Statistique Inférentielle, M1 Info Lyon 2, 2021, TD2

Guillaume Metzler

Exercice 1

Commençons par charger les données

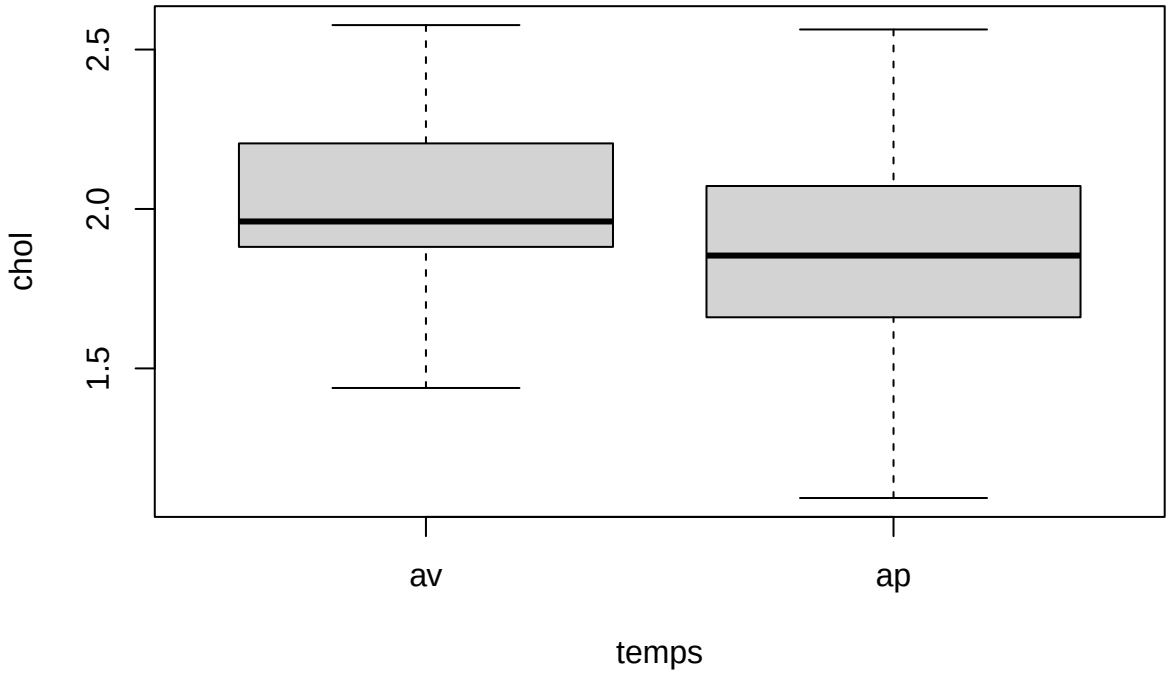
```
cholesterol <- read.csv2("~/Desktop/A exporter/Cours/Lyon2/Statistiques_Master/cholesterol.csv")
```

On va mettre en forme le data.frame de sorte à avoir une colonne par variable, qui sont: le taux de cholesterol, le temps auquel il est mesuré, le sexe de la personne.

```
data=cbind(rep(cholesterol$id,2),stack(cholesterol[,2:3]),rep(cholesterol$sexe,2))
colnames(data)=c('id','chol','temps','genre')
```

On peut commencer par représenter les données:

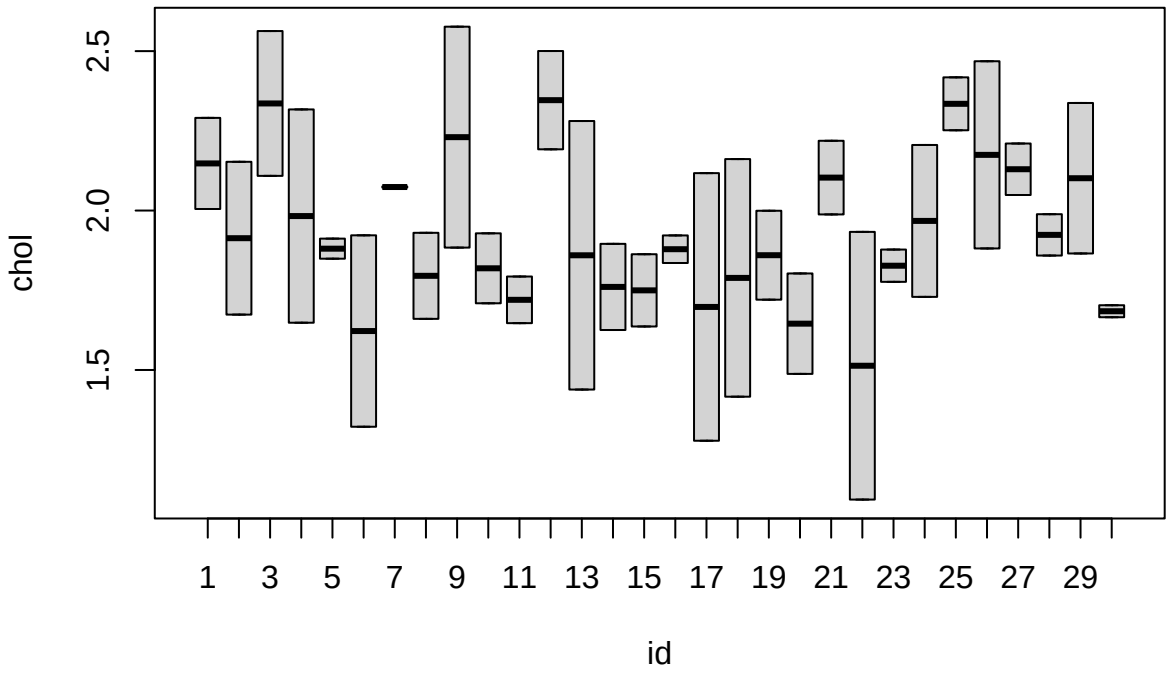
```
boxplot(chol~temps,data=data)
```



Il semble que le taux de cholesterol soit plus faible après le traitement. La question est: est-ce significatif ou dû au hasard ? Pour répondre à cette question, nous allons réaliser un test statistique de comparaison de deux grands échantillons ($n \geq 20$) : un test de Student.

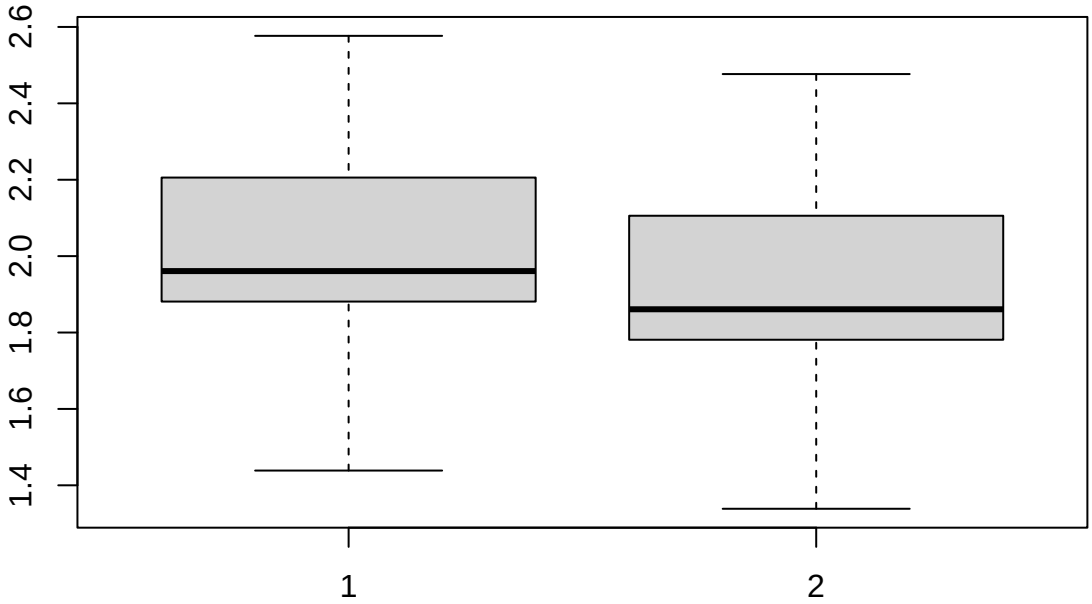
Mais attention, les données sont **appariées** : ce sont les mêmes individus qui sont mesurés avant et après le traitement. On a vu qu'il y a de grandes différences entre les individus :

```
boxplot(chol~id,data=data)
```



Ces grandes différences peuvent éventuellement masquer l'effet du médicament. Prenons un exemple pour illustrer ceci : créons artificiellement des taux de cholesterol après prise du médicament, de sorte que chaque individus ait un taux qui ait diminué de 0.1 :

```
avant=cholesterol$av
après=avant-0.1
boxplot(avant,après)
```




```
t.test(avant,après)
```

```
##
##  Welch Two Sample t-test
##
## data:  avant and après
```

```
## t = 1.4282, df = 58, p-value = 0.1586
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04015648  0.24015648
## sample estimates:
## mean of x mean of y
##  1.995673  1.895673
```

Le test de Student n'est pas significatif, alors que tous les patients ont un taux qui a diminué de 0.1 : c'est parceque les variabilités inter-individus masquent les variabilité intra-individu.

Il est donc primordial de prendre en compte le fait que ce soit des mesures répétées, et techniquement ont fait cela en travaillant sur la différence avant-après.

Revenons à nos vraies données, avec une alternative greater pour montrer que le médicament diminue le taux de cholesterol :

```
t.test(chol~temps,data=data,paired=TRUE,alternative="greater")
```

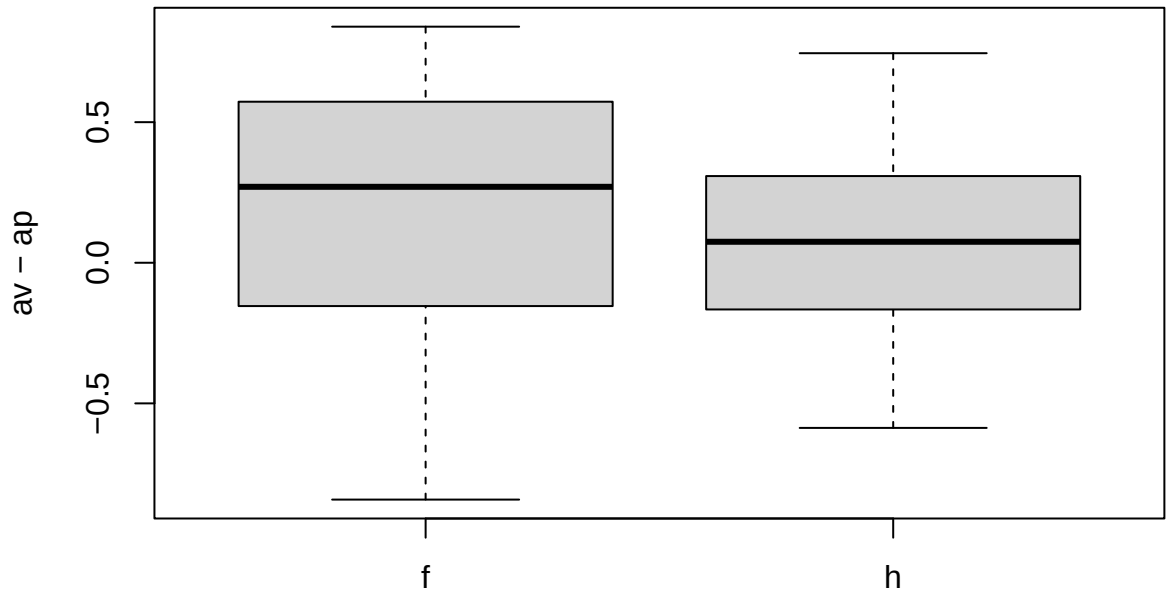
```
##
##  Paired t-test
##
## data:  chol by temps
## t = 1.6917, df = 29, p-value = 0.05071
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0005856456           Inf
## sample estimates:
## mean of the differences
##               0.1335859
```

On obtient une pvalue de 0.0507, on est pile à la limite du risque qui nous permet généralement de rejetter H_0 (on rejette si la p-value est <0.05), et donc de conclure à l'effet du médicament: on ne peut donc pas conclure ici... Il faudrait plus d'individus dans l'étude pour faire infléchir dans un sens ou dans l'autre la p-value.

L'efficacité diffère-t-elle selon le genre ?

Cherchons à représenter graphiquement ceci:

```
boxplot(av-ap~sexe,data=cholesterol)
```



Visuellement, on *dirait* que l'efficacité est un peu plus grande chez les femmes. Mais est-ce significatif ?

Ici, on ne peut pas utiliser le test de Student car on n'a moins de 30 hommes et 30 femmes : quand on veut comparer 2 échantillons, ici les hommes et les femmes, il faut pour utiliser le test de Student que chacun des échantillons soit gaussien, ou qu'ils soient grands (plus de 30 ind). Quand ce n'est pas le cas, on utilise une alternative non paramétrique, à savoir le **test de wilcoxon**

```
wilcox.test(av-ap~sexe,data=cholesterol,alternative="greater")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  av - ap by sexe
## W = 130, p-value = 0.1842
## alternative hypothesis: true location shift is greater than 0
```

La p-value de 0.1842 n'est pas significative... On ne peut pas conclure à une différence homme femme (soit il n'y en a pas, soit on n'a pas assez de données pour le montrer).

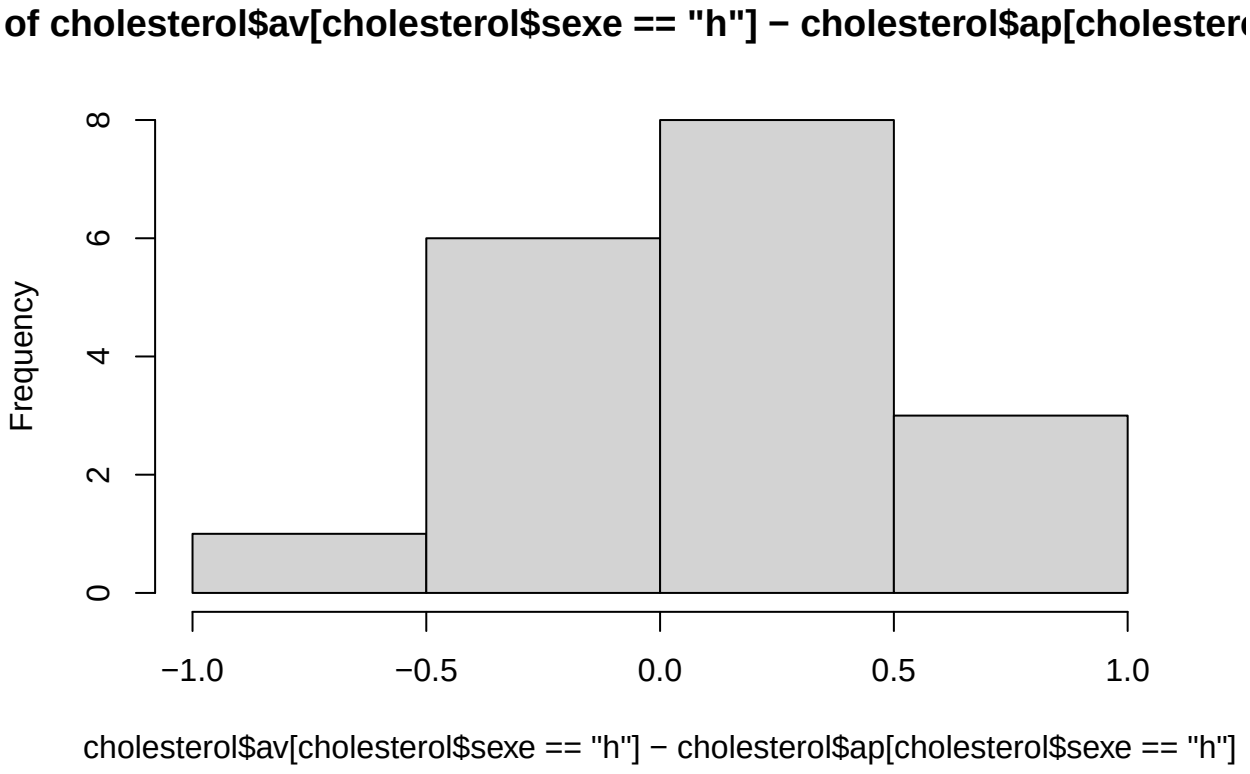
Le test de Student étant plus puissant (risque de seconde espèce plus faible pour un α fixé), on peut chercher à l'utiliser néanmoins en vérifiant si les distributions des données peuvent être considérées comme gaussiennes. On peut faire pour cela un test de Shapiro, qui lui même demandera d'avoir au moins une dizaine d'observations (sinon sa puissance sera trop faible).

```
summary(cholesterol$sexe)
```

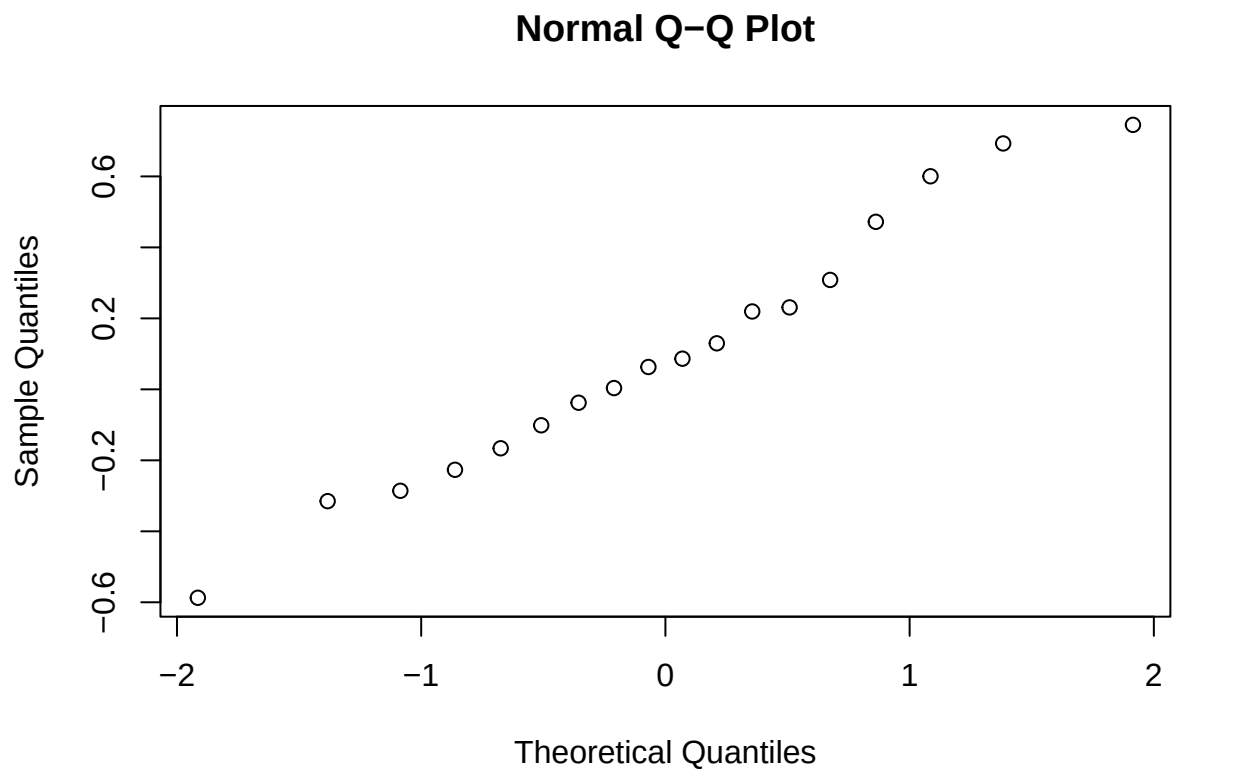
```
##    Length     Class      Mode
##        30 character character
```

On a 12 femmes et 18 hommes, on peut faire un test de Shapiro par échantillons. Attention, on veut tester la normalité de la différence avant - arpès :

```
hist(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"], breaks=4)
```




```
qqnorm(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"])
```




```
shapiro.test(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"])
```

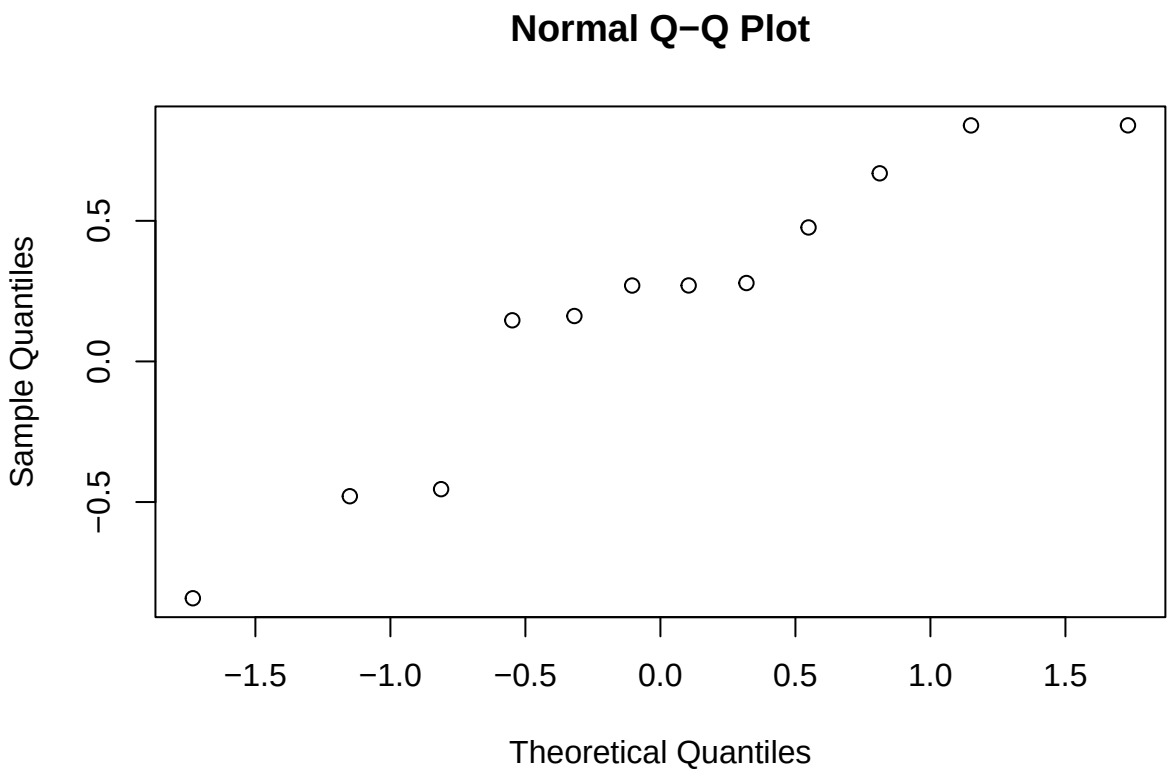
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  cholesterol$av[cholesterol$sexe == "h"] - cholesterol$ap[cholesterol$sexe == "h"]
## W = 0.97629, p-value = 0.9039
```

La p-value est grande, le qqnorm proche d'une droite, l'échantillon peut être considéré comme gaussien.

```
hist(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
```


```
qqnorm(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
```




```
shapiro.test(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cholesterol$av[cholesterol$sexe == "f"] - cholesterol$ap[cholesterol$sexe == "f"]
## W = 0.91513, p-value = 0.2481
```

Idem pour les femmes.

On peut donc réaliser un test de Student, mais dont la conclusion est la même que pour le test de Wilcoxon: pas de différence significative entre hommes et femmes

```
t.test(av-ap~sexe,data=cholesterol,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  av - ap by sexe
## t = 0.45173, df = 17.889, p-value = 0.3284
## alternative hypothesis: true difference in means between group f and group h is greater than 0
## 95 percent confidence interval:
##  -0.2257321        Inf
## sample estimates:
## mean in group f mean in group h
##       0.1812758       0.1017926
```

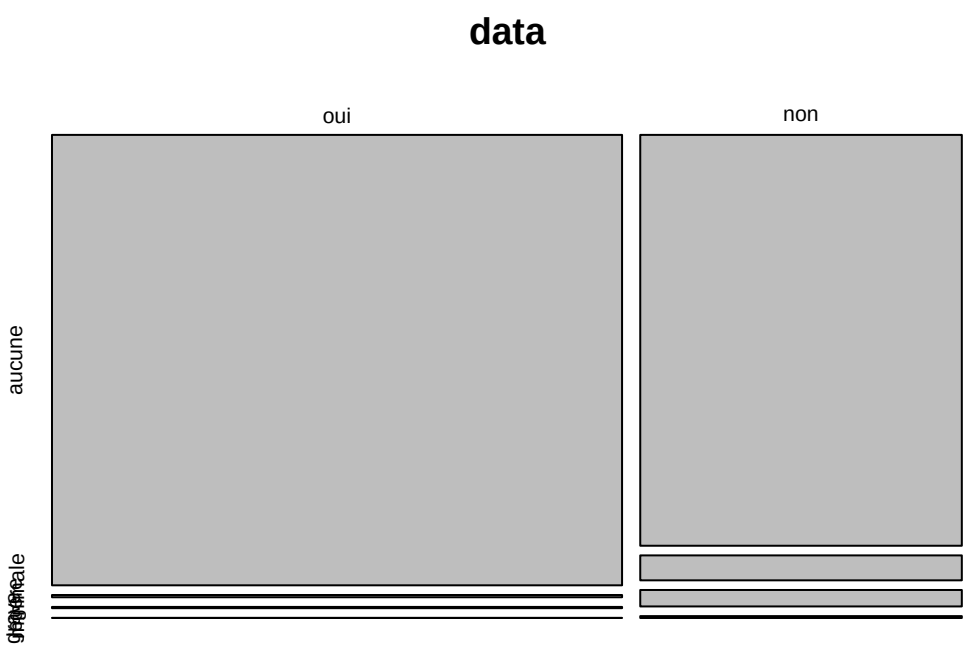
Exercice 2

Nous sommes en présence de deux variables catégorielles. Commençons par créer le tableau de contigence :

```
tab=rbind(c(128213,647,359,42),c(65963,4000,2642,303))
colnames(tab)=c('aucune','minimale','legere','grave')
rownames(tab)=c('oui','non')
```

On peut le représenter en utilisant la fonction plot spécifique pour les tableaux de contingences. Il faut alors définir que la matrice que l'on vient de créer est un tableau de contigence :

```
data=as.table(tab)
plot(data)
```



Afin de tester un lien entre ces deux variables catégorielles, on va utiliser un test du chi2 :

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 9326.7, df = 3, p-value < 2.2e-16
```

La p-value est très significative, le port de la ceinture a clairement une influence sur la gravité des blessures...

Exercice 3

L'étude porte sur l'utilisation du fichier GermanCredit.data qui est disponible sur la page de votre enseignant. Commençons donc par charger les données.

```
library(readr)
GermanCredit <- read_table2("http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/GermanCredit.data",
    col_names = FALSE)
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
## ! curl package not installed, falling back to using `url()`
```

```
##
```

```
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   X2 = col_double(),
##   X5 = col_double(),
##   X8 = col_double(),
##   X11 = col_double(),
##   X13 = col_double(),
##   X16 = col_double(),
##   X18 = col_double(),
##   X21 = col_double()
## )
## i Use `spec()` for the full column specifications.
```

Parmi ces 21 variables, la 21ème indique la qualité du client. Quelles sont, parmis les 20 autres variables, celles sont qui sont liées à la qualité du client ?

On va donc créer une fonction qui va parcourir l'ensemble des variables et qui va, selon la nature de la variable testée, effectuer le test statistiques correspondant.

Nous sommes dans un cas où nous disposons de grands échantillons, donc on, selon la nature de la variable, effectuera une analyse de corrélations de type

- quali - quanti :Student car la variable cible n'a que deux modalités
- quali - quali : Khi-deux

Remarque : la variable 'X21' ne prenant que deux valeurs, nous pourrions effectuer, au choix, un test de Student ou une ANOVA, les résultats seront identiques.

```
# Conversion des données pour nous permettre de récupérer le type des variables.
data <- data.frame(GermanCredit)

# Quelques tests

test_quali_quali <- function(i,data){

  # On commence par créer notre table de contingence à l'aide de la fonction test
  # On effectue le test
  test_khi <- chisq.test(table(as.factor(data[,21]),data[,i]))
  p_value <- test_khi$p.value
  return(p_value)

}

test_quali_quanti <- function(i,data){

  # On a simplement à effectuer le test
  test_stud <- t.test(data[,i]~X21, data=data)
  # On extrait la p-value que l'on pourra comparer au risque d'erreur alpha
  p_value <- test_stud$p.value
  return(p_value)

}
```

On va ensuite boucler sur l'ensemble des variables afin de regarder, quelles sont les variables qui sont donc corrélées à notre variable 21 à l'aide des fonctions précédentes

```r
list_corr = NULL
list_uncorr = NULL

for (i in 1:(dim(data)[2]-1)){

  print(i)

  # On commence par extraire la nature de variable et on regarde si elle est de type "numérique"
  # ou "charactère"
  nature_variable <- typeof(data[,i])

  if (nature_variable == "character"){

    nature_test <- "Khi_deux"
    p_value <- test_quali_quali(i,data)

  }

  if (nature_variable == "double"){

    nature_test <- "Student"
    p_value <- test_quali_quanti(i,data)

  }

  # On compare la p-value extraite au risque d'erreur de alpha = 5% pour tirer des conclusions quant
  # quant à la corrélation entre X21 et la variable testée

  if (p_value < 0.05) {

    print(paste(sprintf("Nous avons effectué un test du %s. Ce test nous a conduit à une p_value égale a

    list_corr <- c(list_corr,i)

  } else {

        print(paste(sprintf("Nous avons effectué un test du %s. Ce test nous a conduit à une p_value éga

     list_uncorr <- c(list_uncorr,i)

  }
}
```

```
## [1] 1
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.218902e-
## [1] 2
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 2.404081e-1
## [1] 3
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.279187e-
## [1] 4
## Warning in chisq.test(table(as.factor(data[, 21]), data[, i])): Chi-squared
## approximation may be incorrect
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.157491e-
```

```
## [1] 5
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 2.477713e-
## [1] 6
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 2.761214e
## [1] 7
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.045452e
## [1] 8
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 2.033538e-
## [1] 9
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 2.223801e
## [1] 10
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 3.605595e
## [1] 11
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 9.249780e-
## [1] 12
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 2.858442e
## [1] 13
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 3.788491e-
## [1] 14
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.629318e
## [1] 15
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.116747e
## [1] 16
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 1.416126e-
## [1] 17
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 5.965816e
## [1] 18
## [1] "Nous avons effectué un test du Student. Ce test nous a conduit à une p_value égale à 9.239934e-
## [1] 19
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 2.788762e
## [1] 20
## [1] "Nous avons effectué un test du Khi_deux. Ce test nous a conduit à une p_value égale à 1.583075e
```

```
# La liste des variables corrélées à la variable est donnée par
print(paste("X",list_corr,sep=""))
```

```
##  [1] "X1"  "X2"  "X3"  "X4"  "X5"  "X6"  "X7"  "X8"  "X9"  "X10" "X12" "X13"
## [13] "X14" "X15" "X20"
```

```
# La liste des variables non corrélées à la variable est donnée par
print(paste("X",list_uncorr,sep=""))
```

```
## [1] "X11" "X16" "X17" "X18" "X19"
```

La correction est incomplète ici. En effet, pour effectuer le test de Student, il faudrait vérifier que les conditions soient réunies :

- taille des différents groupes ≥ 30 ou gaussiens (test de shapiro)
- vérfier, à l'aide d'un test de Fisher, si les variances sont égales ou non

De la même façon, pour le test du Chi-deux, il faudrait vérifier que les effectifs de nos différentes tables de contigence soient au minimum égaux à 5. C'est le cas pour la variable référencée dans la quatrième colonne de notre jeu de données où il est nécessaire de regrouper trois colonnes ensembles afin de vérifier cette condition.