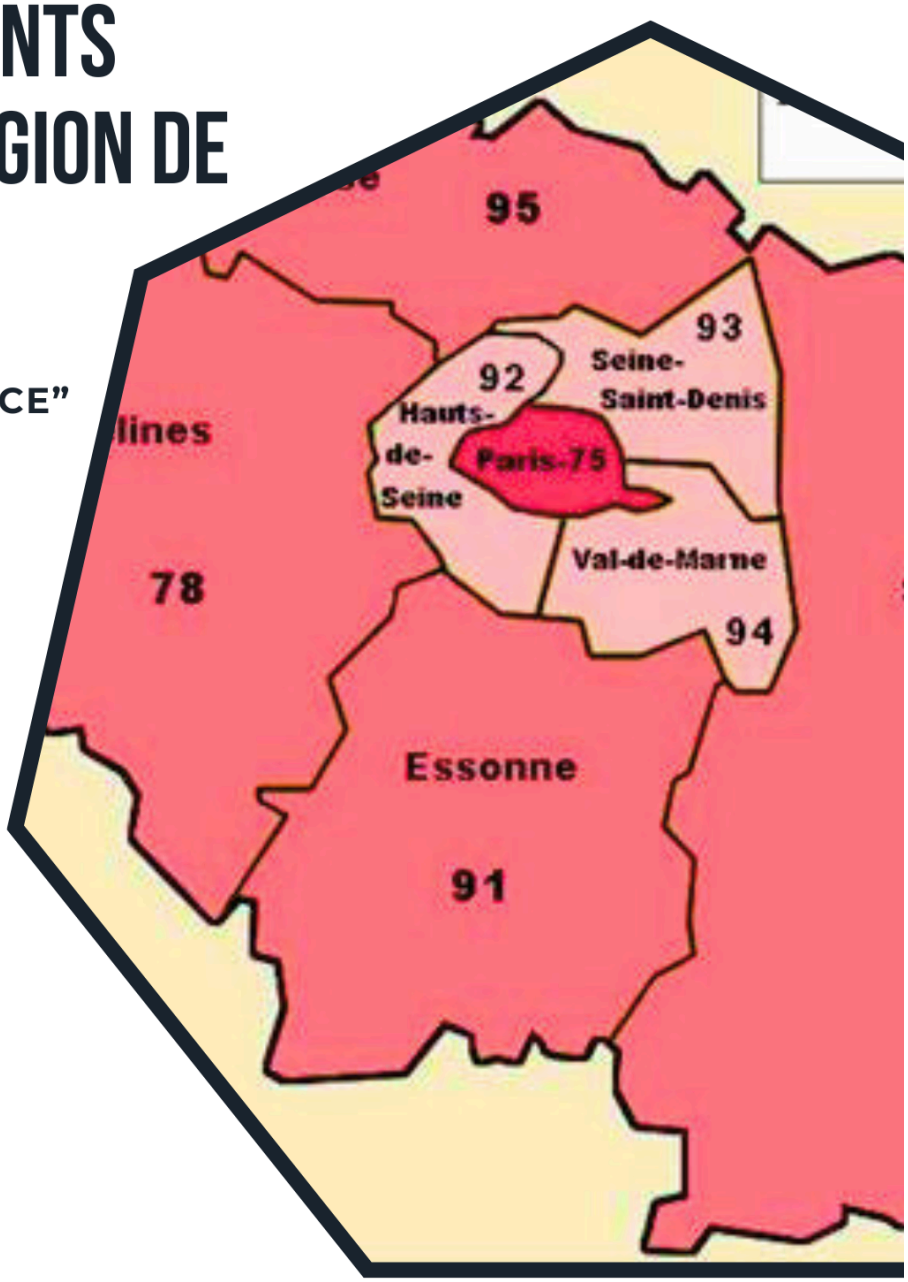
ESTIMATION DU NOMBRE D'HABITANTS D'UNE RÉGION DE FRANCE

AUDIBERT
CARRACO

"ILE-DE-FRANCE"



îledeFrance

MAI 2025

L'objectif de cette SAE est de manipuler des jeux de données avec R. Dans le cadre de la première partie « Estimation du nombre d'habitants d'une région de France ». Nous avons choisi de nous intéresser à la région Île-de-France. Celle-ci sera donc l'objet de notre étude. Pour se faire, nous avons récupéré le jeu de données contenant les données concernant la population des communes françaises. Une fois la table initiale chargée, nous créons une autre table ne contenant que les données de l'Île-de-France et en ne gardant que les variables « code département », « commune » et « population totale ». Les données sont désormais prêtes pour notre étude.

```r
# 1 Import des fichiers à notre disposition
table <- read.csv2(file="population_francaise_communes.csv",sep=";",dec=",",header=TRUE)

# Filtre
donnees <- table[table$Nom.de.la.région == "Île-de-France",c("Code.département","Commune","Population.totale") ]
head(donnees)
```

Nous commençons par calculer le nombre de communes en île de France, la population totale de la région. En résulte une population totale de 12 384 734 habitants répartis sur 1287 communes.

```r
# 2 Variable U
U = donnees$Commune
head(U)

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

Partie 1.2 Échantillonnage aléatoire simple

On se propose maintenant d'estimer la population totale de l'île-de-France à partir d'un échantillon aléatoire de 100 communes.

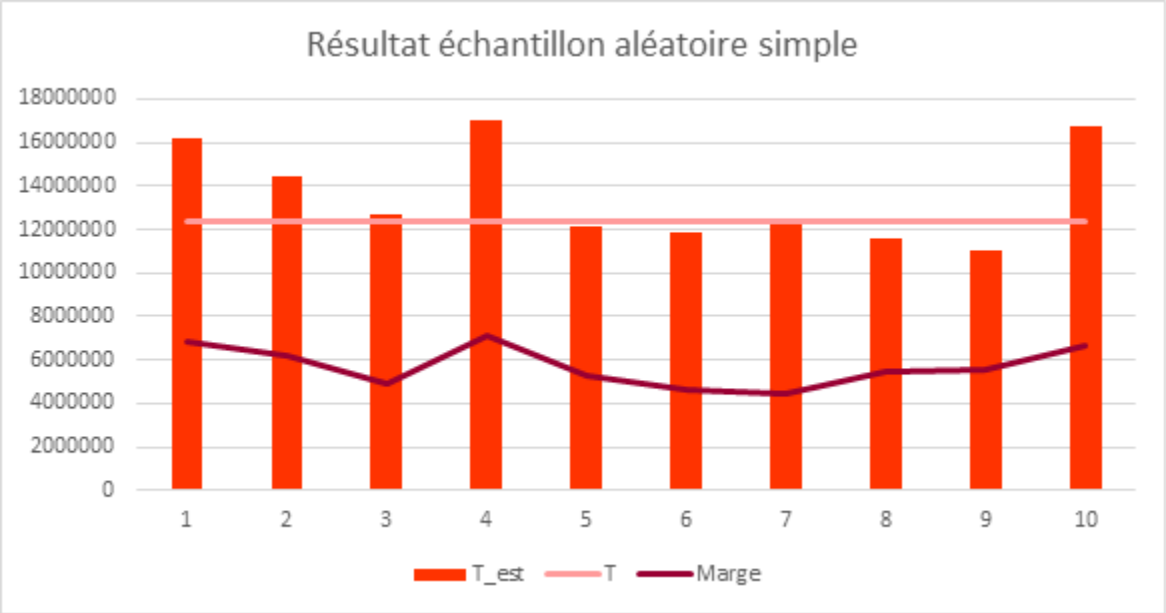
```r
# 4 Tirage aléatoire
n = 100
E = sample(U,n)
length(E)
head(E)
```

Pour ce faire,nous calculons le nombre moyen d'habitants dans l'échantillon, nous formons un Intervalle de confiance à 95 % ainsi que la marge d'erreur.

```r
# 5 Donnee tout echantillon
donnees1 = donnees[donnees$Commune %in% E, ]
head(donnees1)

# Moyenne echantilon
xbar = mean(donnees1$Population.totale)

# 6 IDC pour Mu
idcmoy = t.test(donnees1$Population.totale)$conf.int
idcmoy

# Estimation
T_est = N*xbar
T_est

# IDC de T
idcT = idcmoy*N
idcT

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

Nous répétons ces opérations une dizaine de fois dans le but de former un tableau à l'aide d'Excel qui résumerait les résultats de nos 10 expériences.

	T	T_est	IDC		Marge
			Borne inf	Borne sup	
1	12384734	16216946	9354688	23079205	6862259
2	12384734	14408360	8178756	20637964	6229604
3	12384734	12692922	7821796	17564048	4871126
4	12384734	17036392	9892254	24180530	7144138
5	12384734	12162613	6877075	17448152	5285539
6	12384734	11860052	7211239	16508866	4648813
7	12384734	12226054	7810272	16641836	4415782
8	12384734	11601107	6175653	17026562	5425454
9	12384734	11035401	5449581	16621220	5585820
10	12384734	16773716	10111860	23435571	6661856



On observe que la majorité des estimations obtenues varient autour de la valeur réelle, avec parfois une surestimation (cas des expériences 1, 2,3, 4 et 10) et parfois une sous-estimation (cas des expériences 5 à 9). Notons également que la valeur des marges d'erreur est relativement stable mais non négligeable, traduisant une incertitude due à l'échantillonnage aléatoire. Cette incertitude s'exprime par l'intervalle de confiance à 95 %, qui n'inclut pas systématiquement la valeur réelle, ce qui peut être attribué à des fluctuations liées au hasard de l'échantillonnage.

La méthode d'échantillonnage aléatoire simple que nous avons utilisée présente l'avantage d'être facile à mettre en œuvre et d'assurer une certaine objectivité, puisque chaque commune a la même probabilité d'être tirée. Toutefois, les résultats obtenus à partir des 10 échantillons montrent une variabilité non négligeable dans les estimations de la population totale. Bien que l'intervalle de confiance couvre souvent la vraie valeur, la marge d'erreur reste relativement importante, ce qui limite la précision de notre estimation.

Ce constat peut s'expliquer par l'hétérogénéité des communes d'Île-de-France : certaines sont très peuplées (comme Paris ou Boulogne-Billancourt), tandis que d'autres comptent moins de 1 000 habitants. Cette forte disparité entre les unités statistiques rend l'échantillonnage aléatoire simple peu efficace pour représenter fidèlement l'ensemble de la région.

Pour améliorer la précision de l'estimation tout en conservant un échantillon de taille raisonnable, nous pourrons utiliser une méthode différente. C'est pourquoi, dans la partie suivante, nous allons nous intéresser à l'échantillonnage aléatoire stratifié**,** qui permet de tenir compte de l'hétérogénéité de la population en répartissant les unités dans des strates plus homogènes avant le tirage. Cette approche devrait nous permettre d'obtenir des estimations plus précises et plus fiables de la population totale de l'Île-de-France.

Partie 1.2 Échantillonnage aléatoire stratifié

Dans cette deuxième approche, nous avons adopté un échantillonnage aléatoire stratifié, ce qui permet une meilleure représentativité de la population, surtout lorsqu'elle est hétérogène. Les strates ont été définies à partir des quantiles de la population totale des communes d'Île-de-France, découpant ainsi l'ensemble en six strates de taille à peu près équivalente.

```
# 1 Paramètres pour les strates
k <- 7
bornes <- quantile(donnees$Population.totale, probs = seq(0, 1, length.out = k + 1), na.rm = TRUE)
donnees$strate <- cut(donnees$Population.totale, breaks = bornes, labels = 1:k, include.lowest = TRUE)

# 2 Préparation de la table ordonnée et calcul des effectifs
datastrat <- donnees[, c("Code.département","Commune", "Population.totale", "strate")]
data <- datastrat[order(datastrat$strate), ]
Nh <- table(data$strate)
Nh
```

Nous avons utilisé la fonction cut() en nous basant sur les quantiles de la variable Population.totale afin d'attribuer à chaque commune une strate comprise entre 1 et 7. Le tableau « datastrat » obtenu contient donc les colonnes suivantes : Code.département, Commune, Population.totale et strate. Ces strates permettent de répartir la population selon le niveau de population des communes, des plus petites aux plus grandes.

```r
# 3 Calcul des poids des strates et des tailles d'échantillon pour chaque strate
n <- 100
gh <- Nh / N
nh <- round(n * Nh / N)
fh <- nh / Nh

# Estimation des paramètres stratifiés sans boucle
st <- strata(data, stratanames = c("strate"), size = nh, method = "srswr")
data1 <- getdata(data, st)
head(data1)
length((data1$Commune))
```

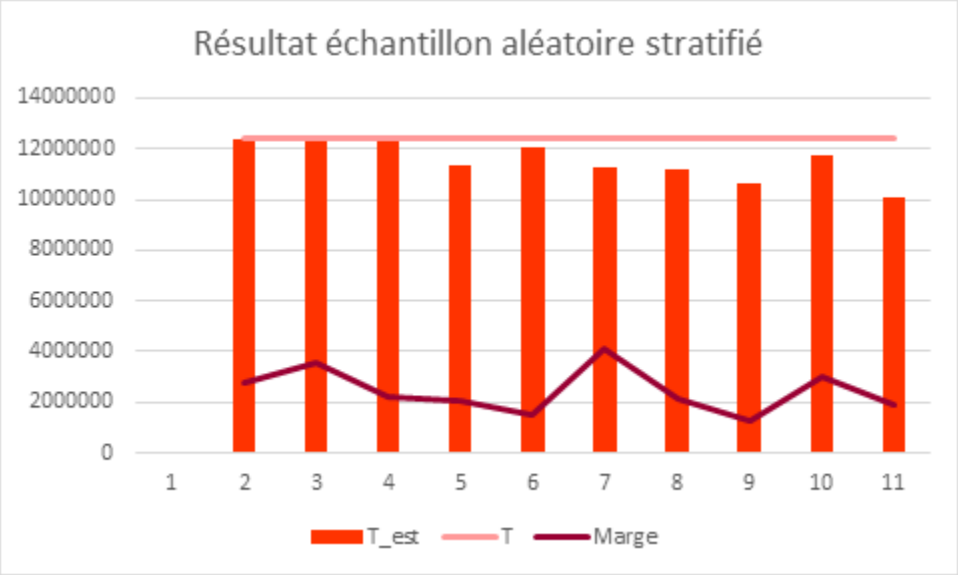
Nous avons ensuite réalisé un tirage aléatoire stratifié sans remise (srswr) avec une taille totale de n = 100 communes.

À partir de l'échantillon obtenu, nous avons extrait les sept strates et calculé la moyenne et la variance de la population totale pour chacune. Ces valeurs ont permis de calculer la moyenne pondérée stratifiée et une estimation de la variance de cette moyenne.
Nous avons ensuite pu construire un intervalle de confiance à 95 % pour le nombre moyen d'habitants par commune, et l'étendre à une estimation du total de la population T.

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

#estim du total T
Tstr=N*Xbarst
Tstr

#estimation par IDC du total T
binf=idcmoy[1]*N
binf
bsup=idcmoy[2]*N
bsup
idcT=c(binf, bsup)
idcT

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

Nous répétons ces opérations une dizaine de fois dans le but de former un tableau à l'aide d'Excel qui résumerait les résultats de nos 10 expériences.

	T	T_est	IDC		Marge
			IDC inf	IDC sup	
1	12384734	12374621	9604295	15144947	2770326
2	12384734	12536317	9012142	16060492	3524175
3	12384734	12341716	10117272	14566159	2224443
4	12384734	11314884	9255960	13373808	2058924
5	12384734	12042814	10496537	13589090	1546276
6	12384734	11233685	7142318	15325053	4091367
7	12384734	11159267	9043726	13274808	2115541
8	12384734	10637118	9377108	11897128	1260010
9	12384734	11727685	8700352	14755019	3027333
10	12384734	10075353	8136001	12014704	1939352



Par rapport à la méthode utilisée dans la partie 1.1, l'approche stratifiée présente plusieurs avantages notables. Tout d'abord, elle réduit la variabilité entre les échantillons, ce qui se traduit par une marge d'erreur plus faible. Ensuite, cette méthode permet d'obtenir une estimation plus précise de la quantité totale T. Enfin, le découpage de la population en strates offre une meilleure prise en compte des disparités démographiques entre les communes, rendant l'échantillonnage plus représentatif de la diversité observée sur le terrain.

Dans cette SAE, nous avons exploré différentes méthodes d'échantillonnage afin d'estimer la population totale de la région Île-de-France à partir des données communales. Après avoir extrait et préparé les données nécessaires, nous avons d'abord appliqué un échantillonnage aléatoire simple. Cette méthode, bien que facile à mettre en œuvre, a montré certaines limites en raison de la forte hétérogénéité entre les communes franciliennes. Les estimations obtenues

présentaient une variabilité notable et des marges d’erreur parfois importantes, traduisant une précision limitée.

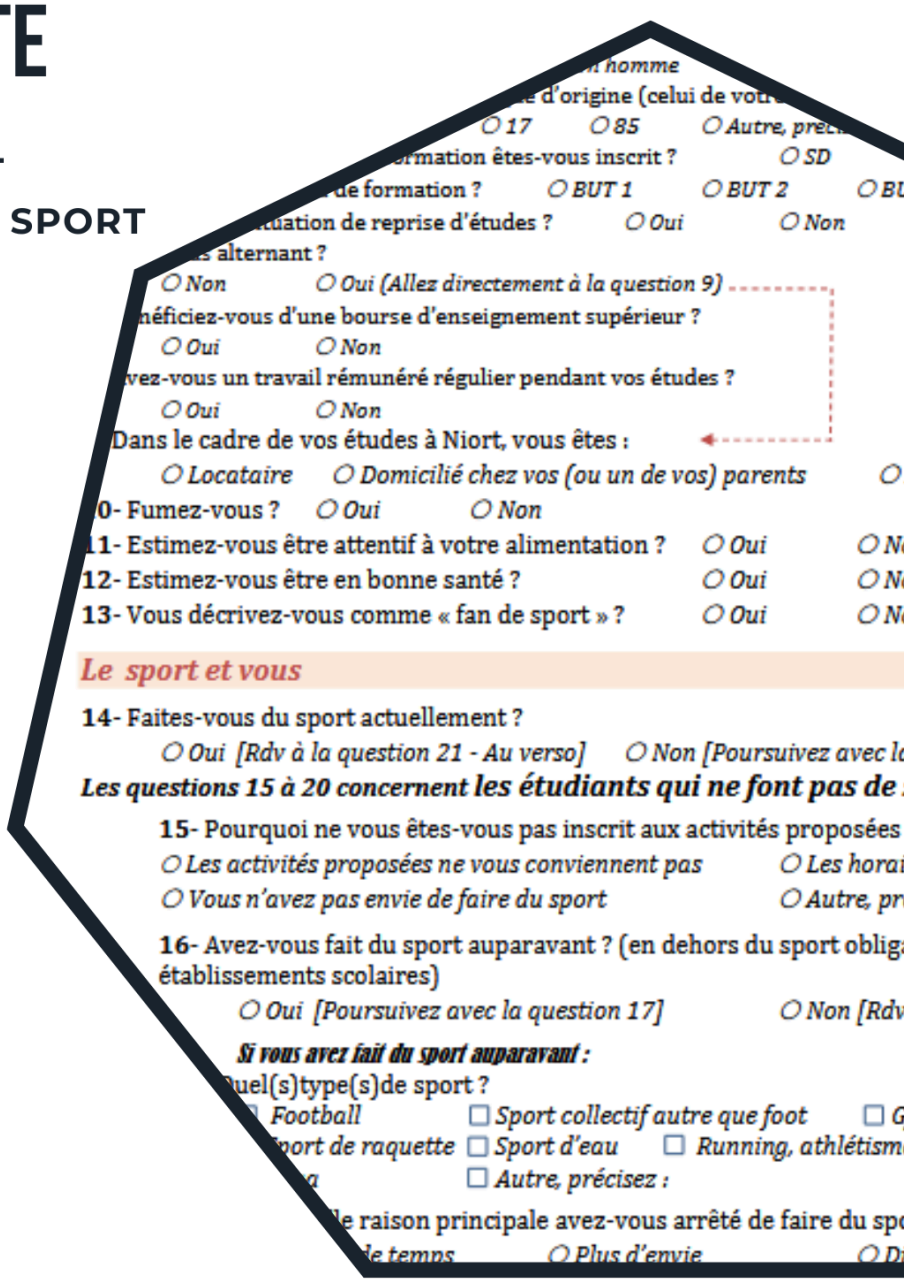
Face à ce constat, nous avons ensuite mis en œuvre un échantillonnage aléatoire stratifié, en répartissant les communes en strates homogènes selon leur taille démographique. Cette approche a permis de réduire la variabilité des estimations et d’obtenir des résultats globalement plus précis et plus fiables. Le recours à la stratification s’est ainsi révélé pertinent pour mieux représenter les disparités internes à la région.

En définitive, cette étude nous souligne l’intérêt d’adapter la méthode d’échantillonnage à la structure de la population étudiée, en vue d’améliorer la qualité des estimations.

TRAITEMENT DE DONNÉES D'ENQUÊTE

AUDIBERT
CARRACO

ÉTUDIENTS ET PRATIQUE DU SPORT



*homme
d'origine (celui de votr
○ 17 ○ 85 ○ Autre, prec
rmation êtes-vous inscrit ? ○ SD
de formation ? ○ BUT 1 ○ BUT 2 ○ BU
tuation de reprise d'études ? ○ Oui ○ Non
s alternant ?
○ Non ○ Oui (Allez directement à la question 9)
néficiez-vous d'une bourse d'enseignement supérieur ?
○ Oui ○ Non
vez-vous un travail rémunéré régulier pendant vos études ?
○ Oui ○ Non
Dans le cadre de vos études à Niort, vous êtes :
○ Locataire ○ Domicilié chez vos (ou un de vos) parents ○
0- Fumez-vous ? ○ Oui ○ Non
11- Estimez-vous être attentif à votre alimentation ? ○ Oui ○ N
12- Estimez-vous être en bonne santé ? ○ Oui ○ N
13- Vous décrivez-vous comme « fan de sport » ? ○ Oui ○ N

Le sport et vous

14- Faites-vous du sport actuellement ?
○ Oui [Rdv à la question 21 - Au verso] ○ Non [Poursuivez avec l
Les questions 15 à 20 concernent les étudiants qui ne font pas de
15- Pourquoi ne vous êtes-vous pas inscrit aux activités proposées
○ Les activités proposées ne vous conviennent pas ○ Les horai
○ Vous n'avez pas envie de faire du sport ○ Autre, pr
16- Avez-vous fait du sport auparavant ? (en dehors du sport oblig
établissements scolaires)
○ Oui [Poursuivez avec la question 17] ○ Non [Rdv
Si vous avez fait du sport auparavant :
uel(s)type(s)de sport ?
☐ Football ☐ Sport collectif autre que foot ☐ G
port de raquette ☐ Sport d'eau ☐ Running, athlétism
☐ Autre, précisez :
e raison principale avez-vous arrêté de faire du spo
e temps ○ Plus d'envie ○ D*

BUT Science des données Niort

MAI 2025

L'objectif de cette SAE est de manipuler des jeux de données avec R. Dans le cadre de la seconde partie, nous avons choisi d'analyser les résultats d'une enquête menée auprès des étudiants sur leur pratique du sport.

Partie 2 : Traitement de données d'enquête

L'objectif est d'étudier les liens éventuels entre la pratique du sport et d'autres variables qualitatives (comme le sexe, le statut d'alternant, le département de formation, etc.).
Premièrement, on commence par importer le fichier EnqueteSportEtudiant2024.csv contenant les réponses de l'enquête.

```
139 # 1 Import des fichiers à notre disposition
140 tablesport = read.csv2("EnqueteSportEtudiant2024.csv",sep = ";", dec = ";", header = TRUE)
```

La table contient une ligne par étudiant ayant répondu à l'enquête. Chaque ligne correspond donc à un individu. Les variables sont essentiellement qualitatives (ex. : sport, sexe, alternance, niveau d'études, logement, alimentation...).

```
145 # 4 Croisement de la variable sport avec le sexe/alternance/département de formation
146 TCD_Sexe = table(tablesport$sport, tablesport$sexe)
147 TCD_Alternant = table(tablesport$sport, tablesport$alternant)
148 TCD_Dept = table(tablesport$sport, tablesport$deptformation)
149 TCD_niveau = table(tablesport$sport, tablesport$niveau)
150 TCD_logement = table(tablesport$sport, tablesport$logement)
151 TCD_fumer = table(tablesport$sport, tablesport$fumer)
152 TCD_alimentation = table(tablesport$sport, tablesport$alimentation)
153 TCD_sante = table(tablesport$sport, tablesport$sante)
```

On croise la variable sport avec différentes autres variables qualitatives :

- Sexe
- Statut d'alternant
- Département de formation
- Niveau d'études
- Type de logement
- Fait de fumer ou non
- Type d'alimentation
- État de santé ressenti

Les tableaux croisés obtenus permettent de visualiser les répartitions et d'identifier rapidement d'éventuelles différences notables. Pour chaque tableau croisé, on effectue un test du khi² afin de déterminer si la variable *sport* est liée de façon significative à l'autre variable testée.

```r
# 5 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

khideux_Dept= chisq.test(TCD_Dept)
khideux_Dept
#p-value = 0.004557 --> 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_fumer)
khideux_fumer
#p-value = 0.6666 --> très élevé
```

La variable sexe, alimentation et département de formation ressortir avec une p valeur très faible. On calcule donc le V de Cramer, qui permet de mesurer l'intensité du lien.

```r
# 6 V de Cramer
n<-dim(tablesport)[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
# bien

n<-dim(tablesport)[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
# bien et mieux

n<-dim(tablesport)[1]
p <- nrow(TCD_Dept)
q <- ncol(TCD_Dept)
m <- min(p-1, q-1)
V_Dept =sqrt(khideux_Dept$statistic/(n*m))
V_Dept
# bien
```

Par la suite nous décidons de regrouper toutes ces informations dans un tableau excel.

	Variable	V de Cramer	Signification
1	V_Sexe	0.198274	bien
2	V_alimentation	0.2107832	bien et mieux
3	V_Dept	0.1582276	bien

On en conclut que les variables sexes, alimentation et département de formation ont un lien significatif avec la pratique du sport.

Cela signifie que c'est la variable alimentation qui a la plus grande association avec la pratique du sport parmi celles testées.

Les tests montrent que la pratique du sport chez les étudiants varie selon certaines caractéristiques, en particulier le sexe, le département d'études et surtout le type d'alimentation.

Ces résultats peuvent orienter des campagnes de sensibilisation ou des actions ciblées pour promouvoir le sport chez certaines catégories d'étudiants.

En revanche, d'autres variables comme le logement, la santé perçue ou le fait de fumer ne semblent pas liées à cette pratique.

Cette seconde partie nous a permis d'analyser les résultats d'une enquête portant sur la pratique du sport chez les étudiants, en explorant les liens éventuels entre cette pratique et diverses variables qualitatives. À travers des croisements de données et des tests statistiques (khi² et V de Cramer), nous avons mis en évidence que certaines caractéristiques, notamment le sexe, le département de formation et surtout le type d'alimentation, sont significativement liées à la pratique sportive.

Parmi ces facteurs, l'alimentation ressort comme étant la variable la plus fortement associée au sport, suggérant que les habitudes alimentaires peuvent influencer, ou être influencées par, un mode de vie plus actif. En revanche, d'autres variables comme le type de logement, la santé perçue ou le fait de fumer ne montrent pas de lien statistique significatif avec la pratique sportive.