Les tests statistiques

Thibault LAURENT

Mise à jour: 05 décembre 2021

Contents

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 librairies 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 librairies 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 lien https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf écrit par les auteurs du site Wikistat (http://wikistat.fr/). 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 Interprétation de la valeur-p

Il y a souvent confusion dans l'interprétation de la valeur-p associée à un test statistique. C'est pourquoi on rappelle ici la démarche associée à un test statistique. Cela peut se résumer en 3 étapes :

- construction d'une hypothèse nulle et d'une hypothèse alternative. Par exemple, on est en présence d'un jeu de données qu'on suppose être issue d'une loi gaussienne. On souhaite vérifier cette supposition. Pour cela, on va construire l'hypothèse nulle suivante H_0 : {l'échantillon est distibué selon une loi gaussienne} et l'hypothèse alternative (la négation de H_0), H_1 : {l'échantillon n'est pas distibué selon une loi gaussienne}.
- construction d'une statistique de test. Pour chaque test que nous allons voir, nous allons systématiquement construire à partir de l'échantillon de données, une statistique de test θ, qui sous l'hypothèse nulle, sera supposée se comporter comme étant issue d'une loi de distribution connue.
- construction de la valeur-p. Cette valeur est la probabilité pour que la statistique de test appartienne à la loi de distribution dont elle est supposée être issue. Valeur arbitraire de 5% : si la valeur-p est supérieure à 5%, dans ce cas, cela indique que la statistique de test est bien issue de la loi de distribution et dans ce cas, l'hypothèse nulle ne peut être rejetée. Si elle est inférieure à 5%, c'est que la statistique de test est une valeur "anormale" et donc on ne peut pas accepter l'hypothèse nulle.

2 Problèmes à un échantillon

2.1 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()*.

2.1.1 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 au vecteur **u** suivant, réalisation d'une variable aléatoire issue d'une loi uniforme $\mathcal{U}(16, 24)$.

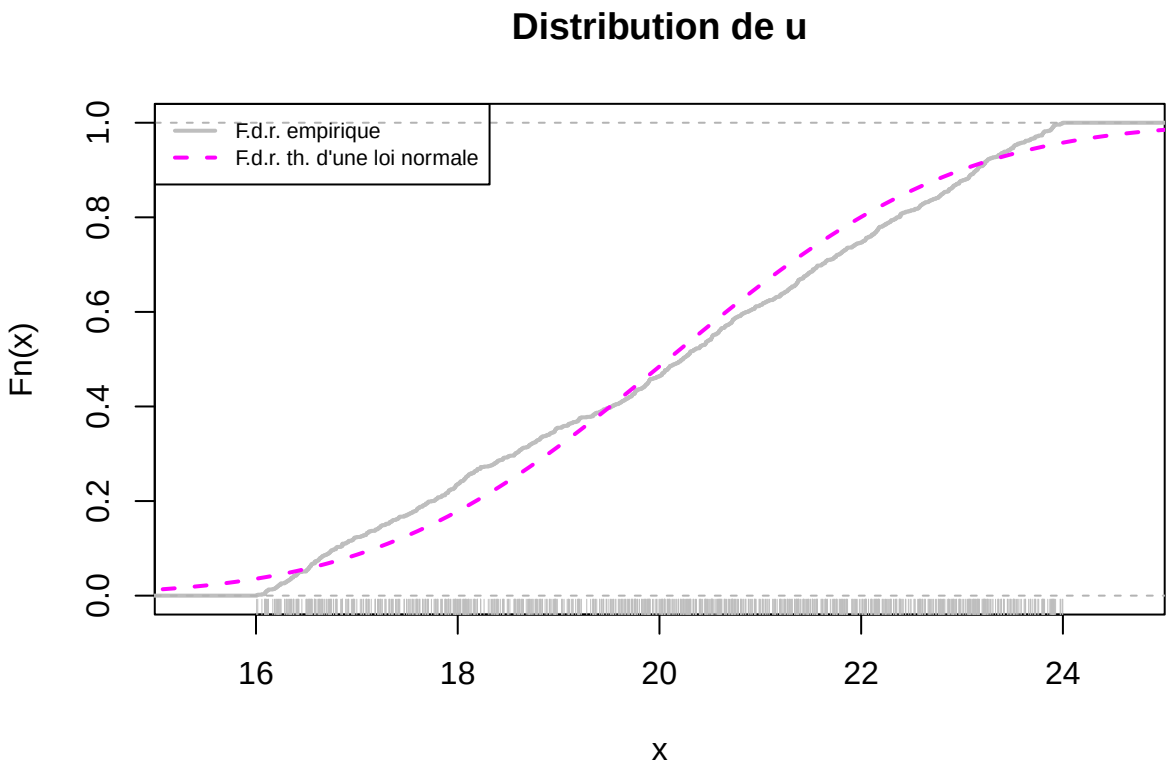
```
u <- runif(1000, 16, 24)
```

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. L'hypothèse nulle est donc ici $H_0 : \{F_n = F_{th}\}$. L'hypothèse alternative étant sa négation $H_1 : \{F_n \neq F_{th}\}$.

Dans un premier temps, on va se donner une idée intuitive de ce qu'on cherche exactement à faire. Pour cela, on représente graphiquement les deux fonctions de répartitions. La fonction de répartition empirique se fait avec la fonction générique *plot()* appliquée à un objet créé par la fonction *ecdf()*. Pour représenter la fonction de répartition théorique, cela se fait avec la fonction *pnorm()* (voir chapitre précédent) :

```
fdr.u <- ecdf(u)
plot(fdr.u, main = "Distribution de u", lwd = 2, col = "grey")
x <- seq(10, 30, 0.1)
lines(x, pnorm(x, mean(u), sd(u)), lty = 2, lwd = 2, col = "magenta")
rug(u, col = "grey")
```

```
legend("topleft", c("F.d.r. empirique", "F.d.r. th. d'une loi normale"),
       lty = 1:2, lwd = 2, col = c("grey", "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} (pour une valeur de x fixée), 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 (à x fixé). Pour plus d'informations, le lecteur pourra consulter la p.11 du document http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf

La fonction *ks.test()* prend comme argument d'entrée l'échantillon **x** correspondant à la 1ère distribution et l'argument **y** qui est soit un second échantillon, soit le nom d'une distribution connue (suivie des paramètres qui définissent cette distribution). Dans l'exemple suivant, on compare la distribution de l'échantillon **u** avec celle d'une loi théorique normale de paramètres la moyenne et écart-type de **u** :

```
(res.ks <- ks.test(x = u,
                   y = "pnorm", mean = mean(u), sd = sd(u)))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  u
## D = 0.066052, p-value = 0.0003247
## alternative hypothesis: two-sided
```

Interprétation : la fonction retourne d'une part la statistique de test D (l'écart le plus grand observé entre les deux courbes) ainsi que la valeur-p associée. Si l'hypothèse nulle est vraie (ici H_0 : $\{F_n = F_{th}\}$), cela implique que la valeur de D n'est pas trop éloignée de 0, 0 étant la valeur pour laquelle il n'y a aucune différence entre la fonction de répartition empirique et théorique. Pour vérifier cela, on suppose que sous l'hypothèse H_0, la statistique de test D devrait se comporter comme une valeur issue d'une loi de Kolmogorov, dont on connaît la distribution théorique (voir https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test). Ici, on constate que la valeur-p est très faible (0.0003). Autrement dit, la valeur de D ne peut pas être

considérée comme étant issue d'une loi de Kolmogorov 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.

Remarque : dans l'exemple ci-dessus, on a comparé la distibution de l'échantillon **u** avec celle d'une loi normale, mais on aurait pu la comparer avec n'importe quelle autre distribution en remplaçant **"pnorm"** par une autre distribution connue.

2.1.2 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'hypthèse nulle H_0 est {l'échantillon est distibuée selon une loi gaussienne}.

Exemple : pour illustrer l'utilisation de cette fonction, nous allons nous intéresser au vecteur **v** suivant, réalisation d'une variable aléatoire de loi normale $\mathcal{N}(20, 4)$.

```
set.seed(123)
v <- rnorm(1000, 20, 2)
```

Le résultat du test statistique est :

```
(res.sh <- shapiro.test(v))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  v
## W = 0.99838, p-value = 0.4765
```

Interprétation : la formule de la statistique de test W retournée est donnée dans cette page https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk. Sous H_0, plus W est grand, plus la compatibilité avec la loi normale est crédible. Ici, la valeur-p est égale à 0.4764686, ce qui implique que la statistique de test n'est pas une valeur "anormale", et donc on ne peut pas rejeter l'hypothèse nulle H_0.

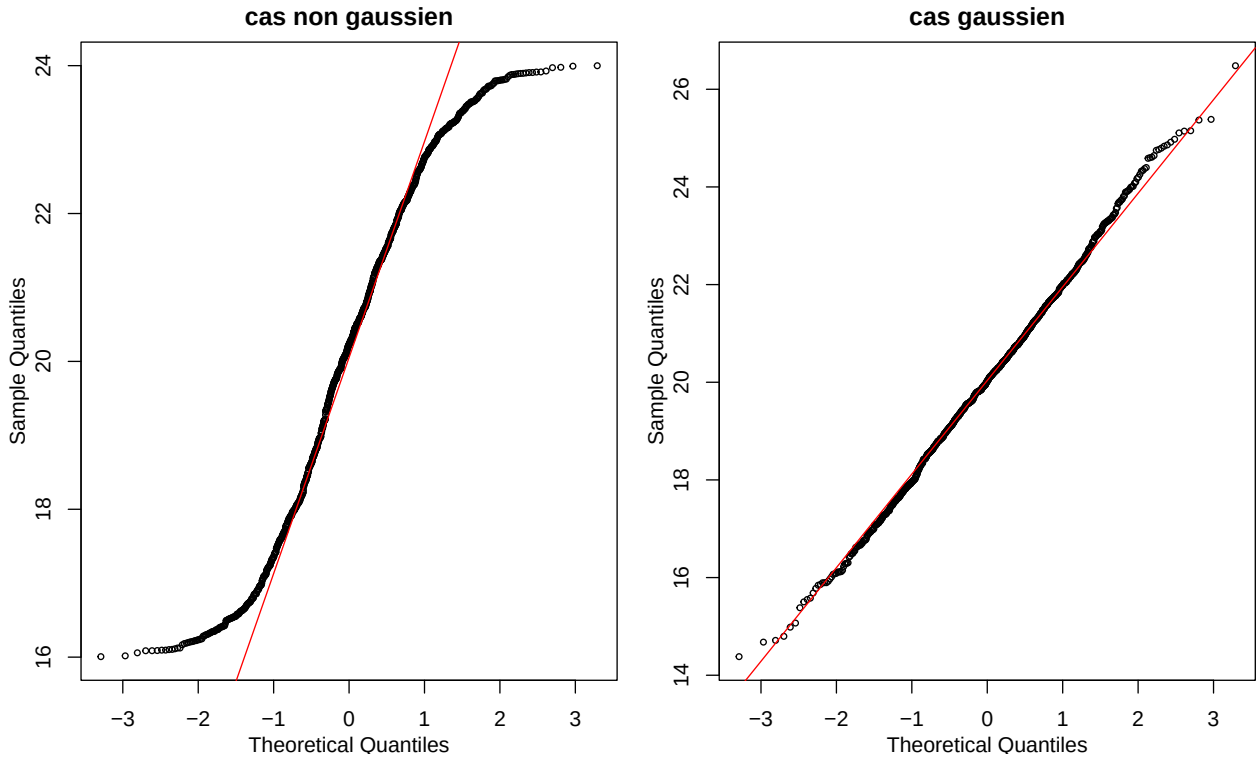
2.1.3 Le QQ (Quantile-Quantile) plot "gaussien"

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 *qnorm()*.

Exemple sur les variables **u** et **v** :

```
op <- par(mfrow = c(1, 2), mar = c(3, 3, 2, 1),
          mgp = c(2, 1, 0))
qqnorm(u, cex = 0.6, main = "cas non gaussien")
qqline(u, col = "red")
qqnorm(v, cex = 0.6, main = "cas gaussien")
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. Ceci peut se voir dans les figures ci-dessus qui confirment que l'échantillon représenté à gauche n'est pas issue d'une loi normale, ce qui n'est pas le cas de l'échantillon à droite.

2.1.3.1 Pour mieux comprendre un qqplot

On rappelle d'abord la définition de quartiles et déciles avant de généraliser à la définition des quantiles.

- Wikipedia "Un quartile est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population".
- Insee "Si on ordonne une distribution de salaires, de revenus, de chiffre d'affaires..., les déciles sont les valeurs qui partagent cette distribution en dix parties égales".

Comparaison entre quartiles empiriques et quartiles théoriques.

Pour calculer les valeurs empiriques des quartiles sur l'échantillon **v**, il suffit de trouver les valeurs de **v** qui séparent en 4 part égales cet échantillon. Pour cela, on peut utiliser la fonction **rank** qui donne le classement des observations. Comme on a 1000 observations, on va sélectionner les rangs 250, 500 et 750 pour les quartiles. Cela donne :

```
v.rank <- rank(v)
v[v.rank == 250] # 1er quartile
```

```
## [1] 18.74084
```

```
v[v.rank == 500] # 2ème quartile
```

```
## [1] 20.01458
```

```
v[v.rank == 750] # 3ème quartile
```

```
## [1] 21.32883
```

On peut également sélectionner la fonction quantile en renseignant les probabilités correspondantes, à savoir 0.25, 0.5 et 0.75 :

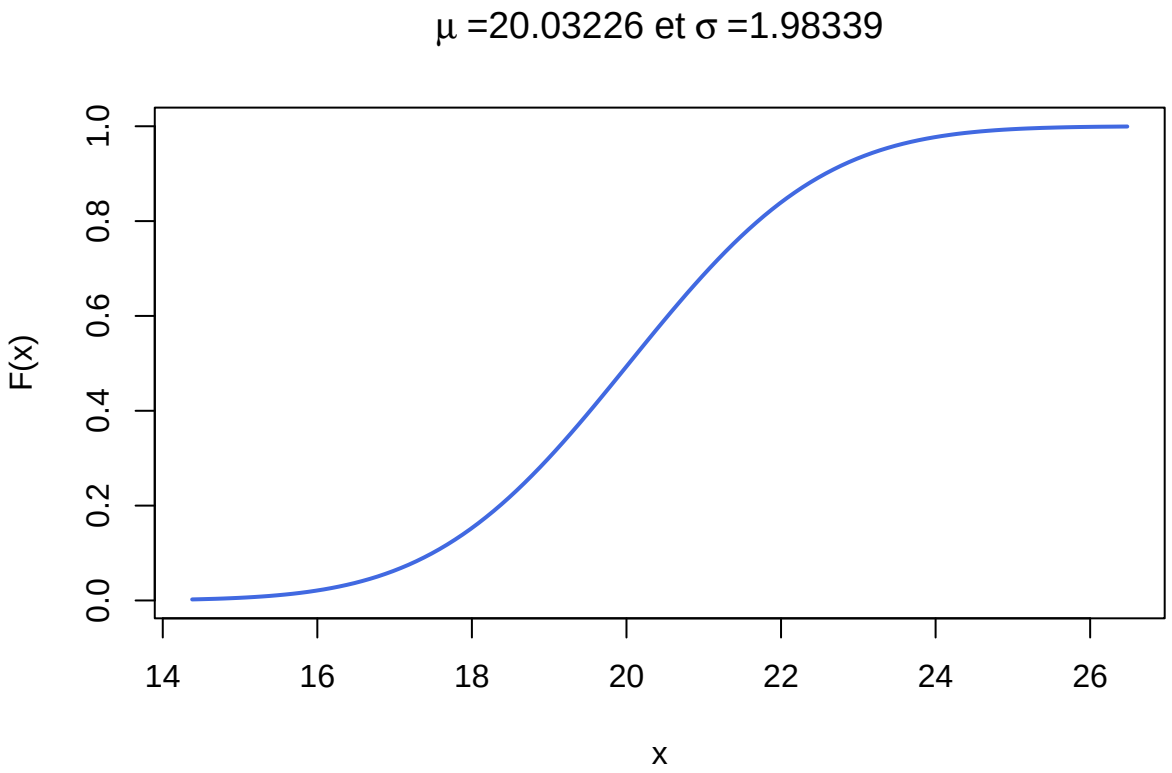
```
quantile(v, c(0.25, 0.5, 0.75))
```

```
##      25%      50%      75%
## 18.74335 20.01842 21.32920
```

Remarque: les valeurs ne sont pas exactement les mêmes que la fonction *quantile()* car cette dernière propose des algorithmes plus complexes (voir l'aide de la fonction pour plus de précisions) pour calculer les quantiles.

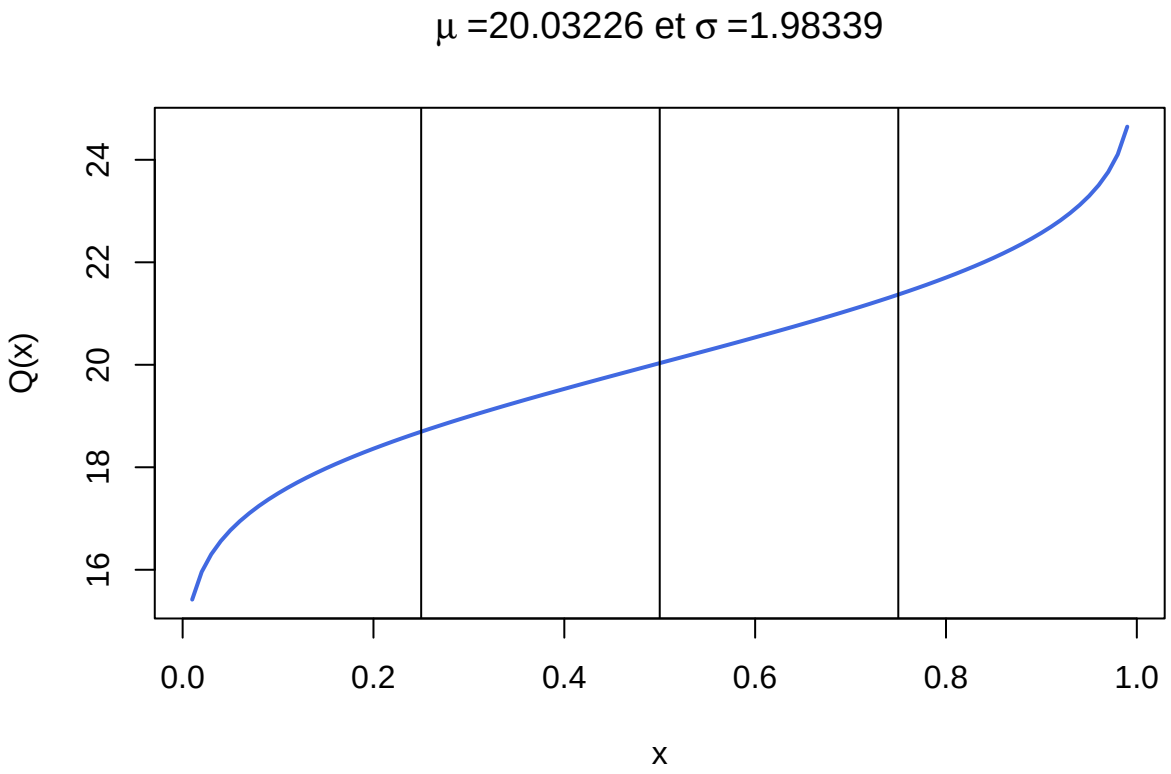
Maintenant, comment connaître les quartiles théoriques d'une loi normale de paramètres μ et σ ? On a vu dans le chapitre précédent la fonction *pnorm()* qui représente la fonction de répartition empirique de la loi gaussienne :

```
x <- seq(min(v), max(v), 0.1)
plot(x, pnorm(x, mean(v), sd(v)), type = "l", lwd = 2,
         ylab = "F(x)", col = "royalblue",
         main = bquote(mu~paste("=",.(mean(v)),"  et")~sigma~paste("=",.(sd(v)))))
```



Pour obtenir les quartiles, on voit bien qu'il s'agit des valeurs x t.q $F(x) = 0.25, 0.5, 0.75$. Pour avoir directement ces valeurs, on a donc besoin de connaître la fonction $Q(p) = F^{-1}(p)$, définie pour $p \in]0;1[$ qui est appelée la fonction quantile et qui s'obtient avec **R** avec la fonction *qnorm*() :

```
x <- seq(0.01, 0.99, 0.01)
plot(x, qnorm(x, mean(v), sd(v)), type = "l", lwd = 2,
         ylab = "Q(x)", col = "royalblue",
         main = bquote(mu~paste("=",.(mean(v)),"  et")~sigma~paste("=",.(sd(v)))))
abline(v = c(0.25, 0.5, 0.75))
```



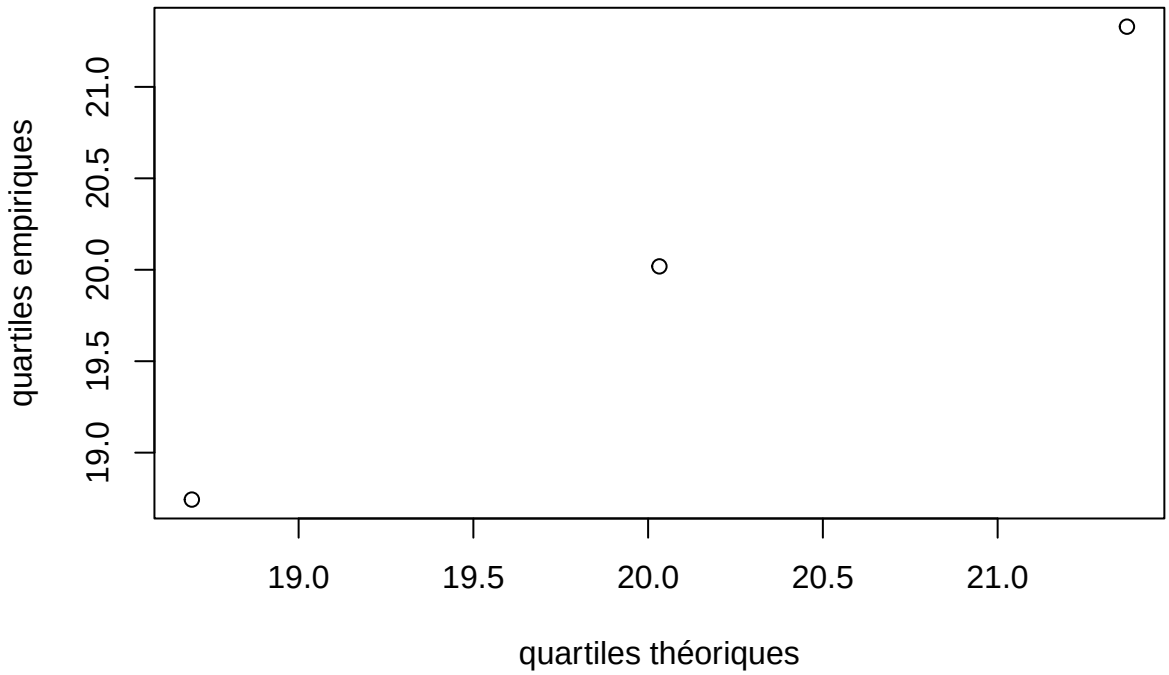
Pour obtenir les quartiles théoriques de l'échantillon **v**:

```
qnorm(c(0.25, 0.5, 0.75), mean(v), sd(v))
```

```
## [1] 18.69448 20.03226 21.37003
```

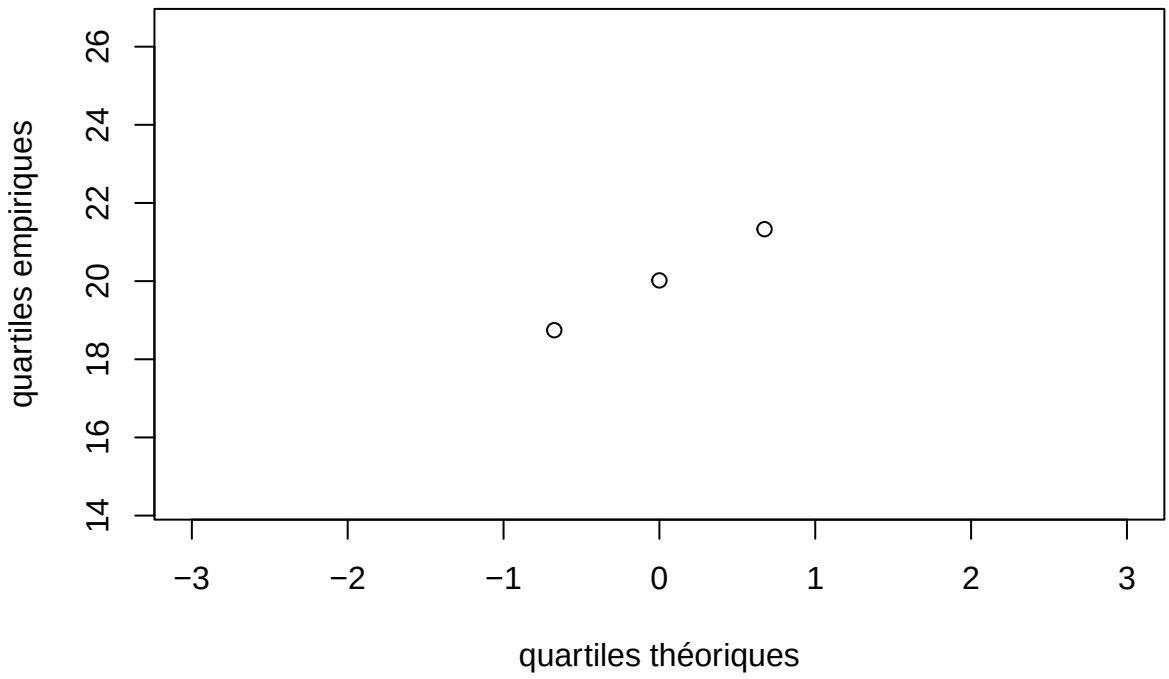
A présent, on peut représenter dans un graphique les quartiles théoriques en abscisses et les quartiles empiriques en ordonnées :

```
plot(qnorm(c(0.25, 0.5, 0.75), mean(v), sd(v)),
     quantile(v, c(0.25, 0.5, 0.75)),
     xlab = "quartiles théoriques",
     ylab = "quartiles empiriques")
```



Pour des soucis de lecture, on représente les quartiles théoriques de la loi gaussienne centrée et réduite en abcisses. Sur **R**, la fonction *qqline()* représente justemment la droite qui passe par les points correspondant aux quartiles 0.25 et 0.75.

```
plot(qnorm(c(0.25, 0.5, 0.75)),
     quantile(v, c(0.25, 0.5, 0.75)),
     xlab = "quartiles théoriques",
     ylab = "quartiles empiriques",
     xlim = c(-3, 3),
     ylim = range(v))
```



A présent, pour obtenir le **qqplot**, on représente le nuage de points des quantiles empiriques VS quantiles théoriques. On rappelle la définition du quantile :

- Wikipedia "En statistiques et en théorie des probabilités, les quantiles sont les valeurs qui divisent un jeu de données en intervalles contenant le même nombre de données. Il y a donc un quantile de moins que le nombre de groupes créés."

Pour réaliser le **qqplot**, on calcule autant de quantiles qu'il y a d'observations. Cela revient donc à représenter les quantiles théoriques d'ordre $Q(i/n)$, où $i = 1, ..., n$ d'une loi gaussienne centrée et réduite en abscisses et les quantiles observées d'ordre i/n en ordonnées, qui correspondent aux valeurs observées triées : $q_{obs,1/n} = x_{(1)}, ..., q_{obs,1} = x_{(n)}$. Pour calculer les ordres i/n, $i = 1, ..., n$, on pourra faire cela avec la fonction *qqpoints()*. Celle-ci ne calcule pas exactement les valeurs i/n pour $i = 1, ..., n$ pour des raisons techniques, mais c'est elle fait pratiquemment la même chose. Finalement, pour réaliser un graphique **qqplot** avec les commandes de base, on peut faire fait :

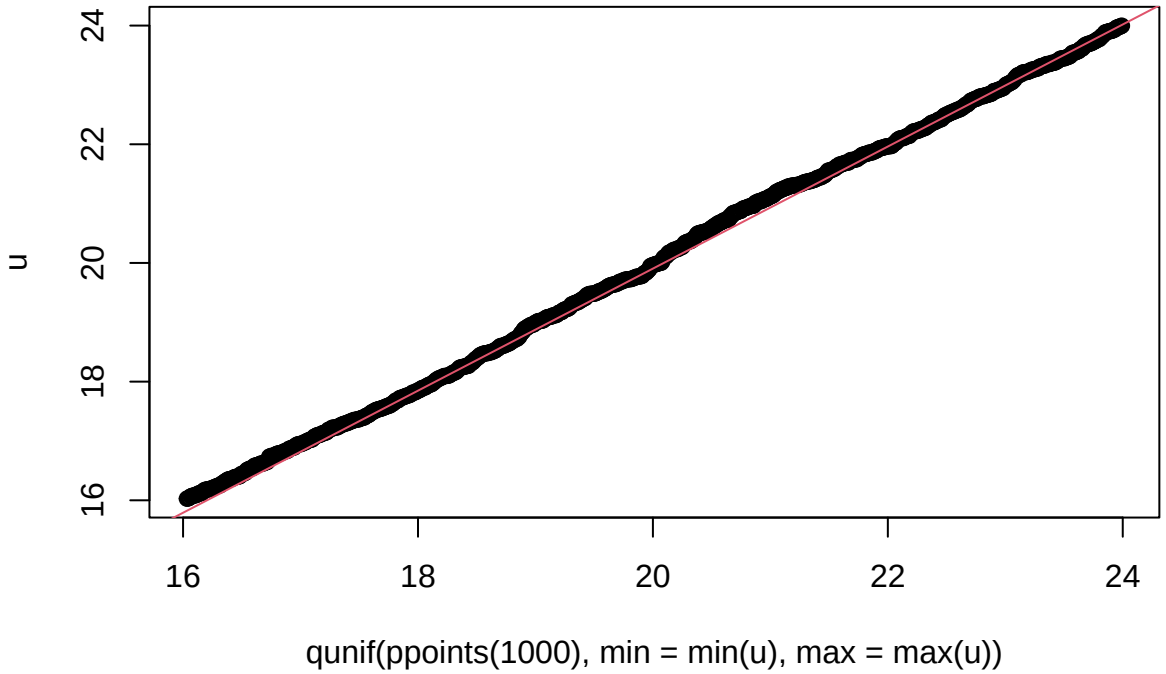
```
plot(qnorm(ppoints(length(v)), 0, 1), sort(v))
abline(mean(v), sd(v), col = "red")
```

Remarque : dans l'exemple ci-dessus, on a ajouté la droite d'équation $y = ax + b$ où **a=mean(u)** et **b=sd(u)**. On a vu que la fonction *qqline()* représentait la droite qui passe par les points correspondant aux quartiles 0.25 et 0.75.

2.1.4 Le QQ (Quantile-Quantile) plot non "gaussien"

Pour comparer un échantillon à une distribution quelconque, on utilisera la fonction *qqplot()*. Le premier argument correspondra aux quantiles théoriques de la loi de distribution qui nous intéresse et le second argument à l'échantillon observée. Par exemple, pour comparer l'échantillon **u** à une loi uniforme :

```
u <- runif(1000, 16, 24)
qqplot(qunif(ppoints(1000), min = min(u), max = max(u)), u)
qqline(u, distribution = function(p) qunif(p, min = min(u), max = max(u)), col = 2)
```



Remarque : la fonction *qqline()* permet d'ajuster la droite de régression à n'importe quelle famille de distribution connue. L'interprétation est la même que pour le cas gaussien c-à-d qu'on vérifie que les points s'ajustent bien à la droite de régression tracée.

2.2 Test d'égalité du paramètre μ d'un échantillon gaussien de loi $N(\mu, \sigma^2)$

Si la distribution d'un échantillon est gaussien, pour vérifier le test d'égalité du paramètre de moyenne μ à une certaine valeur, on utilise alors le test de Student, effectué par la fonction *t.test()*.

2.2.1 Le test de Student

Ici, il s'agit du cas où on sait que la distribution de la variable X est normale, de moyenne μ inconnue et de variance σ^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 0.2 et de variance 1 :

```
set.seed(113)
u <- rnorm(100, 0.2, 1)
```

Supposons qu'on se retrouve avec cette échantillon sans en connaître les paramètres (mais en sachant que la distribution est normale), et qu'on souhaite vérifier que le paramètre μ est différent d'une valeur μ_0 (par exemple 0). Pour cela, on va choisir l'hypotèse nulle suivante $H_0 : \{\mu = \mu_0\}$ contre $H_1 : \{\mu \neq \mu_0\}$.

Remarque : nous avons choisi comme hypothèse nulle, l'hypothèse que l'on souhaite rejeter. C'est en général ce que l'on fait de telle sorte qu'on ait un test plus puissant.

La formule de la statistique de test t retournée par la fonction *t.test()* est donnée dans https://fr.wikipedia.org/wiki/Test_de_Student. Elle vaut :

$$t = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

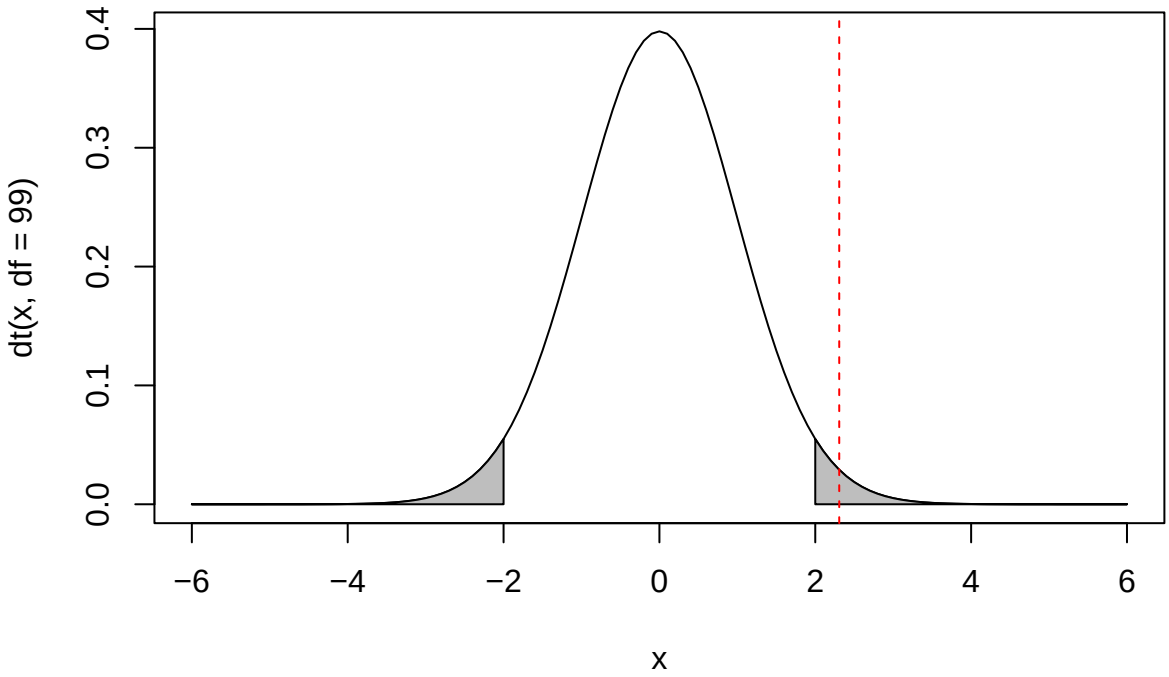
Pour la calculer sous **R** :

```
(stat_t <- (mean(u) - 0)/(sd(u) / sqrt(length(u))))
```

```
## [1] 2.307538
```

Une fois qu'on a calculé cette valeur, on va regarder si celle-ci peut-être issue d'une loi de distribution de Student à $n - 1$ degrés de liberté. Pour vérifier cela visuellement, on représente dans un graphique la loi de Student, la valeur calculée de t et enfin les zones à "exclure", c'est-à-dire les zones pour lesquelles on va considérer que la valeur de t est anormale. Comme il s'agit d'un test bilatéral (ici, l'hypothèse alternative est $H_1 : \{\mu \neq 0\}$, autrement dit on regarde si $\mu > 0$ ou $\mu < 0$), en choisissant un niveau de confiance à 5%, il faut représenter à la fois à droite et à gauche de la courbe les valeurs pour lesquelles on ne pourra pas accepter l'hypothèse nulle. Il s'agit des valeurs $Q_{t_{99}}(0.025)$ et $Q_{t_{99}}(0.975)$ où $Q_{t_{99}}$ est la fonction quantile d'une loi de Student à 99 dégrés de liberté.

```
x <- seq(-6, 6, 0.1)
plot(x, dt(x, df = 99), type = "l")
# représentation des zones exclues à gauche
x_lim <- x[x < qt(0.025, df = 99)]
polygon(c(x_lim, x_lim[length(x_lim)]),
        c(dt(x_lim, df = 99), 0), col = "grey")
# représentation des zones exclues à droite
x_lim <- x[x > qt(0.975, df = 99)]
polygon(c(x_lim[1], x_lim, x_lim[length(x_lim)]),
        c(0, dt(x_lim, df = 99), 0), col = "grey")
# représentation de la valeur t
abline(v = stat_t, lty = 2, col = "red")
```



Dans la figure ci-dessus, on constate que la valeur de la statistique de test t est une valeur "anormale" si on la compare à une loi de distribution de Student à 99 dégrés de liberté. Autrement dit, on ne pourra pas accepter l'hypothèse nulle.

En utilisant la fonction *t.test()* (en précisant le vecteur contenant l'échantillon et la moyenne à comparer, argument **mu**), on obtient les résultat suivants :

```
(res.ttest <- t.test(u, mu = 0, alternative = "two.sided"))
```

```
##
##  One Sample t-test
##
## data:  u
## t = 2.3075, df = 99, p-value = 0.02311
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.03508886 0.46576908
## sample estimates:
## mean of x
##  0.250429
```

On retouve la valeur de la statistique de test t, le nombre de degrés de liberté et la p-value. Comme l'hypotèse alternative est bilatérale (option **alternative = "two.sided"** qui est celle par défaut), la p-value correspond à $1 - Pr(-t < X < t)$ qui est égale à $Pr(X < -t) + (1 - Pr(X < t))$. Pour obtenir cette valeur sur **R**, on aurait pu faire :

```
pt(-stat_t, 99) + (1 - pt(stat_t, 99))
```

```
## [1] 0.02310682
```

Parmi les autres résultats qui sont affichés par la fonction, on trouve la valeur de la moyenne calculée sur l'échantillon ($\bar{x} = 0.250429$) ainsi que l'intervalle de confiance associé à μ avec un niveau de confiance égal à 95%. Cet intervalle de confiance est de la forme :

$$\bar{x} - t_{n-1,1-\alpha/2} s/\sqrt{n} < \mu < \bar{x} + t_{n-1,1-\alpha/2} s/\sqrt{n}$$

et peut se calculer à un niveau $\alpha = 5\%$ sous **R** de la façon suivante :

```
mean(u) - qt(0.975, df = 99) * sd(u) / sqrt(length(u))
```

```
## [1] 0.03508886
```

```
mean(u) + qt(0.975, df = 99) * sd(u) / sqrt(length(u))
```

```
## [1] 0.4657691
```

Pour interpéter l'intervalle de confiance : si on avait choisi de tester μ égal à une valeur comprise dans cet intervalle de confiance, on n'aurait pas pu rejeter l'hypothèse nulle.

Remarque 1 : la fonction rappelle quelle est l'hypothèse alternative. Ici, on peut lire **alternative hypothesis: true mean is not equal to 0**.

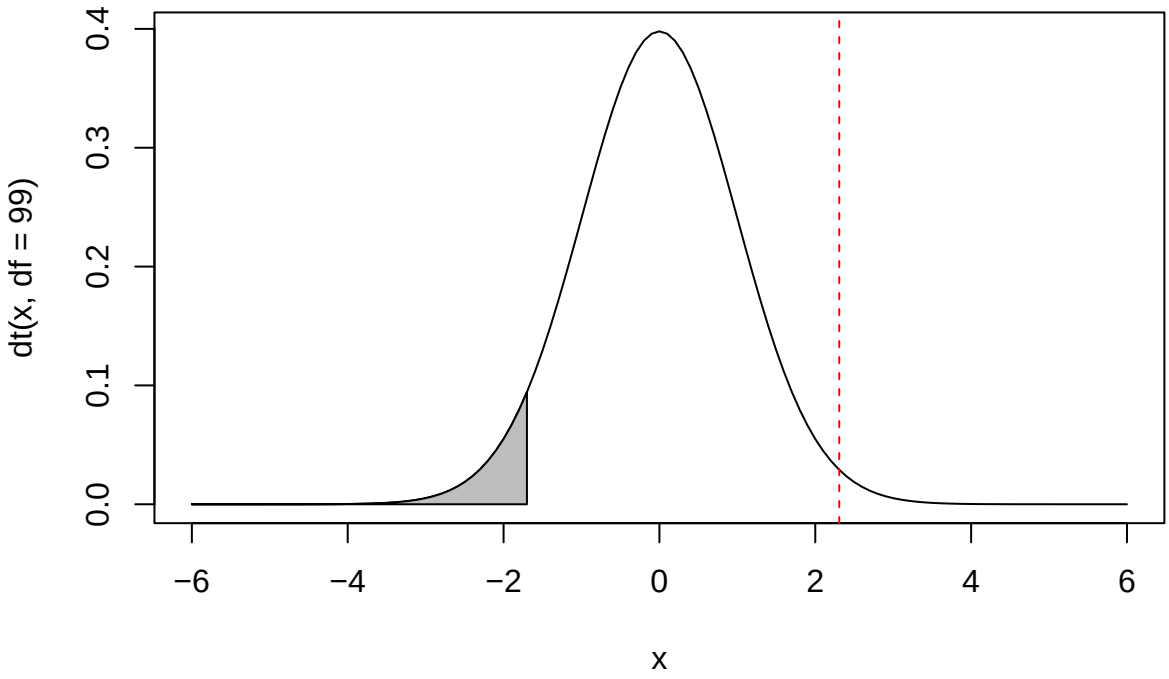
Remarque 2 : dans les modèles de régression linéaire, on utilise un test de Student pour tester la nullité de chacun des paramètres β_i.

2.2.2 Test bilatéral VS test unilatéral

Dans l'exemple précédent, on a considéré l'hypothèse alternative $H_1 : \{\mu \neq 0\}$, dite bilatérale. On aurait pu choisir une hypothèse alternative unilatérale, c'est-à-dire de la forme $H_1 : \{\mu > 0\}$ ou $H_1 : \{\mu < 0\}$.

La statistique de test reste la même, mais c'est au moment de calculer la p-value que le calcul va être différent. Par exemple, si on considère l'hypothèse alternative $H_1 : \{\mu < 0\}$, pour construire la zones à exclure, il s'agira uniquement des valeurs de t qui seront anormalement petites, à savoir inférieures au quantile $Q_{t_{99}}(0.05)$. Dans ce cas-là, le graphique ci-dessous ainsi que la fonction *t.test()* avec l'option **alternative="less"** montre qu'on ne peut pas rejeter l'hypothèse nulle.

```
x <- seq(-6, 6, 0.1)
plot(x, dt(x, df = 99), type = "l")
# représentation des zones exclues à gauche
x_lim <- x[x < qt(0.05, df = 99)]
polygon(c(x_lim, x_lim[length(x_lim)]),
        c(dt(x_lim, df = 99), 0), col = "grey")
# représentation de la valeur t
abline(v = stat_t, lty = 2, col = "red")
```




```
(res.ttest <- t.test(u, mu = 0 ,
                     alternative = "less"))
```

```
##
##  One Sample t-test
##
## data:  u
## t = 2.3075, df = 99, p-value = 0.9884
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf 0.4306254
## sample estimates:
## mean of x
##  0.250429
```

2.3 Test non paramétrique sur un échantillon de loi quelconque

2.3.1 Le test des signes et rangs de Wilcoxon

Dans la section précédente, on a supposé que l'échantillon suivait une loi gaussienne $N(\mu, \sigma^2)$. Il était donc possible de faire un test sur le paramètre μ.

Dans le cas où ne fait aucune hypothèse sur la loi de distribution de l'échantillon, on va construire une hypothèse sur un paramètre de position, ici la médiane. Soit un échantillon aléatoire $(X_1, X_2, ..., X_n)$ de loi parente une loi continue de fonction de répartition F_X dont la médiane est notée η.

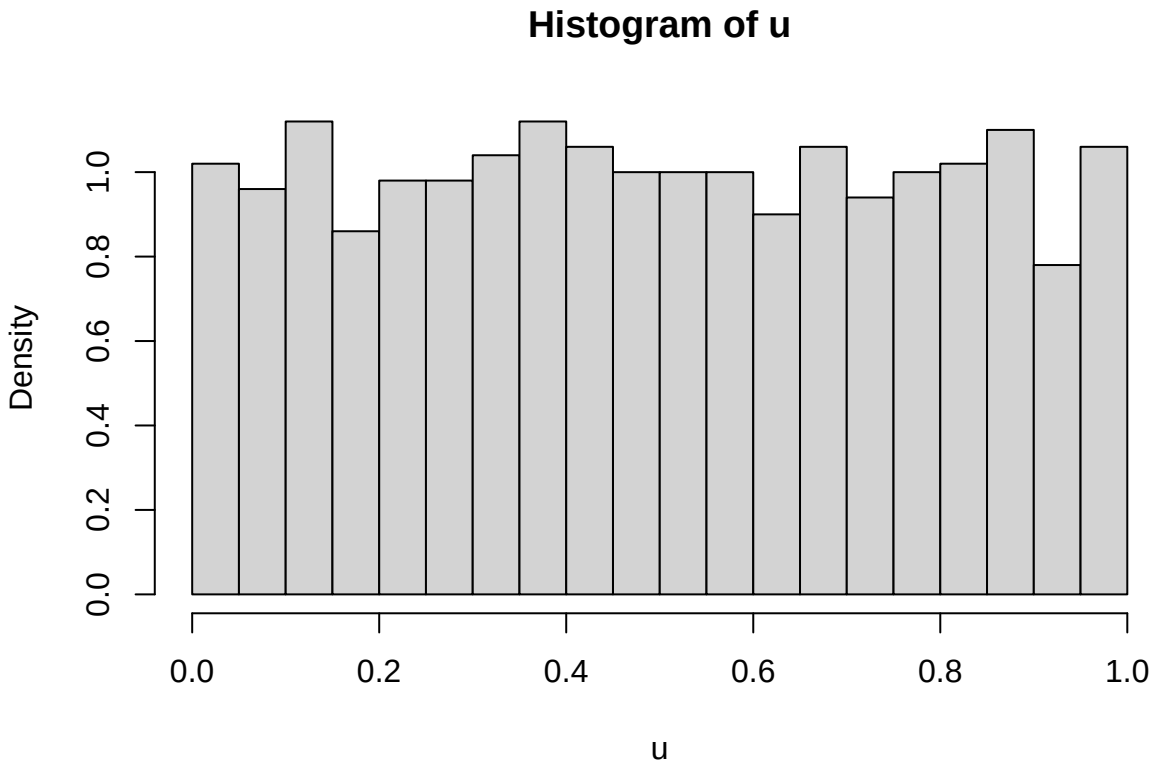
Le test des signes et rangs de Wilcoxon teste l'hypothèse nulle que la distribution de l'échantillon est symétrique par rapport à $\eta = \eta_0$ (cela implique donc d'avoir vérifié la symétrie avec un histogramme par exemple). L'hypothèse nulle est donc ici $H_0 : \{\eta = \eta_0\}$. L'hypothèse alternative par défaut sous **R** étant sa négation $H_1 : \{\eta \neq \eta_0\}$.

Pour calculer la statistique de test, on considère d'une part le vecteur $abs(X_1 - \eta_0, X_2 - \eta_0, ..., X_n - \eta_0)$ qu'on va ordonner en lui donnant des rangs de 1 à n. Ensuite on fera la somme des rangs seulement sur les valeurs qui vérifient $X_i > \eta_0$, pour obtenir la statistique de test V dont la loi de distribution est connue. Prenons par exemple un échantillon **u** issu d'une loi uniforme de paramètre 0 et 1 :

```r
set.seed(123)
u <- runif(1000, 0, 1)
```

On va s'intéresser à l'hypothèse nulle suivante $H_0 : \{\eta = 0.5\}$. On vérifie d'abord la symétrie autour de 0.5, ce qui semble le cas au vue de l'histogramme :

```r
hist(u, probability = T, nclass = 15)
```



Pour calculer la statistique de test :

```r
rank.u <- rank(abs(u - 0.5))
sum(rank.u * (u > 0.5))
```

```
## [1] 247519
```

C'est la valeur qui est également retournée par la fonction *wilcox.test()* :

```r
(res.wilco <- wilcox.test(u, mu = 0.5, conf.int = TRUE))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  u
## V = 247519, p-value = 0.765
## alternative hypothesis: true location is not equal to 0.5
## 95 percent confidence interval:
##  0.4788820 0.5155417
## sample estimates:
## (pseudo)median
##      0.4972185
```

Sous $H_0 : \{\eta = 0.5\}$, la probabilité qu'une observation soit supérieure à la valeur de la statistique de test $V = 2.47519 \times 10^5$ est 0.7650262, ce qui n'est donc pas un fait rare. On ne peut pas rejeter l'hypothèse nulle

et on admet donc que **u** est issu d'une distribution symétrique dont la médiane vaut 0.5.

2.4 Test de comparaison à une proportion

Lors d'un sondage réalisé auprès de 1000 individus, 521 d'entre eux ont affirmé apprécié la marque blablavelo. On peut représenter ce resultat sous forme de table de contingence :

	x
apprecie_Oui	521
apprecie_Non	479
nombre_total	1000

L'enseigne souhaiterait vérifier que plus de 50% de la population addhère à la marque.

Une façon de représenter le problème statistique est de supposer que l'échantillon observée (1000 individus qui ont répondu oui ou non) est distribuée selon une loi binomiale de paramètre $B(n, \pi)$ avec $n = 1000$ connu, et π inconnu. Dans ce cas, pour répondre à la problématique, cela revient à tester l'hypothèse $H_0 : \pi = 0.5$. Il s'agit de l'hypothèse qu'on souhaiterait rejeter au profit de l'hypothèse alternative $H_1 : \pi > 0.5$.

2.4.1 Test exact

Pour de faibles échantillons, il est conseillé d'utiliser un test exact à l'aide de la fonction *binom.test()*. Dans ce cas, ce n'est pas la peine de constuire une statistique de test. En effet, l'idée est de vérifier si un tirage aléatoire suivant une loi binomiale et comportant 521 succès peut avoir le paramètre 0.5. C'est ce que fait la fonction *binom.test()*. On précise ici l'argument **aternative="greater"** (qui spécifie que pour l'hypothèse alternative, on regarde $H_1 : \pi > 0.5$). :

```
binom.test(521, 1000, p = 0.5,
           alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  521 and 1000
## number of successes = 521, number of trials = 1000, p-value = 0.09738
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4944947 1.0000000
## sample estimates:
## probability of success
##                  0.521
```

Ici, on ne peut pas rejeter l'hypothèse nulle. Par ailleurs, l'intervalle de confiance nous indique qu'on n'aurait pas pu rejeter l'hypothèse nulle pour n'importe quelle valeur de π comprise entre 0.49 et 1. L'enseigne devra encore revoir sa communication pour améliorer l'image de sa marque pour convaincre "significativement" plus de la moitié de la population.

2.4.2 Test basé sur la statistique du χ^2

Une alternative à la fonction *binom.test()* est la fonction *prop.test()* qui donne des résultats très similaires. La statistique de test calculée est celle du χ^2. Cette statistique consiste à comparer les effectifs empiriques à ceux dits théoriques sous l'hypothèse nulle. Ici, la statistique de test sera égale à :

```
(521 - 500) ^ 2 / 500 + (479 - 500) ^ 2 / 500
```

```
## [1] 1.764
```

Elle s'obtient directement au moyen de la fonction *prop.test()*.

```
(res.prop <- prop.test(521, 1000, p = 0.5, 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 statistique de test est supposée être issue d'une loi de χ^2 à 1 degré de liberté. Ici, la valeur-p (0.0973958) est supérieure à 5%, autrement dit la valeur de la statistique de test n'est pas "anormale" (1.681) et on ne peut donc pas rejeter l'hypothèse nulle.

Remarque : la p-value est légèrement différente de celle du test exact, mais reste du même ordre de grandeur.

3 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.

3.1 Tests de comparaison de variance

3.1.1 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()*. L'hypothèse nulle est la suivante : $H_0 : \{\frac{\sigma_1}{\sigma_2} = 1\}$ (égalité des variances).

Simulons 2 échantillons **u1** et **u2** 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. C'est ce que l'on va vérifier :

```
(res.var <- var.test(u1, u2))
```

```
##
##  F test to compare two variances
```

```
##
## data:  u1 and u2
## F = 0.85857, num df = 149, denom df = 119, p-value = 0.3767
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6077063 1.2049672
## sample estimates:
## ratio of variances
##          0.8585703
```

Interprétation : la statistique de test F est basée sur le rapport des variances empiriques ($\frac{s^2_{u1}}{s^2_{u2}}$). Sous H_0, F est supposée être issue d'une loi de Fisher à 149 (taille du 1er échantillon moins 1) et 119 (taille du 2ème échantillon moins 1) degrés de libertés. On regarde donc si cette statistique est très éloignée de la valeur 1. Dans notre exemple, la valeur-p n'étant pas un événement rare ($p - value = 0.3767188$) la statistique de test $F = 0.8585703$ est donc bien issue d'une loi de Fisher. On ne peut donc pas rejeter l'hypothèse d'égalité des variances. Une autre façon de déterminer l'issue du test est de regarder si la valeur 1 est comprise dans l'intervalle de confiance à 95%. Dans cet exemple, c'est bien le cas ce qui suggère que le rapport des variances n'est pas significativement différent de 1.

Remarque : la fonction *var.test()* ne peut s'appliquer qu'à deux échantillons, ce qui n'est pas le cas de la fonction suivante.

3.1.2 Test de Barlett (sous hypothèse de normalité)

Le test de comparaison de variance de Barlett rend possible, lui, la comparaison de plusieurs échantillons de tailles différentes, tous issus d'une loi normale. Simulons ici 3 échantillons **u1**, **u2** et **u3** 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)
```

L'hypothèse nulle est donc ici $H_0 : \{\sigma_1 = \sigma_2 = \sigma_3\}$ (contre H_1 il existe au moins un échantillon pour lequel le paramètre de variance est différent des deux autres). 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 **u1**, **u2** et **u3** 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 = 76.913, df = 2, p-value < 2.2e-16
```

Interprétation: la formule de la statistique de test qui permet d'obtenir la valeur retournée par la fonction est donnée dans cette page https://fr.wikipedia.org/wiki/Test_de_Bartlett.

Sous H_0, la statistique de test est supposée être issue d'une loi de χ^2 à 2 degrés de libertés. Dans cet exemple, la valeur-p est très faible, ce qui sous-entend que la statistique de test est anormalement grande $K = 76.9131616$ par rapport à une loi du χ^2 à 2 degrés de libertés. On rejette donc bien l'hypothèse nulle : toutes les variances ne sont pas égales.

Remarque : pour utiliser la fonction *bartlett.test()*, on peut également utiliser la syntaxe de type **formula** à partir d'un **data.frame**. Pour cela, on créé d'abord le **data.frame** :

```
test_groupe <- data.frame(x = x, groupe = groupe)
head(test_groupe)
```

```
##          x groupe
## 1 14.98865     G1
## 2 26.06853     G1
## 3 34.11716     G1
## 4 15.44004     G1
## 5 32.51665     G1
## 6 46.90078     G1
```

Ensuite, on utilise la syntaxe suivante :

```
(res.bart <- bartlett.test(x ~ groupe, data = test_groupe))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  x by groupe
## Bartlett's K-squared = 76.913, df = 2, p-value < 2.2e-16
```

3.1.3 Tests d'égalité de variance sans hypothèses sur les distributions

Si les échantillons ne sont pas gaussiens, pour vérifier l'hypothèse d'égalité des variances, on pourra utiliser un de ces trois tests non paramétriques :

- Ansari-Bradley Test à partir de la fonction *ansari.test()* (comparaison de deux groupes seulement)
- Mood Two-Sample Test à partir de la fonction *mood.test()* (comparaison de deux groupes seulement)
- Fligner-Killeen Test à partir de la fonction *fligner.test()*

On ne rentre pas dans le détail de la construction de ces tests qui sont nettement moins répandus que les autres que nous avons présentés jusqu'à maintenant. Pour les utiliser :

```
ansari.test(u1, u2)
mood.test(u2, u3)
fligner.test(x ~ groupe, data = test_groupe)
```

3.2 Tests de comparaison de moyennes pour échantillons 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 mentioner 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 vu précédemment ne peut être rejeté :

```
var.test(debout, abattu)
```

```
##
##  F test to compare two variances
##
## data:  debout and abattu
## F = 0.89819, 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
```

3.2.1 Cas de données indépendantes

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 = 0.67994, num df = 1, denom df = 22, p-value = 0.4185
```

La statistique de test F est celle utilisée dans un modèle d'analyse de variance (on pourra voir la p.9 du document suivant : http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf) pour plus de précisions sur sa définition. Cette statistique de test est supposée être issue d'une loi de Fisher à 1 degré de liberté au numérateur (le nombre de groupes moins 1) et 22 au dénominateur (la taille totale de l'échantillon moins le nombre de groupes).

Remarque : comme les données sont non appariées, cela revient à dire qu'elles sont indépendantes. Dans ce cas, la taille de l'échantillon est $12 + 12 = 24$.

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 évenement rare ($p - value = 0.4184573$), c'est donc qu'on ne ne peut pas rejeter l'hypothèse nulle, autrement dit les deux échantillons ont une moyenne qui n'est pas significativement différente.

Remarque : dans ce cas particulier, on aurait pu faire un test de Student avec la fonction *t.test()* dans la mesure où nous avons seulement deux échantillons. Dans ce cas, l'hypothèse nulle est $H_0 : \{\mu_1 - \mu_2 = 0\}$ et la formule de la statistique de test est donnée à la p. 7 de http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf. Cela nous aurait conduit exactement à la même conclusion, les deux tests étant ici équivalents.

3.2.2 Cas de données appariées

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 permetant d'utiliser l'option **paired = TRUE**. On garde l'hypothèse d'égalité des variances. Dans ce cas, l'hypothèse nulle est $H_0 : \{\mu_1 - \mu_2 = 0\}$.

```
(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 : la formule de la statistique de test est donnée à la p. 8 de http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf. Pour la construire on construit un échantillon de la forme $(X - Y)$. Cette statistique de test est supposée suivre une loi de Student dont le degré de liberté vaut la taille de l'échantillon moins 1 (ici 11).

Remarque : comme les données sont appariées, cela revient à dire qu'on considère que même si une observation a été observée deux fois (debout et abbatu), elle ne va compter que pour une autre observation (contrairement au cas indépendant). Dans ce cas, la taille de l'échantillon est 12.

Dans ce cas, la valeur-p étant inférieure au seuil de 5%, cela implique que la statistique de test n'est pas issue d'une loi de Student à 11 degrés de liberté. L'hypothèse H_0 ne peut donc être acceptée. Autrement dit, la moyenne n'est pas la même dans les deux groupes. Cette conclusion n'est pas la même que celle où on a supposé les données indépendantes, d'où l'importance de bien faire la distinction entre données appariées et données indépendantes.

3.3 Tests de comparaison d'échantillons issus 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 Mann-Whitney (pour données indépendantes), celui des rangs signés de Wilcoxon (pour données appariées) et celui de Kruskall-Wallis dans le cas de plus de deux échantillons. Dans ces tests, on s'intéresse à l'hypothèse H_0 selon laquelle les échantillons sont identiquement positionnés.

3.3.1 Test non paramétrique de Mann-Whitney sur données non appariées

Le test de Mann-Whitney est utilisée lorsque l'on ne peut faire l'hypothèse de normalité pour les distributions de deux échantillons non appariés. Dans la littérature anglo-saxonne, il est connu sous le nom de **Wilcoxon rank-sum test** (à ne pas confondre avec le test **Wilcoxon signed rank test** qui est pour données appariées). Il est réalisé dans **R** par la fonction *wilcox.test()*

Remarque : il s'agit de la même fonction que nous avons vue pour le test de comparaison à une médiane dans le cas d'un échantillon simple. Toutefois, selon les arguments qu'on lui donne, cette dernière n'effectue pas les mêmes tests.
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)))
```

Question: est-ce que les deux échantillons peuvent être considérés comme étant issu d'une même loi de distribution ?

D'abord, on vérifie l'égalité des variances au moyen d'un test non paramétrique vu ci-dessus :

```
ansari.test(ech.1, ech.2)
```

```
##
##  Ansari-Bradley test
##
## data:  ech.1 and ech.2
## AB = 17, p-value = 0.961
## alternative hypothesis: true ratio of scales is not equal to 1
```

Ensuite, on va construire la statistique de test W de la façon suivante :

Etape 1

- 1. Classer toutes les observations par ordre croissant
- 2. Affecter son rang à chaque observation
- 3. Calculer la somme des rangs des deux échantillons.

Ici, on calcule les statistiques U_1 et U_2 :

```
(u1 <- sum(rank(x)[groupe == "G1"]))
```

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

```
(u2 <- sum(rank(x)[groupe == "G2"]))
```

```
## [1] 41
```

Etape 2

La statistique de test utilisée dans la fonction *wilcox.test()* consiste à construire la statistique de test W obtenue en prenant le minimum de : $U_1 - n_1(n_1+1)/2 = 10$ et $U_2 - n_2(n_2+1)/2 = 20$. Ici, on choisit donc $W = U_1 = 10$. Cela correspond au nombre total de fois où un élément de l'échantillon 1 dépasse un élément de l'échantillon 2. Autrement dit, plus W est petit, plus cela implique que l'échantillon 1 aurait de petites valeurs comparées à l'échantillon 2.

On retrouve ce résultat avec la fonction *wilcox.test()*. Dans ce cas, ceci revient à tester l'hypothèse nulle H_0 : { la position des deux échantillons diffère par un paramètre $\mu = 0$}, ce qui revient à tester plus ou moins que la distribution des deux échantillons est identique.

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

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

Interprétation : il n'y a pas de lois connus pour W, mais elle a été tabulée (autrement dit, pour chaque valeur de la statistique de test, est associée une valeur-p). Sous H_0, la valeur-p étant égale à 0.4286 cela signifie que la statistique de test $W = 10$ n'est pas anormale. On ne peut donc pas rejeter l'hypothèse nulle.

3.3.2 Test non paramétrique de Wilcoxon sur données appariées

On reprend l'exemple des arbres pour lesquels on a mesuré la hauteur avant et après abattage. Il s'agit donc de données appariées; si on suppose que la distribution de ces échantillons n'est pas issue d'une loi gaussienne, on utilise un test non paramétrique. Il s'agit du test de Wilcoxon des rangs signés (*Wilcoxon signed rank test*). Il s'obtient avec la fonction *wilcox.test()* en ajoutant l'option **paired=TRUE**. Cela revient à faire un

test d'égalité de médiane (voir section précédente) sur l'échantillon $(X_1 - X_2)$ en prenant comme hypothèse nulle $H_0 : \{\eta = 0\}$.

```
wilcox.test(debout, abattu, paired = TRUE)
```

```
## Warning in wilcox.test.default(debout, abattu, paired = TRUE): cannot compute
## exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  debout and abattu
## V = 8.5, p-value = 0.01856
## alternative hypothesis: true location shift is not equal to 0
```

Ce qui revient à faire :

```
wilcox.test(debout - abattu, mu = 0)
```

```
## Warning in wilcox.test.default(debout - abattu, mu = 0): cannot compute exact p-
## value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  debout - abattu
## V = 8.5, p-value = 0.01856
## alternative hypothesis: true location is not equal to 0
```

Dans ce cas, on rejette l'hypothèse nulle et les deux échantillons ne sont donc pas issus d'une même distribution.

3.3.2.1 Test non paramétrique de Kruskall-Wallis

Il est possible de comparer les distributions de plusieurs échantillons avec le test non paramétrique de Kruskall-Wallis. Il s'agit d'une généralisation du test de Mann-Whitney. Ce dernier est utile lorsque l'on se trouve face à plusieurs échantillons dont on ne connait 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)))
```

Dans ce cas, l'hypothèse nulle est $H_0 : \{\eta_1 = \eta_2 = \eta_3\}$ contre l'hypothèe alternative qu'il existe au moins un paramètre de location différent de celui des autres.

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 = 11.359, df = 2, p-value = 0.003414
```

La formule de la statistique de test est donnée à la p. 13 de https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf. Sous H_0, la valeur-p étant égale à 0.0034145, cela implique que la statistique de test est anormalement grande ($K = 11.3594771$). On ne peut donc pas ici accepter l'hypothèse nulle.

4 Tests de corrélation

Les tests de corrélation peuvent se baser, au moins, sur trois "statistiques" différentes : le coefficient de corrélation linéaire de Pearson, le τ de Kendall et le ρ 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 vue 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. L'hypothèse nulle est donc : $H_0 : \{r = 0\}$ contre $H_1 : \{r \neq 0\}$. On utilise 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
```

La formule de la statistique de test t donnée dans https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Testing_using_Student.27s_t-distribution est supposée suivre une loi de Student à $(n - 2)$ degrés de liberté (ici 148). La valeur-p qui est au-dessus du seuil de 5% indique que la valeur de la statistique de test n'est pas "anormale" et peut donc être assimilée à une loi de Student à 148 degrés de liberté. On ne peut donc pas rejeter l'hypothèse nulle, et donc affirmer que le coefficient de corrélation linéaire de Pearson n'est pas significativement différent de 0. Il n'existe pas de "lien linéaire" significatif entre les deux variables.

Remarque : on constate que la valeur 0 est comprise dans l'intervalle de confiance à 95%, ce qui confirme que le coefficient de corrélation linéaire de Pearson n'est pas significativement différent de 0.

5 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 ne comportant que peu d'observations (le seuil est à 5), le test utilisé sera différent. Dans

le premier cas, le test du χ^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

Remarque : si on avait disposé du jeu de données original (cf table ci-dessous), aurait pu utiliser directement la fonction *table()* pour obtenir la table de contingence.

Table 4: Extrait du jeu de données original

fumeur	probleme
TRUE	TRUE
TRUE	TRUE
FALSE	FALSE
FALSE	TRUE
FALSE	FALSE
FALSE	TRUE
TRUE	TRUE

L'hyothèse nulle est ici H_0 : { X et Y sont indépendantes}.

5.1 Test d'indépendance du χ^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 = 0.02649
```

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 χ^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. C'est pourquoi ici le degré de liberté n'est pas disponible (NA i.e. Non Available).
La valeur-p étant très faible (0.0264868), cela indique que la statistique de test est une valeur peu probable, et on ne peut donc pas accepter l'hypothèse d'indépendance des deux variables. Le fait de fumer peut donc entraîner des complications pendant le traitement.

5.2 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 de façon efficace.

```
(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 χ^2, la valeur-p (0.0263478) nous conduit à rejeter l'hypothèse nulle d'indépendance de X et Y.