TP R 1: Tests statistiques

BINET Thomas

Statistiques biomédicales

Table des matières

1. Tests : quelques fonctions utiles

1.1 Tests de normalité

De nombreux tests ou IC sont valables sous l'hypothèse que l'échantillon est généré selon une loi normale. Il convient alors de tester la normalité de l'échantillon.

Il est possible d'installer la fonction `lillie.test` du package `nortest` à installer à l'aide de la commande `install.packages('nom_du_package')` ou manuellement dans l'onglet `Tools-> Install packages` de la barre de contrôle (ci-dessus). On peut aussi simplement utiliser la fonction `shapiro.test`.

1. Parmi les échantillons suivants, lesquels semblent de loi normale d'après les tests de Lilliefors et Shapiro-Wilks ?

```
###Exemple
library(nortest)
x1=c(12.672656, 13.762919,  7.386138,  9.936208, 11.758227, 13.718379, 12.543564, 14.635996, 15.682780,
x2=c(7, 8, 9, 10, 11, 12, 13, 14, 15, 16)
x3=c(2, 5, 4, 3, 4, 16, 15, 16, 15, 17)
x4=c(0,1,1,1,0,1,1,0,0,1)
x5=rnorm(50,0,1)
x6=round(rnorm(50,0,1))
x7=round(rnorm(50,0,1),1)
x8=c(rnorm(50,20,1.5),rnorm(20,42,4))

x9=1:100
```

```
lillie.test(x9)
shapiro.test(x9)
```

Après Test, on remarque que seulements les échantillions x1, x2, x5, x7. Pour x9 le doute persiste car pour le test de Lilliefors la pvalue est de 0.47 alors que pour le test de Shapiro-Wilk la pvalue est de 0.0017.

1.2 Tests pour échantillon(s) de loi normale(s)

Installer manuellement le package `OneTwoSamples` (le plus récent dans les archive du CRAN) ; puis utiliser la librairie associée `library(OneTwoSamples)`. Cela donne accès à `mean_test1` et `mean_test2` : des procédures de tests utilisant la loi normale avec variance connue ou inconnue. Voir aussi les fonctions `var_test1` et `var_test2`.

2. Commentez les résultats des trois lignes de codes suivantes en considérant un seuil de rejet à 5%. Qu'est ce qui était prévisible(s) ?

```
# Ligne 1:
mean_test1(rnorm(20,0,1),mu=1,side=-1,sigma=1)
# Ligne 2:
mean_test1(rnorm(20,0,1),mu=1,side=-1,sigma=4)
# Ligne 3:
mean_test1(rnorm(20,0,1),mu=1,side=1)
```

Le Test 1 rejette l'hypothèse H_0 :mu=1 et sigma>=1 contre l'hypothèse H_1 : mu<1 Le Test 2 n'est pas rejetté sous l'hypothèse H_0 : mu>=1 et sigma=4 contre H_1 : mu<1 Le Test 3 n'est pas rejetté sous l'hypothèse H_0 : mu<=1 contre H_1 : mu>1

3. L'échantillon suivant correspond aux âges de 20 étudiants. Peut-on rejeter l'hypothèse H_0 que l'âge moyen des étudiants est inférieur ou égal à 20 ans (on vérifiera d'abord l'hypothèse de normalité) ?

```
AGE<-c(18,25,22,21,23,19,20,17,19,20,22,23,20,28,21,23,19,20,22,21)
lillie.test(AGE)
shapiro.test(AGE)
```

Le Test de Normalité donne une pvalue >20% on peut alors accepter la normalité des données. ON peut alors appliquer le test de la moyenne.

```
mean_test1(AGE,mu=20,side=1)
```

On rejette l'hypothèse H_0 : mu<=20 contre H1 : mu>20. (pvalue=0.0276539).

4. Expliquer (ou prévoir) les résultats des lignes suivantes. On s'intéressera en particulier aux p-valeurs et aux dégrés de liberté. Que signifie `Z` et `T` ?

```
x=rnorm(10, mean = 10, sd = 1);
y=rnorm(20, mean = 10, sd = 1);
mean_test2(x, y, sigma = c(0.1, 0.1), side = 0)
mean_test2(x, y, var.equal = TRUE, side = 0)
```

Z et T sobt les statistique de test utilisé (T car on estime s'ils ont la meme variance et Z pour estimer la moyenne). On remarquera alors que le second test passera contrairement au premier.

5. Expliqer le résultat du code suivant et modifier les paramètres de la fonction pour qu'elle renvoie un résultat censé.

```
x1=rnorm(20, mean=10, sd = 4)
x2=rnorm(10, mean=20, sd = 4)
var_test2(x1,x2,mu=c(10,20))
```

Desc : Compute the two sided or one sided test of hypothesis of sigma1^2 and sigma2^2 of two normal samples when the population means are known or unknown COmme on a mis 10 et 10 en moyenne et que x2 est normalement à 20 de moyenne le seul moyen d'expliquer cette différence serait de jouer sur sigma avec un sigma bien plus grand (ce qui explique que le test rejette l'hypothèse de variance égale).

On doit donc changer le mu connu et mettre 20 pour x2 car en faussant l'information de la moyenne le test est biasé de base.

6. On dispose de deux échantillons de temps de révision pour le partiel de mathématiques, l'un en L1 biologie, l'autre en L1 psychologie. Les temps sont donnés en heures. On cherche à rejeter l'hypothèse H_0 que les étudiants de en psychologie révisent plus les maths que ceux de la filière biologie. Qu'en concluez-vous ? On prendra toutes les garanties pour savoir si il est possible d'appliquer la fonction `mean_test2`. Expliquez le degré de liberté renvoyé par la fonction (aide de `R`, wikipédia ...)

```
Temps_SD=c(2.1,0.9,2.6,2.9,3.9,1.3,0.2,3.6,1.0,1.4,2.3,4.1,1.2,4.1,3.6,2.6,2.4,1.4,3.1,2.8)
Temps_INFO=c(1.6, 0.8, 2.1, 1.4, 2.7, 1.1, 2.2, 1.1, 2.8, 0.9, 1.0, 1.3, 3.1, 1.8, 1.9, 1.8, 0.7, 2.5, 
shapiro.test(Temps_INFO)
shapiro.test(Temps_SD)
```

Les tests de normalités sont valides, on va donc pouvoir faire un test de moyenne.

```
mean_test2(Temps_INFO,Temps_SD,side = 1)
```

On accepte H0 : mu1<mu2 (les infos révisent moins que les SD en math). df=31.16698, le degrès de liberté

7. Une fonction alternative à `mean_test1` et `mean_test2` est `t.test`. Cependant cette fonction ne permet pas de traiter le cas de la variance connue. De même, `var.test` est une alternative à `var_test1` et `var_test2`, mais sans le cas de la moyenne connue. Peut-on rejeter au seuil de 5% que les sardines de Bretagne sont plus petites que celle de la mer méditérranée ? On utilisera les fonctions `t.test` et `var.test`.

```
#taille en cm des sardines de Bretagne
x1=c(10.8,9.6,11.9,13.2,17.0,15.9,12.1,9.6,10.7,15.4)
#taille en cm des sardine de mer méditérranée
x2=c(10.4,8.9,8.5,13.1,9.0,8.6,8.7,7.9,11.3,10.6,8.6,11.9,11.2,10.6,8.3)
```

```
shapiro.test(x1)
shapiro.test(x2)
```

Les tests acceptent l'hypothèse de normalité des deux jeux de données.

1.3 Tests asymptotiques basés sur le Théorème Central Limite

8. Bien que dédié aux échantillons de lois normales, il est possible d'utiliser les fonctions `mean_test1` et `mean_test2` pour effectuer des tests asymptotiques à l'aide du TCL. Expliquer comment faire. En particulier, on traitera le cas suivant. Peut-on rejeter l'hypothèse H_0 que le médicament 1 est plus efficace que le médicament 2 ?

```r
# On simule l'efficacité de deux médicaments
# 1, pour efficace, 0 sinon

# médicament 1
x1=rbinom(80,1, prob= 0.6)
#médicament 2
x2=rbinom(120,1, prob= 0.7)
```

1.4 Tests du χ^2

La fonction chisq.test permet de réaliser les tests du χ^2 de conformité, d'indépendance et d'homogénéité.

- Le test du χ^2 de conformité permet de savoir si il y a correspondance entre une répartition théorique et une répartition observée. L'hypothèse H_0 est "la répartition est celle donnée par la théorie".
- Le test du χ^2 d'indépendance permet d'étudier l'indépendance entre 2 critères susceptibles d'être associés à une différence de répartition. L'hypothèse H_0 est "la répartition ne dépend pas du critère".
- Le test du χ^2 d'homogénéité permet d'étudier la correspondance entre les répartitions de différents échantillons. L'hypothèse H_0 est "les répartitions sont identiques".

9. Ici, la fonction à utiliser est `chisq.test` . Cette année, en M2, il y a 8 filles et 15 garçons. Peut-on rejeter au seuil de 10% qu'il y autant de filles que de garçons dans cette filière ? On effectuera un test de conformité standard. Est-il possible d'utiliser le tests du χ^2 standard ou faut-il évaluer la p-valeur par simulation de Monte Carlo ?

11. Exemple du test d'indépendance : compléter le code ci-dessous (matrice) pour savoir si il est possible de rejeter l'hypothèse d'indépendance entre le genre et la rémunération.

```r
# Dans une entreprise, les salaires on été étudiés en fonction du genre. Il se répartissent en trois cl
# Créations des vecteurs correspondant aux 2 catégories :
hommes = c(110,80,50)
femmes = c(120,90,30)

# Création d'une matrice comparative. On veut que la ligne des valeurs pour les hommes correspondent à

# Réalisation du test khi-deux - les résultats sont sauvegardés dans "khi_test"
khi_test = chisq.test(M)

khi_test # affiche le résultat du test
```

12. Les élèves de 3 classes de CP apprennent à lire selon trois méthodes différentes. La première valeur correspond au nombre de réussite à un test de lecture et la seconde au nombre d'échec. Compléter le code suivant à l'aide de `matrix` ou `rbind`. Les résultats des trois méthodes doivent constituer les trois lignes.

```r
meth1 = c(13,17)

meth2 = c(18, 9)

meth3 = c(20, 6)

# Réalisation du test khi-deux - les résultats sont sauvegardés dans "khi_test"
khi_test = chisq.test(M)
khi_test # affiche le résultat du test
```

1.5 ANOVA (Analysis of Variance)

Un premier exemple

L'analyse de la variance (ANOVA stands for **An**alysis **o**f **Va**riance) a pour but de tester l'égalité des moyennes entre plusieurs échantillons. C'est un test qui se base sur la normalité des échantillons et l'homogénéité des variances (homoscedasticité). Il faut donc contrôler que les échantillons peuvent être considérés comme issus de lois normales ('shapiro.test') et que leur variance semblent égales (`var.test`).

13. Le fichier `Anova.Rdata` contient 4 échantillons `x1`, ..., `x4`. Identifier un groupe d'échantillons vérifiant les hypothèses d'application de l'anova au seuil de 5%. Dans la suite, on travaille sur ces échantillons.

```
load(file='Anova.Rdata')
```

14. Représenter les boxplots des échantillons sélectionnés (function `boxplot`).

Rappel : ANOVA à un facteur Soit $x_1, \dots, x_p$ des échantillons. L'échantillon x_i comprend n_i variables aléatoires de lois normales : $x_i = (x_{i,1}, \dots, x_{i,n_i})$.

On définit la somme des carrés entre groupe (SSB, between groups) et à l'intérieur des groupes (SSW, within groups) :

$$SSB = \sum_{i=1}^{p} n_i (\overline{x_i} - \overline{x})^2,$$

et

$$SSW = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x_i})^2,$$

avec $\overline{x}$ la moyenne de tous les échantillons réunis et $\overline{x_i}$ la moyenne de l'échantillon x_i. La statistique F suit une loi de Fisher :

$$F = \frac{\frac{SSB}{p-1}}{\frac{SSW}{n-p}} \sim \mathcal{F}(p-1, n-p)$$

15. Ecrire une fonction qui renvoit SSB, SSW , la statistique de tests et la p-valeur pour la loi de Fisher correspondant au test ANOVA à 1 facteur. Comparer les résultats obtenus avec ceux de la fonction 'oneway.test'.

```
n1 <- n2 <- n3 <- 100
x1 <- rnorm(n1,2.65)
x2 <- rnorm(n2, 2.55)
x3 <- rnorm(n3, 3.2)
x <- c(x1,x2,x3)
m <- mean(x)

# mise en forme (regarder attentivement la forme que prennent ces données)
tab <- data.frame(value = x, group = c(rep(1,n1),rep(2,n2),rep(3,n3)))
#
oneway.test(value~group, tab, var.equal = TRUE)
```

16. Appliquer l'ANOVA au jeu de données suivant : pour des décés de personnes retraités survenus entre 2010 et 2020, âge de décés et classe sociale. On réalisera les contrôles d'usages.

```
data=read.csv("Life_expectancy.csv")
```

2. Exercices

```
library(nortest)
library(OneTwoSamples)
```

Exercice 1 : Analyse de data frame

a. Naissance :

C'est une base de donnée contenant la taille (en cm) et le sexe (1=masculin, 2=féminin) des nouveau-nés français en 2019.

Naît-il plus de filles ou de garçons ? Y a-t-il une différence de taille à la naissance ?

```
data=read.csv("Naiss_fr_2019.csv")
```

```
# Compter le nombre de garçons et de filles
sex_counts <- table(data$SEXE)

# Afficher les résultats
print(sex_counts)
```

```
##
##      1      2
## 385038 368345
```

```
prop.test(sex_counts, sum(sex_counts), p =0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sex_counts, null probability 0.5
## X-squared = 369.83, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5099492 0.5122081
## sample estimates:
##         p
## 0.5110787
```

```
prop.test(sex_counts, sum(sex_counts), p =0.5,alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sex_counts, null probability 0.5
## X-squared = 369.83, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5101307 1.0000000
## sample estimates:
##         p
## 0.5110787
```

```
# Calculer la moyenne de la taille pour les garçons et les filles
boy_mean <- mean(data$TAILLE[data$SEXE == 1])
girl_mean <- mean(data$TAILLE[data$SEXE == 2])

# Afficher les résultats
print(paste("Moyenne de la taille des garçons :", boy_mean))
```

```
## [1] "Moyenne de la taille des garçons : 50.0054168679455"
```

```
print(paste("Moyenne de la taille des filles :", girl_mean))
```

```
## [1] "Moyenne de la taille des filles : 48.9982815023959"
```

```
# Effectuer un test t pour vérifier si la différence de taille est significative
t.test(data$TAILLE ~ data$SEXE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$TAILLE by data$SEXE
## t = 102.42, df = 734293, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.987862 1.026409
## sample estimates:
## mean in group 1 mean in group 2
##        50.00542        48.99828
```

On remarque alors que la pvalue du test de binomialité sur le sexe Homme/Femme est de 10^-16, donc on rejette l'hypothèse d'avoir la meme probabilité de naissance d'une fille ou d'un garcon contre celle d'avoir une plus grosse probabilité de naissance de garcon. De plus la différence de taille est significative selon le sexe du nouveau né

b. Espérance de vie :

Le jeu de données suivant rassemble des informations sur des décès. On notera que la variable `SEXE` a été équilibrée ("autant" de femmes, que d'hommes dans chaque catégorie social). Il n'est donc pas possible d'étudier la répartition des hommes et des femmes dans ces catégories.

- `AGE` : âge de la personne décédée ;
- `CLASSE` : la catégorie sociale du décédé ;
- `SEXE` : 1, pour les hommes, 2, pour les femmes ;
- `DATE` : date approximative du décès (3 classes).

```
data=read.csv("Life_expectancy_large.csv")
```

Calculer l'âge moyen des personnes décédées par catégorie sociale :

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data %>% group_by(CLASSE) %>%  summarize(mean_age = mean(AGE))
```

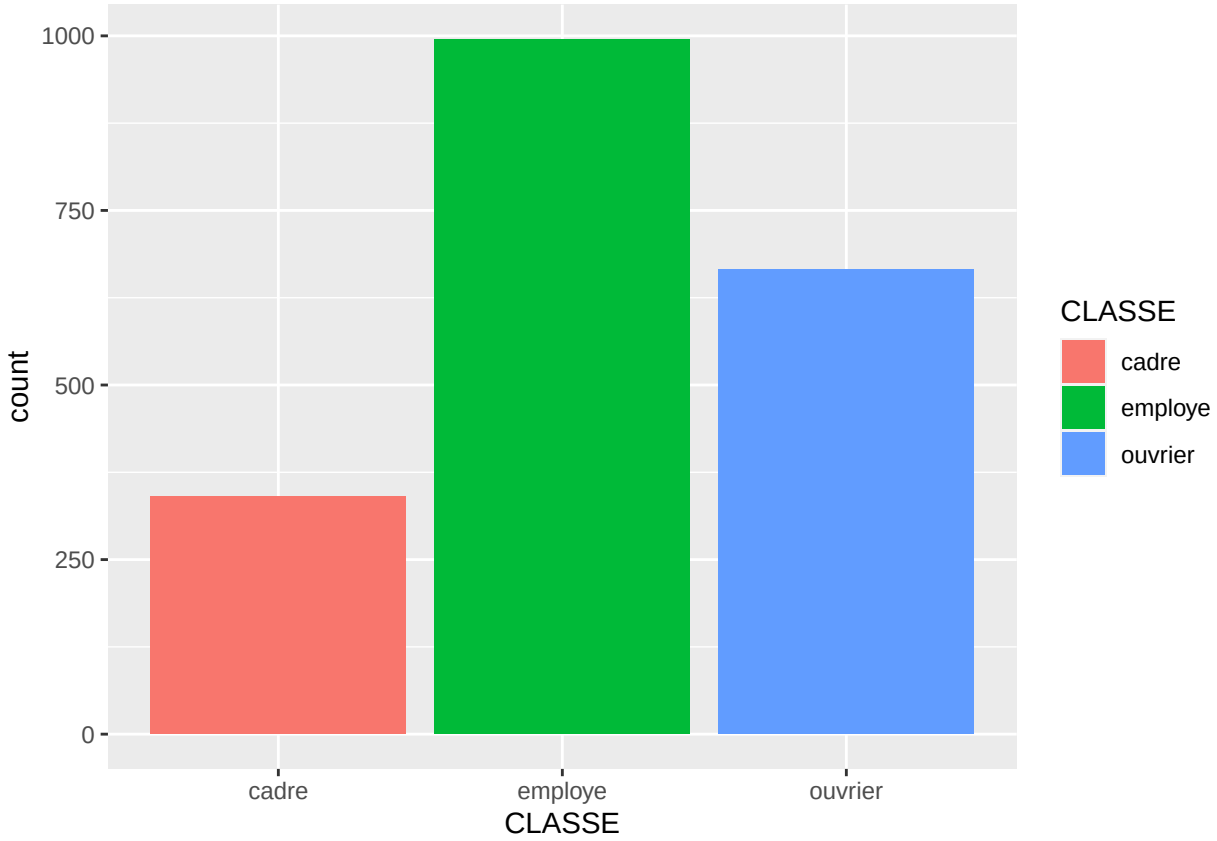
```
## # A tibble: 3 x 2
##   CLASSE  mean_age
##   <chr>      <dbl>
## 1 cadre       82.3
## 2 employe     79.3
## 3 ouvrier     78.2
```

Visualiser la répartition des décès par catégorie sociale à l'aide d'un diagramme à barres :

```
library(ggplot2)
```

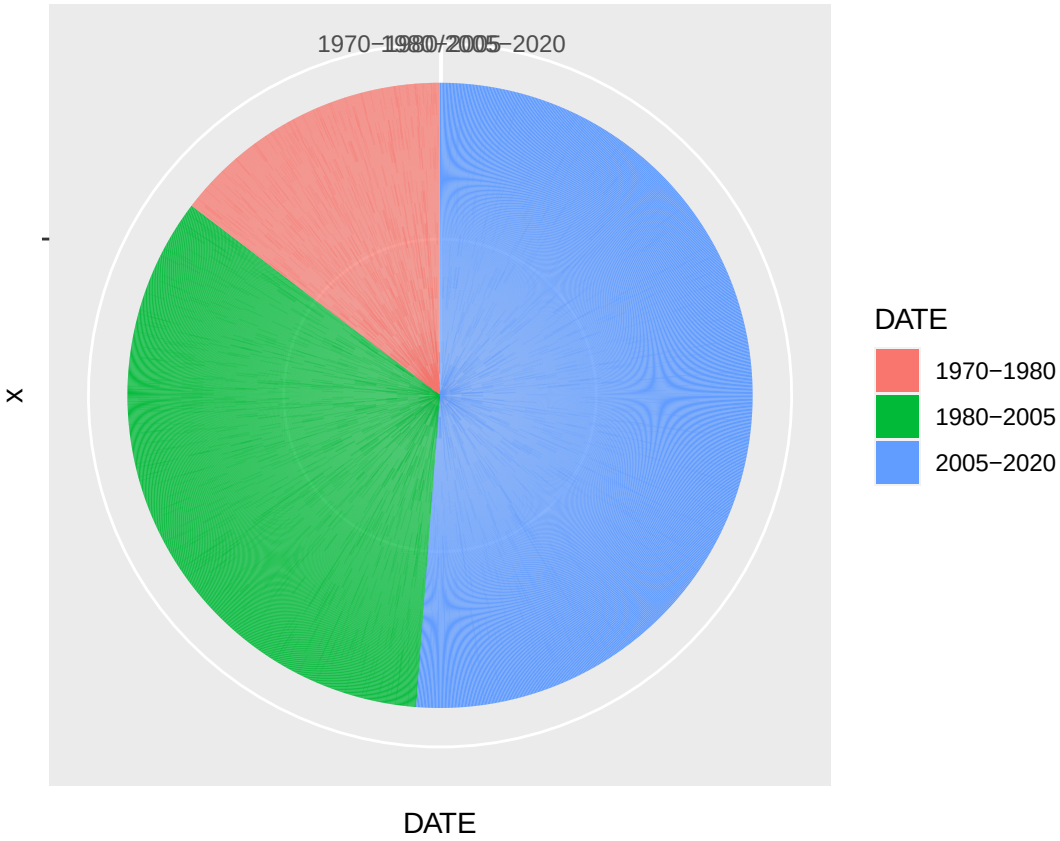
```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
ggplot(data, aes(x=CLASSE)) + geom_bar(aes(fill=CLASSE))
```



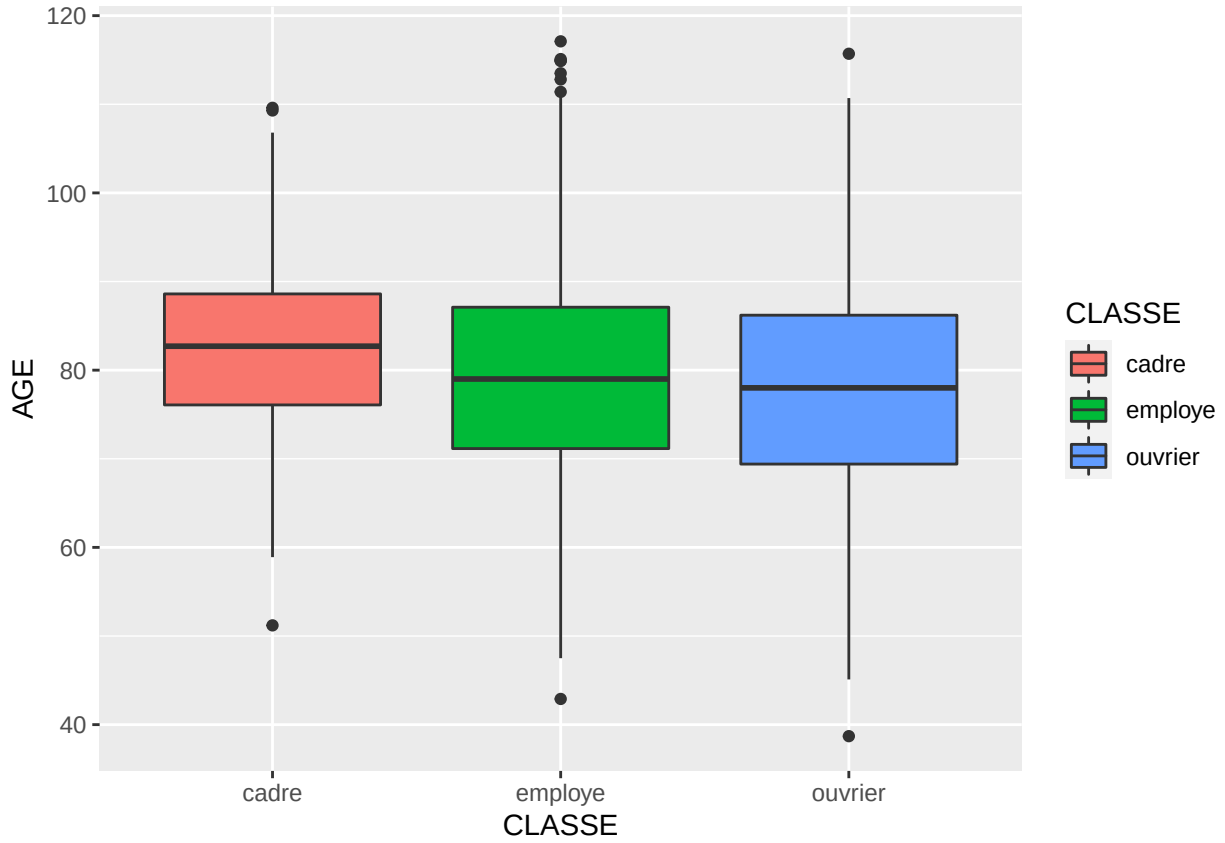
Visualiser la répartition des décès par date approximative à l'aide d'un diagramme à secteurs :

```
ggplot(data, aes(x="", y=DATE, fill=DATE)) + geom_bar(width = 1, stat = "identity") + coord_polar("y", s
```



Analyse de la répartition des décès en fonction de l'âge et de la classe sociale :

```
ggplot(data, aes(x=CLASSE, y=AGE, fill=CLASSE)) + geom_boxplot()
```



Test Statistique On va essayer de faire des Test statistiques.

```
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.5
```

```
## Loading required package: grid
```

```
result <- chisq.test(table(data$SEXE, data$CLASSE))
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$SEXE, data$CLASSE)
## X-squared = 1.5971, df = 2, p-value = 0.45
```

Le test chi-deux de Pearson indique qu'il n'y a pas d'association significative entre les variables SEXE et CLASSE dans les données. La p-valeur est de 0,45, ce qui est supérieur au seuil couramment utilisé de 0,05, ce qui signifie qu'il n'y a pas assez d'éléments pour rejeter l'hypothèse nulle selon laquelle il n'y a pas d'association entre les variables. La statistique chi-deux est de 1,5971 et le degré de liberté est de 2. Il est important de garder à l'esprit qu'un résultat non significatif ne signifie pas qu'il n'y a pas d'association entre les variables, cela signifie simplement que nous n'avons pas assez d'éléments pour suggérer qu'il y a une association en se basant sur les données d'échantillon. D'autres facteurs tels que la taille de l'échantillon, la distribution des données et les erreurs de mesure peuvent également affecter les résultats. Il est également

important de vérifier l'indépendance des observations ainsi que les fréquences attendues, si certaines d'entre elles sont trop faibles, cela pourrait affecter les résultats.

```
result <- chisq.test(table(data$AGE, data$SEXE))
```

```
## Warning in chisq.test(table(data$AGE, data$SEXE)): Chi-squared approximation may
## be incorrect
```

```
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$AGE, data$SEXE)
## X-squared = 682.17, df = 508, p-value = 3.501e-07
```

Ici le Test montre qu'il y a une association significative entre les variables AGE et SEXE, ce qui signifie que l'age de mort est bien associé au sexe.

Exercice 2 : Tests exacts de Fisher.

Le data frame suivant est composé des données d'un test clinique pour un traitement atténuant les ronflements. Celui-ci est testé contre un placebo.

- `T` : 1 pour le traitement, 0 pour le placebo ;
- `X` : moyenne du volume sonore (db) du ronflement sur la semaine précédent le test ;
- `Y` : volume sonore (db) du ronflement durant la nuit du test.

```
data=read.csv("Snoring_treatment.csv")

# Sélectionner les premiers 10 patients
patients_10 <- data[1:10,]

# Effectuer un test exact de Fisher pour comparer le traitement et le placebo
fisher.test(patients_10$T, patients_10$Y, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  patients_10$T and patients_10$Y
## p-value = 1
## alternative hypothesis: two.sided
```

1. Effectuer un test exact de Fisher sur les 10 premiers patients (fonction `combn`).

2. Mettre en place un test exact de Fisher pour l'ensemble du data frame. Afin d'approximer la p-valeur par simulations de Monte Carlo, on pourra permuter l'ordre des patients en conservant l'ordre de traitement (fonction `sample.int`).

```
# Effectuer un test exact de Fisher pour comparer le traitement et le placebo
result <- fisher.test(data$T, data$Y, alternative = "two.sided", simulate.p.value = TRUE)

print(result$p.value)
```

```
## [1] 0.5122439
```

```r
# Initialiser un compteur pour stocker le nombre de permutations qui donnent un résultat plus extrême q
counter <- 0

# Boucle pour effectuer les permutations avec 1000 essaies car sinon c'est trop long
for(i in 1:1000) {
    # Permuter l'ordre des patients en conservant l'ordre de traitement
    data_permuted <- data[sample.int(nrow(data), replace = TRUE, size = nrow(data)),]
    result_permuted <- fisher.test(data_permuted$T, data_permuted$Y, alternative = "two.sided",simulate
    if(result_permuted$p.value <= result$p.value) {
        counter <- counter + 1
    }
}

# Calculer la p-valeur approximative par Monte Carlo
p_value <- counter/1000

# Afficher les résultats
print(paste("P-value approximative par Monte Carlo :", p_value))
```

```
## [1] "P-value approximative par Monte Carlo : 1"
```

Exercice 3 : Intervalles de confiance et tests pour les proportions.

Soit $(X_1, \ldots, X_n)$ un n- échantillon de loi de bernoulli $\mathcal{B}(p)$. On rappelle que :

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Le code suivant définis la fonction int.p permettant de calculer l'intervalle de confiance au niveau α (alpha) pour un n-échantillon de loi de Bernoulli.

```r
int.p = function(vector, conf.level, na.rm=T) {
     if (length(vector)==0) { cat("Erreur ! Le vecteur ",substitute(vector),"est vide.\n")}
      else {
      n = length(vector)-sum(is.na(vector))
      proba = (1-conf.level)*100 ; proba = (100-proba/2)/100
      q_norm = qnorm(proba,0,1) # quantile
      moyenne = mean(vector,na.rm=na.rm)
      dist_max = q_norm * sqrt(moyenne*(1-moyenne)/n)
      intervalle = c(moyenne-dist_max, moyenne+dist_max)
      return(list(intervalle=intervalle, moyenne=moyenne, dist_max)) }}
```

La ligne qui suit permet d'essayer la fonction int.p sur un 100-échantillon de loi de Bernoulli $\mathcal{B}(0.2)$.

```r
x = rbinom(100,1,0.2) ; int.p(x,conf.level = 0.9)
```

On charge les deux dataframes suivants :

```r
load(file="df_trait1.Rdata")
load(file="df_trait2.Rdata")
```

0. Appliquer la fonction int.p sur les données contenues dans ces dataframes.

```
int.p(df1$Resp,conf.level = 0.9)
```

```
## $intervalle
## [1] 0.7283551 0.8156449
##
## $moyenne
## [1] 0.772
##
## [[3]]
## [1] 0.04364488
```

```
int.p(df2$Resp,conf.level = 0.9)
```

```
## $intervalle
## [1] 0.7973605 0.8826395
##
## $moyenne
## [1] 0.84
##
## [[3]]
## [1] 0.04263948
```

Ces dataframes sont les réponses positives ou non de patient à un traitement. Le laboratoire pharmaceutique disposait du traitement 1 et à tenter de l'améliorer (traitement 2). Chacun des traitements 1 et 2 a été essayé sur deux groupes de 250 et 200 patients. Les résultats pour les traitements 1 et 2 sont rangés dans les dataframe df1 et df2.

On rappelle que :

$$\frac{D_n - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

1. En vous inspirant de la fonction int.p, réaliser une fonction int.diff.p permettant de calculer l'intervalle de confiance au niveau α (alpha) pour la différence de deux n-échantillons de loi de Bernoulli. L'intervalle de confiance au niveau 95% obtenu pour la différence $p_2 - p_1$ est-il strictement positif ?

```
int.diff.p = function(x1, x2, alpha) {
  n1=length(x1);
  n2=length(x2);
  p1 <- sum(x1)/n1;
  p2 <- sum(x2)/n2;
  se <- sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2);
  z <- qnorm(1-alpha/2);
  lower <- p1-p2 - z*se;
  upper <- p1-p2 + z*se;
  return(c(lower, upper));
}
```

```
int.diff.p(df1$Resp,df2$Resp,0.95)
```

```
## [1] -0.07032613 -0.06567387
```

L'intervalle de confiance à 95% donne une borne superieur négative donc on a $p_1 - p_2 \leq 0$ donc on a bien la strict positivité de $p_2 - p_1$.

2. Les investigateurs du projet trouvent que l'intervalle de confiance est trop grand. Ils hésitent à rajouter 150 patients pour le traitement 1 et 100 patients pour le traitement 2. En supposant que les estimations des variances pour chaque groupe sont correctes, quelle serait la nouvelle longueur de l'intervalle ?

La nouvelle longeur de l'intervalle serait :

$$L_{Intervalle} = 2 \cdot Z \sqrt{\frac{p'_1(1-p'_1)}{n_1+150} + \frac{p'_2(1-p'_2)}{n_2+100}}$$

Avec p_1 et p_2 les nouvelles valeurs de moyenne empirique.

On aurait alors un gain :

$$Gain = \frac{L_{Test_1} - L_{Test_2}}{L_{Test_1}} = 1 - \frac{\sqrt{(p'_1 \cdot (1-p'_1)/(n_1+150) + p'_2 \cdot (1-p'_2)/(n_2+100)}}{\sqrt{(p_1 \cdot (1-p_1)/(n_1) + p_2 \cdot (1-p_2)/(n_2)}}$$

Avec Z la valeur critique de la loi normale standardisée correspondant au niveau de confiance choisi.