

# Les tests statistiques

Thibault LAURENT

1er Décembre 2014

Ce document a été généré directement depuis RStudio en utilisant l'outil Markdown. La version .pdf se trouve [ici](#).

## Résumé

L'objet de ce chapitre est de vous présenter les moyens offerts par R pour réaliser quelques tests statistiques parmi les plus répandus. Les bibliothèques installées par défaut fournissent à l'utilisateur la possibilité de réaliser une gamme assez large de tests. Sauf indication contraire, les fonctions que nous aborderons dans cette partie proviendront des bibliothèques de base. Le but de cette partie n'est pas de présenter les tests statistiques, mais de décrire le moyen de les mettre en oeuvre dans R. Pour une description approfondie de ces tests, nous renvoyons le lecteur sur ce [document](#) écrit par les auteurs du site [Wikistat](#). Le livre de Saporta : "Probabilités, analyse des données et statistique" est également une référence incontournable en ce qui concerne les tests statistiques.

**Rappel** : on parlera parfois dans ce chapitre de test paramétrique et de test non paramétrique. On rappelle ici les définitions :

- **Test paramétrique** : les hypothèses nulle et alternative du test portent sur un paramètre statistique (moyenne ou variance par exemple). Ces tests nécessitent généralement des conditions de validité (distribution normale des données par exemple).
- **Test non paramétrique** : un test non paramétrique porte globalement sur la répartition des données sans hypothèse sur leur distribution (*free distribution* en anglais)

## 1. Problèmes à un échantillon

### 1.1 Test de normalité d'une variable

Les 2 tests "classiques" de normalité d'une variable sont le test de Kolmogorov-Smirnov et le test de Shapiro-Wilk, tous les deux implémentés dans R par le biais des fonctions `ks.test()` et `shapiro.test()`.

#### Test de Kolmogorov-Smirnov : `ks.test()`

Cette fonction est un peu plus générale que `shapiro.test()`. En effet, son principe est de comparer 2 distributions. En cela, `ks.test()` permet de réaliser des tests bidimensionnels. Pour le cas unidimensionnel, il suffit juste de comparer la distribution qui nous intéresse à la distribution théorique d'une loi normale de paramètres estimés  $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  et  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ .

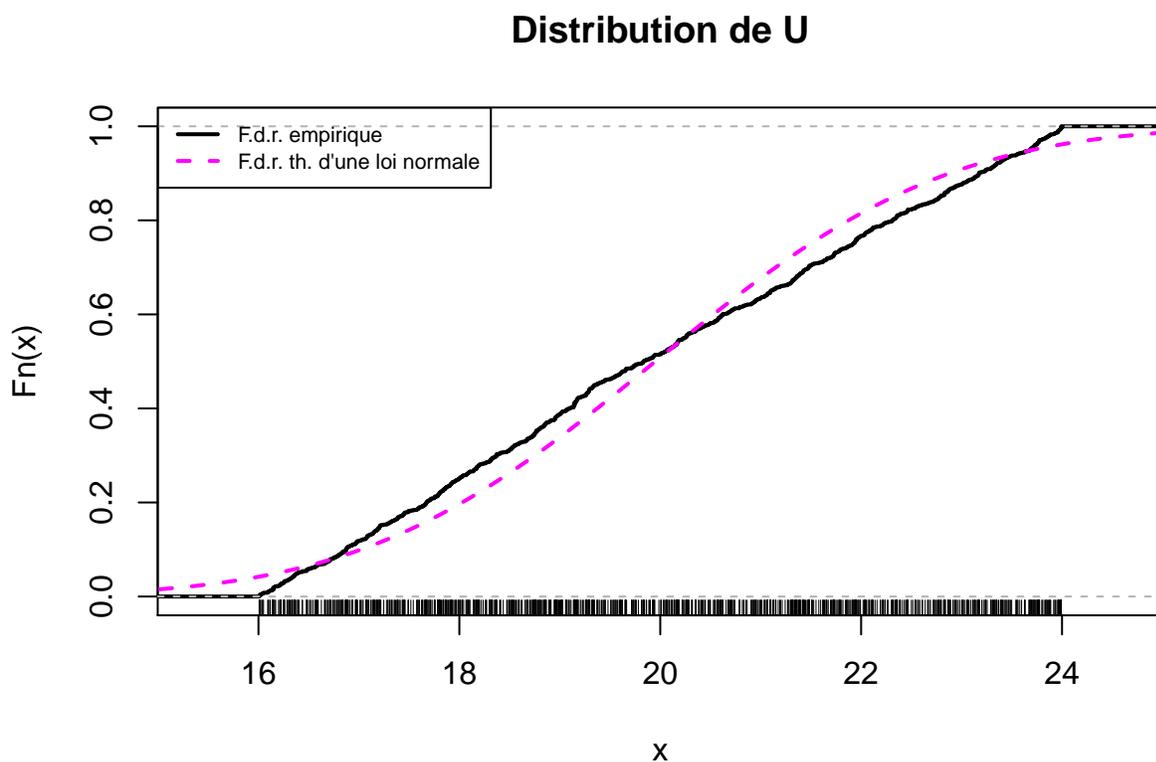
Le résultat de `ks.test()` indique également dans la composante alternative, le type de test (unilatéral ou bilatéral) effectué.

**Exemple** : pour illustrer l'utilisation de cette fonction, nous allons nous intéresser à la variables  $U$  issue d'une  $\mathcal{U}(16, 24)$ .

```
u<-runif(1000,16,24)
```

Le principe du test de Kolmogorov-Smirnov est de comparer 2 distributions au moyen de la fonction de répartition. Dans ce cas, on va comparer la fonction de répartition empirique  $F_n$  avec la fonction de répartition théorique  $F_{th}$  d'une loi normale dont les paramètres sont estimés à partir de l'échantillon. On représente graphiquement les deux fonctions de répartitions :

```
# f.d.r. empirique de U
fdr.u=ecdf(u)
test=plot(fdr.u, main="Distribution de U", lwd=2)
x=seq(10,30,0.1)
# estimation de la f.d.r théorique d'une gaussienne
# de paramètre bar(u) et hat(sigma[u])
lines(x, pnorm(x, mean(u), sd(u)), lty=2, lwd=2, col="magenta")
# représentation des valeurs de U
rug(u)
# légende
legend("topleft",c("F.d.r. empirique","F.d.r. th. d'une loi normale"),lty=1:2,
      lwd=2,col=c("black", "magenta"),cex=0.7)
```



A l'évidence, les deux courbes ne sont pas identiques. De plus, on constate que l'écart entre  $F_n$  et  $F_{th}$  peut être parfois très grand. Le test de Kolmogorov-Smirnov consiste à donner comme statistique de test, l'écart le plus important entre les deux courbes.

```
(res.ks=ks.test(u, "pnorm", mean(u), sd(u)))
```

```
##  
## One-sample Kolmogorov-Smirnov test  
##  
## data: u  
## D = 0.0591, p-value = 0.001854  
## alternative hypothesis: two-sided
```

**Interprétation** : si l'hypothèse nulle est vraie (ici  $H_0 : \{F_n = F_{th}\}$ ), la probabilité de voir la statistique de test ( $D = 0.0590918$ ) autant éloignée de la valeur 0 est inférieure à 0.0019, soit une faible probabilité. Autrement dit, la valeur de D est anormalement grande et donc on ne peut pas accepter l'hypothèse nulle. Ceci confirme donc que la loi de probabilité de  $U$  n'est pas celle d'une loi normale.

### Test de Shapiro-Wilk : *shapiro.test()*

L'utilisation de *shapiro.test()* ne requiert quant à elle que le vecteur de valeurs numériques sur lequel on va tester l'hypothèse de normalité. Ici, l'hypothèse nulle  $H_0$  est {l'échantillon est distribuée selon une loi gaussienne}.

**Exemple** : pour illustrer l'utilisation de cette fonction, nous allons nous intéresser à la variables  $V$  issue d'une  $\mathcal{N}(20, 4)$ .

```
v<-rnorm(1000,20,2)
```

Le résultat du test statistique est :

```
(res.sh=shapiro.test(v))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: v  
## W = 0.999, p-value = 0.8877
```

**Interprétation** : si la valeur-p est inférieure à un seuil  $\alpha$  qu'on s'est donné (en général, 5%), alors on rejette  $H_0$ . Ici, la valeur-p est égale à 0.8876675, ce qui nous conduit (bien entendu) à accepter  $H_0$ .

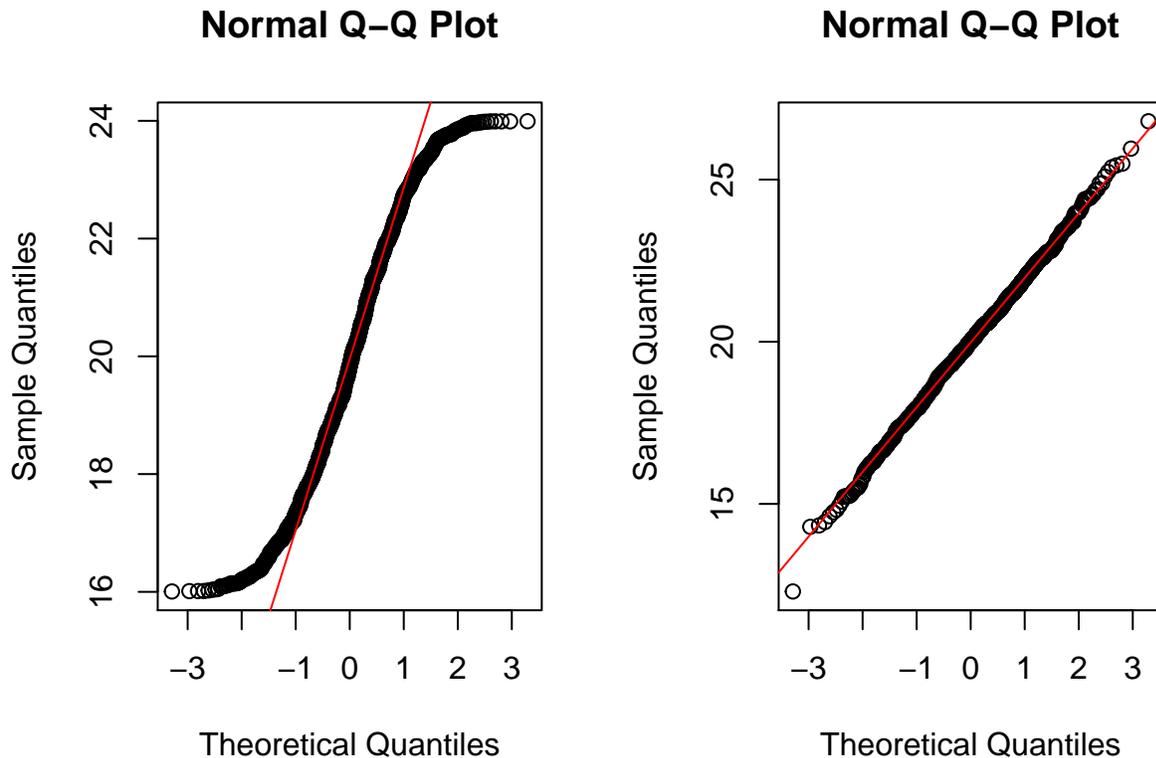
### Le QQ (Quantile-Quantile) plot "gaussien"

Le principe du QQ-plot "général" est de comparer la position de certains quantiles dans la population observée avec leur position dans la population théorique.

Le QQ-plot gaussien est un outil graphique qui permet d'apprécier visuellement la normalité d'une variable quantitative. La fonction à utiliser est la fonction *qqnorm()*.

**Exemple** sur les variables  $U$  et  $V$  :

```
op<-par(mfrow=c(1,2))  
qqnorm(u)  
qqline(u,col="red")  
qqnorm(v)  
qqline(v,col="red")
```



```
par(op)
```

**Interprétation statistique** : si les points sont alignés autour de la droite tracée par la fonction `qqline()`, c'est que la distribution de la variable étudiée est celle d'une loi normale.

**Remarque** : pour mieux comprendre ce que représente la fonction `qqnorm()`, voici une autre façon d'obtenir le même graphique.

```
plot(qnorm(ppoints(length(v)),0,1), sort(v))
abline(lm(quantile(v,c(0.25,0.75))-qnorm(c(0.25,0.75))),col="blue")
```

## 1.2 Test de comparaison à une moyenne

Pour comparer la moyenne d'un échantillon à une valeur fixée a priori, il existe plusieurs méthodes, selon que l'on puisse supposer ou non la normalité de l'échantillon en notre possession. Si une telle supposition peut être faite, on utilise alors le test de Student, effectué par la fonction `t.test()`. Dans le cas contraire, on a recours au test non paramétrique des rangs signés de Wilcoxon (*Wilcoxon signed-rank test* en anglais), effectué par la fonction `wilcox.test()`.

### a. Le test de Student

Ici, il s'agit du cas où on sait que la distribution de la variable  $X$  est normale, de moyenne  $\mu$  inconnue et de variance  $\sigma^2$  inconnue. Le test de Student est mis en oeuvre dans R par la fonction `t.test()`. Prenons ici un échantillon  $U$  de taille 100, issu d'une loi normale de moyenne 30 et de variance 25 :

```
u<-rnorm(100,30,5)
```

Supposons que vous vous retrouviez avec cette échantillon sans en connaître les paramètres (mais en sachant que la distribution est normale), et que vous souhaitiez vérifier si la moyenne est différente de 32. En précisant le vecteur contenant l'échantillon et la moyenne à comparer (argument **mu**), vous obtiendrez le résultat du test par la commande :

```
(res.ttest=t.test(u, mu=32))
```

```
##  
## One Sample t-test  
##  
## data: u  
## t = -4.3366, df = 99, p-value = 3.486e-05  
## alternative hypothesis: true mean is not equal to 32  
## 95 percent confidence interval:  
## 28.69260 30.76911  
## sample estimates:  
## mean of x  
## 29.73085
```

Sous  $H_0 : \{\mu = 32\}$ , la probabilité qu'une observation soit supérieure à la valeur de la statistique de test en valeur absolue  $|t| = 4.3365768$  (il s'agit d'un test bilatéral par défaut) est inférieure à 0.0000349, ce qui est très faible. Autrement dit, on ne peut pas accepter l'hypothèse nulle : la valeur de  $\mu$  n'est pas égale à 32. A noter qu'il est possible de rendre le test unilatéral en précisant pour l'argument **alternative**, le paramètre **less** ou **greater**.

Parmi les autres résultats qui sont affichés par la fonction, on trouve la valeur de la moyenne calculée sur l'échantillon ( $\bar{x} = 29.7308541$ ) ainsi que l'intervalle de confiance associé. En d'autres termes, si on avait choisi une valeur de  $\mu$  comprise dans cet intervalle de confiance, on n'aurait pas pu rejeter l'hypothèse nulle.

**Remarque** : dans un test statistique, il est absolument essentiel de repérer quelle est l'hypothèse nulle et l'hypothèse alternative. Ici, on peut lire **alternative hypothesis: true mean is not equal to 32**. Autrement dit, on construit bien la statistique de test  $t$  sous l'hypothèse nulle que la moyenne de l'échantillon est égale à 32.

## b. Le test de Wilcoxon

Pour ce test, on se place dans le cas d'un échantillon de faible taille et/ou dont on ne connaît pas la distribution. On sait seulement qu'elle est symétrique. Prenons par exemple un échantillon  $S$  issu d'une loi de Student à 5 degrés de liberté :

```
s<-rt(120,5)  
mean(s)
```

```
## [1] 0.1000655
```

La moyenne calculée sur cet échantillon est proche de 0. On veut donc savoir si elle l'est significativement.

```
(res.wilco=wilcox.test(s, mu=0))
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data:  s
## V = 4053, p-value = 0.2685
## alternative hypothesis: true location is not equal to 0
```

Sous  $H_0 : \{\mu = 0\}$ , la probabilité qu'une observation soit supérieure à la valeur de la statistique de test  $V = 4053$  est 0.2685222, ce qui n'est donc pas un fait rare. On ne peut rejeter l'hypothèse nulle et on admet donc que  $S$  est issu d'une distribution dont l'espérance vaut 0.

**Remarque :** pour retrouver la statistique de test  $V$ , on effectue le calcul suivant.

```
rank=1:length(s)
min(sum(rank[s[order(abs(s))>0]]),
    sum(rank[s[order(abs(s))<0]]))
```

```
## [1] 3207
```

### 1.3 Test de comparaison à une proportion

Lors d'un sondage réalisé auprès de 1000 individus, 521 d'entre eux ont affirmé apprécié la marque X. L'enseigne souhaiterait vérifier que plus de la moitié de la population adhère à la marque. Pour cela, une façon de faire est de tester l'hypothèse  $H_0 : \pi < 0.5$  (on ne prend pas  $H_0 : \pi > 0.5$ , car on formule  $H_0$  de façon à ce qu'elle soit rejetée). Ceci s'effectue au moyen de la fonction `prop.test()`. On précise ici l'argument `alternative="greater"` (qui spécifie donc bien l'hypothèse alternative  $H_1 : \pi > 0.5$ ).

```
(res.prop=prop.test(521,1000, alternative="greater"))
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  521 out of 1000, null probability 0.5
## X-squared = 1.681, df = 1, p-value = 0.0974
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.4944942 1.0000000
## sample estimates:
##      p
## 0.521
```

Sous  $H_0$ , la valeur de la statistique de test n'est pas "anormale" (1.681) ce qui est validé par la valeur-p (0.0973958). On ne peut donc pas rejeter l'hypothèse nulle et l'enseigne devra encore revoir sa communication pour améliorer l'image de sa marque.

## 2. Problèmes à deux ou plusieurs échantillons

Lorsque l'on dispose de deux ou plusieurs échantillons, la question se pose de savoir s'ils proviennent de la même population. On verra qu'avant d'effectuer un test, il faudra déterminer si les 2 échantillons sont indépendants ou appariés, dont on rappelle les deux définitions suivantes :

- **Données indépendantes** : les observations sont indépendantes à l'intérieur de chaque échantillon et d'un échantillon à l'autre. Exemple : résultats scolaires filles et garçons, dosage d'un produit chez 2 groupes de patients ayant reçu une molécule ou un placebo.
- **Données appariées** : les mêmes individus sont soumis à 2 mesures successives d'une même variable. Exemple : notes de copies soumises à une double correction, dosage d'un produit avant et après un traitement chez les mêmes individus.

Comme pour le paragraphe précédent, les tests ne seront pas les mêmes suivant que l'on ait ou non pu faire l'hypothèse de la normalité des 2 distributions. Dans le premier cas, le test se décompose en deux phases en comparant successivement la variance et la moyenne.

## 2.1 Tests de comparaison de variance

### a. Test de Fisher (sous hypothèse de normalité)

La comparaison des variances de deux échantillons est réalisée par le test de Fisher. à l'aide de la fonction `var.test()`. Simulons 2 échantillons  $U_1$  et  $U_2$  issus respectivement d'une  $\mathcal{N}(\mu_1 = 20, \sigma_1^2 = 25)$  et d'une  $\mathcal{N}(\mu_2 = 25, \sigma_2^2 = 25)$ .

```
u1<-rnorm(150,20,5)
u2<-rnorm(120,25,5)
```

Dans ces conditions, on doit s'attendre à ne pas rejeter l'hypothèse  $H_0 : \{\frac{\sigma_1}{\sigma_2} = 1\}$  d'égalité des variances.

```
(res.var=var.test(u1, u2))
```

```
##
## F test to compare two variances
##
## data:  u1 and u2
## F = 1.0147, num df = 149, denom df = 119, p-value = 0.9383
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7182408 1.4241362
## sample estimates:
## ratio of variances
##          1.014734
```

**Interprétation** : sous  $H_0$ , la statistique de test  $F = 1.0147339$  n'est ni trop grande, ni trop petite, la valeur-p n'étant pas un événement rare ( $p - value = 0.9382665$ ). On ne peut donc pas rejeter l'hypothèse d'égalité des variances.

### b. Test de Barlett (sous hypothèse de normalité)

Le test de comparaison de variance de Barlett rend possible, lui, la comparaison de plusieurs échantillons de taille différentes, tous issus d'une loi normale. Simulons ici 3 échantillons  $U_1$ ,  $U_2$  et  $U_3$  issus respectivement de lois  $\mathcal{N}(\mu_1 = 30, \sigma_1^2 = 81)$ ,  $\mathcal{N}(\mu_2 = 30, \sigma_2^2 = 25)$  et  $\mathcal{N}(\mu_3 = 30, \sigma_3^2 = 16)$ .

```
u1<-rnorm(100,30,9)
u2<-rnorm(80,30,5)
u3<-rnorm(110,30,4)
```

De par la définition des 3 variances (81, 25 et 16), le test devrait conduire à rejeter l'hypothèse nulle d'égalité des variances. La fonction `bartlett.test()`, qui réalise ce test admet deux arguments principaux : un vecteur numérique contenant toutes les observations (ici la concaténation de  $U_1$ ,  $U_2$  et  $U_3$  de taille  $n = 290$ ) et un objet de type **factor** (de taille  $n = 290$ ) indiquant l'échantillon d'appartenance.

```
x<-c(u1,u2,u3)
groupe<-as.factor(paste("G",c(rep(1,100),rep(2,80),rep(3,110)),sep=""))
(res.bart=bartlett.test(x,groupe))

##
## Bartlett test of homogeneity of variances
##
## data:  x and groupe
## Bartlett's K-squared = 78.6216, df = 2, p-value < 2.2e-16
```

**Interprétation:** l'hypthèse alternative n'est pas rappelée dans la sortie. On regarde dans l'aide de la fonction et on peut lire : "Performs Bartlett's test of the null that the variances in each of the groups (samples) are the same". L'hypothèse nulle est donc ici  $H_0 : \{\sigma_1 = \sigma_2 = \sigma_3\}$ .

Sous  $H_0$ , la statistique de test est anormalement grande  $K = 78.6215525$ , la valeur-p étant très faible. On rejette donc bien l'hypothèse nulle : toutes les variances ne sont pas égales.

## 2.2 Tests de comparaison des moyennes

### a. Echantillons supposés de lois gaussiennes

La principale fonction permettant de réaliser un test paramétrique de comparaison de moyennes d'échantillons issus de lois normales est `oneway.test()` (basé sur la statistique de test de Fisher). La fonction `t.test()` (basée sur la statistique de test de Student) le permet également, mais est restreinte à la comparaison de seulement 2 échantillons. Toutefois, la fonction `t.test()` permet de mentionner le fait que les échantillons sont appariés ou non.

**Exemple :** on a mesuré la hauteur (en mètres) de 12 arbres selon deux méthodes différentes, avant et après la coupe de l'arbre. Ces données sont extraites de D. Chessel et A.B. Dufour (Biométrie et Biologie Evolutive, Université de Lyon 1), [Pratique des tests élémentaires](#).

```
debout = c(20.4,25.4,25.6,25.6,26.6,28.6,28.7,29.0,29.8,30.5,30.9,31.1)
abattu = c(21.7,26.3,26.8,28.1,26.2,27.3,29.5,32.0,30.9,32.3,32.3,31.7)
arbres = c(debout, abattu)
groupe<-as.factor(c(rep("debout",12),rep("abbatu",12)))
```

Dans un premier temps, on traite les deux échantillons comme s'ils étaient **indépendants**, de lois normales et on vérifie que le test d'égalité des variances ne peut être rejeté :

```
var.test(debout,abattu)

##
## F test to compare two variances
##
## data:  debout and abattu
## F = 0.8982, num df = 11, denom df = 11, p-value = 0.8619
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
## 0.2585696 3.1200517
## sample estimates:
## ratio of variances
## 0.8981929
```

Une fois ceci vérifié, on peut ensuite faire le test de comparaison des moyennes en utilisant une des fonctions `oneway.test()` ou `t.test()` et en précisant que les variances des échantillons sont les mêmes (option `var.equal=TRUE`)

```
(res.F=oneway.test(arbres~groupe, var.equal=TRUE))
```

```
##
## One-way analysis of means
##
## data: arbres and groupe
## F = 0.6799, num df = 1, denom df = 22, p-value = 0.4185
```

Sous  $H_0 : \{\mu_1 = \mu_2\}$ , la statistique de test de Fisher  $F = 0.6799435$  n'est pas anormalement grande, la valeur-p ne reflétant pas un événement rare ( $p - value = 0.4184573$ ). On aurait pu essayer de faire le test de Student avec la fonction `t.test()` qui nous aurait conduit à la même conclusion. La différence est que la statistique de test de Student est égal à  $t = \sqrt{F} = 0.8245869$ .

A présent, nous allons considérer les deux échantillons comme étant **appariés**, ce qui est effectivement le cas puisqu'il s'agit des mêmes individus (les arbres) qui ont été mesurés selon deux façons différentes. On va donc utiliser la fonction `t.test()`, celle-ci nous permettant d'utiliser l'option `paired = TRUE`. On garde l'hypothèse d'égalité des variances.

```
(res.ttest=t.test(arbres~groupe, paired=TRUE, var.equal=TRUE))
```

```
##
## Paired t-test
##
## data: arbres by groupe
## t = 3.2343, df = 11, p-value = 0.007954
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.3434464 1.8065536
## sample estimates:
## mean of the differences
## 1.075
```

**Interprétation** : le constat est sans appel ! L'hypothèse  $H_0$  ne peut être acceptée (statistique de test anormalement élevée et valeur-p inférieure au seuil de 5% en général utilisé comme valeur de référence), dans le cas où les échantillons sont considérés comme étant appariés. On peut donc affirmer qu'on commet des erreurs lorsqu'on mesure les arbres debouts.

## b. Echantillons de lois quelconques

Si aucune hypothèse n'a pu être faite sur la distribution des variables, nous pouvons avoir recours à un test non paramétrique : le test de Kruskal-Wallis ou celui de Mann-Whitney.

**Test non paramétrique de Mann-Whitney** Le test de Mann-Whitney est utilisée lorsque l'on ne peut faire l'hypothèse de normalité pour les distributions des deux échantillons. Dans la littérature anglo-saxonne, il est connu sous le nom de *Wilcoxon rank-sum test*. Il est réalisé dans R par la fonction `wilcox.test()`. Il s'agit de la même fonction utilisée pour le test de comparaison à une moyenne. En fait, selon les arguments qu'on lui donne, cette dernière n'effectue pas les mêmes tests. Dans le cas qui nous intéresse ici, en indiquant le nom des deux vecteurs, nous obtiendrons le résultat de la comparaison entre la différence des moyennes et l'argument `mu`, égal à 0 par défaut.

**Exemple** : la concentration d'un produit est mesurée sur 2 échantillons indépendants de tailles respectives  $n_1 = 5$  et  $n_2 = 6$ . Il est à noter que lorsque les échantillons sont de petites tailles (ce qui n'est pas rare, notamment dans les expériences biologiques), les test non paramétriques sont à privilégier. Voici les mesures :

```
ech.1=c(1.31, 1.46, 1.85, 1.58, 1.64)
ech.2=c(1.49, 1.32, 2.01, 1.59, 1.76, 1.86)
x=c(ech.1, ech.2)
groupe=factor(c(rep("G1",5), rep("G2",6)))
```

On souhaite savoir si les données sont significativement différentes dans les 2 groupes. Pour cela, on utilise la fonction `wilcox.test()` :

```
wilcox.test(ech.1, ech.2, mu=0)
```

```
##
## Wilcoxon rank sum test
##
## data: ech.1 and ech.2
## W = 10, p-value = 0.4286
## alternative hypothesis: true location shift is not equal to 0
```

**Interprétation** : sous  $H_0$  (hypothèse sous laquelle les moyennes des deux échantillons sont égales), la statistique de test  $W = 10$  n'est pas anormalement petite, la valeur-p étant égale à 0.4286 ce qui n'est pas un événement rare. On ne peut donc pas rejeter l'hypothèse nulle.

**Remarque** : comment peut-on retrouver manuellement la statistique de test  $W$  renvoyée par R ?

### Etape 1

- 1. Classer toutes les observations par ordre croissant
- 2. Affecter son rang à chaque observation
- 3. Calculer la somme des rangs d'un échantillon (en général celui de plus petite taille).

Ici, on calcule la statistique  $W_1$  de l'échantillon le plus petit :

```
(w1=sum(rank(x)[groupe=="G1"]))
```

```
## [1] 25
```

**Etape 2** La statistique de test utilisée dans la fonction `wilcox.test()` consiste à construire la statistique de test  $W$  obtenue à partir de  $W_1$  :  $W = W_1 - n_1(n_1 + 1)/2 = 10$ . Elle correspond au nombre total de fois où un élément de l'échantillon 1 dépasse un élément de l'échantillon 2. On comprend donc que plus  $W$  est petit, plus cela implique que l'échantillon 1 aurait de petites valeurs comparées à l'échantillon 2.

**Test non paramétrique de Kruskal-Wallis** Il est possible de comparer les moyennes de plusieurs échantillons avec le test non paramétrique de Kruskal-Wallis. Ce dernier est utile lorsque l'on se trouve face à plusieurs échantillons dont on ne connaît pas la distribution. Il est mis en oeuvre par la fonction `kruskal.test()`. On reprend l'exemple précédent et on considère un troisième échantillon :

```
ech.3=c(2.49, 2.32, 3.01, 2.59, 2.76, 2.86)
x=c(ech.1, ech.2, ech.3)
groupe=factor(c(rep("G1",5), rep("G2",6), rep("G3",6)))
```

On utilise la fonction `kruskal.test()` ainsi :

```
(res.krus=kruskal.test(x~groupe))

##
## Kruskal-Wallis rank sum test
##
## data: x by groupe
## Kruskal-Wallis chi-squared = 11.3595, df = 2, p-value = 0.003414
```

Sous  $H_0$  (hypothèse sous laquelle les moyennes de tous les échantillons sont égales), la statistique de test est anormalement grande ( $K = 11.3594771$ ), la valeur-p étant égale à 0.0034145. On ne peut donc ici accepter l'hypothèse nulle.

### 3. Test de corrélation

Les tests de corrélation peuvent se baser, au moins, sur trois “statistiques” différentes : le coefficient de corrélation linéaire de Pearson, le  $\tau$  de Kendall et le  $\rho$  de Spearman. Ces trois méthodes sont disponibles dans la fonction `cor.test()`. Pour réaliser un test de corrélation entre deux variables, il suffit donc d'indiquer, en argument de cette fonction, les noms des deux vecteurs contenant les observations, ainsi que la méthode choisie (option **method**). Pour “affiner” le test, comme la plupart des fonctions de tests vues précédemment, il est également possible d'en préciser le type unilatéral ou bilatéral (option **alternative**) et le niveau de confiance (option **conf.level**).

**Remarque** : ce test peut être vu comme un test appliqué à deux échantillons appariés, dans la mesure où les variables  $X$  et  $Y$  ont été observées sur la même population.

**Exemple** : on considère la matrice des corrélations des variables quantitatives du jeu de données **iris**. On rappelle que la cellule  $(i, j)$  représente le coefficient de corrélation linéaire de Pearson  $r_{ij}$  entre la variable  $X_i$  et  $X_j$ .

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Au vu de la matrice des corrélations, on se pose la question si le coefficient de corrélation linéaire de Pearson entre les variables **Sepal.Length** et **Sepal.Width** est significativement différent de 0. On utilise donc ici

l'option par défaut `method = pearson`.

```
with(iris, cor.test(Sepal.Length, Sepal.Width))

##
## Pearson's product-moment correlation
##
## data: Sepal.Length and Sepal.Width
## t = -1.4403, df = 148, p-value = 0.1519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.27269325  0.04351158
## sample estimates:
##          cor
## -0.1175698
```

Sous  $H_0 : \{r = 0\}$ , la valeur de la statistique de test n'est pas anormale, ce qui est confirmé par la valeur-p qui est au-dessus du seuil de 5% usuel. On peut donc affirmer que le coefficient de corrélation linéaire de Pearson n'est pas significativement différent de 0. Autrement dit, le "lien linéaire" entre les deux variables n'est pas significatif.

## 4. Test d'indépendance de deux caractères

Les tests d'indépendance entre deux caractères reposent sur une table de contingence. Selon l'existence ou non de cellules de cette taille ne comportant que peu d'observations (le seuil est à 5), le test utilisée sera différent. Dans le premier cas, le test du  $\chi^2$  sera utilisé. Dans le second, ce sera le test non paramétrique exact de Fisher. Pour illustrer ces tests, nous allons simuler un cas classique : on cherche à savoir si le fait de présenter une caractéristique particulière (exemple : le fait de fumer) a une influence sur la réponse à un traitement. On a donc deux vecteurs  $X$  et  $Y$  avec pour variable, deux modalités ("nonfumeur" et "fumeur") pour  $X$  et ("pas de problème" et "problème") pour  $Y$ . Ici, on n'observe pas directement les variables  $X$  et  $Y$ , mais la table de contingence qui se construit, on le rappelle, avec la commande `table()` :

```
m=matrix(c(14,10,20,45),2,2,byrow=TRUE)
rownames(m)=c("non fumeur","fumeur")
colnames(m)=c("pas pbme","pbme")
tab=as.table(m)
```

On obtient la table de contingence suivante :

	pas pbme	pbme
non fumeur	14	10
fumeur	20	45

### a. Test d'indépendance du $\chi^2$

Pour ces 2 variables, nous nous retrouvons dans le cas où les cellules contiennent au moins 5 observations. Le test est donc valable. Pour utiliser la fonction `chisq.test()`, on a le choix de mettre comme argument directement les variables qualitatives, ou bien directement la table de contingence. Ici, on indique donc la

table de contingence :

```
(res.chisq=chisq.test(tab, simulate.p.value = TRUE))
```

```
##  
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##  
## data: tab  
## X-squared = 5.6411, df = NA, p-value = 0.02749
```

L'objet créé par la fonction contient les effectifs observés (`res.chisq$observed`), les effectifs théoriques (`res.chisq$expected`), et les résidus de Pearson (`res.chisq$residuals`). On peut retrouver ainsi la valeur de la statistique de test en utilisant la formule :

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{ij}^{th})^2}{n_{ij}^{th}}$$

où  $n_{ij}$  est la valeur observée dans la ligne  $i$  et colonne  $j$  de la table de contingence et  $n_{ij}^{th} = \frac{n_{i.}n_{.j}}{n}$  est la fréquence théorique sous l'hypothèse que les variables  $X$  et  $Y$  sont indépendantes.

```
sum((res.chisq$observed-res.chisq$expected)^2/res.chisq$expected)
```

```
## [1] 5.64106
```

On a choisi ici l'option `simulate.p.value = TRUE` qui signifie que la valeur-p n'est pas obtenue en utilisant une table théorique d'une loi de  $\chi^2$  à  $(I-1) \times (J-1)$  degrés de libertés ( $I$  et  $J$  sont les nombres de modalités de  $X$  et  $Y$ ), mais par une méthode de simulation de Monte Carlo.

Sous  $H_0 : \{X \text{ et } Y \text{ sont indépendantes}\}$ , on a trouvé qu'il était peu probable de trouver une valeur supérieure à 5.6410599 en faisant de la simulation de Monte Carlo. La p-value étant très faible (0.0274863), on ne peut accepter l'hypothèse d'indépendance des deux variables. Le fait de fumer peut donc entraîner des complications pendant le traitement.

## b. Test exact de Fisher

Ce test est particulièrement utile lorsque des cellules de la table de contingence à étudier sont inférieures à 5. Les calculs utilisés pour construire la statistique de test ne sont pas triviaux (ils font appel à des factorielles qui peuvent être compliqués à calculer), mais la fonction `fisher.test()` fait appel à un algorithme qui permet de faire ces calculs.

```
(res.F=fisher.test(tab))
```

```
##  
## Fisher's Exact Test for Count Data  
##  
## data: tab  
## p-value = 0.02635  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 1.076182 9.323975  
## sample estimates:  
## odds ratio  
## 3.106185
```

Comme pour le test d'indépendance du  $\chi^2$ , la valeur-p (0.0263478) nous conduit à rejeter l'hypothèse nulle d'indépendance de  $X$  et  $Y$ .