Projet sur l’analyse univariée et les tests statistiques

La problématique est l’étude du surpoids et de l’obésité et leurs conséquences sur la santé

Océane ETOUBLEAU-ETIENNE et KOSINSKI Justine

★ Introduction

Cette étude s'intéresse à une population de 60 individus vivant aux Etats-Unis d'Amérique , représentatifs de la population américaine. Elle vise à examiner la prévalence du surpoids et de l'obésité, ainsi que leurs conséquences sur la santé. Les données utilisées dans cette étude proviennent du département de la santé et des services humains américains, ce qui garantit une source fiable et officielle pour notre analyse. En ce qui concerne la mise en contexte, nous ne sommes pas sans savoir que le surpoids et l'obésité sont des problèmes de santé publique majeurs aux Etats-Unis, avec des implications considérables sur la santé individuelle et le système de soins de santé. C'est pourquoi l'Organisation Mondiale de la Santé (OMS) a établi l'Indice de Masse Corporelle (IMC) comme une norme internationale pour évaluer les risques associés au surpoids. L'OMC a également défini des intervalles standard d'IMC en se basant sur la relation statistique entre l'IMC et le taux de mortalité. De plus, un tour de taille excessif est reconnu comme un facteur prédictif de maladies cardiaques et du diabète de type 2. Ainsi en comprenant la distribution de l'IMC, du tour de la taille et d'autres paramètres liés à la santé, nous pouvons cerner les risques et maladies associées au surpoids dans cette population américaine.

Nom des variables avec leur descriptif, leur unité et leur type :

- SEXE : variable qualitative nominale à 2 modalités (codé 0 pour les hommes et 1 pour les femmes)
- AGE : variable quantitative continue en années (36 valeurs différentes > 15 valeurs différentes)
- TAILLE : variable quantitative continue en cm (50 valeurs différentes)
- TTAILLE : variable quantitative continue en cm (51 valeurs différentes)
- CHOL : variable quantitative continue en mg/dl (58 valeurs différentes)
- IMC : variable quantitative continue en kg/m^2 (53 valeurs différentes)

★ Transformations et créations de variables

Création de la variable IMCC

La variable notée IMCC correspond à la variable IMC discrétisée en trois classes en utilisant les standards de l'OMS. Ces standards sont les suivants :

- IMC < 25 : corpulence normale
- 25 <= IMC < 30 : surpoids
- IMC >= 30 : obésité

Pour cela, nous avons utilisé la fonction `cut`.

```r
#Importation des données
setwd("/Users/justinekosinski/Desktop/MIAGE/L3/S1/Statistiques/Projet1")
data_imc_TP2_groupe1=read.table("data_imc_TP2_groupe1.txt",header=TRUE)
is.data.frame(data_imc_TP2_groupe1)
View(data_imc_TP2_groupe1)
attach(data_imc_TP2_groupe1)

#Question 2
#(a) IMCC correspond à la variable IMC discrétisée en trois classes en utilisant les standards de l'OMS
IMCC <- cut(data_imc_TP2_groupe1$IMC, breaks = c(-Inf, 25, 30, Inf), labels = c("Corpulence normale IMC<25", "Surpoids 25<=IMC<30", "Obésité IMC >=30"))
IMCC
```

```
> IMCC <- cut(data_imc_TP2_groupe1$IMC, breaks = c(-Inf, 25, 30, Inf), labels = c("Corpulence normale IMC<25", "Surpoids 25<=IMC<30", "Obésité IMC >=30"))
> IMCC
 [1] Corpulence normale IMC<25 Surpoids 25<=IMC<30       Corpulence normale IMC<25
 [4] Obésité IMC >=30          Surpoids 25<=IMC<30       Surpoids 25<=IMC<30
 [7] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Obésité IMC >=30
[10] Surpoids 25<=IMC<30       Obésité IMC >=30          Surpoids 25<=IMC<30
[13] Corpulence normale IMC<25 Corpulence normale IMC<25 Obésité IMC >=30
[16] Corpulence normale IMC<25 Surpoids 25<=IMC<30       Obésité IMC >=30
[19] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Obésité IMC >=30
[22] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Surpoids 25<=IMC<30
[25] Corpulence normale IMC<25 Surpoids 25<=IMC<30       Corpulence normale IMC<25
[28] Corpulence normale IMC<25 Corpulence normale IMC<25 Surpoids 25<=IMC<30
[31] Corpulence normale IMC<25 Corpulence normale IMC<25 Surpoids 25<=IMC<30
[34] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Corpulence normale IMC<25
[37] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Corpulence normale IMC<25
[40] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Obésité IMC >=30
[43] Corpulence normale IMC<25 Surpoids 25<=IMC<30       Surpoids 25<=IMC<30
[46] Surpoids 25<=IMC<30       Corpulence normale IMC<25 Surpoids 25<=IMC<30
[49] Corpulence normale IMC<25 Surpoids 25<=IMC<30       Corpulence normale IMC<25
[52] Corpulence normale IMC<25 Corpulence normale IMC<25 Corpulence normale IMC<25
[55] Corpulence normale IMC<25 Obésité IMC >=30          Corpulence normale IMC<25
[58] Corpulence normale IMC<25 Obésité IMC >=30          Corpulence normale IMC<25
Levels: Corpulence normale IMC<25 Surpoids 25<=IMC<30 Obésité IMC >=30
```

Transformation du taux de cholestérol en g/l

Le taux de cholestérol est en mg/dl, unité que l'on n'utilise pas en France. Afin d'obtenir le taux de cholestérol en g/l il suffit simplement de diviser par 100.

```r
#(b) Transformation appropriée pour obtenir le taux de cholestérol en g/l
CHOL <- data_imc_TP2_groupe1$CHOL / 100
CHOL
```

```
> #(b) Transformation appropriée pour obtenir le taux de cholestérol en g/l
> CHOL <- data_imc_TP2_groupe1$CHOL / 100
> CHOL
 [1]  1.76  0.31  1.27  6.90  1.38  2.50  4.16  3.39  2.73  2.65  2.72  6.49  1.39  6.13  4.66  7.40  7.02
[18]  9.72  6.38  3.16  2.88  6.56  2.65  0.75  5.22  7.62  9.57  5.78  5.90 12.52  1.81  9.20  1.12  0.89
[35]  2.37  2.07  1.73  1.30  1.26  1.25  1.49  2.93  2.67  3.09  2.54  1.75  0.62  4.47  0.08  1.49  0.94
```

Labels appropriés

La variable « SEXE » est caractérisé par un codage. En effet, elle est codée 0 pour les hommes et 1 pour les femmes. Pour que R identifie ces variables comme des variables qualitatives on utilise la fonction `factor`.

```r
#(c) La variable sexe est de type integer, or sexe est une qualitative
# donc on va transformer en "factor" ces variables qualitatives
SEXE=factor(data_imc_TP2_groupe1$SEXE,labels=c("Homme","Femme"))
SEXE
```

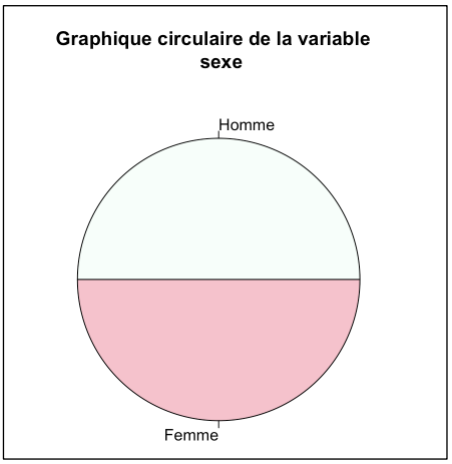
```
> SEXE=factor(data_imc_TP2_groupe1$SEXE,labels=c("Homme","Femme"))
> SEXE
 [1] Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme
[18] Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme Femme Femme Femme Femme
[35] Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme
[52] Femme Femme Femme Femme Femme Femme Femme Femme Femme
Levels: Homme Femme
```

★ Étude descriptive univariée

Variable qualitative : On commence par la variable « *SEXE* ».

```
table(SEXE) # tableau de distribution en effectifs de la variable sexe
prop.table(table(SEXE)) # tableau de distribution en fréquences
pie(prop.table(table(SEXE)),main="Graphique circulaire de la variable  \n sexe",
    col=c("mintcream","pink"))
```

```
> #Question 3 : Analyse univariée
> #Variables qualitatives
> #Commençons par la variable "SEXE"
> table(SEXE) # tableau de distribution en effectifs de la variable sexe
SEXE
Homme Femme
   30    30
> prop.table(table(SEXE)) # tableau de distribution en fréquences
SEXE
Homme Femme
  0.5   0.5
```

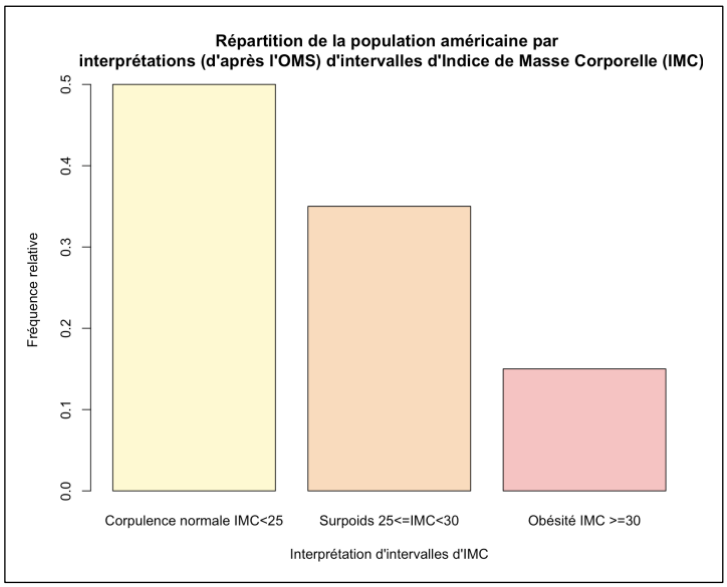


La répartition des sexes dans l'échantillon est équilibrée, avec 50% d'hommes et 50% de femmes. Cela indique une représentation égale des deux genres au sein de l'échantillon ce qui permet de prendre en compte les caractéristiques des deux groupes de manière équilibrée.

Vient ensuite la variable « *IMCC* » (IMC en classes).

```
table(IMCC)
prop.table(table(IMCC))
barplot(prop.table(table(IMCC)),main="Répartition de la population américaine par \n interprétations (d'après l'OMS) d'intervalles d'Indice de Masse Corporelle (IMC)",
        ylab="Fréquence relative", xlab="Interprétation d'intervalles d'IMC",col=c("lemonchiffon","peachpuff","rosybrown1"))
```

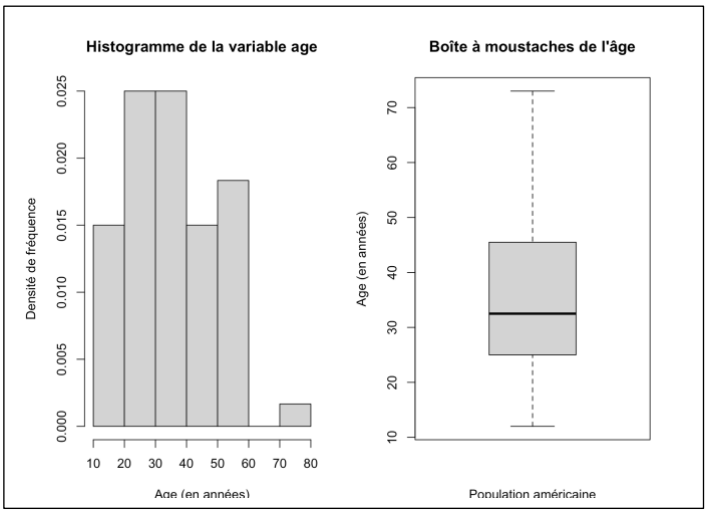
```
> #Passons à la variable "IMCC"
> table(IMCC)
IMCC
Corpulence normale IMC<25       Surpoids 25<=IMC<30          Obésité IMC >=30
                       30                        21                         9
> prop.table(table(IMCC))
IMCC
Corpulence normale IMC<25       Surpoids 25<=IMC<30          Obésité IMC >=30
                     0.50                      0.35                      0.15
```



La majorité des individus se situent dans la catégorie "Corpulence normale" (IMC<25) : ils représentent 50% de la population américaine. 35% de cette même population est considérée en "Surpoids" avec un IMC compris entre 25 et 30 (exclu). Les personnes obèses sont peu nombreuses : environ 15% de la population américaine.

Variable quantitative : Débutons par la variable « *AGE* ».

```
par(mfrow=c(1,2))
#Essentiel des résumés numériques
summary(AGE)
cv=sd(AGE)/mean(AGE)# coefficient de variation : permet de mesurer la dispersion
#Histogramme
hist(AGE,freq=FALSE,xlab="Age (en années)",
     ylab="Densité de fréquence",main="Histogramme de la variable age")
#Boîte à moustaches
boxplot(AGE,xlab="Population américaine",ylab="Age (en années)",
        main="Boîte à moustaches de l'âge")
```



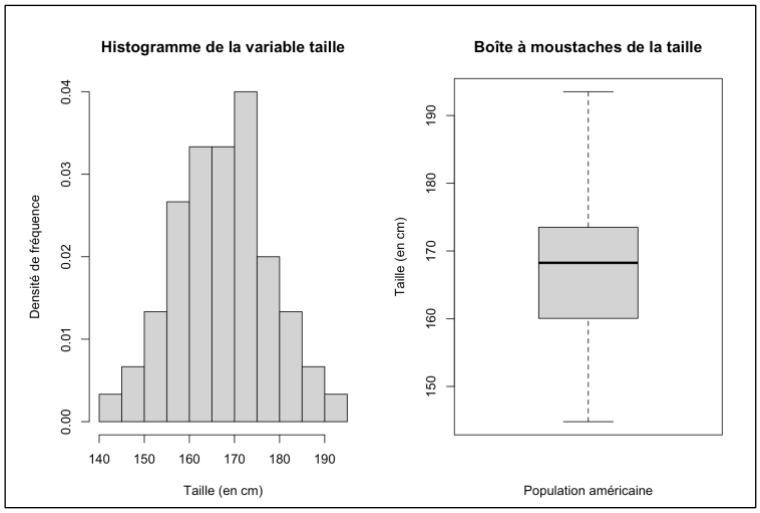

```
> #Variables quantitatives continues
> #On étudie d'abord la variable "AGE"
> par(mfrow=c(1,2))
> #Essentiel des résumés numériques
> summary(AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   25.00   32.50   35.52   45.25   73.00
```

Data	
data_imc_TP2_groupe1	60 obs. of 6 variables
Values	
CHOL	num [1:60] 1.76 0.31 1.27 6.9 1.38 2.5 4.16 3
cv	0.382046610905055
IMCC	Factor w/ 3 levels "Corpulence normale IMC<25
SEXE	Factor w/ 2 levels "Homme","Femme": 1 1 1 1 1

On peut voir que la médiane est inférieure à la moyenne, ce qui suggère une légère asymétrie positive, on remarque également une concentration autour des âges de 29 et 32 ans dans la distribution. Cela signifie que quelques individus ont des âges élevés, ce qui fait augmenter la moyenne par rapport à la médiane. Ce que nous confirme la boîte à moustaches. La plage d'âges va de 12 à 75 ans, ce qui montre la diversité de l'âge dans votre échantillon. Toutes les personnes de notre échantillon, ont en moyenne 35.52 ans. La médiane indique que la moitié des individus ont un âge inférieur ou égal à 32.5 ans, tandis que l'autre moitié a un âge supérieur. Le coefficient de variation, égal à 0.38 (>0.25) révèle que l'âge des individus n'est pas fortement dispersé par rapport à la moyenne, cependant il existe tout de même une certaine variabilité.

Passons à la variable « *TAILLE* ».

```
#Essentiel des résumés numériques
summary(TAILLE)
cv=sd(TAILLE)/mean(TAILLE)# coefficient de variation : permet de mesurer la dispersion
#Histogramme
hist(TAILLE,freq=FALSE,xlab="Taille (en cm)",
     ylab="Densité de fréquence",main="Histogramme de la variable taille")
#Boîte à moustaches
boxplot(TAILLE,xlab="Population américaine",ylab="Taille (en cm)",
        main="Boîte à moustaches de la taille")
```



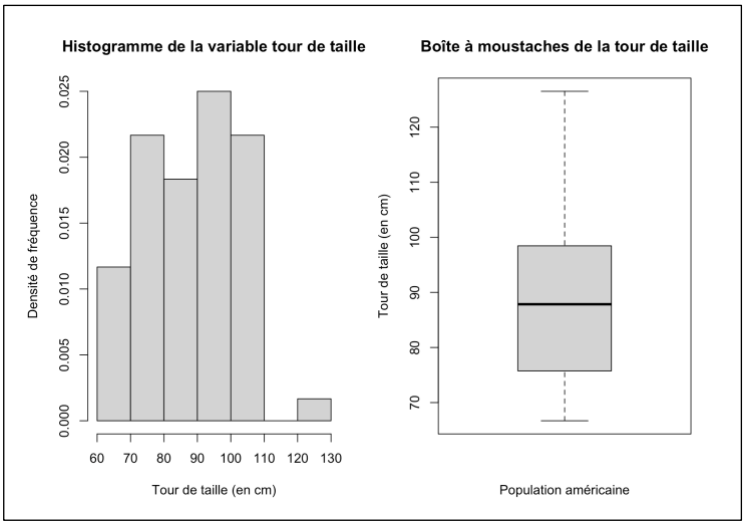

```
> #Essentiel des résumés numériques
> summary(TAILLE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  144.8   160.2   168.2   167.2   173.5   193.5
```

Data	
data_imc_TP2_groupe1	60 obs. of 6 variables
Values	
CHOL	num [1:60] 1.76 0.31 1.27 6.9 1.38 2
cv	0.0619030712364866
IMCC	Factor w/ 3 levels "Corpulence norma
SEXE	Factor w/ 2 levels "Homme","Femme":

On constate que la médiane est proche de la moyenne, ce qui suggère une distribution relativement symétrique des tailles avec quelques individus présentant des tailles extrêmes.La plage des tailles va de 144.8 cm à 193.5 cm, montrant une variabilité considérable dans les données. Le coefficient de variation étant tres faible (0.06) indique une distribution de données relativement homogène où la plupart des valeurs sont proches de la moyenne et peu de valeurs s'éloignent de manière significative de cette moyenne.La moyenne de taille de la population américaine est de 167.2 cm. On remarque que 25% de la population américaine mesure une taille inférieure ou égale à 160.2 cm.

Intéressons-nous à la variable « *TTAILLE* ».

```
#Essentiel des résumés numériques
summary(TTAILLE)
cv=sd(TTAILLE)/mean(TTAILLE)# coefficient de variation : permet de mesurer la dispersion
#Histogramme
hist(TTAILLE,freq=FALSE,xlab="Tour de taille (en cm)",
     ylab="Densité de fréquence",main="Histogramme de la variable tour de taille")
#Boîte à moustaches
boxplot(TTAILLE,xlab="Population américaine",ylab="Tour de taille (en cm)",
        main="Boîte à moustaches de la tour de taille")
```



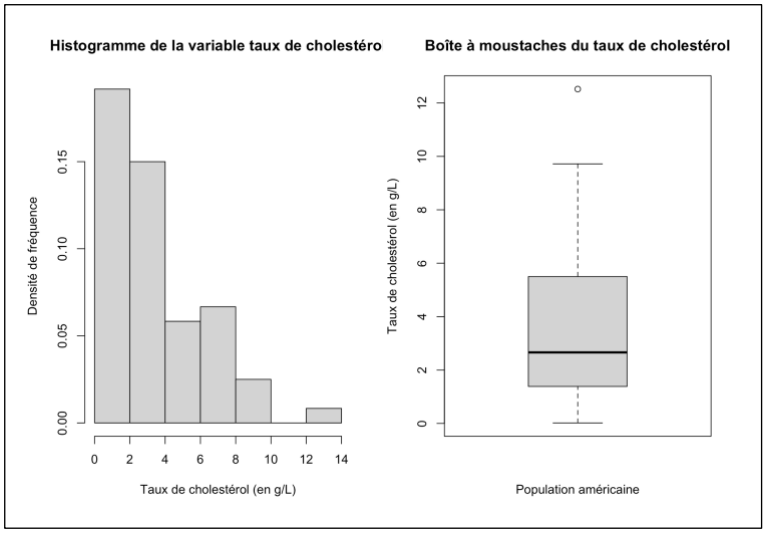

```
> summary(TTAILLE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  66.70   75.83   87.85   88.07   98.22  126.50
```

Data	
data_imc_TP2_groupe1	60 obs. of 6 variables
Values	
CHOL	num [1:60] 1.76 0.31 1.27 6.9 1.38 2.5 4.16 3.
cv	0.151762440887361
IMCC	Factor w/ 3 levels "Corpulence normale IMC<25"
SEXE	Factor w/ 2 levels "Homme","Femme": 1 1 1 1 1

La distribution du tour de taille est étendue, avec une variabilité considérable dans les données. Vous pouvez voir que la moyenne est légèrement supérieure à la médiane, suggérant une légère asymétrie positive. La concentration de la population autour des valeurs proches de la médiane montre une cohérence relative, mais il y a tout de même des individus avec des mesures de tour de taille plus éloignées de la moyenne, ce qui explique le coefficient de variation (0.15). L'histogramme et la boîte à moustaches, nous permette de mieux visualiser cette distribution. En effet, on constate que la moustache supérieure est plus longue que la moustache inférieure. La moyenne du tour de taille est d'environ 88.07 cm. 75% des américains a un tour de taille inférieur ou égal à 98.22 cm.

On regarde maintenant la variable « *CHOL* ».

```
#Essentiel des résumés numériques
summary(CHOL)
cv=sd(CHOL)/mean(CHOL)# coefficient de variation : permet de mesurer la dispersion
#Histogramme
hist(CHOL,freq=FALSE,xlab="Taux de cholestérol (en g/L)",
     ylab="Densité de fréquence",main="Histogramme de la variable taux de cholestérol")
#Boîte à moustaches
boxplot(CHOL,xlab="Population américaine",ylab="Taux de cholestérol (en g/L)",
        main="Boîte à moustaches du taux de cholestérol")
```




```
> summary(CHOL)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.020   1.387   2.660   3.453   5.360  12.520
```

Data	
data_imc_TP2_groupe1	60 obs. of 6 variables
Values	
alpha	0.05
CHOL	num [1:60] 1.76 0.31 1.27 6
cv	0.800166085338242
IMCC	Factor w/ 3 levels "Corpule

On identifie 1 valeur extrême du côté élevé (12.52 g/L) de la distribution du taux de cholestérol. La moyenne du taux de cholestérol qui vaut 3.453 g/L est influencée par la valeur extrême et est légèrement plus élevée que la médiane, ce qui suggère une distribution légèrement asymétrique positive. En d'autres termes, la valeur extrême tire la moyenne vers le haut. La population étudiée présente une diversité importante en termes de taux de cholestérol. Le cv égal à 0.8 indique que les données de taux de cholestérol dans notre échantillon sont dispersées sur une plage relativement large.

★ Test unilatéral

Dans cette partie nous allons effectuer un test unilatéral en utilisant la fonction t.test de R avec l'argument `alternative` défini sur « `greater` ». On va comparer la moyenne au taux maximal recommandé qui est de 2 g/l. On stock le résultat dans une variable resultat_test.

```r
#Question 4 : Test unilatéral
# Niveau de signification (alpha)
alpha <- 0.05
resultat_test <- t.test(CHOL, alternative = 'greater', mu = 2)
cat("Statistique de test :", resultat_test$statistic, "\n")
cat("Valeur-p :", resultat_test$p.value, "\n")
if (resultat_test$p.value < alpha) {
  cat("La population a une proportion significativement plus élevée de personnes avec un taux de cholestérol supérieur à 2 g/L (p < alpha).\n")
} else {
  cat("Il n'y a pas suffisamment de preuves pour conclure que la population a une proportion significativement plus élevée de personnes avec un
      taux de cholestérol supérieur à 2 g/L (p >= alpha).\n")
}
```

```
> #Question 4 : Test unilatéral
> # Niveau de signification (alpha)
> alpha <- 0.05
> resultat_test <- t.test(CHOL, alternative = 'greater', mu = 2)
> cat("Statistique de test :", resultat_test$statistic, "\n")
Statistique de test : 4.072929
> cat("Valeur-p :", resultat_test$p.value, "\n")
Valeur-p : 7.016801e-05
> if (resultat_test$p.value < alpha) {
+   cat("La population a une proportion significativement plus élevée de personnes avec un taux de cholestérol supérieur à 2 g/L (p < alpha).\n")
+ } else {
+   cat("Il n'y a pas suffisamment de preuves pour conclure que la population a une proportion significativement plus élevée de personnes avec un
+       taux de cholestérol supérieur à 2 g/L (p >= alpha).\n")
+ }
La population a une proportion significativement plus élevée de personnes avec un taux de cholestérol supérieur à 2 g/L (p < alpha).
```

Data	
data_imc_TP2_groupe1	60 obs. of 6 variables
resultat_test	List of 10
Values	
alpha	0.05
CHOL	num [1:60] 1.76 0.31 1.27 6.9 1
cv	0.800166085338242
IMCC	Factor w/ 3 levels "Corpulence
SEXE	Factor w/ 2 levels "Homme","Fem

On obtient une p-value = 7.016801e-05 (< alpha) ce qui nous amène à conclure que la population étudiée a un taux de cholestérol strictement supérieur au taux maximal recommandé qui est de 2g/l.

★ Liaisons entre les variables de l'étude

Pour répondre à ces questions, nous allons réaliser des analyses bivariées. Elles vont nous permettre d'étudier la liaison entre variables à l'aide de tests d'hypothèses.

L'IMC (en classes) dépend-il du sexe ?

Les variables de ce test sont toutes deux qualitatives. Nous allons donc effectuer un test du χ^2 d'indépendance, avec `R : chisq.test`. Pour cela, nous commençons par établir un tableau de contingence qui va présenter simultanément et de manière croisée l'IMC et le sexe. Il met en évidence la notion de dépendance entre les variables.

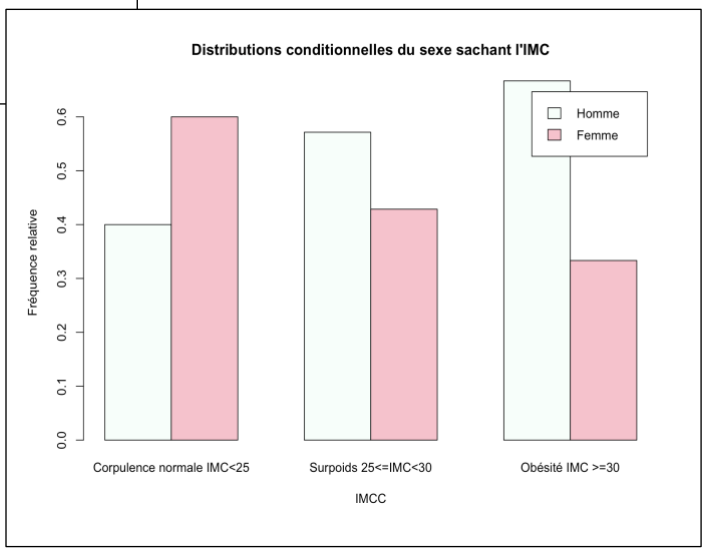
```r
par(mfrow=c(1,1))
#(a) L'IMCC dépend-il du sexe ?
#Tableau de contingence du couple (IMCC, sexe)
tableauIS = table(IMCC, SEXE)
tableauIS
addmargins(tableauIS)
#Tableau des profils-lignes : on a plutôt envie de comparer comment l'IMC varie en fonction du sexe
round(prop.table(tableauIS,1),digits=3)
#60% de la population américaine ayant une corpulence normale (IMC<25) sont des femmes

#Diagrammes en colonnes
barplot(t(prop.table(tableauIS,1)),beside=TRUE,col=c("mintcream","pink"),
        main="Distributions conditionnelles du sexe sachant l'IMC",xlab= "IMCC",ylab="Fréquence relative"
        legend.text=TRUE)
```

```
> #Question 5
> par(mfrow=c(1,1))
> #(a) L'IMCC dépend-il du sexe ?
> #Tableau de contingence du couple (IMCC, sexe)
> tableauIS = table(IMCC, SEXE)
> tableauIS
                          SEXE
IMCC                       Homme Femme
  Corpulence normale IMC<25    12    18
  Surpoids 25<=IMC<30          12     9
  Obésité IMC >=30              6     3
> addmargins(tableauIS)
                          SEXE
IMCC                       Homme Femme Sum
  Corpulence normale IMC<25    12    18  30
  Surpoids 25<=IMC<30          12     9  21
  Obésité IMC >=30              6     3   9
  Sum                          30    30  60
```

Une fois le tableau de contingence établit, nous passons à l'étape des profils-lignes et profils-colonnes ainsi que le graphique associé. Suivant notre contexte, il nous a semblé plus intéressant de montrer le profils-lignes c'est-à-dire les distributions conditionnelles de l'IMC sachant le sexe.

```
> #Tableau des profils-lignes : on a plutôt envie de comparer comment l'IMC varie en fonction du sexe
> round(prop.table(tableauIS,1),digits=3)
                          SEXE
IMCC                       Homme Femme
  Corpulence normale IMC<25 0.400 0.600
  Surpoids 25<=IMC<30       0.571 0.429
  Obésité IMC >=30          0.667 0.333
```



On constate qu'il y a des différences notables dans la répartition des catégories d'IMC (corpulence normale, surpoids, obésité) entre les hommes et les femmes. La catégorie "Surpoids" comporte une proportion plus élevée d'hommes (57%) par rapport aux femmes (43%). L'obésité est plus prévalente chez les hommes, en effet 67% de la population étant obèse sont des hommes. Les femmes quant à elles représentent 60% de la population américaine ayant une corpulence normale.Cela incite à penser que l'IMC dépend du sexe.

On peut à présent passer au test. Il se déroule en plusieurs étapes :

1- On définit notre hypothèse nulle et notre hypothèse alternative

2- On décide de la valeur alpha = 0.05, cela implique qu'on prend un risque de 5% de conclure que les deux variables sont indépendantes alors qu'en réalité elles ne le sont pas.
3- On vérifie les conditions de validité
4- On effectue le test et on en tire nos conclusions

On pose H0 : indépendance de l'IMC et du sexe puis H1 : pas indépendance (= les variables sont liées). Les conditions de validités du test nous imposent que tous les effectifs attendus sous H0 (= e_ij) doivent dépasser 5. Ce n'est pas le cas ici, en effet on remarque un effectif = 3 mais comme nous n'avons pas d'alternative, nous faisons quand même le test du khi-deux d'indépendance.

```
khi2IS=chisq.test(tableauIS)
khi2IS
```

```
> khi2IS

        Pearson's Chi-squared test

data:  tableauIS
X-squared = 2.6286, df = 2, p-value = 0.2687
```

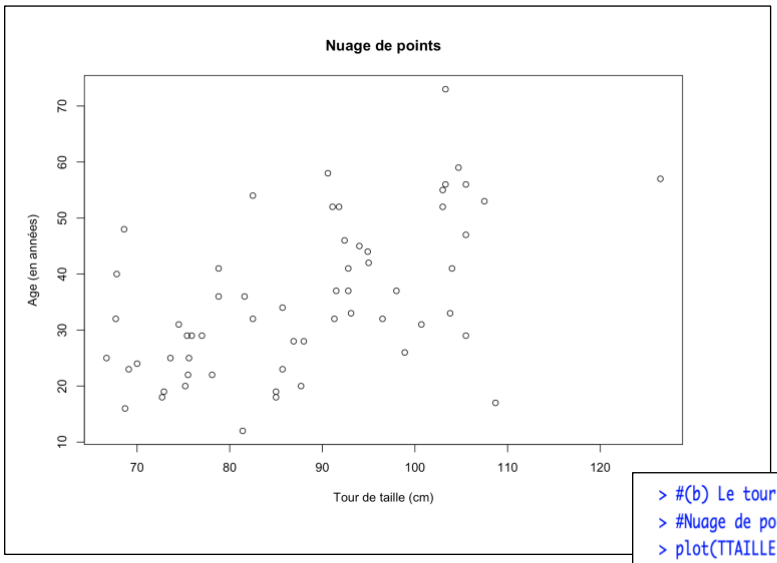
On peut maintenant conclure. La p-valeur est supérieure à 5 % donc on ne rejette pas H0. Au risque de 5 %, on peut conclure qu'il n'y a pas de liaison entre l'IMC et le sexe. Le tableau des profils-lignes indique la distribution de l'IMC parmi les hommes et les femmes, montrant que, par exemple, un pourcentage plus élevé d'hommes est en surpoids par rapport aux femmes. Cela suggère une association apparente entre l'IMC et le sexe. Cependant, le test du chi-deux examine si cette association est statistiquement significative ou si elle pourrait simplement être due au hasard. En d'autres termes, bien que le tableau des profils-lignes montre des différences apparentes, ces différences ne sont peut-être pas suffisamment importantes pour être statistiquement significatives dans le contexte de votre échantillon. Cela peut être dû à la variabilité naturelle des données ou à la taille de l'échantillon.

Le tour de taille augmente-t-il avec l'âge ?

Cette fois nous allons utiliser la fonction `cor.test`, car il s'agit de deux variables quantitatives. La première étape est de générer un nuage de points. On observe ensuite le coefficient de corrélation linéaire empirique.

```
#Nuage de points
plot(TTAILLE,AGE,main="Nuage de points", xlab="Tour de taille (cm)",
     ylab="Age (en années)")

#Coefficient de corrélation linéaire empirique (arrondi à deux décimales)
round(cor(TTAILLE,AGE),digits=2)
```




```
> #(b) Le tour de taille augmente-t-il avec l'age?
> #Nuage de points
> plot(TTAILLE,AGE,main="Nuage de points", xlab="Tour de taille (cm)",
+      ylab="Age (en années)")
> #Coefficient de corrélation linéaire empirique (arrondi à deux décimales)
> round(cor(TTAILLE,AGE),digits=2)
[1] 0.55
```

Le nuage semble à peu près réparti autour d'une droite croissante. Il semble que lorsque l'âge augmente, le tour de taille augmente. Un coefficient de corrélation de 0,55 suggère qu'il y a une relation positive, mais modérée, entre les deux variables. La corrélation indique que cette relation n'est pas due au hasard.

De manière analogue au test du χ^2, nous effectuons le test du coefficient de corrélation linéaire avec les hypothèses suivantes :

- H0 : absence de liaison linéaire entre le tour de taille et l'âge.
- H1 : liaison linéaire qu'on peut aussi écrire : rho différent de 0.

Ce test nécessite les variables aléatoires X et Y gaussiennes, ce qui n'est pas trop vrai ici mais le test est valide quand même, car n = 60 grand (condition : n supérieur à 30).

```
#nom : test du coefficient de corrélation linéaire
cor.test(TTAILLE,AGE)
```

```
        Pearson's product-moment correlation

data:  TTAILLE and AGE
t = 4.9767, df = 58, p-value = 6.112e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3403893 0.7032693
sample estimates:
      cor
0.5470323
```

On obtient une p-value de 6.112e-06, on rejette H0. Ce qui indique une relation statistiquement significative entre le tour de taille et l'âge dans notre échantillon. On conclut donc, au risque de 5 % de se tromper que ces deux variables sont corrélées. Ainsi, sans surprise, avec l'âge il y a une tendance à l'accumulation de graisse autour

de la taille. En moyenne, les personnes plus âgées ont tendance à avoir un tour de taille plus grand que les personnes plus jeunes.

Le taux de cholestérol est-il le même en moyenne pour les deux sexes ?

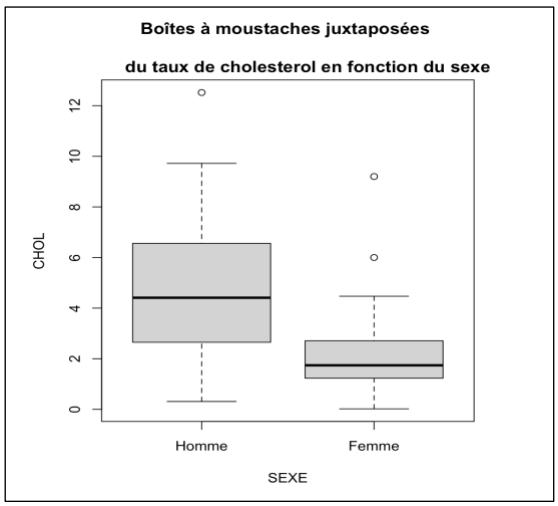
Nous avons affaire à une variable quantitative, le taux de cholestérol et une variable qualitative, le sexe. Pour ce faire, dans un premier temps nous allons nous faciliter la comparaison visuelle grâce aux boîtes à moustaches juxtaposées, en effet les distributions du taux de cholestérol et des sexes apparaissent côte à côte sur le même graphique. Dans un second temps, nous établissons les moyennes et variances par groupe.

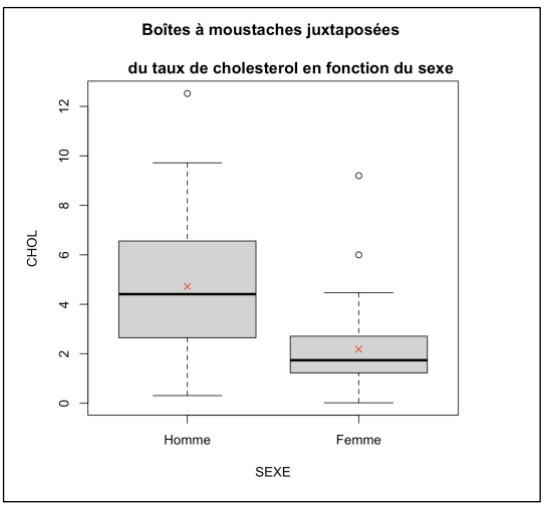
```
#Boîtes à moustaches juxtaposées
#attach(data_imc_TP2_groupe1)
boxplot(CHOL~SEXE,main="Boîtes à moustaches juxtaposées \n
        du taux de cholesterol en fonction du sexe")

#Moyennes par groupe
moycond=tapply(CHOL,SEXE,mean)
moycond

#Variances par groupe
tapply(CHOL,SEXE,var)
points(moycond,col="red",pch=4,cex=1) #on rajoute les moyennes conditionnelles aux bàm juxtaposées
```

```
> #(c) Le taux de cholestérol est-il le meme en moyenne pour les deux sexes?
> #Boîtes à moustaches juxtaposées
> #attach(data_imc_TP2_groupe1)
> boxplot(CHOL~SEXE,main="Boîtes à moustaches juxtaposées \n
+         du taux de cholesterol en fonction du sexe")
> #Moyennes par groupe
> moycond=tapply(CHOL,SEXE,mean)
> moycond
   Homme     Femme
4.719000 2.186333
> #Variances par groupe
> tapply(CHOL,SEXE,var)
   Homme     Femme
8.845058 3.365438
> points(moycond,col="red",pch=4,cex=1) #on rajoute les moyennes conditionnelles aux bàm juxtaposées
```





On observe que la moyenne et la variance du taux de cholestérol chez les hommes sont bien plus élevés que chez les femmes. Cela suggère une différence significative entre les deux groupes. On voit que le troisième quartile du taux de cholestérol des femmes est égal au premier quartile du taux de cholestérol des hommes, ce qui signifie que 75% des femmes ont des taux de cholestérol qui s'étendent jusqu'au niveau où se situent les taux de cholestérol minimums que touchent 75% des hommes. Les deux distributions présentent des valeurs extrêmes (dans les hauts taux de cholestérol).

Enfin, nous pouvons nous attaquer au test. Comme précédemment, il y a des hypothèses à poser et des conditions de validité du test. On veut tester H_0 : mu_F=mu_H contre H_1 : mu_F différent de mu_H. Dans ce cas on utilise le test de Student ou le test de Welch. Le test de comparaison des espérances de Student nécessite l'égalité des variances donc on commence par tester cette égalité. Les conditions de validité sont la loi normale (=gaussienne) dans chaque groupe ou n_F et n_H grand (supérieurs à 30 pour fixer les idées). On travaille sur des taux de cholesterol donc pas gaussien, ce que l'on peut vérifier par un graphique quantile-quantile normal.

```
par(mfrow=c(1,2)) # pour avoir les deux histogrammes côte à côte
tapply(CHOL,SEXE,qqnorm)
table(SEXE)
```

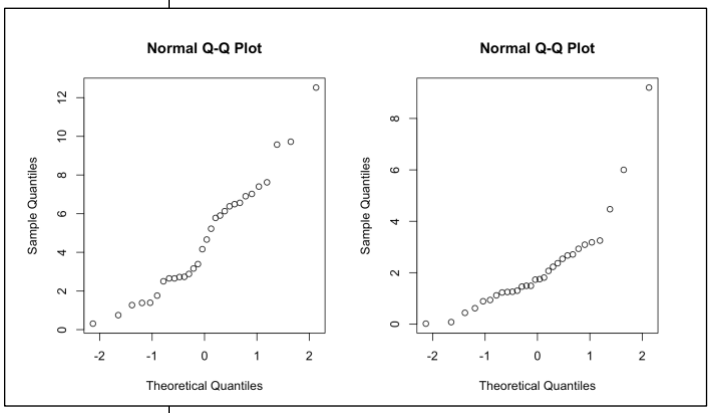
```
> #Conditions de validité : loi normale (= gaussienne) dans chaque groupe
> # ou n_F et n_H grand (supérieurs à 30 pour fixer les idées)
> #Ce sont des taux de cholesterol donc pas gaussien, ce que l'on peut vérifier par un graphique quantile-quantile normal
> par(mfrow=c(1,2)) # pour avoir les deux histogrammes côte à côte
> tapply(CHOL,SEXE,qqnorm)
$Homme
$Homme$x
 [1] -0.9027348 -2.1280452 -1.3829941  0.7835004 -1.1918162 -0.7835004
 [7] -0.0417893 -0.1256613 -0.3853205 -0.6744898 -0.4770404  0.5729675
[13] -1.0364334  0.3853205  0.0417893  1.0364334  0.9027348  1.6448536
[19]  0.4770404 -0.2104284 -0.2967378  0.6744898 -0.5729675 -1.6448536
[25]  0.1256613  1.1918162  1.3829941  0.2104284  0.2967378  2.1280452

$Homme$y
 [1]  1.76  0.31  1.27  6.90  1.38  2.50  4.16  3.39  2.73  2.65  2.72  6.49
[13]  1.39  6.13  4.66  7.40  7.02  9.72  6.38  3.16  2.88  6.56  2.65  0.75
[25]  5.22  7.62  9.57  5.78  5.90 12.52


$Femme
$Femme$x
 [1]  0.1256613  2.1280452 -0.7835004 -1.0364334  0.3853205  0.2104284
 [7] -0.0417893 -0.3853205 -0.4770404 -0.5729675 -0.2104284  0.7835004
[13]  0.5729675  0.9027348  0.4770404  0.0417893 -1.1918162  1.3829941
[19] -1.6448536 -0.1256613 -0.9027348 -2.1280452  0.6744898  0.2967378
[25] -1.3829941  1.6448536 -0.6744898  1.1918162  1.0364334 -0.2967378

$Femme$y
 [1] 1.81 9.20 1.12 0.89 2.37 2.07 1.73 1.30 1.26 1.25 1.49 2.93 2.67 3.09 2.54
[16] 1.75 0.62 4.47 0.08 1.49 0.94 0.02 2.71 2.23 0.44 6.00 1.23 3.25 3.18 1.46


> #On voit que les distributions ne sont pas gaussiennes donc il faut avoir des effectifs suffisants pour que le test soit
> #valide. C'est le cas ici, ce que l'on vérifie avec le tableau de distribution en effectifs
> table(SEXE)
SEXE
Homme Femme
   30    30
```



On voit que les distributions ne sont pas gaussiennes donc il faut avoir des effectifs suffisants pour que le test soit valide. C'est le cas ici, ce que l'on vérifie avec le tableau de distribution en effectifs. Comme les distributions ne sont pas normales et qu'on utilise le fait que les deux effectifs sont grands, on ne peut pas faire de test pour comparer les variances (le test de Fisher n'est pas valide). On fait donc directement le test de Welch avec variances inégales.

```
t.test(CHOL~SEXE,var.equal=FALSE)
```

```
> t.test(CHOL~SEXE,var.equal=FALSE)

        Welch Two Sample t-test

data:  CHOL by SEXE
t = 3.9698, df = 48.277, p-value = 0.0002384
alternative hypothesis: true difference in means between group Homme and
group Femme is not equal to 0
95 percent confidence interval:
 1.250115 3.815218
sample estimates:
mean in group Homme mean in group Femme
           4.719000            2.186333
```

On rejette l'égalité des moyennes, et à l'aide des "moycond" calculées plus haut, on conclut que les taux de cholestérol des hommes sont en moyenne plus élevés que ceux des femmes (4.719 versus 2.186333 g/L).

Vérifier que l'IMC (en classes) et le tour de taille sont liés. Si oui, préciser quelles classes diffèrent. On supposera pour cette question que la variable TTAILLE suit une loi normale dans chacune des classes d'IMC ?

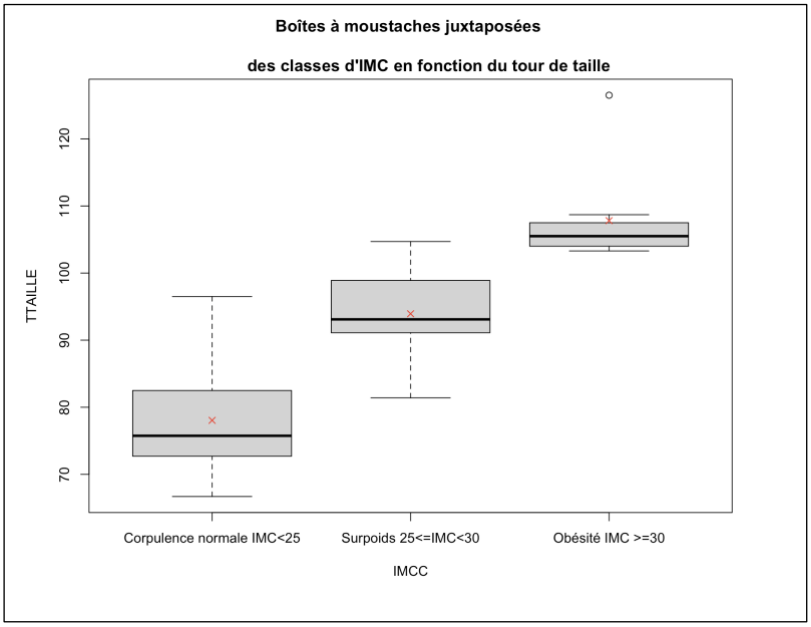
Dans ce cas nous avons une variable qualitative à 3 modalités qui est l'IMC (corpulence normale, surpoids, obésité) et une variable quantitative : le tour de taille. On procède dans le même ordre que la question précédente. En commençant par les boîtes à moustaches juxtaposées, les moyennes et variances par groupe puis le test statistique. Pour ce dernier on pose nos hypothèses, on vérifie les conditions de validité qui sont : loi gaussienne pour le tour de taille dans chacune des classes d'IMC (vérifié ici), on test l'homoscédasticité des variances avec le test de Brown-Forsythe `leveneTest`, on test ensuite si les moyennes diffèrent `oneway.test` et si c'est le cas on continue pour savoir quelles moyennes diffèrent deux à deux avec les tests de comparaisons multiples `pairwise.t.test`.

```
#Boites à moustaches juxtaposéeef
par(mfrow=c(1,1))
boxplot(TTAILLE~IMCC,main="Boites à moustaches juxtaposées \n
        des classes d'IMC en fonction du tour de taille")

#Moyennes par groupe
moycond=tapply(TTAILLE,IMCC,mean)
moycond

#Variances par groupe
tapply(TTAILLE,IMCC,var)
#La variance empirique est beaucoup plus importante dans le groupe des corpulences normales donc plus grande dispersion dans ce groupe.
points(moycond,col="red",pch=4,cex=1) #on rajoute les moyennes conditionnelles aux bàm juxtaposées
```

```
> #(d) Vérifier que l'IMC (en classes) et le tour de taille sont liés. Si oui, préciser quelles classes
différent. On supposera pour cette
> #question que la variable TTAILLE suit une loi normale dans chacune des classes d'IMC.
> #Boites à moustaches juxtaposéeef
> par(mfrow=c(1,1))
> boxplot(TTAILLE~IMCC,main="Boites à moustaches juxtaposées \n
+         des classes d'IMC en fonction du tour de taille")
> #Moyennes par groupe
> moycond=tapply(TTAILLE,IMCC,mean)
> moycond
Corpulence normale IMC<25       Surpoids 25<=IMC<30          Obésité IMC >=30
                78.04667                  93.93333                 107.81111
> #Variances par groupe
> tapply(TTAILLE,IMCC,var)
Corpulence normale IMC<25       Surpoids 25<=IMC<30          Obésité IMC >=30
                68.41085                  44.79333                  52.14361
> #La variance empirique est beaucoup plus importante dans le groupe des corpulences normales donc plus
grande dispersion dans ce groupe.
> points(moycond,col="red",pch=4,cex=1) #on rajoute les moyennes conditionnelles aux bàm juxtaposées
```



Les résultats montrent clairement que les moyennes du tour de taille varient significativement entre les différentes classes d'IMC, ce qui suggère que le tour de taille est lié à l'IMC. Les variances du tour de taille varient également entre les classes d'IMC, ce qui indique que la dispersion des données de tour de taille peut être différente dans chaque classe. Les obèses ont des tours de taille bien plus élevés que les personnes en surpoids qui à leur tour ont des tours de taille plus élevés que les personnes en corpulence normale. On peut voir sur les boites à moustaches juxtaposées que 75% des personnes ayant une corpulence normale ont des tours de taille qui n'atteignent pas le tour de taille minimum qui touchent 75% des personnes en surpoids. Ces mêmes personnes ont des tours de taille qui n'atteignent pas le tour de taille minimum qui touchent 75% des obèses. En effet, le 3e quartile des corpulences normales est inférieur au 1er quartile des surpoids. Même chose pour le 3e quartile des surpoids qui est inférieur au 1er quartile des obèses. On observe une valeur extrême chez les obeses. La dispersion des salaires des obèses est très faible (mais il faut dire qu'ils ne sont pas très nombreux).

Réalisons maintenant le test statistique avec toutes ses étapes. Après avoir posé nos hypothèse (H0 : mu_1=mu_2=mu_3 contre H1 : au moins deux moyennes diffèrent), nous vérifions l'homogénéité des variances par groupe pour choisir quel test faire. Il existe 2 tests possibles pour comparer les moyennes : le test classique de l'analyse de variance ou sa version modifiée en cas de variances inégales.

```
library(car)    # nécessite le package "car"
leveneTest(TTAILLE,IMCC)
```

```
> leveneTest(TTAILLE,IMCC)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.1367  0.328
      57
```

On ne rejette pas l'égalité des variances (p-valeur 0.328 > 0.05) il faut donc utiliser la variante de l'analyse de la variance pour variances égales avec `var.equal = TRUE`.

```
oneway.test(salaire~stat_pro,var.equal=TRUE)
```

```
> oneway.test(salaire~stat_pro,var.equal=TRUE)

        One-way analysis of means

data:  salaire and stat_pro
F = 434.26, num df = 2, denom df = 470, p-value < 2.2e-16
```

Comme p-value < 0.05 on rejette H0 et on conclut que les moyennes diffèrent globalement. On continue donc pour savoir quelles moyennes diffèrent deux à deux avec les tests de comparaisons multiples : test LSD (pas d'ajustement du risque alpha) `pairwise.t.test`.

```
pairwise.t.test(TTAILLE,IMCC,p.adjust.method= "none",pool.sd=TRUE)
```

```
> pairwise.t.test(TTAILLE,IMCC,p.adjust.method= "none",pool.sd=TRUE)

        Pairwise comparisons using t tests with pooled SD

data:  TTAILLE and IMCC

                    Corpulence normale IMC<25 Surpoids 25<=IMC<30
Surpoids 25<=IMC<30 8.6e-10                   -
Obésité IMC >=30    1.3e-14                   2.6e-05

P value adjustment method: none
```

★ Conclusion

En résumé, les analyses statistiques que nous avons réalisées sur l'échantillon de population que nous avons étudié ont permis de mettre en évidence plusieurs observations et tendances significatives. L'échantillon est caractérisé par une répartition équilibrée des sexes, avec une représentation égale d'hommes et de femmes. La majorité des individus de l'échantillon ont une corpulence normale, tandis qu'une proportion importante est en surpoids, et une proportion plus faible est obèse. L'échantillon présente une diversité d'âges allant de 12 à 75 ans, avec une distribution des tailles étendue et une variabilité importante. La distribution du tour de taille montre une grande variabilité dans les données, et une valeur extrême influente affecte la moyenne du taux de cholestérol.

De plus, des tests statistiques ont été effectués pour évaluer les relations entre certaines variables, notamment le tour de taille et l'âge, ainsi que le taux de cholestérol et les sexes. Les résultats de ces tests ont permis de conclure que le tour de taille est corrélé à l'âge, que les hommes ont en moyenne des taux de cholestérol plus élevés que les femmes, et que les tours de taille des personnes obèses sont significativement plus élevés que ceux des personnes en surpoids, qui sont à leur tour plus élevés que ceux des personnes ayant une corpulence normale.

Ces conclusions fournissent des informations essentielles sur les caractéristiques et les relations au sein de l'échantillon étudié. Cependant, il est important de noter que ces résultats sont spécifiques à cet échantillon et ne peuvent pas nécessairement être généralisés à l'ensemble de la population américaine. Pour parvenir à des conclusions plus générales, des études plus vastes et plus représentatives seraient nécessaires.