Statistiques Inférentielles L3 - TIC

Guillaume Metzler

11/24/2021

Fiche de TD numéro 3

Exercice 1 : Etude de deux populations indépendantes

On peut commencer par considérer que les deux populations sont Gaussiennes. On note la loi des garçons $\mathcal{N}(\mu_1, \sigma_1^2)$ et la loi des filles $\mathcal{N}(\mu_2, \sigma_2^2)$. Comparer les populations consiste alors a comparer leurs distributions et comme ce qui distingue deux Gaussiennes sont les espérances et les variances, on doit donc les comparer l'une après l'autre.

Comparaison des variances

On commence traditionnellement par comparer les variances. Les hypothèses sont :

- H_0: les variances sont égales
- H_1: les variances ne sont pas égales.

Pour faire ce test, on suppose que H_0 est vraie. Comme on sait que

$$(n_1 - 1)S_1^2/\sigma_1^2 \sim \chi^2_{n_1-1} \quad \text{et} \quad (n_2 - 1)S_2^2/\sigma_2^2 \sim \chi^2_{n_2-1},$$

on a par définition de la loi de Fisher

$$\frac{\frac{(n_1-1)S_1^2/\sigma_1^2}{n_1-1}}{\frac{(n_2-1)S_2^2/\sigma_2^2}{n_2-1}} \sim F_{n_1-1,n_2-1}$$

ce qui se simplifie en

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Maintenant, sous H_0, on a $\sigma_1^2 = \sigma_2^2$, et on obtient

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}.$$

Par principe, nous respecterons ce que vous avez vu en cours et on mettra au numérateur la plus grande des deux variance, et ainsi on sera amener à faire un test unilatéral (car la statistique sera forcément plus grande que 1). Dans notre cas, on considérera donc la statistique

$$\frac{S_2^2}{S_1^2} \sim F_{n_2-1,n_1-1}.$$

Le quantile associé au risque $1-\alpha$ dans la table de la loi de Fisher avec degrés $n_2-1=60$ et $n_1-1=40$ est $f_{60,40,.95}=1.64$.On calcule maintenant le rapport des variances corrigées. On a $s_1^2 = 41/40 \cdot 0.08 = 0.082$ et $s_2^2 = 61/60 \cdot 0.09 = 0.0915$. Le rapport est

$$\frac{s_2^2}{s_1^2} = 1.12.$$

Comme la statistique de test ne dépasse pas ce quantile, on ne peut pas rejeter H_0. Il est raisonnable de supposer les variances égales (même si toutefois le risque de se tromper, le risque de second espèce, n'est pas connu).

Comparaison des moyennes

On se tourne maintenant vers la comparaison des espérances. On a

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n2}\right).$$

Sous l'hypothèse que les variances sont 'egales, c'est à dire que $\sigma_1 = \sigma_2 =: \sigma^2$, on obtient

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n2}\right)\right).$$

En remplaçant les variances par les valeurs empiriques, on estime σ^2 traditionnellement par

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \quad \text{numériquement} \quad S^2 = \frac{(41-1)\times 0.082 + (61-1)\times 0.0915}{41+61-2} = 0.0877$$

et on a le résultat suivant: la variable T donnée par

$$T = \frac{\bar{X}_1 - \mu_1 - (\bar{X}_2 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

suit une loi de Student à n_1+n_2-2 degrés de liberté. Pour faire le test de savoir si les deux espérances sont égales, on définit les deux hypothèses complémentaires:

- H_0: les espérances μ_1 et μ_2 sont égales.
- H_1: les espérances μ_1 et μ_2 sont différentes.

On suppose alors que H_0 est vraie et sous H_0, on a :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

On calcule alors les quantiles de Student aux niveau 0.025 et 0.975, c'est a dire $t_{n_1+n_2-2,.025} = -1.98$ et $t_{n_1+n_2-2,.975} = 1.98$ on obtient l'intervalle de décision

$$I_\alpha = [-1.98, 1.98] \text{ (bilatéral)} \quad \text{ou} \quad I_\alpha = [-\infty, 1.66] \text{ (unilatéral).}$$

On calcule alors la valeur réalisée par T sur nos données et on obtient

$$\mathsf{t} = \frac{1.07 - 1.03}{\sqrt{0.0877 \cdot (\frac{1}{41} + \frac{1}{61})}} = 0.502.$$

Comme $\mathtt{t} \in I_\alpha$, on en déduit que la taille moyenne entre les filles et les garçons n'est pas significativement différente.

On nous demandait si la taille moyenne des garçons était supérieure à celle des filles. Nous aurions donc du effectuer un test unilatéral et notre intervalle de confiance aurait donc était de la forme

$$I_\alpha = [-\infty, t_{100,0.95}] = [-\infty, 1.66].$$

Donc on ne rejette toujours pas l'hypothèse H_0.

Remarques :

- Un test unilatéral supérieur correspond à une zone de rejet "à droite", i.e. de la forme $[u_{1-\alpha}, +\infty]$
- Un test unilatéral inférieur correspond à une zone de rejet "à gauche", i.e. de la forme $[-\infty, u_{1-\alpha}]$

Nous aurions également pu calculer la p-value, ce que l'on se propose de faire avec R, à l'aide de la fonction "pt".

```
# Calcul de la pvalue pour le test bilatéral
pvalue_bi = 2*(1-pt(0.502,100))
paste("La pvalue est égale à",pvalue_bi)
```

```
## [1] "La pvalue est égale à 0.616771062051632"
```

```
# Calcul de la pvalue pour le test unilatéral supérieur

pvalue_uni = (1-pt(0.502,100))
paste("La pvalue est égale à",pvalue_uni)
```

```
## [1] "La pvalue est égale à 0.308385531025816"
```

Exercice 2 : Etude de deux populations (échantillons) indépendantes

Dans la cas présent, les échantillons sont trop petits pour qu'il soit raisonnable de faire l'hypothèse gaussienne pour les données. Pour tester si l'alcool a une influence sur le temps de réaction on va donc faire un test de Wilcoxon-Mann-Whitney.

Une approche non-paramétrique

Pour cela, on calcule les rangs de chaque donnée du groupe 1 (sans) dans l'ensemble des données et on trouve :

Sans	0.68	0.64	0.68	0.82	0.58	0.80	0.72	0.65	0.84	0.73	0.65	0.59	0.78	0.67	0.65
Rang	10.0	4.0	10.0	25.0	1.0	23.5	13.5	6.0	26.0	15.5	6.0	2.0	20.0	8	6

On rappelle que si l'on a des ex-aequos dans notre échantillon, on associe le même rang à chaque donnée de l'échantillon et ce rang est égales à la moyenne des rangs de notre échantillon.

La somme des rangs ainsi calculée est alors notée R. On doit comparer les deux hypothèses suivantes :

- H_0 : les deux populations ont même loi
- H_1 : les deux populations n'ont pas la même loi

Sous H_0, l'espérance de la variable aléatoire R ainsi que sa variance (dans le cas où nous avons des ex-aequos, on doit donc la corriger) sont connues :

$$\mu_R = \mathbb{E}[R] = \frac{n_1(n_2+n_1+1)}{2} \quad \text{et} \quad \sigma_R^2 = Var(R) = \frac{n_1 n_2 (n_1+n_2+1)}{12} - \frac{n_1 n_2 \sum_{j=1}^{e} D_j^3 - D_j}{12(n_1+n_2)(n_1+n_2+1)}.$$

Ainsi la variable aléatoire Z définie par

$$Z = \frac{R - \mu_R}{\sigma_R}$$

est approximativement gaussienne (en outre, les tables ne sont disponibles que lorsque n_1 et n_2 sont inférieures à 10) et on peut utiliser un intervalle de décision (ou intervalle de confiance) de la forme

$$I_\alpha = \left[z_{\alpha/2}, z_{1-\alpha/2}\right].$$

Si la valeur z de la statistique de test réalisée par Z tombe dans cet intervalle, alors on ne rejette pas l'hypothèse H_0. Calculons les quantités qui nous intéressent :

Dans notre cas, les valeurs de D_j (nombre de répétitions de chaque valeur dans l'échantillon) qui sont diff'erentes de 1 sont : 3, 3, 2, 2, 3, 2, ce qui donne $\sum_{j=1}^{e} D_j^3 - D_j = 90$. Ainsi

$$\sigma_R^2 = 579.4355.$$

Nous avons également

$$\mathbb{E}(R) = 232.5,$$

ainsi, notre statistique de test z est égale à

$$z = \frac{176.5 - 232.5}{\sqrt{579.4355}} = -2.32$$

Cette valeur n'appartient à l'intervalle de confiance, on peut donc rejeter l'hypothèse H_0. Confirmons cela avec une analyse sous R

```
# Valeurs de notre échantillon avec et sans prise d'alcool
sans <- c(68,64,68,82,58,80,72,65,84,73,65,59,78,67,65)/100
avec <- c(73,62,76,92,68,87,77,70,88,79,72,80,78,86,78)/100

# Test de Wilcoxon
wilcox.test(avec,sans,alternative = "greater", correct=FALSE)
```

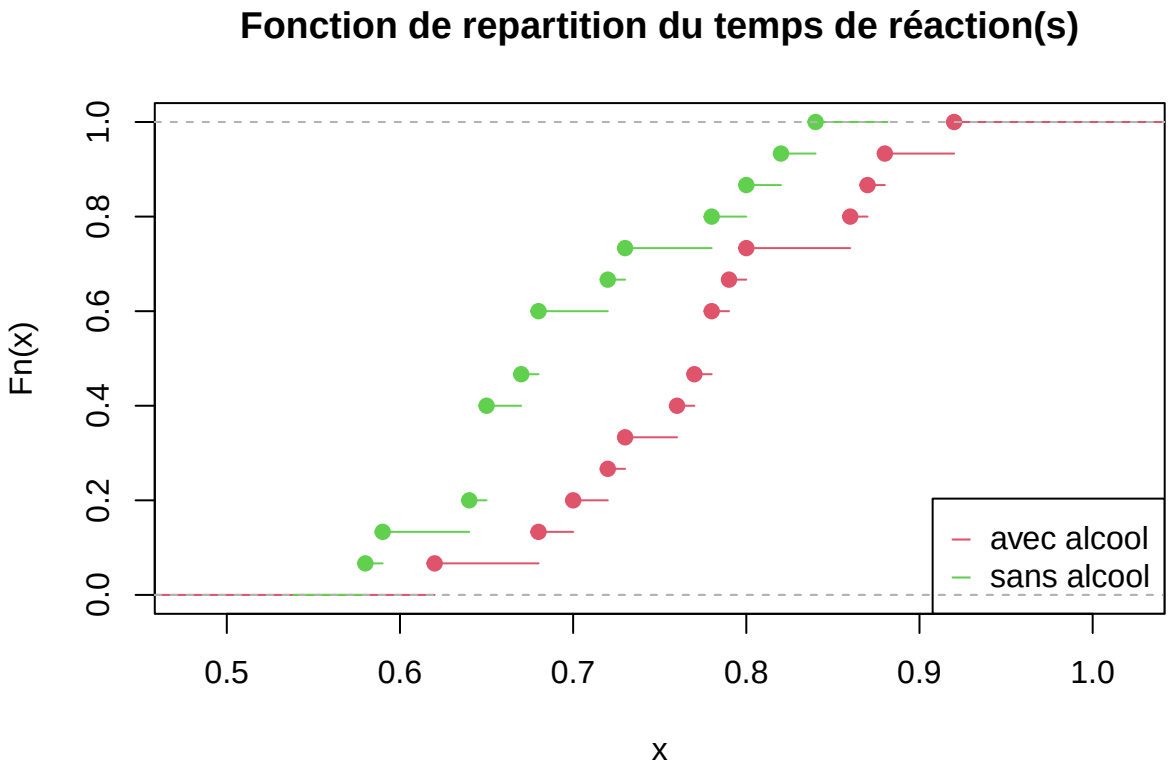
```
## Warning in wilcox.test.default(avec, sans, alternative = "greater", correct =
## FALSE): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test
##
## data:  avec and sans
## W = 168.5, p-value = 0.009992
## alternative hypothesis: true location shift is greater than 0
```

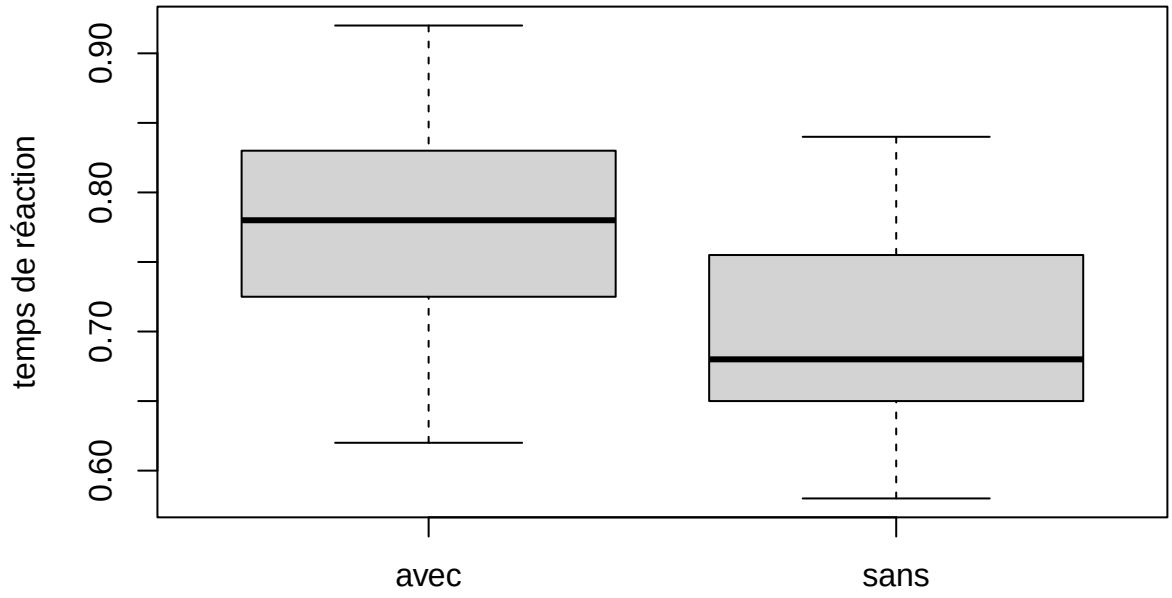
Remarque : les valeurs de la statistique de test sont différentes de celles calculées, nous sommes entrain de regarder pourquoi ...

Un complément graphique pour étudier les deux distributions

```
# Représentation des fonctions de répartition
plot(ecdf(avec),col=2,main="Fonction de repartition du temps de réaction(s)",
     xlim=c(min(avec,sans)-0.1,max(avec,sans)+0.1))
lines(ecdf(sans),col=3)
legend("bottomright",pch='_',c('avec alcool','sans alcool'),col=2:3)
```




```
# Boxplot associé aux deux distributions
boxplot(avec,sans,names=c('avec','sans'),ylab='temps de réaction')
```



Une approche paramétrique

Même si les échantillons sont de tailles < 30, il est tout fait possible de tenter une approche paramétrique basée sur le test de Student, si nos échantillons sont gaussiens !

```
# Vérifions que nos échantillons sont gaussiens

p_sans <- shapiro.test(sans)$p.value
```

```r
p_avec <- shapiro.test(avec)$p.value

if ((p_sans > 0.05)&(p_avec > 0.05)){
  print("Les échantillons sont gaussiens, ok pour l'approche paramétrique")
} else {
  print("Au moins l'une des p_value est inférieure à < 0.05,
        on ne peut donc pas envisager une approche paramétrique")
}
```

```
## [1] "Les échantillons sont gaussiens, ok pour l'approche paramétrique"
```

Pour déterminer le type de test à utiliser, on va commencer étudier les variances de nos deux échantillons :

- si les variances sont égales : on effectuera un test de Student à $n_1 + n_2 - 2$ degrés de libertés
- si les variances ne sont pas égales : on effectue un test de Student basé sur l'approximation d'Aspin-Welch

```r
# Comparaison des variances à l'aide d'un test de Fisher
p_fisher <- var.test(avec,sans)$p.value
if (p_fisher > 0.05){
  print("Les variances sont égales")
} else {
  print("Les variances ne sont pas égales")
}
```

```
## [1] "Les variances sont égales"
```

On va donc effectuer un test de Student à $n_1 + n_2 - 2$ degrés de liberté pour comparer les moyennes deux échantillons

```r
# Test de Student, en précisant que les moyennes sont égales
t.test(avec,sans,var.equal = T,alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  avec and sans
## t = 2.6571, df = 28, p-value = 0.006435
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.02830319        Inf
## sample estimates:
## mean of x mean of y
## 0.7773333 0.6986667
```

On rejette donc l'hypothèse H_0.

Exercice 3 : Etude de deux populations (échantillons) appariées

On souhaite savoir si un régime a un effet sur la masse des individus. On a ici pesé 10 individus avant et après avoir effectué le régime, il y a donc une dépendance entre les valeurs prises par l'échantillon. On ne peut donc pas effectuer un "test de Wilcoxon-Mann-Whitney".

Nous avons à faire à des échantillons appariés. Il va donc falloir s'intéresser aux valeurs de la variable aléatoire D traduisant la différence de masse avant et après le régime et faire le test d'hypothèses suivant :

- H_0 :le régime n'a pas d'impact sur la masse des individus,
- H_1 :le régime a un impact sur la masse des individus.

Trois possiblités s'offrent à nous :

- • un test paramétrique de Student
- • un test non paramétrique : test du signe
- • un test non paramétrique : test du rang signé de Wilcoxon

Commençons par évaluer l'impact du régime sur les individus de notre échantillon :

Avant	67	83	158	78	87	58	79	63	69	72
Après	66	84	121	76	82	58	77	64	68	70
Différence	1	−1	37	2	5	0	2	−1	1	2

```
avant <- c(67, 83, 158, 78, 87, 58, 79, 63, 69, 72)
apres <- c(66, 84, 121, 76, 82, 58, 77, 64, 68, 70)
D_ech <- c( 1 , -1 , 37 , 2 , 5 , 0 , 2 , -1 , 1 , 2)
```

Un test paramétrique ?

Nous pourrions procéder à un "test de Student" dans le cas où nos données sont normalement distribuées. Malheureusement le test de Shapiro ci-dessous, nous montre que cette hypothèse n'est pas valable.

```
# Test de Shapiro
shapiro.test(D_ech)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  D_ech
## W = 0.50333, p-value = 3.979e-06
```

Nous devons donc forcément utiliser un test non paramétrique : "test du signe" ou "test du rang signé de wilcoxon". Ces deux tests nécéssitent d'abord de supprimer les valeurs nulles dans notre échantillon, d'où, ce qui nous donne l'échantillon suivant

Différence	1	−1	37	2	5	2	−1	1	2

Test du signe On va donc effectuer notre "test du signe", pour cela on regarde le nombre de valeurs positives (ou négatives) dans notre échantillon. On rappelle que ce test est le plus puissant lorsque les queues de la distribution sont diffuses, en outre, si on note W la statistique de test associé, nous avons

$$\mathbb{E}[W] = \frac{n}{2} \quad \text{et} \quad \sigma^2(W) = n/4$$

On se concentre ici sur le nombre de différences positives, ici au nombre de 7. Sous l'hypothèse H_0 cette statistique de test suit une loi binomiale $\mathcal{B}(n, 1/2)$.

On réfère ensuite aux tables du test du signe disponibles dans le manuel (pour $n = 9$ et $C = 7$), et on trouve une p-value de 0.1797 ce qui ne permet pas de rejeter l'hypothèse H_0, on ne peut donc pas conclure que le régime a un effet sur les participants.

```
# Vérification sous R

library(BSDA)
```

```
## Le chargement a nécessité le package : lattice
```

```
##
## Attachement du package : 'BSDA'

## L'objet suivant est masqué depuis 'package:datasets':
##
##     Orange
```

```
SIGN.test(D_ech)
```

```
##
##  One-sample Sign-Test
##
## data:  D_ech
## s = 7, p-value = 0.1797
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  -0.6755556  4.0266667
## sample estimates:
## median of x
##         1.5
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.8906  0.0000 2.0000
## Interpolated CI       0.9500 -0.6756 4.0267
## Upper Achieved CI     0.9785 -1.0000 5.0000
```

Ce premier test ne prend en compte que le signe de la différence mais sans tenir compte de l'importance relative de chaque différence.

Test du rang signé de Wilcoxon On se propose de regarder ce que donnnerai, à titre de comparaison, un test du rang signé de Wilcoxon.

Pour rappel, la statistique de tel étudié somme le rang des valeurs positives dans l'échantillon où le rang est calculé selon la valeur absolue des valeurs prises par l'échantillon

$$W = \sum_{i=1}^{n} R_i^+ \times (rang(x_i > 0)).$$

On commence par attribuer les rang correspondants à chaque donnée en fonction de leur valeur absolue

Différence	1	-1	37	2	5	2	-1	1	2
|Différence|	1	1	37	2	5	2	1	1	2
Rang	2.5	2.5	9	5	6	6	2.5	2.5	6

Si on somme les valeurs des rangs des différences positives, nous trouvons

$$W = 2.5 + 2.5 + 6 + 6 + 6 + 9 + 8 = 40$$

Notre échantillon étant de taille faible (<10) on ne peut utiliser un test basée sur une approximation gaussienne, on va donc se fier à votre table fournie en annexe.

La taille de notre échantillon est 9, la table nous indique les bornes de notre intervalle de confiance, il faut donc rejeter l'hypothèse H_0 si nous sommes inférieure ou égale à la borne inférieure ou supérieure ou égale à la borne supérieure.

Pour $n = 9$, l'intervalle de confiance est de la forme $[6, 39]$ or $W = 40$, donc on rejette l'hypothèse H_0 à nouveau. Vérifions cela avec R

```
# Test du rang signé, on précise que les deux échantillons sont appariés
# et que l'on ne souhaite pas effectuer de correction
wilcox.test(avant,apres, paired = TRUE, correct= FALSE)
```

```
## Warning in wilcox.test.default(avant, apres, paired = TRUE, correct = FALSE):
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(avant, apres, paired = TRUE, correct = FALSE):
## cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test
##
## data:  avant and apres
## V = 40, p-value = 0.0358
## alternative hypothesis: true location shift is not equal to 0
```

Le test conduit également au rejet de l'hypothèse H_0.

Fiche de TD numéro 4

Exercice 1

Dans cet exercice, nous étudions deux échantillons différents. Ces derniers correspondent au temps de réponse à un test par un ensemble de 27 candidats : 15 hommes et 12 femmes. Les temps est de réponses sont données ci-dessous :

```
# Temps de réponses
homme <- c(8.6, 10.9, 7.3, 9.2, 8.5, 9.2, 9.1, 8.9, 10.7, 8.2 , 7.1, 9.4, 8.3, 9.7, 9.2)
femme <- c(8.3, 7.2, 8.7, 6.7, 10.3, 6.8, 9.8, 8.9, 9.6, 8.6, 6.7, 7.5)
```

Une étude préalable consisterait à étudier si nos échantillons sont gaussiens ou non à l'aide d'un test de Shapiro, même si pour l'échantillon "femme", la taille fera que l'échantillon ne sera pas assez puissant :

```
# Tests de Shapiro
shapiro.test(homme)$p.value
```

```
## [1] 0.6124982
```

```
shapiro.test(femme)$p.value
```

```
## [1] 0.3158922
```

Les deux tests ne permettant pas de rejeter l'hypothèse de normalité. On va donc pouvoir l'utiliser dans la suite.

(i) Test de comparaison des variances

On souhaite savoir si les variances des temps de réponse des hommes et des femmes sont identiques, il va donc falloir procéder à un test d'égalité des variances (appelée test de comparaison des variances dans votre cours). Pour rappel, ce test repose sur la statistique de Fisher et nous effectuerons ce test au risque de première espèce $\alpha = 0.05$. En outre la statistique de test est la suivante :

$$F = \frac{\frac{n_1 V_1}{n_1 - 1}}{\frac{n_2 V_2}{n_2 - 1}} = \frac{S_1^2}{S_2^2},$$

où V_i désigne la variance biaisée de l'échantillon i et S_i^2 désigne la variance débiaisée de l'échantillon i, elle suit une loi F_{n_1-1,n_2-2} et on supposera que $S_1^2 > S_2^2$.

Regardons cela en pratique :

```
# Par défaut, R calcule les variances débiaisées
v_homme <- var(homme)
v_femme <- var(femme)

# Statistique de Fisher, v_femme > v_homme
f <- v_femme / v_homme

# On compare cette valeur à la valeur critique au risque de 5%
f_critique <- qf(0.95,length(femme)-1,length(homme)-1 )

# Rejet ou non de H_0
paste("Le test me permet-il de rejeter H_0 ?", f>f_critique)
```

```
## [1] "Le test me permet-il de rejeter H_0 ? FALSE"
```

On ne peut donc pas rejeter l'hypothèse H_0. On peut donc pas rejeter l'hypothèse d'égalité des variances. Nous aurions également pu calculer la p-value pour un test unilatéral supérieur (ou à droite)

```
# Calcul de la p-value
p_value <- 1-pf(f,length(femme)-1,length(homme)-1)
p_value
```

```
## [1] 0.2429993
```

R dispose d'une fonction permettant de faire cela automatiquement :

```
# Test d'égalité des variances
# l'option alternative = "greater" indique un test unilatéral supérieur
var.test(femme,homme,alternative = "greater")
```

```
##
##  F test to compare two variances
##
## data:  femme and homme
## F = 1.4767, num df = 11, denom df = 14, p-value = 0.243
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.5756069       Inf
## sample estimates:
## ratio of variances
##           1.476718
```

(ii) Test de comparaison des moyennes

Comme nous avons vu que les variances des groupes sont égales, nous pouvons, afin d'étudier si le sexe du candidat à une infleunce sur la performance moyenne, utiliser un test de Student "standard"(car nos échantillons sont gaussiens). La statistique de test utilisée est la suivante :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 V_1 + n_2 V_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

qui suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté, puis comparer la valeur de cette statistique de test à la valeur critique $t_{0.95,n_1+n_2-2}$.

Regardons cela en pratique :

```
# Statistique de test
lh <- length(homme)
lf <- length(femme)

t = (mean(femme)-mean(homme))/sqrt(( (lh-1)*v_homme + (lf-1)*v_femme)/(lh+lf-2) *(1/lh+1/lf))
t
```

```
## [1] -1.560797
```

```
# Comparaison à la valeur critique : test bilatéral
paste("Le test me permet-il de rejeter H_0 ?", abs(t)>qt(0.975,lh+lf-2 ))
```

```
## [1] "Le test me permet-il de rejeter H_0 ? FALSE"
```

```
# Calcul de la p-value
p_value <- 2*(1-pt(abs(t),lf+lh-2))
p_value
```

```
## [1] 0.1311426
```

On ne peut donc pas rejeter l'hypothèse nulle et on peut donc conclure qu'il n'y a pas de différences de résultats entre les hommes et les femmes.

A nouveau, nous aurions pu faire cela directement comme suit :

```
# Test de comparaison des moyennes de Student
# On précisera que les variances sont égales.
t.test(femme,homme, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  femme and homme
## t = -1.5608, df = 25, p-value = 0.1311
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6120822  0.2220822
## sample estimates:
## mean of x mean of y
##  8.258333  8.953333
```

Exercice 2

On souhaite contrôler le taux de fer dans le foie pour cela, on observe ce taux pour 5 régimes différents. Dans le cas présent, nous étudions 5 échantillons (populations) indépendant(e)s.

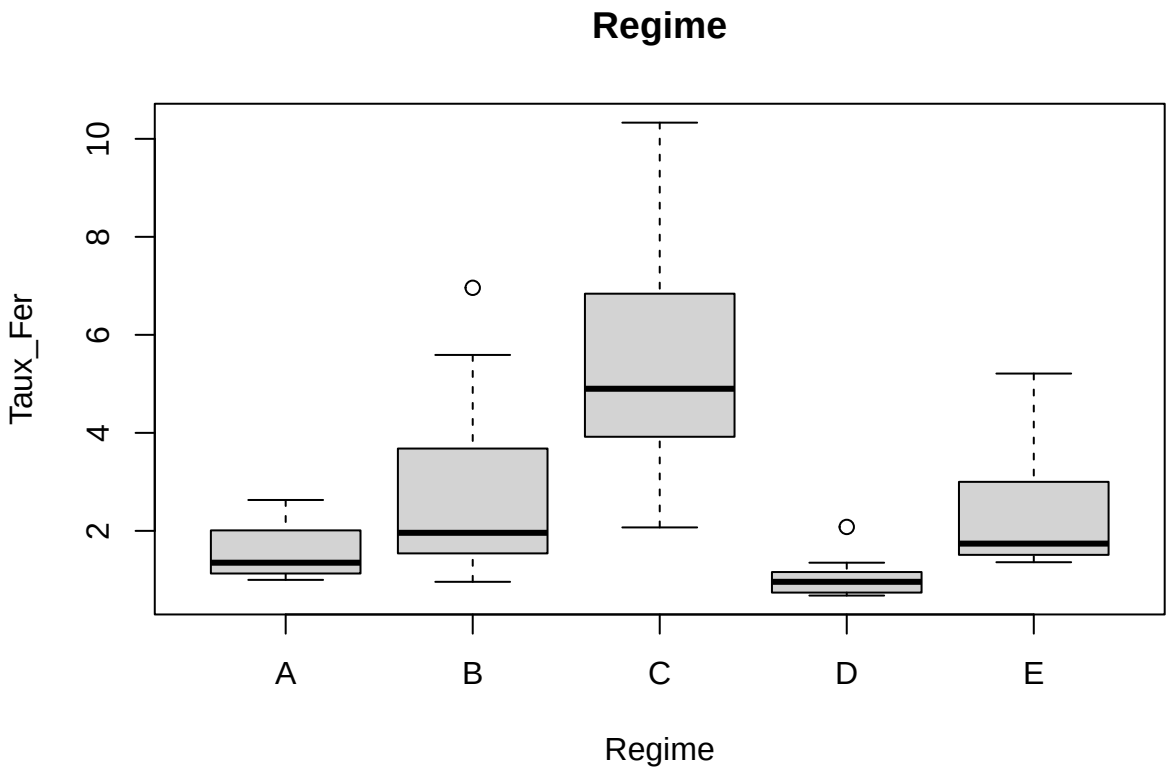
```
# Création de la base de données
data <- rbind(
  data.frame(Regime=rep("A",9),
```

```
            Taux_Fer = c(2.23, 1.14, 2.63, 1.00, 1.35, 2.01, 1.64, 1.13, 1.01)),
  data.frame(Regime=rep("B",9),
            Taux_Fer = c(5.59, 0.96, 6.96, 1.23, 1.61, 2.94, 1.96, 3.68, 1.54)),
  data.frame(Regime=rep("C",9),
            Taux_Fer = c(4.50, 3.92, 10.33, 8.23, 2.07, 4.90, 6.84, 6.42, 3.72)),
  data.frame(Regime=rep("D",9),
            Taux_Fer = c(1.35, 1.06, 0.74, 0.96, 1.16, 2.08, 0.69, 0.68, 0.84)),
  data.frame(Regime=rep("E",9),
            Taux_Fer = c(1.40, 1.51, 2.49, 1.74, 1.59, 1.36, 3.00, 4.81, 5.21))
)
```

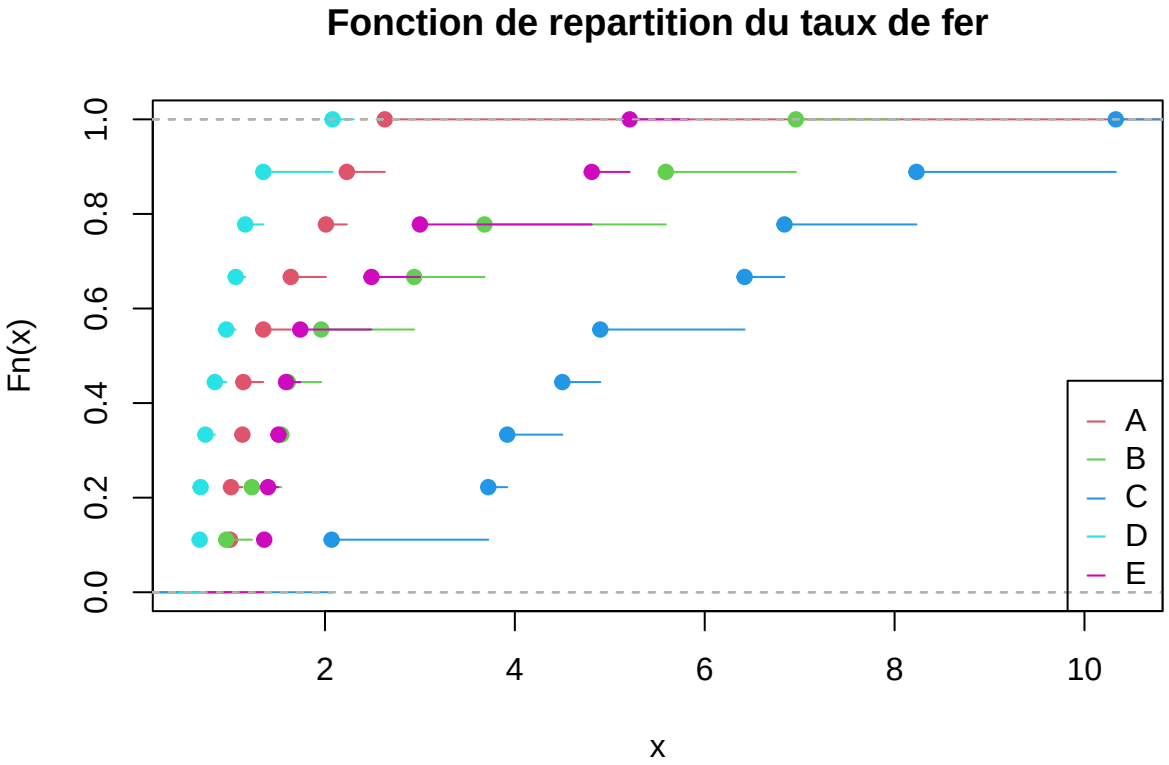
(i) Représentations graphiques

On commence par représenter les 5 boîtes à moustache correspondant aux 5 échantillons et on mettra en dessous les 5 fonctions de répartitions

```
# Boites de Tukey
boxplot(Taux_Fer~Regime,data=data, main='Regime')
```




```
# Fonctions de répartitions
plot(ecdf(data[data$Regime == "A","Taux_Fer"]),
     col=2,
     main="Fonction de repartition du taux de fer",
     xlim=c(min(data[,2])-0.1,max(data[,2])+0.1))
lines(ecdf(data[data$Regime == "B","Taux_Fer"]),col=3)
lines(ecdf(data[data$Regime == "C","Taux_Fer"]),col=4)
lines(ecdf(data[data$Regime == "D","Taux_Fer"]),col=5)
lines(ecdf(data[data$Regime == "E","Taux_Fer"]),col=6)
legend("bottomright",pch='_',c("A","B","C","D","E"),col=2:6)
```



On observe que les distributions sont très différentes entre les différents régimes. Confirmons cela à l'aide d'un test statistique.

(ii) Etude de l'impact des régimes

a) On commence par un premier test paramétrique, on souhaite savoir si une variable qualitative : "le régime" a une influence sur une variable quantitative "le taux de fer". La variable quantitative prenant plus de deux modalités, on va donc effectuer une ANOVA (analyse de variances).

Verifions d'abord l'hypothèse de normalité des données

```
# Test de Shapiro
shapiro.test(data[data$Regime == "A","Taux_Fer"])$p.value
```

```
## [1] 0.1609506
```

```
shapiro.test(data[data$Regime == "B","Taux_Fer"])$p.value
```

```
## [1] 0.08825277
```

```
shapiro.test(data[data$Regime == "C","Taux_Fer"])$p.value
```

```
## [1] 0.8575261
```

```
shapiro.test(data[data$Regime == "D","Taux_Fer"])$p.value
```

```
## [1] 0.04204723
```

```
shapiro.test(data[data$Regime == "E","Taux_Fer"])$p.value
```

```
## [1] 0.01654231
```

Pour cela, on commence par vérifier l'hypothèse de d'homoscédasticité, i.e. que les variances sont homogènes entre les différents groupes à l'aide d'un test de Bartlett

```r
# Test d'homogénéité des variances
bartlett.test(Taux_Fer~Regime, data=data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Taux_Fer by Regime
## Bartlett's K-squared = 27.003, df = 4, p-value = 1.985e-05
```

Ce dernier montre que les variances entre les différents groupes ne sont pas égales ... regardons quand même ce que donnerait notre ANOVA et vérifions que nos résidus sont eux normalement distribués.

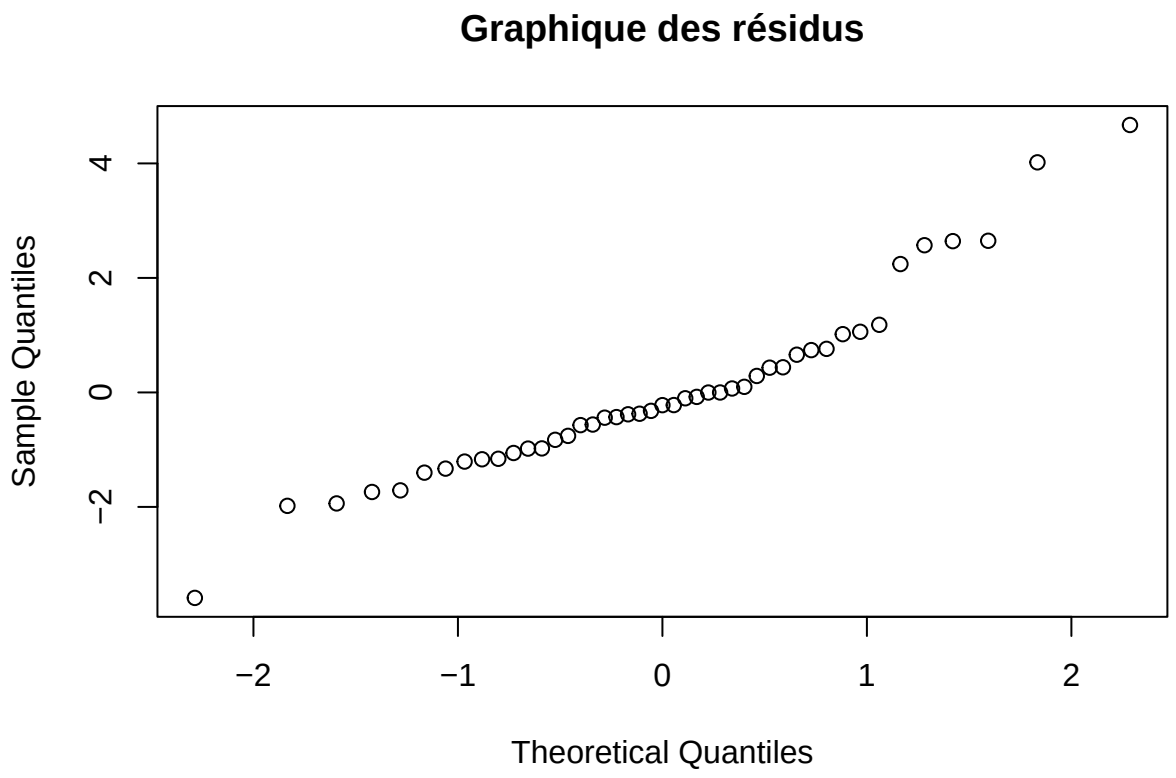
```r
# Analyse de Variances
res=aov(Taux_Fer~Regime, data=data)
print(anova(res))
```

```
## Analysis of Variance Table
##
## Response: Taux_Fer
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Regime     4 114.92 28.7307  10.502 6.535e-06 ***
## Residuals 40 109.44  2.7359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comme nous avons pu le voir graphiquement, le régime a bien un impact sur le taux de fer chez les différents individus, cela se confirme avec le test et une p-value très faible sur le facteur “Regime”.

Vérifions que nos résidus sont normalement distribués.

```r
# Graphe des résidus : idéalement, ils doivent se trouver sur une droite
qqnorm(res$residuals, main = "Graphique des résidus")
```



```r
# Test de Shapiro
shapiro.test(res$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.93213, p-value = 0.01116
```

Le test de Shapiro nous montre que l'on peut rejeter l'hypothèse de normalité pour les résidus de notre modèle.

L'approche paramétrique suppose que nos distributions sont gaussiennes pour les différents groupes ce qui a peu de chance d'être le cas ici (voir débur de la correction), de plus nos échantillons sont trop petits pour que le test de Shapiro soit significatif ici. On va donc se tourner vers une approche non paramétrique de l'ANOVA.

b) Regardons maintenant l'approche non paramétrique reposant sur le test de Kruskal-Wallis, spécialement adapté aux petits échantillons (voir section 3.4.1.2 de votre manuel) et repose sur les rangs. Il s'agit d'une généralisation du test du Wilcoxon. Le test de Kruskal-Wallis va tester l'hypothèse selon laquelle toutes les distributions sont identiques en comparant les fonctions de répartition.

```r
# Test de Kruskl-Wallis
kruskal.test(Taux_Fer~Regime, data=data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Taux_Fer by Regime
## Kruskal-Wallis chi-squared = 24.693, df = 4, p-value = 5.798e-05
```

On observe bien que la variable "Regime" a un impact sur le "Taux_Fer" mesuré.

Exercice 3

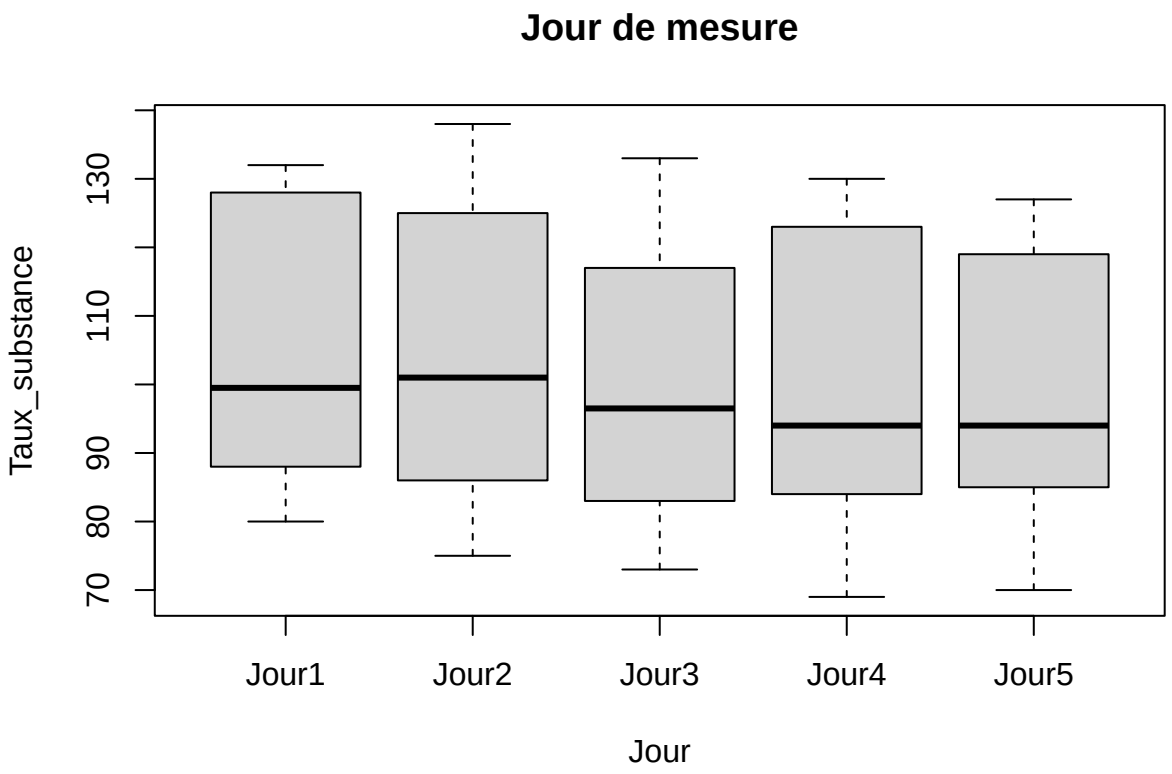
Nous étudions l'évolution du taux d'une substance au cours de 5 jours chez 10 patients différents. Nous sommes donc dans le cas où nous étudions des populations dépendantes ou d'échantillons appariés.

```r
# Construction de la base de données
data <- rbind(
  data.frame(Jour=rep("Jour1",10),
             Patient = c(1:10),
             Taux_substance = c(128, 88, 130, 115, 92, 80, 101, 98, 132, 85)),
  data.frame(Jour=rep("Jour2",10),
             Patient = c(1:10),
             Taux_substance = c(125, 75, 138, 108, 92, 78, 105, 97, 125, 86)),
  data.frame(Jour=rep("Jour3",10),
             Patient = c(1:10),
             Taux_substance = c(117, 73, 133, 108, 92, 74, 101, 92, 124, 83)),
  data.frame(Jour=rep("Jour4",10),
             Patient = c(1:10),
             Taux_substance = c(123, 69, 130, 102, 88, 70, 95, 93, 128, 84)),
  data.frame(Jour=rep("Jour5",10),
             Patient = c(1:10),
             Taux_substance = c(119, 70 , 127, 98, 88, 70, 95, 93, 125, 85))
)
```

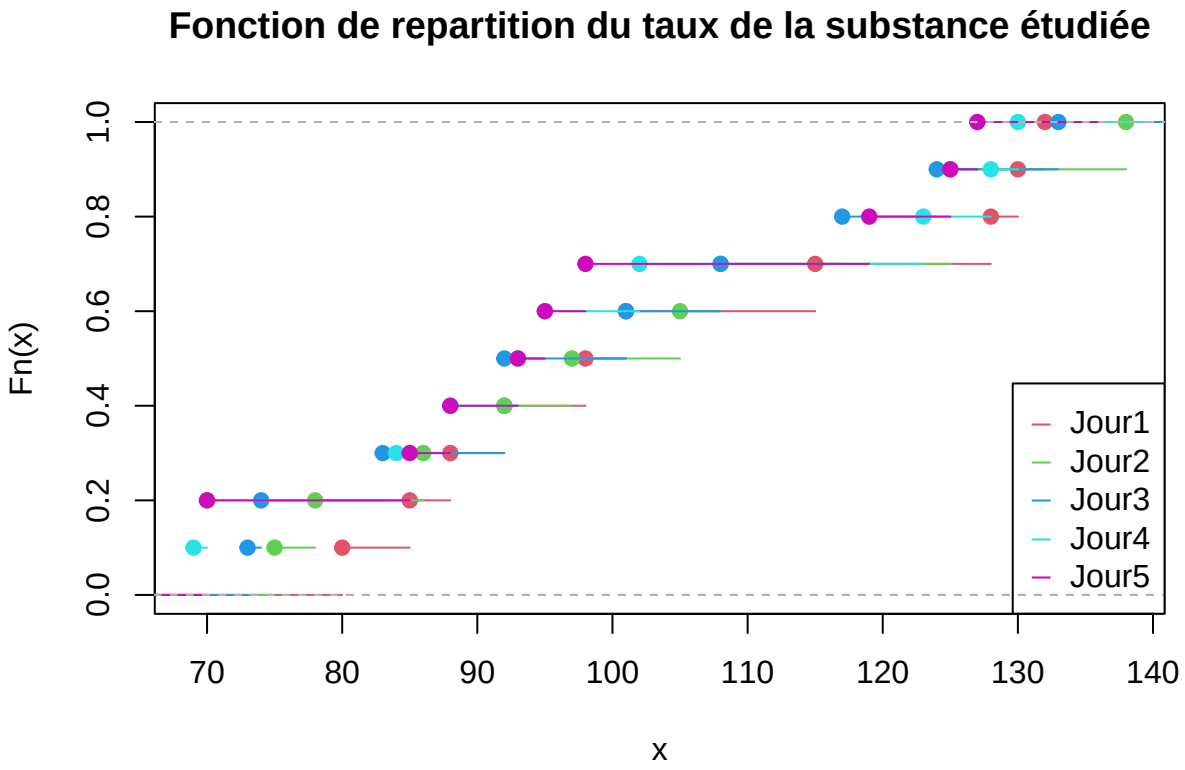
(i) Représentations graphiques

On commence par représenter les 5 boîtes à moustache correspondant aux 5 échantillons et on mettra en dessous les 5 fonctions de répartitions

```
# Boites de Tukey
boxplot(Taux_substance~Jour, data=data, main='Jour de mesure')
```




```
# Fonctions de répartitions
plot(ecdf(data[data$Jour == "Jour1","Taux_substance"]),
     col=2,
     main="Fonction de repartition du taux de la substance étudiée",
     xlim=c(min(data[,3])-0.1,max(data[,3])+0.1))
lines(ecdf(data[data$Jour == "Jour2","Taux_substance"]),col=3)
lines(ecdf(data[data$Jour == "Jour3","Taux_substance"]),col=4)
lines(ecdf(data[data$Jour == "Jour4","Taux_substance"]),col=5)
lines(ecdf(data[data$Jour == "Jour5","Taux_substance"]),col=6)
legend("bottomright",pch='_',c("Jour1","Jour2","Jour3","Jour4","Jour5"),col=2:6)
```



Les distributions semblent peu évoluer au cours des différents jours de mesure. Essayons de vérifier cela avec un test statistique adéquat.

(ii) Etude de la variation du taux au cours du temps

Nous devons regarder si le facteur temps à un impact sur le taux mesuré. Nous disposons à nouveaus d'échantillons de petites tailles et devons étudier l'imapct d'une variable qualitative sur une variable quantiative. Cependant il y a une dépendance entre les différents échantillons, on évalue les même patients sur différents jours, on parle de mesures répétées.

Nous pouvons donc effectuer, au choix, un test de Friedman ou un test de quade. Pour des soucis de complétudes, nous effectuerons les deux tests ci-dessous

Pour le test de Friedman : si on veut etudier l'effet de la variable jour, il faut la mettre dans 'groups' (lire la description du test de friedman)

```
# Test de Friedmann
friedman.test(y = data$Taux_substance, groups = data$Jour, blocks = data$Patient)
```

```
##
##  Friedman rank sum test
##
## data:  data$Taux_substance, data$Jour and data$Patient
## Friedman chi-squared = 22.31, df = 4, p-value = 0.0001738
```

```
# on peut aussi ecrire cel
friedman.test(Taux_substance~Jour|Patient,data=data)
```

```
##
##  Friedman rank sum test
##
## data:  Taux_substance and Jour and Patient
## Friedman chi-squared = 22.31, df = 4, p-value = 0.0001738
```

Ce dernier permet de rejeter l'hypothèse d'indépendance et on peut donc affirmer que le taux mesuré varie au cours du temps.

Pour le test de Quade

```
# Test de Quade
quade.test(Taux_substance~Jour|Patient,data=data)
```

```
##
##  Quade test
##
## data:  Taux_substance and Jour and Patient
## Quade F = 8.376, num df = 4, denom df = 36, p-value = 6.966e-05
```

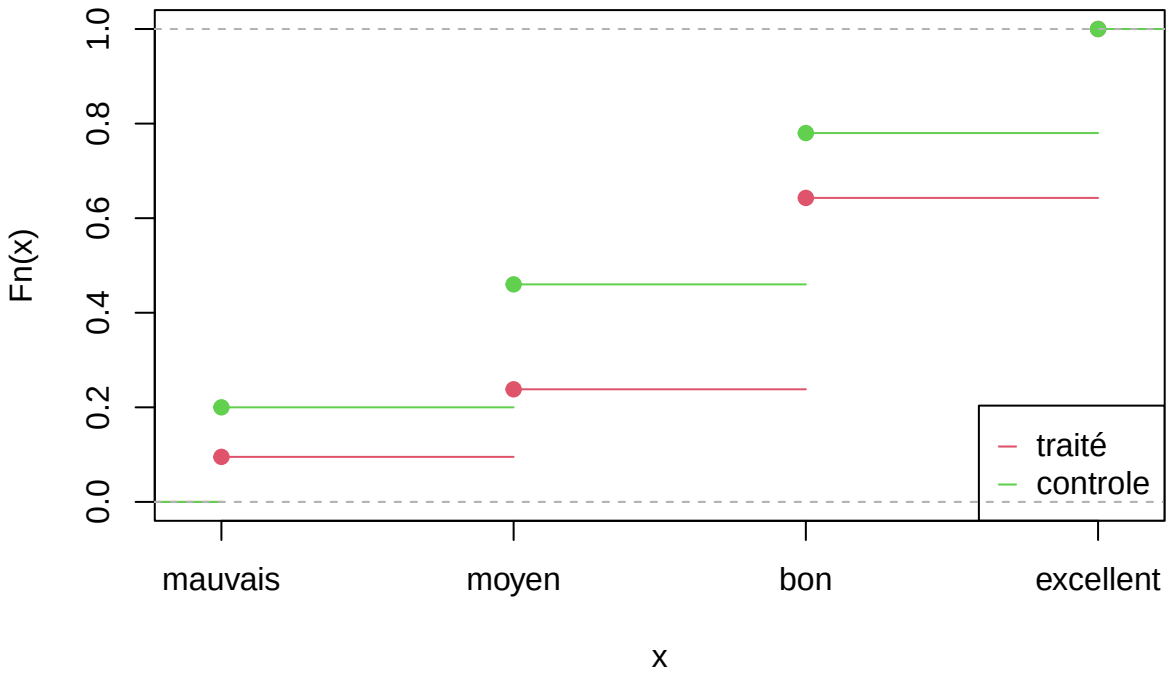
Exercice 4

On va encoder les modalités "Mauvais", "Moyen", "Bon" et "Excellent" en variable numérique afin de pouvoir effectuer une représentation graphique

```
data = rbind(
  data.frame(Groupe = rep("traite", 42),
             Res = c(rep(1, 4),rep(2, 6), rep(3, 17), rep(4, 15))),
  data.frame(Groupe = rep("controle", 50),
             Res = c(rep(1, 10),rep(2, 13), rep(3, 16), rep(4, 11)))
)
```

(i) Représentation graphique

```
# Fonctions de répartitions
plot(ecdf(data[data$Groupe == "traite","Res"]),
     col=2,
     main="Fonction de repartition du taux de la substance étudiée",
     xaxt = "n",
     xlim=c(min(data[,2])-0.1,max(data[,2])+0.1))
axis(1,at=1:4,c('mauvais', 'moyen', 'bon','excellent'))
lines(ecdf(data[data$Groupe == "controle","Res"]),col=3, xaxt= "n")
legend("bottomright",pch='_',c("traité","controle"),col=2:3)
```



Les variables étudiées ici sont ordinales (on évalue la qualité d'un traitement). On ne peut donc pas faire un test de comparaison de moyenne comme un test de Student ou encore un test de Wilcoxon (quelle serait la notion de moyenne sur de telles variables ?).

On peut en revanche effectuer un test basé sur le rang afin de comparer si les deux distributions sont identiques ou non.

Un premier test possible est le test de Wilcoxon

```
wilcox.test(data[data$Groupe == "traite","Res"], data[data$Groupe == "controle","Res"], correct = FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data[data$Groupe == "traite", "Res"] and data[data$Groupe == "controle", "Res"]
## W = 1313.5, p-value = 0.0312
## alternative hypothesis: true location shift is not equal to 0
```

Ce premier test nous permet de confirmer que les distributions diffèrent bien donc que le traitement a bien un effet (inconnu).

On peut également faire un test de Kolmogorov-Smirnov pour comparer les distributions représentées par deux échantillons

```
ks.test(data[data$Groupe == "traite","Res"], data[data$Groupe == "controle","Res"])
```

```
## Warning in ks.test(data[data$Groupe == "traite", "Res"], data[data$Groupe == :
## cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  data[data$Groupe == "traite", "Res"] and data[data$Groupe == "controle", "Res"]
## D = 0.2219, p-value = 0.211
```

```
## alternative hypothesis: two-sided
```

Ce deuxième test conduit par contre au fait que le traitement n'a eu aucun effet.

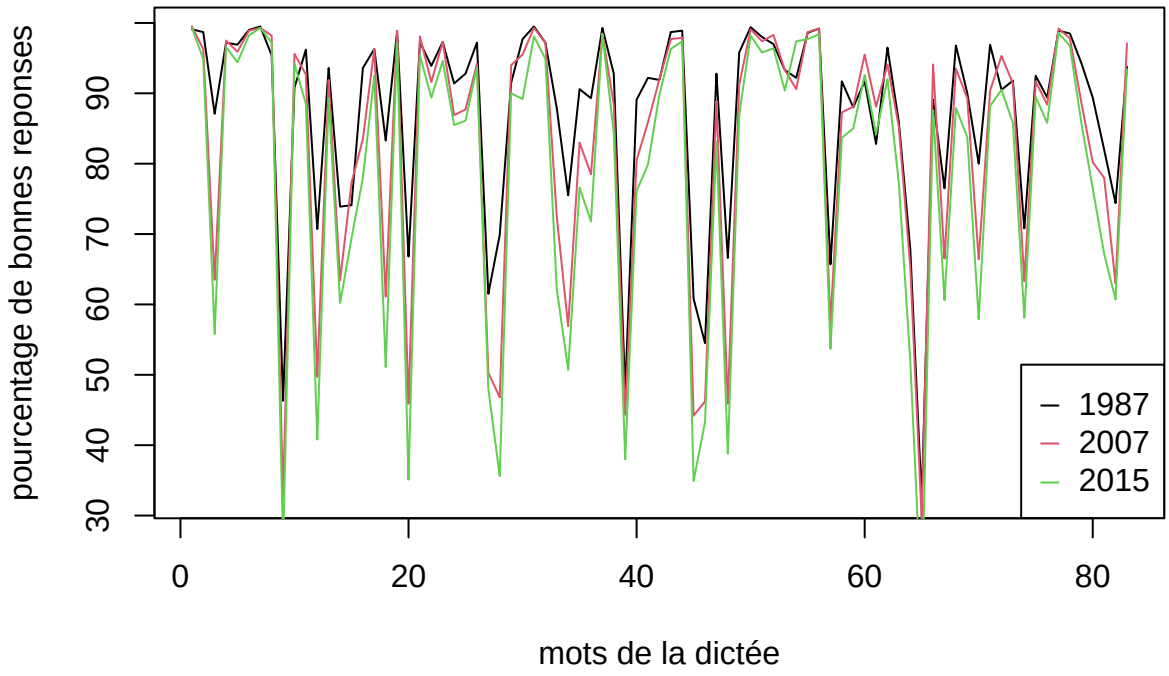
Exercice 5

```
library('xlsx')
library(readxl)
url <- "http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/dictee.xlsx"
destfile <- "dictee.xlsx"
curl::curl_download(url, destfile)
dictee <- read.xlsx("dictee.xlsx", sheetIndex =1, header = TRUE,startRow = 2,endRow = 5)
data=as.matrix(dictee[,2:84])
```

Représentation graphique des données

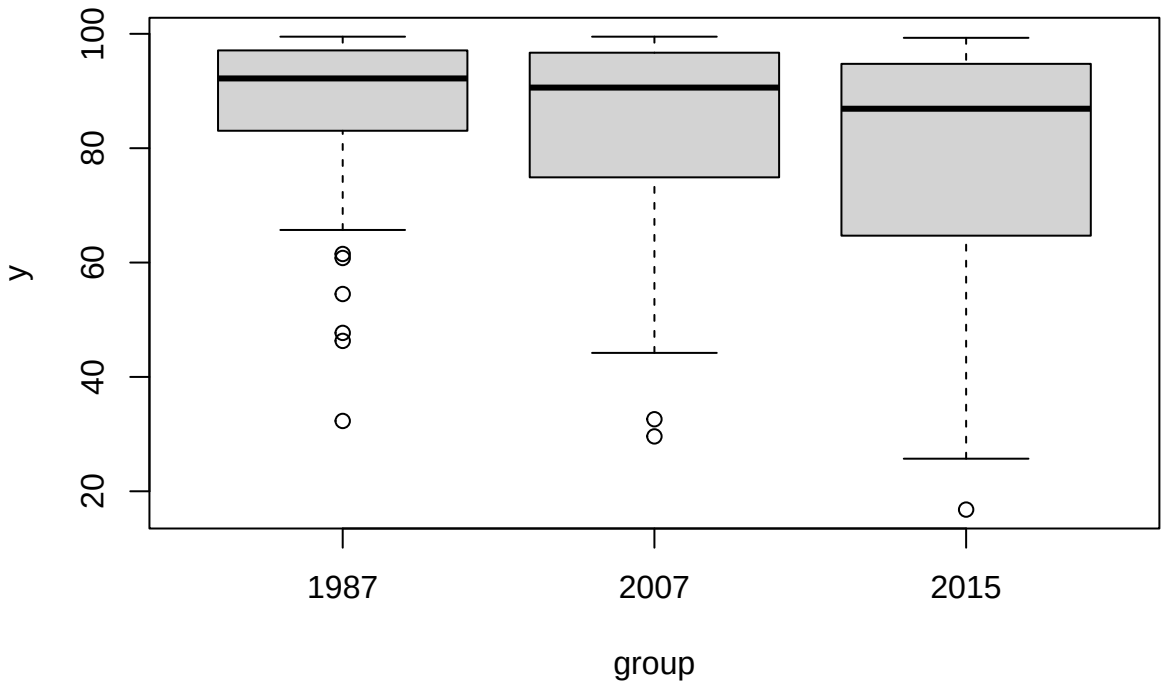
On commencer par visualiser le pourcentage de bonnes réponses selon les différentes années.

```
plot(data[1,],type='l',ylab="pourcentage de bonnes reponses",xlab="mots de la dictée")
lines(data[2,],type='l',col=2)
lines(data[3,],type='l',col=3)
legend("bottomright",legend = dictee[,1],col=1:3,pch="_")
```



Une autre représentation : pourcentage de bonnes réponses à la dictée sur les trois années de l'étude.

```
# Préparation des données
y=c(data[1,],data[2,],data[3,])
group=as.factor(c(rep(dictee[,1],each=83)))
plot(y~group)
```



Etude des données et tests

Comme pour l'exercice 3, on remarque que l'échantillon étudié est le même au cours de différentes périodes. La dictée et les mots utilisés sont identiques au cours des trois années d'études. On ne peut donc pas effectuer un test de Wilcoxon, mais on pourra à nouveau faire un test de Quade ou un test de Friedman.

```
# Test de Friedman
friedman.test(y=y,groups=group,blocks=rep(1:83,3))
```

```
##
##  Friedman rank sum test
##
## data:  y, group and rep(1:83, 3)
## Friedman chi-squared = 112.27, df = 2, p-value < 2.2e-16
```

Ce dernier tend à rejeter l'hypothèse d'indépendance et on remarque que l'année a bien un impact sur les résultats observés.

Exercice 6 : Etude de deux populations (échantillons) indépendantes

Cet exercice se traite de la même façon que l'exercice 2 du TD3

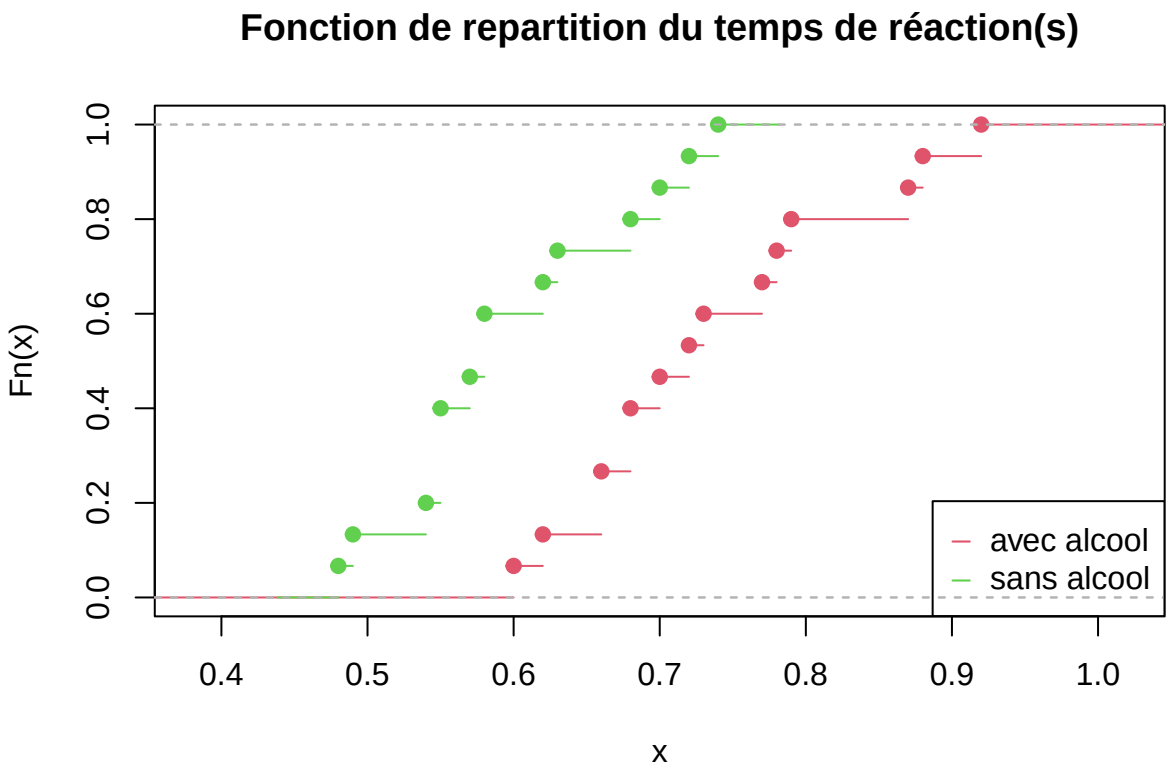
Dans la cas présent, les échantillons sont trop petits pour qu'il soit raisonnable de faire l'hypothèse gaussienne sur l'estimateur de la moyenne $\bar{X}$. Pour tester si l'alcool a une influence sur le temps de réaction on va donc faire un test de Wilcoxon-Mann-Whitney. On pourra aussi regarder si une approche paramétrique est viable, à condition que les hypothèses soient satisfaites.

On se concentrera uniquement sur sa résolution à l'aide de R

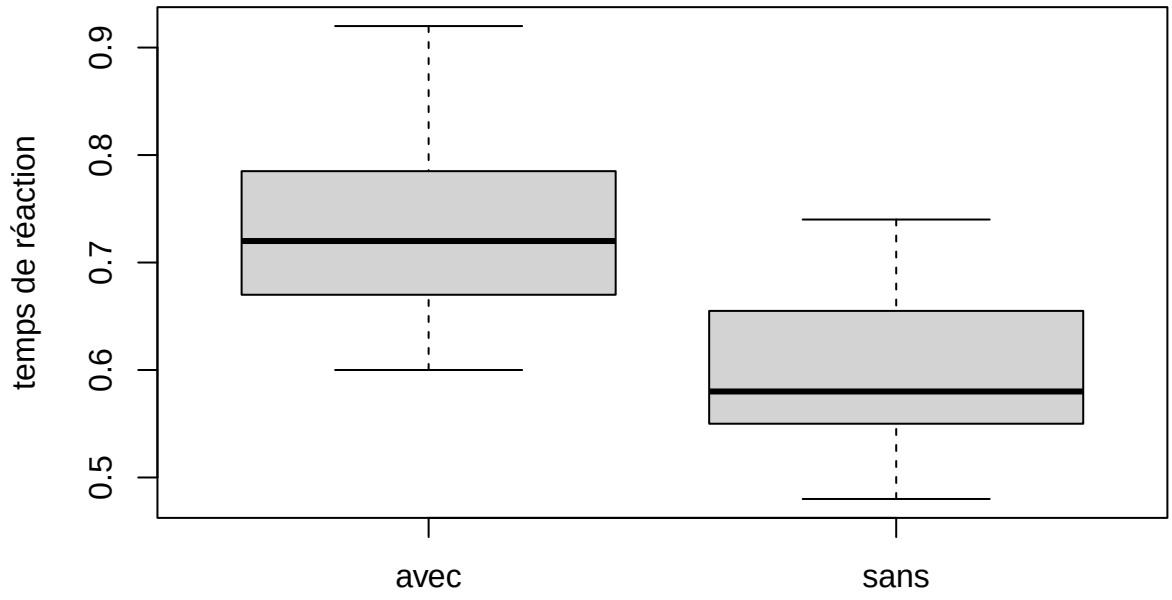
```
# Valeurs de notre échantillon avec et sans prise d'alcool
sans <- c(58,54,58,72,48,70,62,55,74,63,55,49,68,57,55)/100
avec <- c(73,62,66,92,68,87,77,70,88,79,72,60,78,66,68)/100
```

(i) Un complément graphique pour étudier les deux distributions

```
# Représentation des fonctions de répartition
plot(ecdf(avec),col=2,main="Fonction de repartition du temps de réaction(s)",
     xlim=c(min(avec,sans)-0.1,max(avec,sans)+0.1))
lines(ecdf(sans),col=3)
legend("bottomright",pch='_',c('avec alcool','sans alcool'),col=2:3)
```




```
# Boxplot associé aux deux distributions
boxplot(avec,sans,names=c('avec','sans'),ylab='temps de réaction')
```



(ii) Tests Statistiques

Une approche non-paramétrique : test de Wilcoxon Les échantillons sont indépendants, on peut donc faire un test de Wilcoxon pour tester si l'alcool augmente bien le temps de réaction au volant (alternative = "greater", i.e. test unilatéral supérieur)

```
wilcox.test(avec,sans, alternative = "greater")
```

```
## Warning in wilcox.test.default(avec, sans, alternative = "greater"): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  avec and sans
## W = 193.5, p-value = 0.0004132
## alternative hypothesis: true location shift is greater than 0
```

Ce test nous montre bien que l'alcool augmente le temps de réaction au volant.

Une approche paramétrique Même si les échantillons sont de tailles < 30, il est tout fait possible de tenter une approche paramétrique basée sur le test de Student, si nos données sont issues d'une distribution gaussienne !

```
# Vérifions que nos échantillons sont gaussiens

p_sans <- shapiro.test(sans)$p.value
p_avec <- shapiro.test(avec)$p.value

if ((p_sans > 0.05)&(p_avec > 0.05)){
  print("Les échantillons sont gaussiens, ok pour l'approche paramétrique")
} else {
  print("Au moins l'une des p_value est inférieure à < 0.05,
        on ne peut donc pas envisager une approche paramétrique")
}
```

```
## [1] "Les échantillons sont gaussiens, ok pour l'approche paramétrique"
```

Pour déterminer le type de test à utiliser, on va commencer étudier les variances de nos deux échantillons :

- si les variances sont égales : on effectuera un test de Student à $n_1 + n_2 - 2$ degrés de libertés
- si les variances ne sont pas égales : on effectue un test de Student basé sur l'approximation d'Aspin-Welch

```
# Comparaison des variances à l'aide d'un test de Fisher
p_fisher <- var.test(avec,sans)$p.value
if (p_fisher > 0.05){
  print("Les variances sont égales")
} else {
  print("Les variances ne sont pas égales")
}
```

```
## [1] "Les variances sont égales"
```

On va donc effectuer un test de Student à $n_1 + n_2 - 2$ degrés de liberté pour comparer les moyennes deux échantillons

```
# Test de Student, en précisant que les moyennes sont égales
t.test(avec,sans,var.equal = T,alternative = "greater")
```

```
##
```

```
##  Two Sample t-test
##
## data:  avec and sans
## t = 4.2758, df = 28, p-value = 9.989e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.08349791        Inf
## sample estimates:
## mean of x mean of y
## 0.7373333 0.5986667
```

On rejette donc l'hypothèse H_0 et on aboutit à la même conclusion que précédemment.

Exercice 7 : Etude de deux populations (échantillons) appariées

On souhaite savoir si un régime a un effet sur la masse des individus. On a ici pesé 10 individus avant et après avoir effectué le régime, il y a donc une dépendance entre les valeurs prises par l'échantillon. On ne peut donc pas effectuer un "test de Wilcoxon-Mann-Whitney".

Nous avons à faire à des échantillons appariés. Il va donc falloir s'intéresser aux valeurs de la variable aléatoire D traduisant la différence de taux de cholestérol avant et après la prise de médicament et faire le test d'hypothèses suivant :

- H_0 :le médicament n'a un effet sur le taux de cholestérol des individus,
- H_1 :le médicament a un d'effet sur le taux de cholestérol des individus.

Remarque : nous pourrions aussi formuler H_1 comme : le médicament permet de réduire le taux de cholestérol des individus, ce qui nous conduirait à effectuer un test unilatéral "supérieur" –> le taux de cholestérol est plus élevé avant la prise du médicament, ce que l'on va faire dans la suite.

Trois possiblités s'offrent à nous :

- • un test paramétrique de Student
- • un test non paramétrique : test du signe
- • un test non paramétrique : test du rang signé de Wilcoxon

On s'intéressera donc à l'échantillon différence

```
avant <- c(0.1, 0.2, 0.15, 0.3, 0.34, 0.16, 0.09, 0.24, 0.17, 0.29)
apres <- c(0.8, 0.18, 0.12, 0.2, 0.3, 0.21, 0.12, 0.16, 0.17, 0.22)
D_ech <- avant - apres
```

Un test paramétrique ?

Nous pourrions procéder à un "test de Student" dans le cas où nos données sont normalement distribuées. Malheureusement le test de Shapiro ci-dessous, nous montre que cette hypothèse n'est pas valable.

```
# Test de Shapiro
shapiro.test(D_ech)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  D_ech
## W = 0.56057, p-value = 1.881e-05
```

Nous devons donc forcément utiliser un test non paramétrique : “test du signe” ou “test du rang signé de wilcoxon”. Ces deux tests nécéssitent d’abord de supprimer les valeurs nulles dans notre échantillon, d’où, ce qui nous donne l’échantillon suivant

Test du signe On se concentre ici sur le nombre de différences positives, ici au nombre de 6. Sous l’hypothèse H_0 cette statistique de test suit une loi binomiale $\mathcal{B}(n, 1/2)$.

On réfère ensuite aux tables du test du signe disponibles dans le manuel (pour $n = 9$ et $C = 6$), et on trouve une p-value de 0.5078 ce qui ne permet pas de rejeter l’hypothèse H_0, on ne peut donc pas conclure que le médicament a un effet sur les participants.

```
# Vérification sous R

library(BSDA)
SIGN.test(D_ech, alternative = "greater")
```

```
##
##  One-sample Sign-Test
##
## data:  D_ech
## s = 6, p-value = 0.2539
## alternative hypothesis: true median is greater than 0
## 95 percent confidence interval:
##  -0.03213333         Inf
## sample estimates:
## median of x
##       0.025
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.9453 -0.0300    Inf
## Interpolated CI       0.9500 -0.0321    Inf
## Upper Achieved CI     0.9893 -0.0500    Inf
```

Ce premier test ne prend en compte que le signe de la différence mais sans tenir compte de l’importance relative de chaque différence.

Test du rang signé de Wilcoxon Notre échantillon étant de taille faible (<10) on ne peut utiliser un test basée sur une approximation gaussienne.

```
# Test du rang signé, on précise que les deux échantillons sont appariés
# et que l'on ne souhaite pas effectuer de correction
wilcox.test(avant,apres, paired = TRUE, correct= FALSE,alternative = "greater")
```

```
## Warning in wilcox.test.default(avant, apres, paired = TRUE, correct = FALSE, :
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(avant, apres, paired = TRUE, correct = FALSE, :
## cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test
##
## data:  avant and apres
## V = 28.5, p-value = 0.2384
## alternative hypothesis: true location shift is greater than 0
```

Le test ne conduit toujours pas au rejet de l'hypothèse H_0.