TESTS STATISTIQUES AVEC **R**

Amandine Blin
UAR 2700 2AD, Service Analyse de Données, Pôle Analyse de Données

19/10/2022

Table des matières

Chapitre 1

Quelques notions

Des références de livres sont indiquées à la fin du document (Millot (2011), Zar (1984), Frédéric Bertrand (2010), Saporta (2006), Verzani (2005)).

1.1 Quelques définitions

Test statistique

Prendre une décision en fonction d'un échantillon entre deux hypothèses une fois l'étude réalisée. On va vérifier une information et quantifier le risque associé. C'est de la statistique inférentielle.

Comment choisir un test statistique ?

- En fonction de la taille de l'échantillon
- En fonction de la nature des variables quantitatives ou qualitatives : Il existe des tests paramétriques et non paramétriques qui seront détaillés dans la suite du cours.

Les grandes étapes d'un test statistique

- Choix des hypothèses
- Choix du test statistique
- Conditions d'application du test
- Définir la statistique de test
- Choix de la région critique
- Calcul de la valeur observée de la statistique
- Détermination de la zone de rejet
- Conclure par la décision de l'hypothèse retenue et interprétation

Exemple : On mesure le poids de femelles koalas (échantillon). On veut tester si le poids moyen des femelles observées est égal au poids moyen des femelles koalas (valeur de référence égale à 6 kg).

1.2 Les hypothèses statistiques

On distingue deux hypothèses.

L'hypothèse nulle H_0

Dans le cas de cette hypothèse, il y a égalité des paramètres comparés.

L'hypothèse alternative H_1

Dans ce cas, il y a différence ou inégalité des paramètres. C'est l'hypothèse contraire.

Exemple : Dans notre exemple, on pourra prendre :

- H_0 : Le poids est égal à 6 kg
- H_1 : Le poids n'est pas égal à 6 kg.

Décision statistique

Choisir entre H_0 et H_1.

1.3 Statistique de test

La statistique de test (variable de décision) résume l'information de l'échantillon qu'on souhaite tester. Elle est telle qu'on connaît sa loi de probabilité sous l'hypothèse H_0.

1.4 Risques et zone de rejet

Risque de première espèce α

Rejeter H_0 à tort

Risque de seconde espèce β

Ne pas rejeter H_0 alors que H_0 n'est pas vraie.

Niveau de significativité

C'est une erreur de première espèce limitée à un niveau qui est généralement 1, 5 ou 10 %.

Zone de rejet

Ensemble des valeurs de la statistique de test permettant de rejeter l'hypothèse H_0

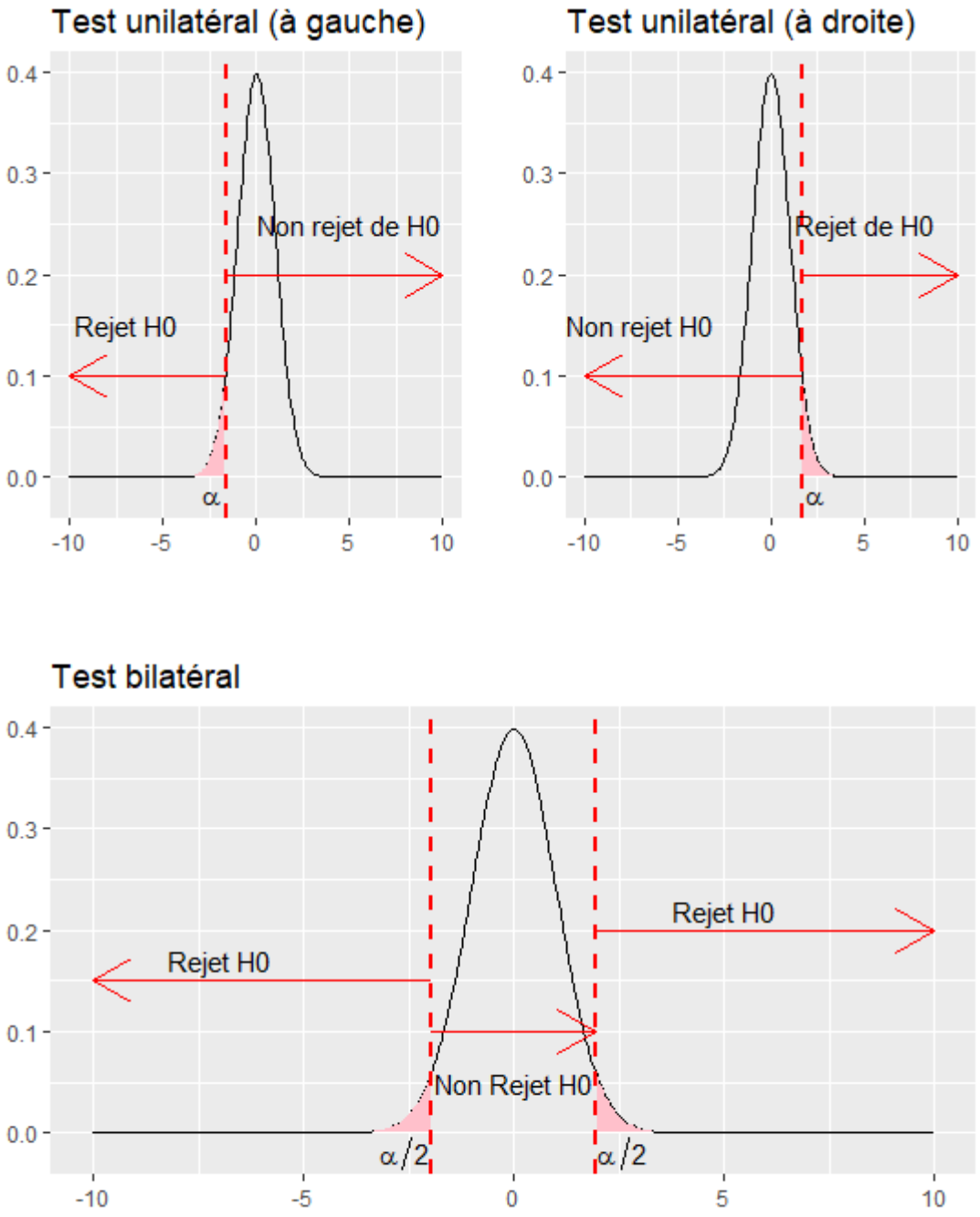
On distingue un test unilatéral d'un test bilatéral.

Test bilatéral

Association à H_1 selon laquelle la différence des signes est inconnue. Dans ce cas, la zone de rejet s'étend de part et d'autre de celle-ci.

Test unilatéral

Association à H_1 selon laquelle la différence des signes est connue. Dans ce cas, la zone de rejet ne s'étend que d'un côté.



1.5 Règles de décision

Règle de décision

- Si la valeur de la statistique calculée est supérieure à la valeur seuil, on rejette H_0 au risque α.
- Si la valeur de la statistique calculée est inférieure ou égale à la valeur seuil, on ne rejette pas H_0.

p-valeur

Probabilité (nombre réel compris entre 0 et 1) sous H_0 d'obtenir une valeur au moins aussi extrême que celle observée. C'est une probabilité critique telle qu'on puisse rejeter H_0 au risque $100\alpha\%$.

Règle de décision

- Si la p-valeur est inférieure ou égale à α, on rejette H_0.
- Si la p-valeur est supérieure à α, on ne rejette pas H_0.

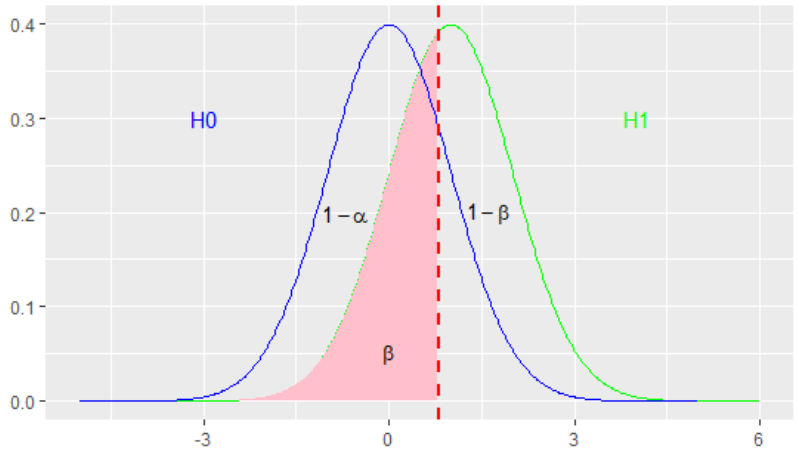
1.6 Risques statistiques

	Réalité	
Décision	Ho décidée	H1 décidée
Ho décidée	1-α	β
H1 décidée	α	1-β

1.7 La puissance du test statistique

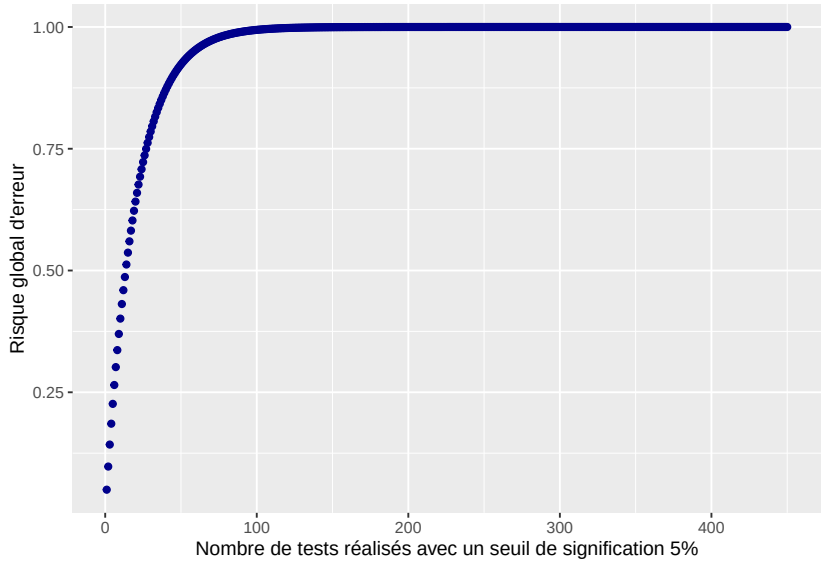
Puissance

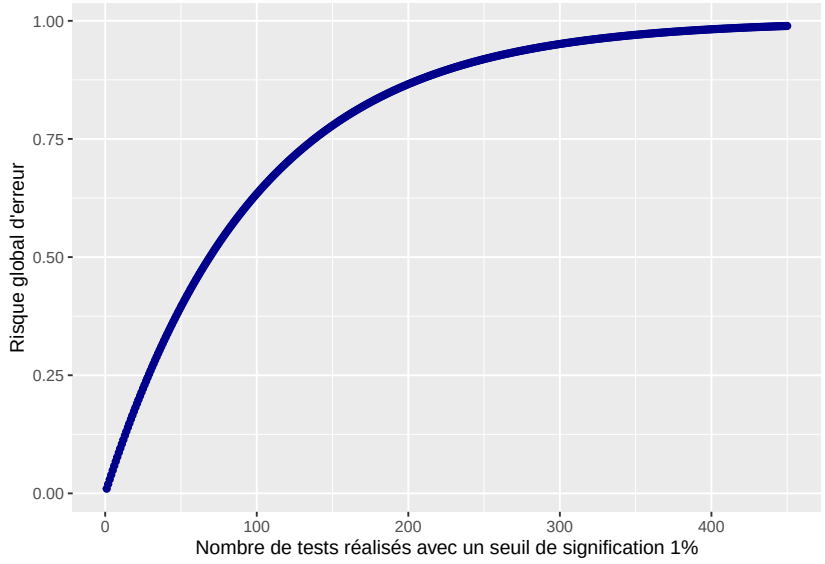
Rejeter H_0 à raison. La puissance est égale à 1-β. Plus la taille de l'échantillon augmente, plus la puissance augmente. De même, si α augmente, la puissance augmente. L'analyse de puissance s'effectue a priori ou a posteriori.



1.8 Introduction aux tests multiples

Lorsqu'un jeu de données est important, il peut être nécessaire d'effectuer de nombreux tests simultanément et le risque global d'erreur de première espèce α augmente. On peut donc obtenir des résultats significatifs au hasard. Lorsqu'on fait un test, la probabilité d'avoir un faux positif est α. Ainsi, la probabilité de ne pas avoir un faux positif est $1 - \alpha$. Le risque global d'erreur de première espèce (et donc de conclure à tort) pour 30 tests est donc $1 - (1 - 0.05)^{30} = 0.78$ soit 78%.





Plusieurs méthodes existent pour ajuster la p-value :

— FWER(Family-wise error rate) : On peut contrôler la probabilité d'avoir au moins un risque de première espèce (Bonferroni, Holm, Hochberg…).

— FDR (Family Discovery Rate) : Le but est de contrôler la proportion de résultats

significatifs qui sont faux (Benjamini-Hochberg…). Cette méthode est moins conservatrice que la précédente notamment concernant la correction de Bonferroni.

Chapitre 2

Les jeux de données et les packages utilisés

2.1 Les jeux de données

Le jeu de données *iris*

Les données qu'on va utiliser ont été collectées par Edgar Anderson (Anderson (1935)). 50 fleurs de 3 espèces (*Species*) différentes d'iris (*iris setosa*, *versicolor*, et *virginica*) ont été mesurées (en cm) :

- La longueur des sépales (*Sepal.Length*)
- La largeur des sépales (*Sepal.Width*)
- La longueur des pétales (*Petal.Length*)
- La largeur des pétales (*Petal.Width*)

Le jeu de données *anorexia* de la librairie *MASS*

Le jeu de données *anorexia* est disponible dans le package *MASS* (Hand et al. (1993), Venables and Ripley (2002)).

On étudie l'effet d'un traitement sur le poids de jeunes filles anorexiques.

Le tableau a 72 lignes (individus) et trois variables :

- La variable *Treat* : le type de traitement
- La variable *Prewt* : le poids avant le traitement
- La variable *Postwt* : le poids après le traitement

Le jeu de données *PlantGrooth*

Il est disponible dans le package de base (Dobson (1983)). Une expérience a été menée pour comparer les rendements (en mesurant le poids de plantes) selon un contrôle et deux conditions de traitement. Le jeux de données a 30 lignes et deux colonnes :

- La variable weight : le poids
- La variable group : condition de traitement (ctrl, trt1, trt2)

Le jeu de données *airquality*

La qualité de l'air à New York a été étudiée de mai à septembre 1973. Les données ont été obtenues auprès du Département de conservation de l'État de New York pour les données sur l'ozone et du National Weather Service pour les données météorologiques.

Le jeu de données a 153 observations et 6 variables :

- La variable *Ozone* : la concentration moyenne d'ozone (île Roosevelt)
- La variable *Solar.R* : radiation solaire (central park)
- La variable *Wind* : vitesse de vent moyenne (mile/heure, aéroport de la Guardian)
- La variable *Temp* : température journalière maximum (Fahrenheit, aéroport de la Guardian)
- La variable *Month* : mois d'observation (numéro) de mai à septembre
- La variable *Day* : jour du mois

Le jeu de données *mtcars*

Les données décrivent la consommation de fuel et le design et la performance de 32 voitures (Motor Trend, 1974, Henderson and Velleman (1981)). Les variables sont les suivantes :

- La variable *mpg* : Consommation (miles/galon)
- La variable *cyl* : Nombre de cylindres
- La variable *disp* : Déplacement
- La variable *hp* : Puissance brute
- La variable *drat* : Essieu arrière
- La variable *wt* : Poids
- La variable *qsec* : Vitesse

— La variable *vs* : Moteur
— La variable *am* : Transmission(automatique /manuelle)
— La variable *gear* : Nombre d'engrenages avant
— La variable *carb* : Nombre de carburateurs

2.2 Les packages utilisés

— Le package *ggpubr* (Kassambara (2020)) : permet de créer des graphiques et des visualisations pour des publications en s'appuyant sur le package *ggplot2*.
— Le package *tidyverse* (Wickham et al. (2019)) : regroupe plusieurs packages tels que *ggplot2*, *dplyr*…
— Le package *pwr* (Champely (2020)) : permet d'effectuer une analyse de puissance.
— Le package *effectsize* (Ben-Shachar, Makowski, and Lüdecke (2020)) : permet de calculer l'effet de la taille d'un échantillon.
— Le package *reshape* (Wickham (2007)) : permet de restructurer des données.
— Le package *car* (Fox and Weisberg (2019)) : regroupe des procédures pour la régression.
— Le package *multcomp* (Hothorn, Bretz, and Westfall (2008)) : permet d'effectuer des tests multiples et de calculer les intervalles de confiance dans le cas de modèles paramétriques.
— Le package *DescTools* (Andri et mult. al. (2020)) : outils pour effectuer de la statistique descriptive.
— Le package *coin* (Hothorn et al. (2008)) : permet de faire des tests de permutation.
— Le package *MASS* (Venables and Ripley (2002)) : on utilisera le package pour charger le jeux de données *anorexia*.
— Le package *PMCMRplus* (Pohlert (2020)) : effectuer des comparaisons multiples deux à deux (non paramétrique)
— Le package *ggstatsplot* (Patil (2018)) : graphiques avec des informations statistiques

Chapitre 3

Tester la normalité et l'homogénéité des variances

Avant d'effectuer certains tests statistiques, il est nécessaire de vérifier certaines conditions d'application comme la normalité et l'homogénéité des variances.

3.1 Le test de normalité

On souhaite vérifier qu'un échantillon est issu d'une population suivant une loi normale. Les hypothèses statistiques (cas d'un test bilatéral) sont les suivantes :

- H_0 : La variable quantitative suit une loi normale.
- H_1 : La variable quantitative ne suit pas une loi normale.

Un des tests de normalité qu'on peut utiliser est le test de Shapiro-Wilk.

Reprenons le jeu de données *iris*. Testons si la longueur des pétales pour chaque espèce d'iris suit une distribution normale.

```
data(iris)
with(iris, tapply(Petal.Length, Species, shapiro.test))
```

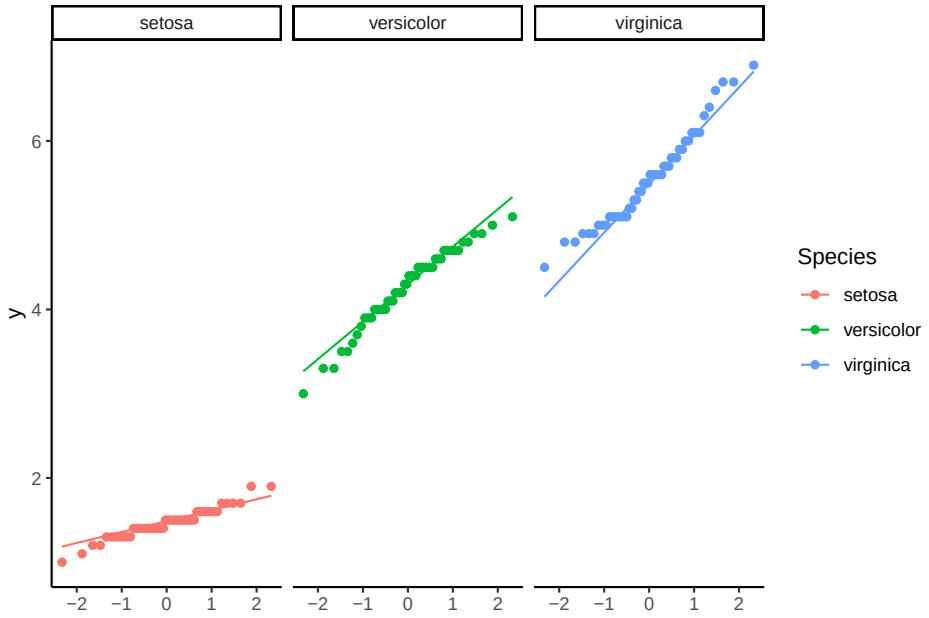
```
## $setosa
##
##  Shapiro-Wilk normality test
```

```
##
## data:  X[[i]]
## W = 0.95498, p-value = 0.05481
##
##
## $versicolor
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.966, p-value = 0.1585
##
##
## $virginica
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96219, p-value = 0.1098
```

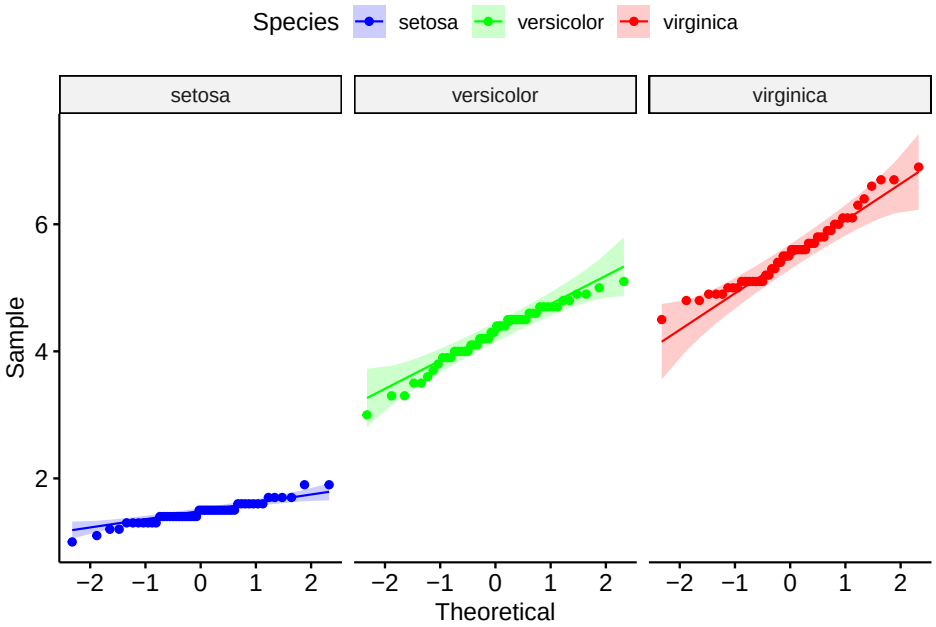
Dans chacun des cas, on ne rejette pas l'hypothèse nulle. Chacun des échantillons suit une loi normale.

Représentation graphique :

```
# On charge la bibliothèque tidyverse dans
# laquelle se trouve le package ggplot2
library(tidyverse)
ggplot(iris, aes(sample=Petal.Length, color=Species)) +
  stat_qq() + stat_qq_line() + xlab("") +
  facet_wrap(~ Species) + theme_classic()
```




```
# Autre manière en utilisant le package ggpubr
library(ggpubr)
ggqqplot(iris, x = "Petal.Length", color="Species",
         palette=c("blue","green","red")) +
  facet_wrap(~ Species)
```



D'autres tests de la normalité existent (Anderson-Darling, Agostino…). On peut utiliser la librairie **nortest** avec la fonction *ad.test* (test d'Anderson-Darling) et la la librairie **fBasics** avec la fonction *dagoTest()* (test d'Agostino).

```
library(nortest)
with(iris, tapply(Petal.Length, Species, ad.test))
```

```
library(fBasics)
with(iris, tapply(Petal.Length, Species, dagoTest))
```

3.2 L'homogénéité des variances

On souhaite tester l'homogénéité des variances entre groupes.

Condition d'application : Les échantillons de chaque groupe suivent une loi normale.

Hypothèses d'application :

- H_0 : Les variances des groupes sont homogènes.
- H_1 : Les variances des groupes ne sont pas homogènes.

Plusieurs tests sont disponibles :

- Le test F de Fisher valable seulement lorsqu'on a deux groupes. Sous **R**, on utilisera la fonction *var.test()*.
- Le test de Bartlett (2 groupes et plus) : On pourra utiliser la fonction *bartlett.test()*.
- Le test de Levene (2 groupes et plus) : On pourra utiliser le package *car* et la fonction *leveneTest()*.

```
with(sleep, tapply(extra,group, shapiro.test))
```

```
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92581, p-value = 0.4079
##
##
## $`2`
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.9193, p-value = 0.3511
```

```
# On charge le package car pour effectuer le test de Levene
library(car)
leveneTest(extra ~ group, data=sleep)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.2482 0.6244
##       18
```

```
bartlett.test(extra ~ group, data=sleep)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  extra by group
## Bartlett's K-squared = 0.10789, df = 1, p-value = 0.7426
```

Conclusion : il y a homogénéité des variances selon les groupes.

Chapitre 4

Les tests paramétriques avec R

Un test paramétrique repose sur le fait que l'on va effectuer une hypothèse paramétrique sur la distribution des données sous l'hypothèse nulle. L'échantillon est issu d'une population suivant une distribution appartenant à une loi connue (normale, Poisson…). Ils sont généralement utilisés lorsque la taille de l'échantillon est supérieure à 30.

4.1 Comparaison de moyennes

Comparaison de moyennes de deux échantillons indépendants : le test de Student et le test de Welch

Reprenons le jeu de données *iris* en sélectionnant que les espèces virginica et versicolor puis comparons les longueurs des pétales. Concrètement, pour comparer deux moyennes, on doit avoir une variable quantitative (comme la variable *Petal.Length*) et une variable qualitative composée de deux classes (espèce Virginica et espèce Versicolor). Le test de Student permet de tester l'égalité de la longueur des pétales selon les deux espèces. On notera m_1 la moyenne de la longueur des pétales pour l'espèce virginica et m_2 la moyenne de la longueur des pétales pour l'espèce versicolor. n_1 est nombre d'observations du premier groupe (espèce virginica) et n_2 le nombre d'observations du second groupe (espèce versicolor).

Conditions d'application du test de Student :

— Indépendance des observations

— La distribution de la variable quantitative pour chacune des classes de la variable qualitative est normale.

— Homogénéité des variances entre les deux groupes de la variable quantitative.

Hypothèses d'application dans le cas d'un test bilatéral

— $H_0 : m_1 = m_2$

— $H_1 : m_1 \neq m_2$

Hypothèses d'application dans le cas d'un test unilatéral (gauche)

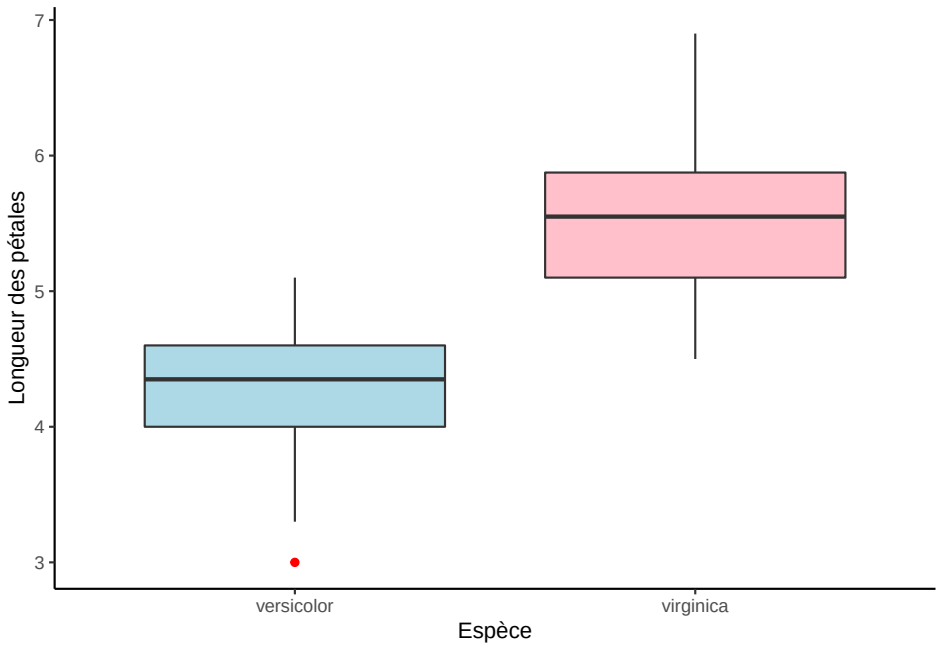
— $H_0 : m_1 = m_2$

— $H_1 : m_1 < m_2$

La variable de test suit une loi de Student à n_1+n_2-2 ddl.

```
indice <- which(iris$Species=="virginica" | iris$Species=="versicolor")
datairis <- iris[indice,]
datairis$Species <- factor(datairis$Species, exclude = NULL)

ggplot(datairis, aes(x=Species, y=Petal.Length)) +
  geom_boxplot(fill=c("lightblue","pink"), outlier.colour="red") +
  xlab("Espèce") + ylab("Longueur des pétales") + theme_classic()
```

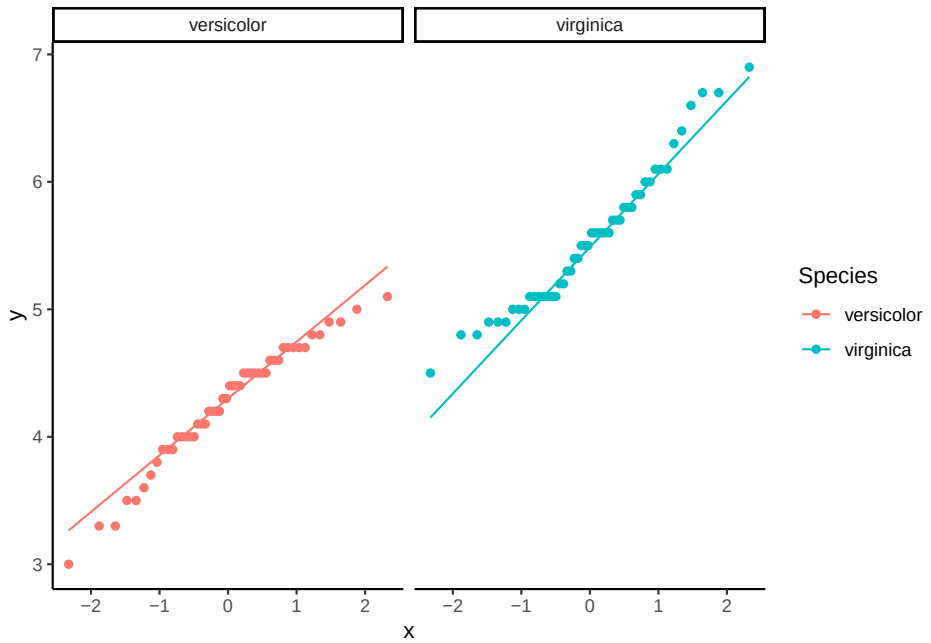


Vérifions les conditions d'application du test.

```
#verification de la normalité
with(datairis, tapply(Petal.Length, Species, shapiro.test))
```

```
## $versicolor
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.966, p-value = 0.1585
##
##
## $virginica
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96219, p-value = 0.1098
```

```
ggplot(datairis, aes(sample=Petal.Length, color=Species)) +
  stat_qq() + stat_qq_line() +
  facet_wrap(~ Species) + theme_classic()
```



```
# Homogeneite des variances
bartlett.test(Petal.Length~Species, data = datairis)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Petal.Length by Species
## Bartlett's K-squared = 1.249, df = 1, p-value = 0.2637
```

```
var.test(Petal.Length~Species, data = datairis)
```

```
##
##  F test to compare two variances
##
## data:  Petal.Length by Species
## F = 0.72497, num df = 49, denom df = 49, p-value = 0.2637
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.411402 1.277530
## sample estimates:
## ratio of variances
##          0.7249678
```

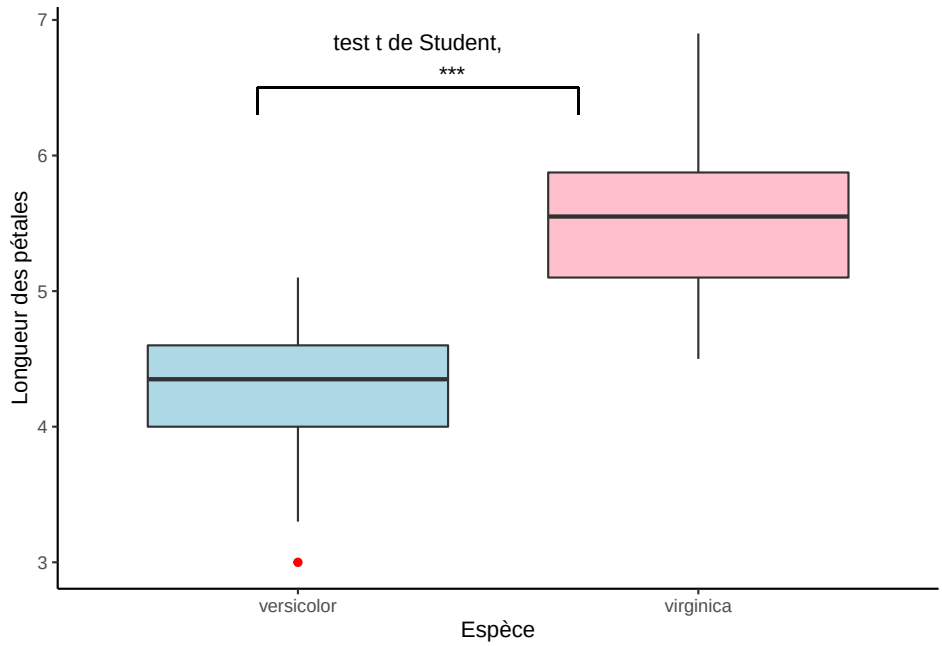
Dans notre cas, l'hypothèse de normalité est respectée ainsi que celle d'homogénéité des variances. Dans le cas où les variances ne sont pas égales, on peut utiliser le test t de Welch si les autres conditions d'application (normalité, observations indépendantes) sont respectées. Si ce n'est pas le cas, il faudra utiliser un test non paramétrique. Dans **R**, la fonction est toujours *t.test()* avec l'option *var.equal=FALSE*. C'est en fait l'option par défaut.

Effectuons le test de Student (bilateral).

```
t.test(Petal.Length~Species, data=datairis,
       alternative="two.sided", paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Petal.Length by Species
## t = -12.604, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group versicolor and group
## 95 percent confidence interval:
##  -1.495426 -1.088574
## sample estimates:
## mean in group versicolor  mean in group virginica
##                    4.260                    5.552
```

```
ggplot(datairis, aes(x=Species, y=Petal.Length)) +
  geom_boxplot(fill=c("lightblue","pink"), outlier.colour="red") +
  xlab("Espèce") + ylab("Longueur des pétales") +
  geom_segment(x=0.9, xend=1.7, y=6.5, yend=6.5, col="black") +
  geom_segment(x=0.9, xend=0.9, y=6.5, yend=6.3, col="black") +
  geom_segment(x=1.7, xend=1.7, y=6.5, yend=6.3, col="black") +
  annotate("text", x = 1.3, y = 6.72, label = "test t de Student,
           ***") + theme_classic()
```



Conclusion : On rejette l'hypothèse nulle (au risque de 5%), les moyennes des longueurs des

pétales ne sont pas égales entre l'espèce versicolor et l'espèce virginica.

Effectuons un test de Student unilatéral (à gauche).

```
t.test(Petal.Length~Species, data=datairis,
       alternative="less", paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Petal.Length by Species
## t = -12.604, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group versicolor and gr
## 95 percent confidence interval:
##       -Inf -1.121779
## sample estimates:
## mean in group versicolor  mean in group virginica
##                    4.260                    5.552
```

Conclusion : On rejette l'hypothèse nulle (au risque de 5%), la moyenne des longueurs des pétales de l'espèce versicolor est inférieure à celle de l'espèce virginica.

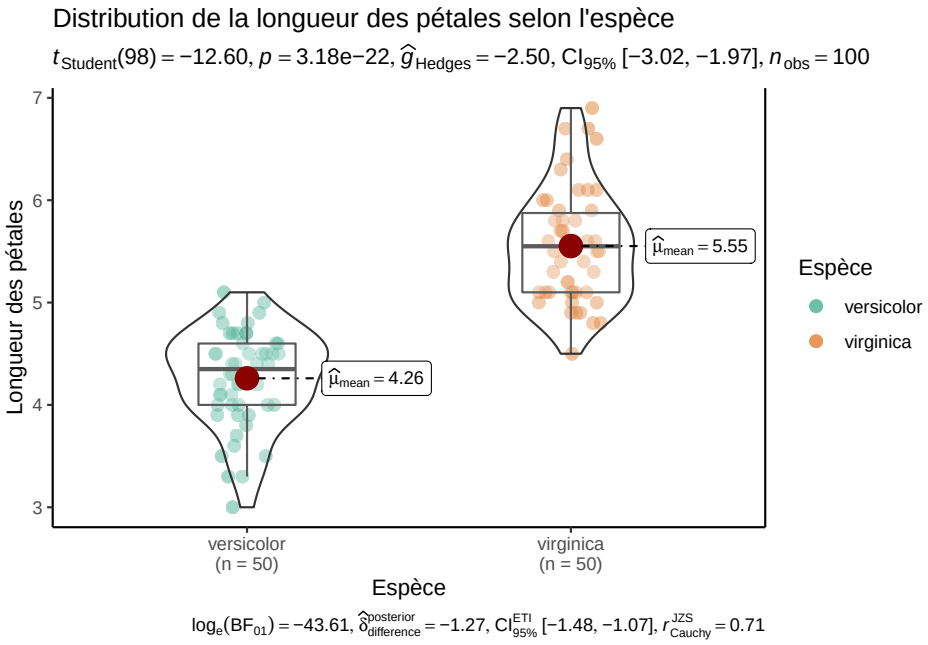
Analysons la puissance du test. Il faut d'abord calculer la mesure de l'effet de la taille (Cohen (2013)) en utilisant l'indice de Cohen. On utilise le package **effectsize** et la fonction *cohens_d*.

```
library(effectsize)
#Calcul de l'indice de Cohen en utilisant le package effsize
indice <-
  cohens_d(Petal.Length ~ Species,
           data=datairis, paired=FALSE)$Cohens_d
```

```
library(pwr)
#Calcul de la puissance en utilisant le package pwr
(puissance <- pwr.t.test(n=50, d=indice, sig.level=0.05,
                         type="two.sample",
                         alternative="two.sided"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 50
##               d = 2.520756
##       sig.level = 0.05
##           power = 1
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Pour information, le package **ggstatsplot** offre des outils intéressants de visualisation de résultats de tests :



Comparaison de moyennes de deux échantillons appariés : le test de Student apparié et le test de Welch apparié

Un échantillon apparié est un groupe d'individus qui ont été mesurés plusieurs fois.

Conditions d'application du test de Student pour échantillons appariés :

- Indépendance des observations
- La variable quantitative est appariée, on fait plusieurs mesures sur le même individu

- La distribution des différences doit suivre une loi normale.
- Chaque individus doit se retrouver dans chacune des deux classes de la variable qualitative.

Hypothèses d'application dans le cas d'un test bilatéral

- $H_0 : m_1 = m_2$
- $H_1 : m_1 \neq m_2$

Prenons le jeu de données *anorexia* et comparons le poids avant et après traitement.

Commençons par vérifier les conditions d'application du test.

```
library(MASS)
# Normalité
diff <- anorexia$Postwt-anorexia$Prewt
shapiro.test(diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff
## W = 0.97466, p-value = 0.1544
```

Les conditions d'application du test sont vérifiées. Effectuons le test de Student pour échantillons appariés (test bilatéral).

```
t.test(anorexia$Postwt,anorexia$Prewt,paired = TRUE,
       alternative="two.sided")
```

```
##
##  Paired t-test
##
## data:  anorexia$Postwt and anorexia$Prewt
## t = 2.9376, df = 71, p-value = 0.004458
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
```

```
##  0.8878354 4.6399424
## sample estimates:
## mean difference
##        2.763889
```

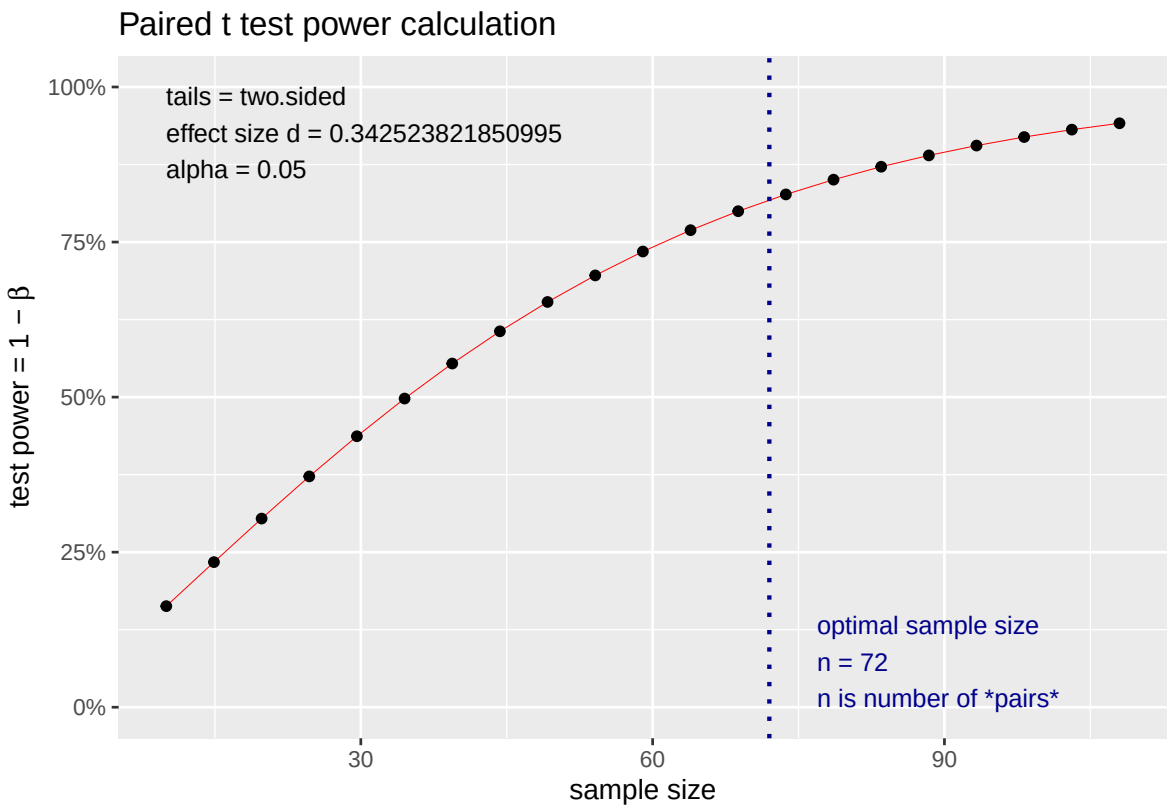
Conclusion : On rejette H_0 au seuil de 5%. La différence de poids avant et après traitement est statistiquement différente.

Calcul de la puissance :

```
#Calcul de l'indice de Cohen (package effectsize)
valeurs <- c(anorexia$Prewt,anorexia$Postwt)
groupes <- c(rep("Prewt",72),rep("Postwt",72))
anorex <- data.frame(groupes,valeurs)
indice <- hedges_g(valeurs~groupes, paired=TRUE, data=anorex)$Hedges_g
#Calcul de la puissance (package pwr)
(puissance <- pwr.t.test(n=72, d=indice,type="paired",
                         alternative = "two.sided"))
```

```
##
##      Paired t test power calculation
##
##               n = 72
##               d = 0.3425238
##       sig.level = 0.05
##           power = 0.8177769
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```
plot(puissance)
```



4.2 Comparaison de proportions d'échantillons indépendants

Comparaison de proportions observées

On dispose de deux variables qualitatives. On va tester si les proportions (probabilité de succès) dans plusieurs groupes sont les mêmes.

Conditions d'application du test de proportion :

- Indépendance des échantillons aléatoires
- L'effectif de chaque modalité de chaque variable qualitative est supérieur ou égal à 5

Prenons un exemple. On cherche à savoir sir la proportion de malades est la même chez les hommes et chez les femmes

```
grippe<-matrix(c(34,50,23,61),2)
dimnames(grippe) <- list(c("homme","femme") ,c("malade","non malade"))
print(grippe)
```

```
##        malade non malade
```

```
## homme      34          23
## femme      50          61
```

Dans ce cas, les hypothèses d'application (test bilatéral) sont :

- H_0 : La proportion de malades est la même chez les hommes et les femmes.
- H_1 : La proportion de malades n'est pas la même chez les hommes et les femmes.

Dans le logiciel **R**, la fonction qu'on utilisera est la fonction *prop.test()*. Par défaut, la fonction applique la correction de Yates (si un des effectifs théoriques est faible).

```
prop.test(grippe, alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  grippe
## X-squared = 2.6553, df = 1, p-value = 0.1032
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02467769  0.31675925
## sample estimates:
##    prop 1    prop 2
## 0.5964912 0.4504505
```

```
prop.test(c(34,50),c(57,111),correct=T)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(34, 50) out of c(57, 111)
## X-squared = 2.6553, df = 1, p-value = 0.1032
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02467769  0.31675925
```

```
## sample estimates:
##    prop 1    prop 2
## 0.5964912 0.4504505
```

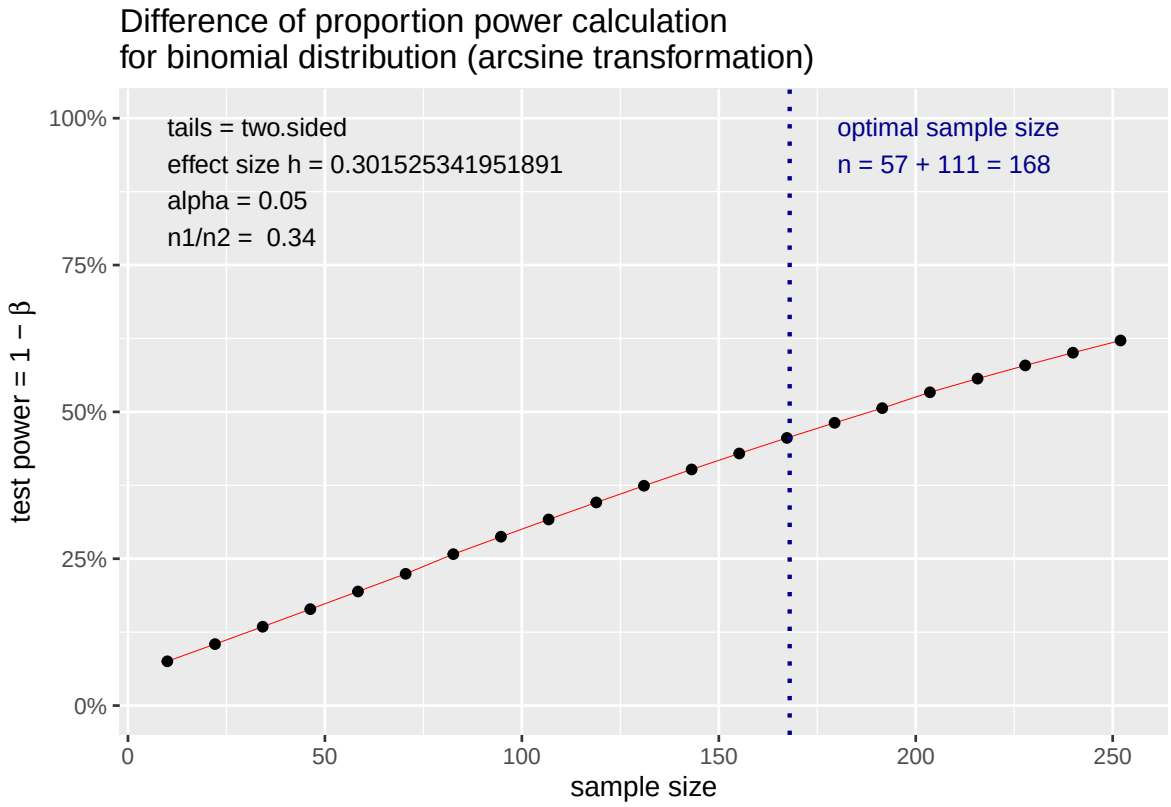
On ne rejette pas H_0. On peut conclure (au seuil de 5%) que la proportion de malades n'est pas significativement différente selon le sexe.

Analysons la puissance.

```
# Calcul de l'effet taille h pour 2 proportions (package pwr)
h<-ES.h(0.60,0.45)
# Calcul de la puissance (package pwr)
(puissance <- pwr.2p2n.test(h, n1=57,n2=111,
              sig.level=0.05,
              alternative="two.sided"))
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsi
##
##               h = 0.3015253
##              n1 = 57
##              n2 = 111
##       sig.level = 0.05
##           power = 0.4564511
##     alternative = two.sided
##
## NOTE: different sample sizes
```

```
plot(puissance)
```



Le test d'indépendance : le test du χ_2

On cherche à tester le lien entre deux variables qualitatives.

Conditions d'application du test du χ_2 :

- Indépendance des échantillons aléatoires
- L'effectif de chaque modalité de chaque variable qualitative est supérieur ou égal à 5

Hypothèses d'application dans le cas d'un test bilatéral

- H_0 : les variables qualitatives sont indépendantes
- H_1 : les variables qualitatives sont liées

Prenons un exemple en prenant le jeu de données Titanic. On souhaite savoir si il existe une liaison entre la classe et le fait d'être survivant au naufrage

```
transfotitanic <- data.frame(Titanic)
# On charge le package reshape pour modifier les
#données avec la fonction untable
```

```
library(reshape)
titanic <- data.frame(untable(transfotitanic[,c(1,2,3,4)],
                              num=transfotitanic[,5]))
# Creation du tableau de contingence
tabletest <- table(titanic$Survived,titanic$Class)
(test <- chisq.test(tabletest))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabletest
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

```
# Comptages observés
test$observed
```

```
##
##       1st 2nd 3rd Crew
##   No  122 167 528  673
##   Yes 203 118 178  212
```

```
# Comptages attendus sous H0
round(test$expected,0)
```

```
##
##       1st 2nd 3rd Crew
##   No  220 193 478  599
##   Yes 105  92 228  286
```

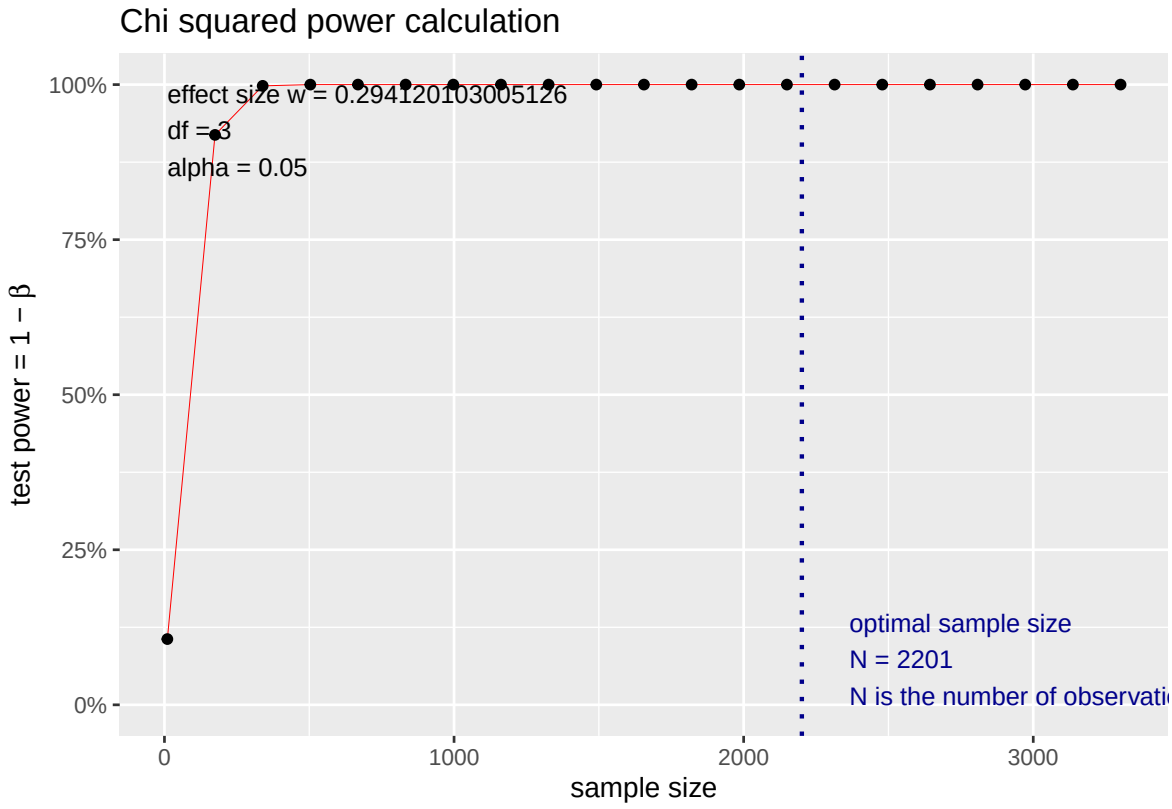
On rejette H_0, il existe une liaison entre la classe et le fait d'être survivant au naufrage.

Effectuons une analyse de puissance.

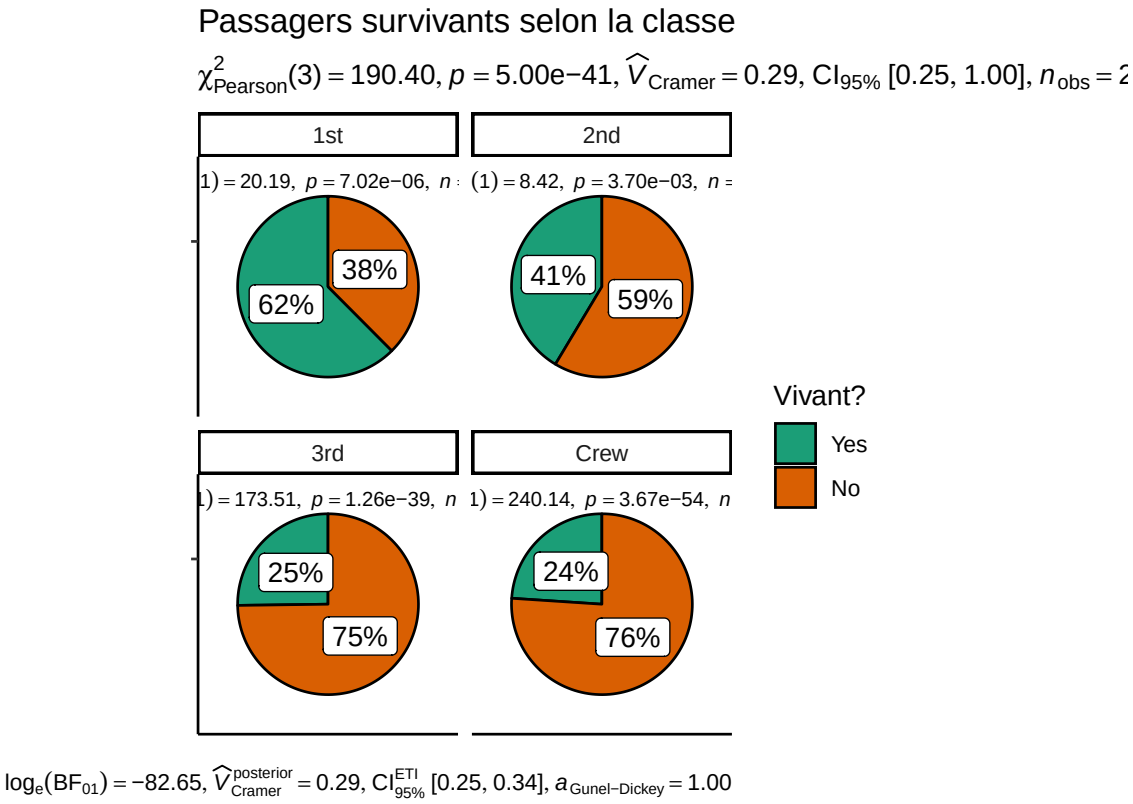
```r
# Calcul de la taille
w <- ES.w2(prop.table(tabletest))
# Calcul de puissance
(puissance <- pwr.chisq.test(w, df=3, N=2201,
                             sig.level=0.05))
```

```
##
##      Chi squared power calculation
##
##               w = 0.2941201
##               N = 2201
##              df = 3
##       sig.level = 0.05
##           power = 1
##
## NOTE: N is the number of observations
```

```r
plot(puissance)
```



De la même manière, le package **ggstatsplot** offre des outils intéressants de visualisation de résultats de tests :



Chapitre 5

Comparaison de plusieurs moyennes : l'ANOVA à un facteur

5.1 Introduction

Prenons les données *PlantGrowth*. On cherche à savoir si le poids des plantes varie selon le traitement (ctrl, trt1 et trt2). La variable qualitative qu'on va utiliser a trois modalités. L'analyse de variance à un facteur consiste à comparer plus de deux moyennes. On dispose d'une variable quantitative et d'une variable qualitative à k classes (>2).

5.2 Les conditions d'application du test

- Indépendance des observations
- La distribution de la variable quantitative pour chacune des classes de la variable qualitative est normale.
- Normalité des résidus (après avoir effectué l'ANOVA)
- Homogénéité des variances entre les groupes de la variable quantitative. Si ce n'est pas le cas, on pourra utiliser une correction de Welch.
- Absence d'autocorrélation des résidus (après avoir effectué l'ANOVA, spécifique aux données temporelles)

5.3 Hypothèses d'application dans le cas d'un test bilatéral

- $H_0 : m_1$=$m_2 = ...m_k$
- H_1 : au moins une moyenne diffère

5.4 Application sous R

Vérifions la première hypothèse d'application à savoir la normalité du poids pour chacun des traitements (ctrl, trt1, trt2).

```
# Vérification de la normalité
with(PlantGrowth, tapply(weight,group,shapiro.test))
```

```
## $ctrl
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95668, p-value = 0.7475
##
##
## $trt1
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.93041, p-value = 0.4519
##
##
## $trt2
##
##  Shapiro-Wilk normality test
##
```

```
## data:  X[[i]]
## W = 0.94101, p-value = 0.5643
```

On conclut que la distribution du poids est normale pour chacun des traitements.

Vérifions l'homogénéité des variances du poids entre les groupes.

```
# Homogénéité des variances
leveneTest(weight ~ group, data=PlantGrowth)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.1192 0.3412
##       27
```

On conclut à une homogénéité des variances.

On cherche à savoir si le poids diffère selon les groupes. Effectuons à présent l'ANOVA. On utilise la fonction *aov()*.

```
# ANOVA
testanova <- aov(weight ~ group, data=PlantGrowth)
summary(testanova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

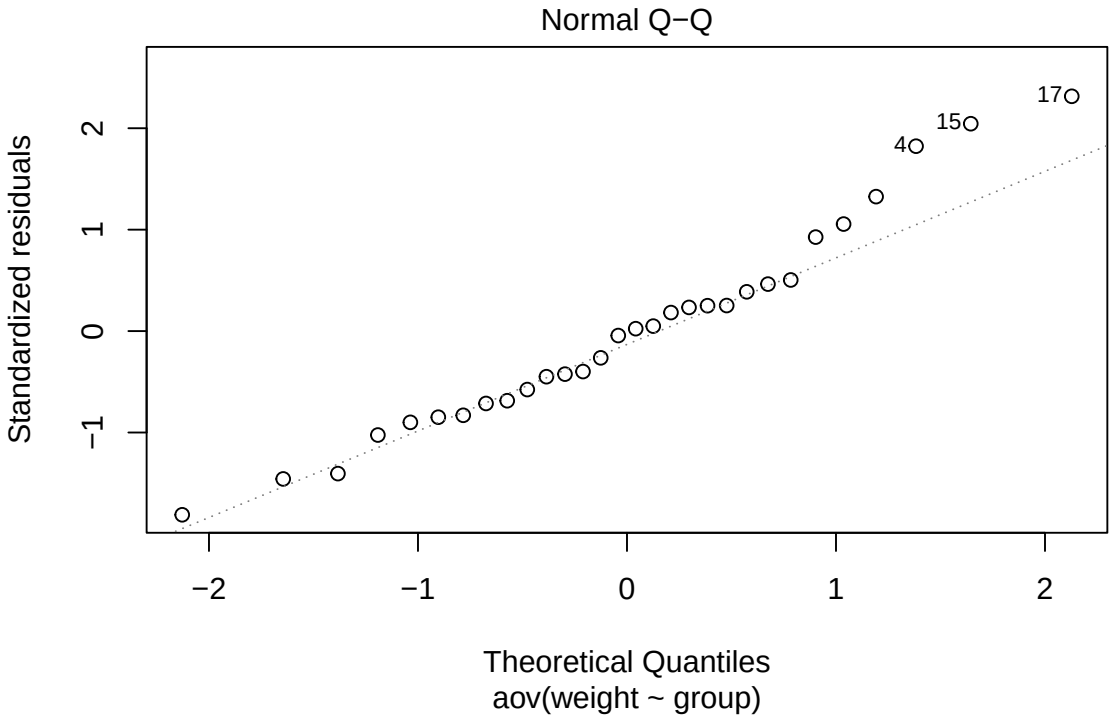
Vérifions la normalité des résidus issus du modèle ANOVA.

```
shapiro.test(testanova$residuals)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  testanova$residuals
## W = 0.96607, p-value = 0.4379
```

```
plot(testanova,2)
```



Vérifions qu'il y a absence d'autocorrélation. Pour cela, on utilise le package **car** et la fonction *durbinWatsonTest*.

```
durbinWatsonTest(testanova)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.3907766       2.70399   0.114
##  Alternative hypothesis: rho != 0
```

Il y a absence d'autocorrélation entre les résidus et le groupe.

On conclut qu'il existe une différence significative de poids de plantes en fonction du traitement.

Dans le cas où il n'y aurait pas égalité des variances, on utilise la fonction *oneway.test()* pour effectuer l'ANOVA.

5.5 Analyse de puissance

```
# Calculer la taille de l'effet
library(effectsize)
f = effectsize(testanova, type="eta")
```

```
## For one-way between subjects designs, partial eta squared is equivalent to eta squa
## Returning eta squared.
```

```
library(pwr)
pwr.anova.test(k=3,n=10,f=0.59, sig.level=0.05, power=NULL)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 10
##               f = 0.59
##       sig.level = 0.05
##           power = 0.7861906
##
## NOTE: n is number in each group
```

5.6 Tests post hoc

Si on souhaite savoir comparer les moyennes 2 à 2, il est nécessaire d'effectuer un test post hoc et des comparaisons multiples (voir chapitre 1 sur les tests multiples).

```
pairwise.t.test(PlantGrowth$weight, PlantGrowth$group,
                p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
```

```
##
## data:  PlantGrowth$weight and PlantGrowth$group
##
##      ctrl  trt1
## trt1 0.583 -
## trt2 0.263 0.013
##
## P value adjustment method: bonferroni
```

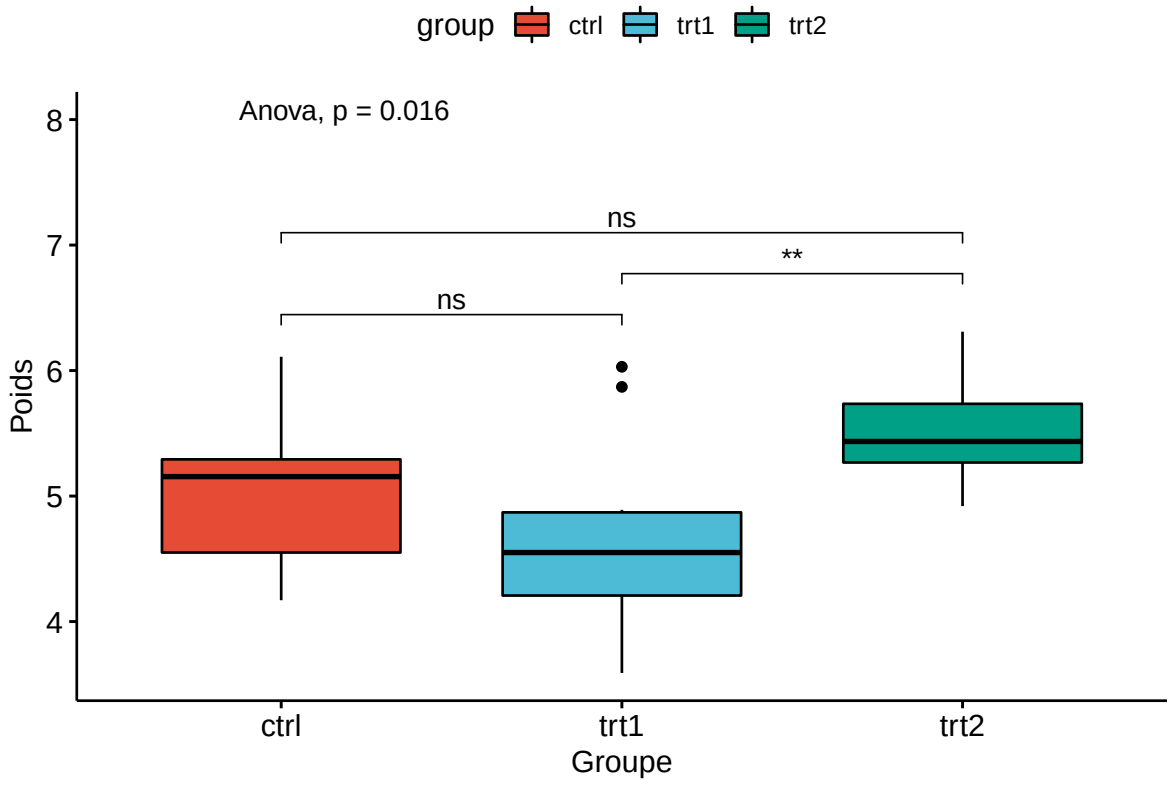
```
TukeyHSD(testanova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ group, data = PlantGrowth)
##
## $group
##             diff        lwr       upr     p adj
## trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
## trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
## trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

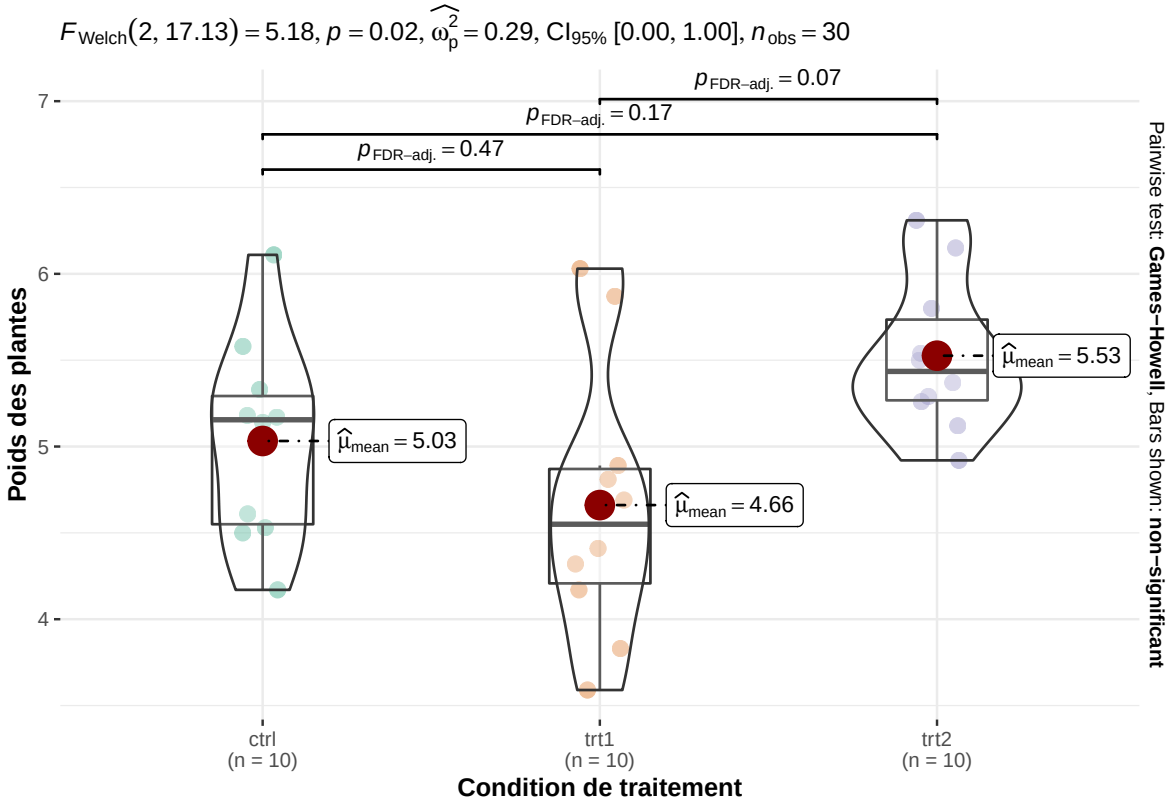
On peut aussi utiliser le package **multcomp** et la fonction *glht*.

On peut effectuer une représentation graphique (package **ggpubr**).

```
comparaisons <- list( c("ctrl","trt1"),
                      c("trt1","trt2"),
                      c("trt2", "ctrl"))
ggboxplot(PlantGrowth, x = "group", y = "weight", fill = "group",
          palette = "npg", ylab="Poids", xlab="Groupe")+
  stat_compare_means(comparisons = comparaisons, label = "p.signif")+
  stat_compare_means(label.y = 8, method="anova")
```



De la même manière, le package **ggstatsplot** offre des outils intéressants de visualisation de résultats de tests :



Chapitre 6

Les tests non paramétriques avec R

Contrairement aux tests paramétriques, les tests non paramétriques ne font aucune hypothèse sur la distribution des données. Ils sont particulièrement utiles lorsqu'on dispose d'un faible échantillonnage. Une alternative serait d'utiliser des tests de permutation (package **coin**).

6.1 Comparaisons de deux échantillons indépendants : le test de Mann-Whitney-Wilcoxon

On dispose d'une variable quantitative (qui peut être qualitative ordinale) et d'une variable qualitative à deux modalités. On notera m_1 la médiane de la variable quantitative du premier groupe et m_2 la médiane de la variable quantitative du second groupe.

Conditions d'application du test de Wilcoxon :

- Indépendance des observations
- La distribution pour chaque groupe est continue
- Même forme de la distribution entre chaque groupe

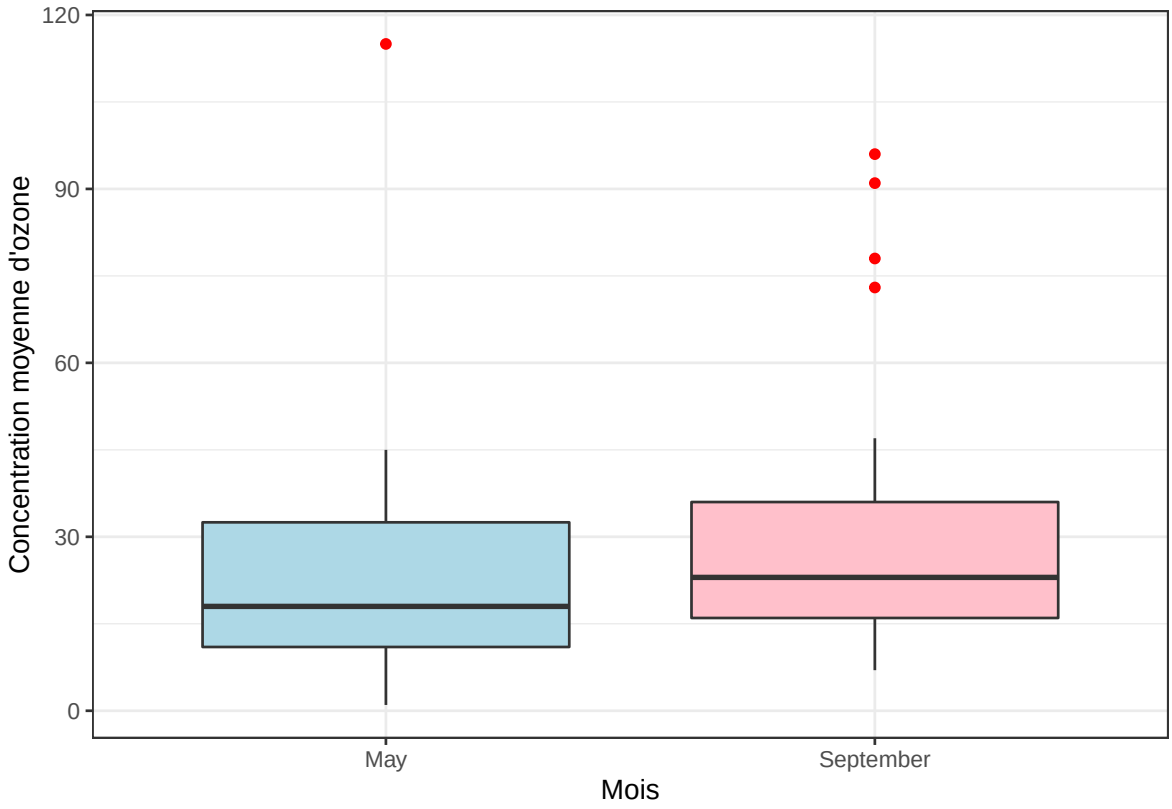
Hypothèses d'application dans le cas d'un test bilatéral

- $H_0 : m_1 = m_2$
- $H_1 : m_1 \neq m_2$

Prenons un exemple avec le jeu de données *airquality*. On souhaiter comparer la concentration moyenne d'ozone entre le mois de mai et le mois de septembre.

```
# Utilisation du package tidyverse (en particulier dplyr)
airquality$Month <- factor(airquality$Month, labels=month.name[5:9])
indice <- which(airquality$Month=="May"| airquality$Month=="September")
datawilcoxon <- airquality[indice,]
datawilcoxon$Month <- factor(datawilcoxon$Month, exclude = NULL)
datawilcoxon <- na.omit(datawilcoxon)

ggplot(datawilcoxon, aes(x=Month, y=Ozone)) +
  geom_boxplot(fill=c("lightblue","pink"), outlier.colour="red") +
  xlab("Mois") + ylab("Concentration moyenne d'ozone") + theme_bw()
```



Commençons par vérifier l'hypothèse de normalité.

```
# verification de la normalité si on effectue un test parametrique
with(datawilcoxon, tapply(Ozone, Month, shapiro.test))
```

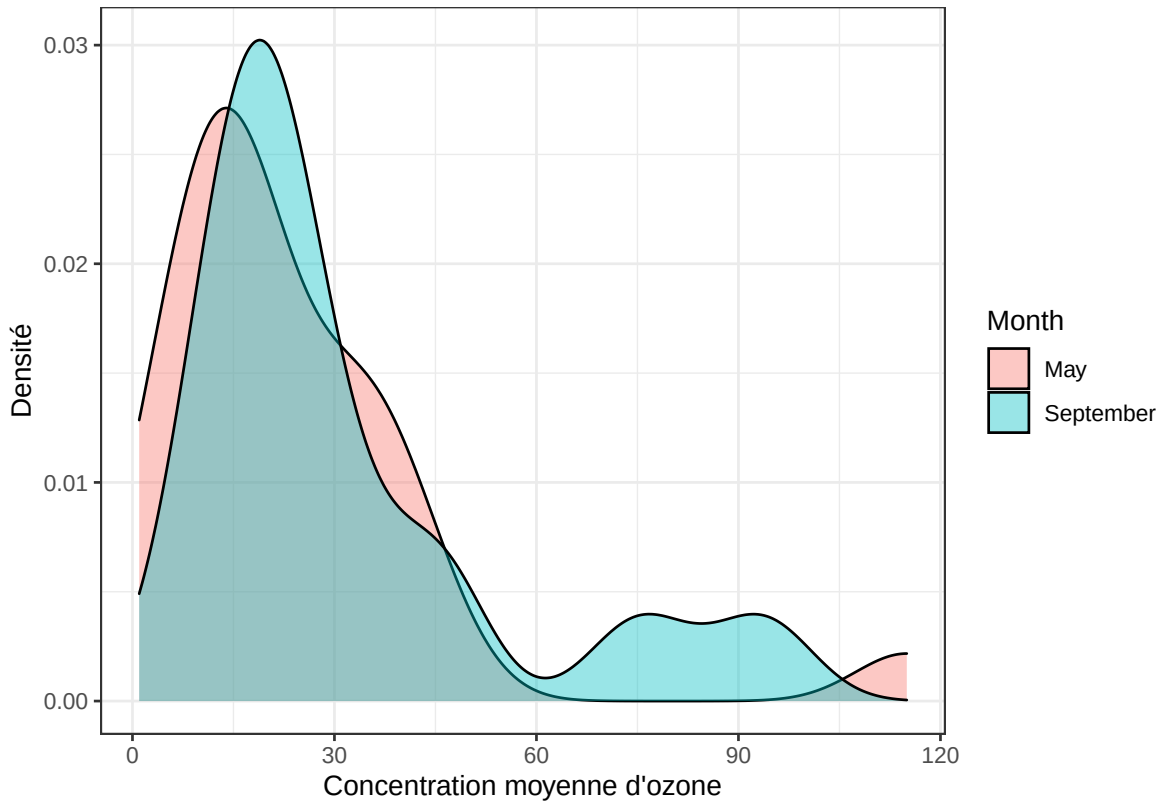
```
## $May
##
##  Shapiro-Wilk normality test
```

```
##
## data:  X[[i]]
## W = 0.71273, p-value = 1.491e-05
##
##
## $September
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.78373, p-value = 4.325e-05
```

L'hypothèse de normalité n'est pas vérifiée.

Regardons la distribution des deux groupes.

```
# Densité
ggplot(datawilcoxon, aes(x=Ozone, fill=Month)) +
  geom_density(alpha=.4) + xlab("Concentration moyenne d'ozone") +
  ylab("Densité") + theme_bw()
```



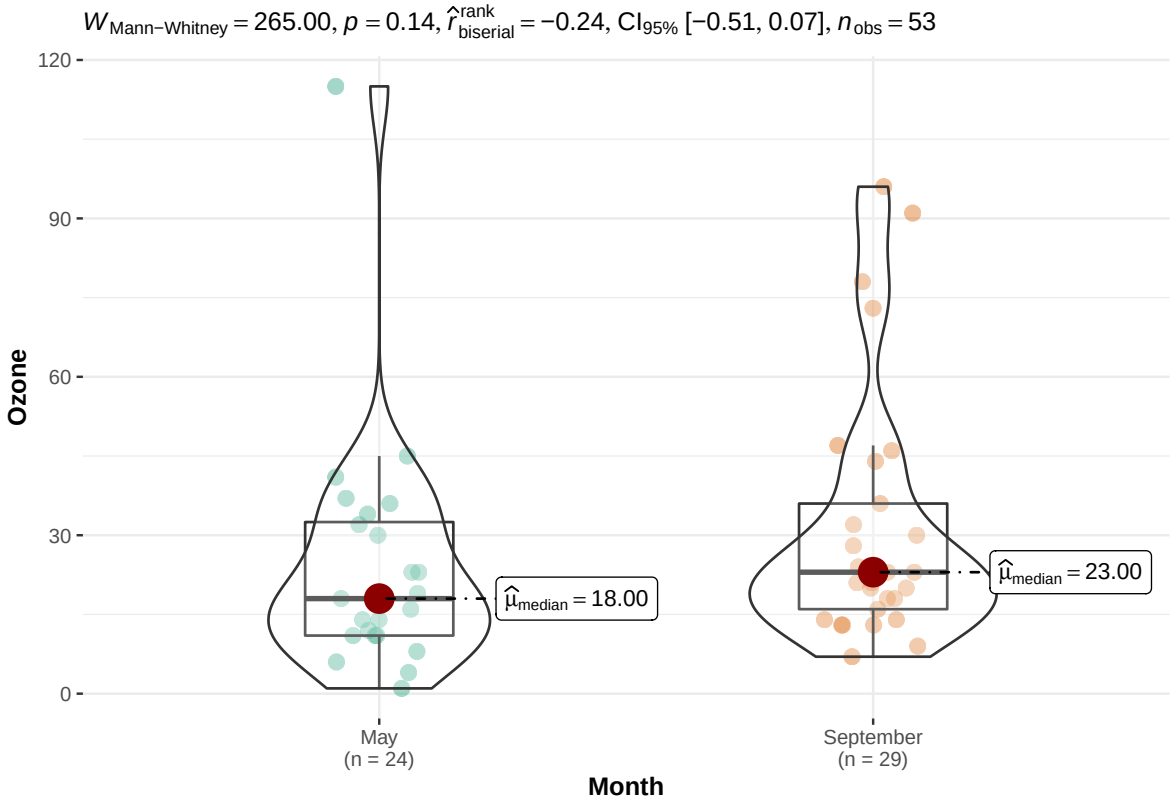
Nous allons donc utiliser un test non paramétrique.

```
wilcox.test(Ozone~Month, data=datawilcoxon, exact=FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Ozone by Month
## W = 265, p-value = 0.1401
## alternative hypothesis: true location shift is not equal to 0
```

On ne rejette pas H_0 au seuil de 5%. Il n'existe pas de différence de concentration d'ozone entre le mois de mai et le mois de septembre.

Le package **ggstatsplot** offre des outils intéressants de visualisation de résultats de tests :



6.2 Comparaisons de deux échantillons appariés : le test des signes de Wilcoxon

On dispose d'une variable quantitative et d'une variable qualitative à deux modalités constituant un série appariée. On notera m_1 la médiane de la variable quantitative du premier groupe et m_2 la médiane de la variable quantitative du second groupe. Prenons un exemple.

Conditions d'application du test de Wilcoxon :

- Indépendance des observations
- La distribution de la différence doit être symétrique.

Hypothèses d'application dans le cas d'un test bilatéral

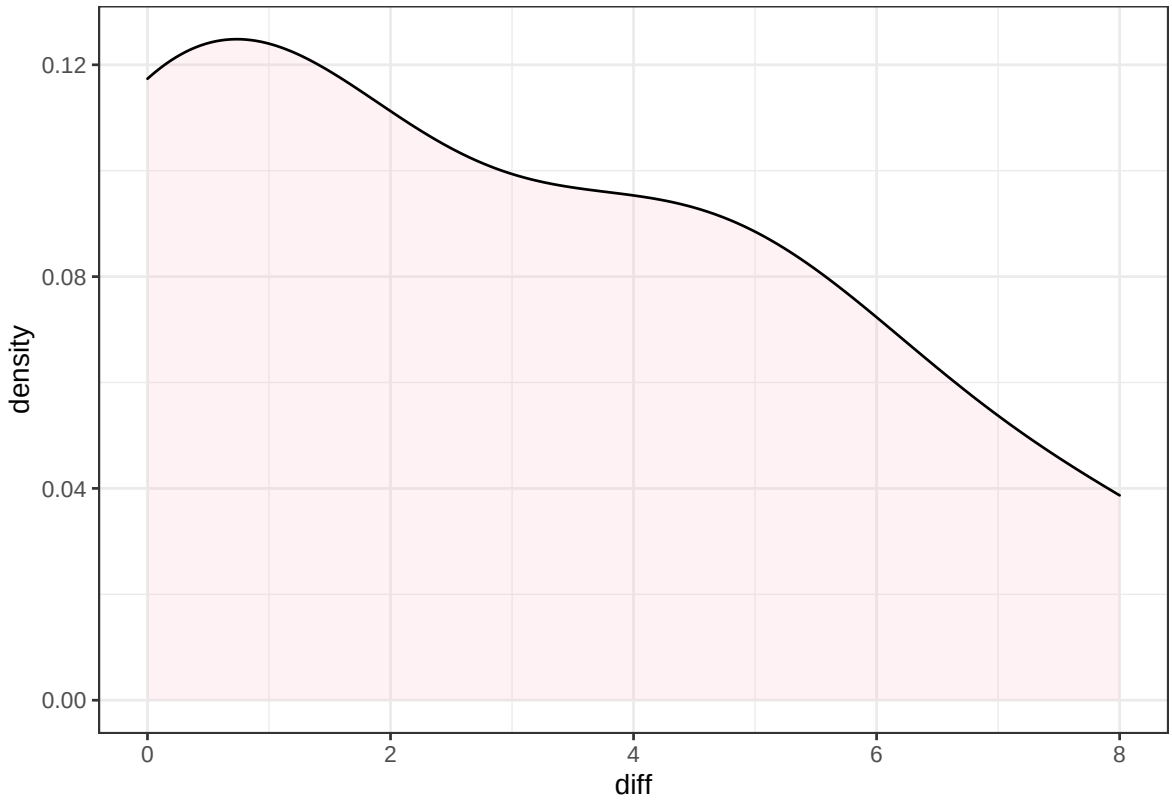
- $H_0 : m_1 = m_2$
- $H_1 : m_1 \neq m_2$

Sous **R**, on utilisera la fonction *wilcox.test()* en spécifiant l'option *paired=TRUE*. Prenons un exemple. Un groupe d'individus est pesé. Après traitement, ils sont à nouveau pesés. Le traitement a-t-il eu un impact sur le poids des individus ? On dispose de deux échantillons appariés.

```
poids <- c(100,90,98,101,103,105,89,90,90,
           100,85,96,100,95, 105,85,90,85)
traitement <- as.factor(c(rep("T1",9), rep("T2",9)))
datasignewilcoxon <- data.frame(poids,traitement)
```

Regardons la distribution de la différence.

```
dataT1 <- datasignewilcoxon[which(traitement=="T1"),]
dataT2 <- datasignewilcoxon[which(traitement=="T2"),]
diff <- data.frame(diff=dataT1$poids-dataT2$poids)
ggplot(diff,aes(x=diff))+ geom_density(alpha=.2, fill="pink") + theme_bw()
```



Nous allons donc utiliser un test des signes qui ne fait pas d'hypothèse sur la symétrie. Pour cela, on utilise la fonction *SignTest()* de la librairie *DescTools*.

```
library(DescTools)
SignTest(dataT1$poids,dataT2$poids)
```

```
##
##  Dependent-samples Sign-Test
##
## data:  dataT1$poids and dataT2$poids
## S = 6, number of differences = 6, p-value = 0.03125
## alternative hypothesis: true median difference is not equal to 0
## 96.1 percent confidence interval:
##  0 5
## sample estimates:
## median of the differences
##                         2
```

Si l'hypothèse sur la symétrie était vérifiée, on aurait utilisé la fonction *wilcox.test()* :

```
# Test des signes de Wilcoxon pour échantillons appariés
wilcox.test(poids~traitement, data=datasignewilcoxon,
            paired=TRUE, exact=FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poids by traitement
## V = 21, p-value = 0.03552
## alternative hypothesis: true location shift is not equal to 0
```

On rejette H_0 au seuil de 5%. On conclut qu'il existe une différence significative de poids avant et après traitement.

6.3 Comparaisons de k échantillons indépendants : le test de Kruskal-Wallis

Le test de Kruskal-Wallis est une alternative à l'ANOVA à un facteur lorsque les conditions d'application ne sont pas appliquées.

On dispose d'une variable quantitative (qui peut être qualitative ordinale) et d'une variable qualitative à k modalités.

Conditions d'application du test de Kruskal-Wallis :

- Indépendance des observations

Hypothèses d'application dans le cas d'un test bilatéral

Si la distribution de la variable quantitative pour chacune des classes de la variable qualitative est identique :

- $H_0 : m_1$=m_2...=m_k
- $H_1 : m_1 = m_2 \neq ... \neq m_2$

Si les distributions de la variable quantitative pour chacune des classes de la variable qualitative sont différentes :

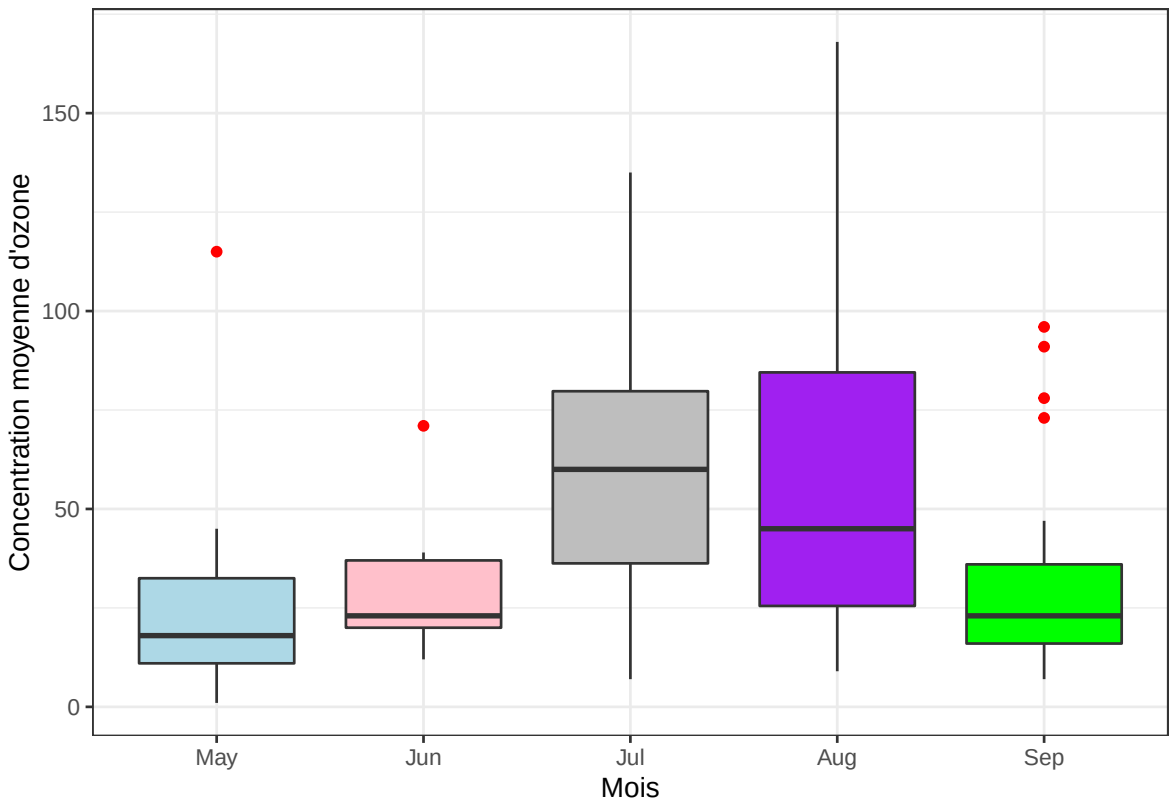
— H_0 : Les différents groupes ont la même distribution.

— H_1 : Au moins un des groupes vient d'une distribution différente.

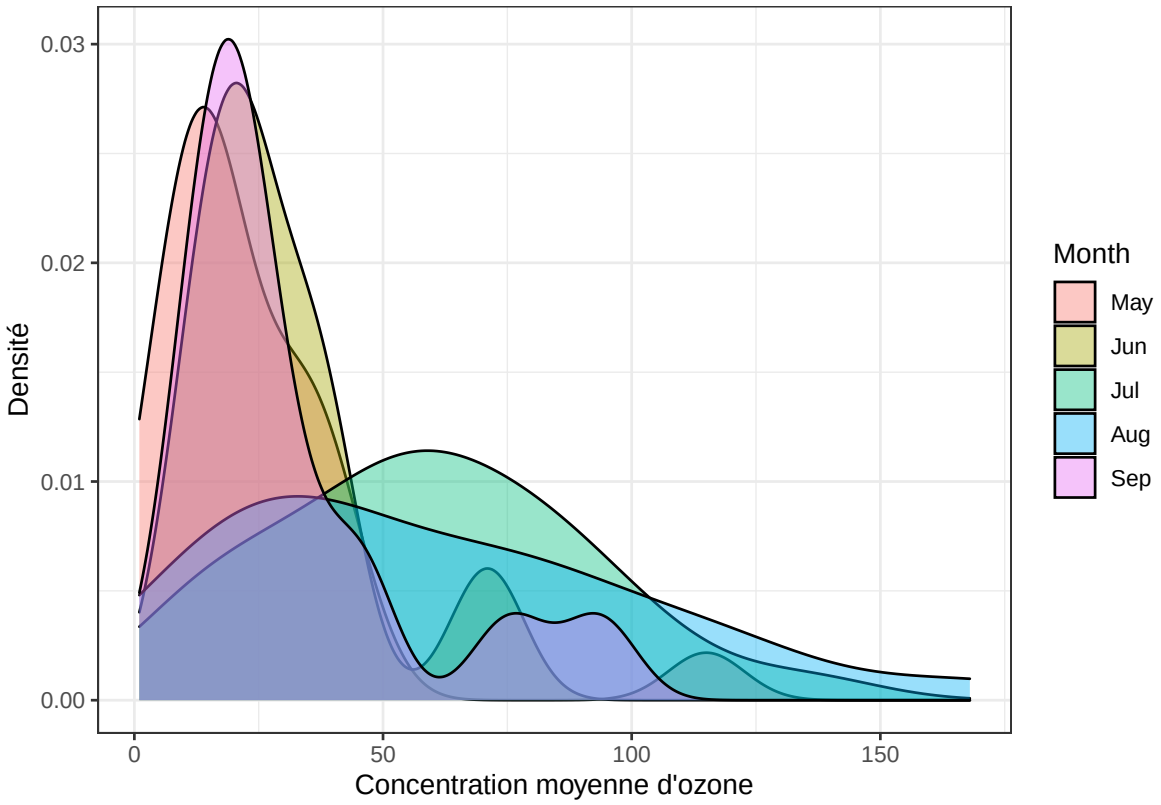
Reprenons le jeu données *airquality*. On cherche à tester si il existe une différence de concentration d'ozone selon le mois.

```
datakruskal <- airquality %>%
  mutate(Month = factor(Month,labels = month.abb[5:9])) %>%
  drop_na()
datakruskal <- na.omit(datakruskal)


ggplot(datakruskal, aes(x=Month, y=Ozone)) +
  geom_boxplot(fill=c("lightblue","pink","grey", "purple","green"),
               outlier.colour="red") +
  xlab("Mois") + ylab("Concentration moyenne d'ozone") + theme_bw()
```




```
ggplot(datakruskal, aes(x=Ozone, fill=Month)) +
  geom_density(alpha=.4) + xlab("Concentration moyenne d'ozone") +
  ylab("Densité") + theme_bw()
```



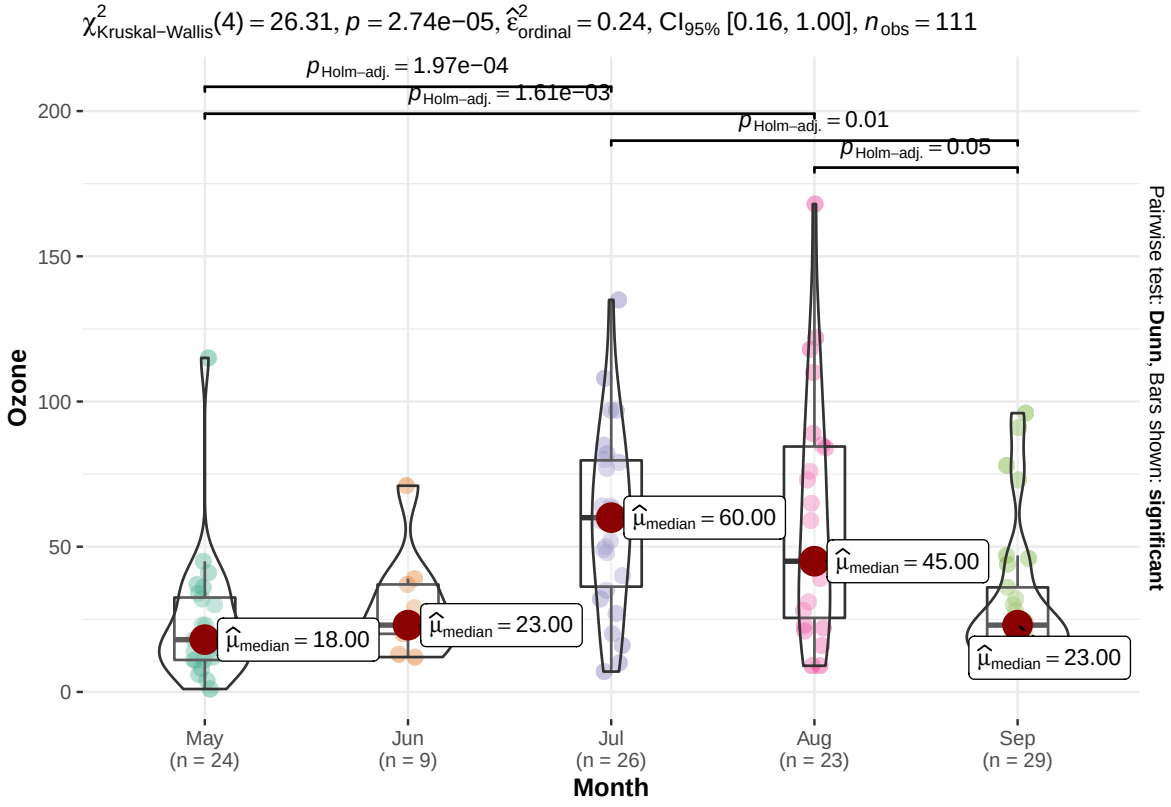
Appliquons le test de kruskal Wallis.

```
kruskal.test(Ozone~Month, data=datakruskal)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Ozone by Month
## Kruskal-Wallis chi-squared = 26.309, df = 4, p-value = 2.742e-05
```

On rejette H_0. Les concentrations d'ozone varie selon le mois.

Le package **ggstatsplot** offre des outils intéressants de visualisation de résultats de tests :



Si on veut effectuer des comparaisons deux à deux, on utilise le test de Dunn (package *PMCMRplus* et la fonction *kwAllPairsDunnTest()*).

6.4 Comparaison d'une proportion observée à une proportion attendue : le test exact binomial

On dispose d'une variable qualitative binaire (type "succès", "échec") et d'une variable quantitative. On souhaite comparer une proportion observée à une des classes de la variable qualitative (notée π_1) à une proportion théorique notée p.

Hypothèses d'application dans le cas d'un test bilatéral

- $H_0 : \pi_1$=p
- $H_1 : \pi_1 \neq$ p

Sous **R**, on utilisera la fonction *binom.test*. Il faut spécifier le nombre de succès correspondant au nombre d'observations qui appartiennent à une catégorie d'intérêt, le nombre d'essais c'est-à-dire le nombre d'observations et la probabilité qu'une observation appartienne à la catégorie d'intérêt. Prenons un exemple. 300 souris sont soumises à un traitement. 120 mâles

et 180 femelles ont été testés. On souhaite savoir si le traitement a plus d'effet sur les femelles que sur les mâles.

```
binom.test(x=180, n=300)
```

```
##
##  Exact binomial test
##
## data:  180 and 300
## number of successes = 180, number of trials = 300, p-value = 0.0006342
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5421336 0.6558710
## sample estimates:
## probability of success
##                    0.6
```

Conclusion : On rejette l'hypothèse H_0. Il y a une différence de réponse au traitement entre les mâles et les femelles.

6.5 Comparaison de proportions : Le test exact de Fisher

On cherche à tester le lien entre deux variables qualitatives dans le cas où l'effectif de chaque modalité de chaque variable qualitative est inférieur à 5.

Hypothèses d'application dans le cas d'un test bilatéral

- H_0 : les variables qualitatives sont indépendantes
- H_1 : les variables qualitatives sont liées

Prenons un exemple connu décrit par Agresti (1990, p61f;2002, p.91) reporté dans l'aide de la fonction *fisher.test()*. L'hypothèse nulle est "il n'existe pas d'association entre l'ordre de versement du lait (avant ou après le thé dans une tasse) et l'affirmation de la femme britannique". Les deux variables sont indépendantes. Ainsi l'hypothèse alternative H_1 est "il existe une association positive (rapport de chance >1)".

```
?fisher.test
```

```
## démarrage du serveur d'aide httpd ... fini
```

```
TeaTasting <-
matrix(c(3, 1, 1, 3), nrow = 2,
       dimnames = list(Guess = c("Milk", "Tea"),
                       Truth = c("Milk", "Tea")))
fisher.test(TeaTasting, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  TeaTasting
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3135693       Inf
## sample estimates:
## odds ratio
##   6.408309
```

On ne rejette pas H_0. La femme ne peut pas distinguer si le lait a été mis avant ou après le thé.

Chapitre 7

Introduction aux tests de permutation

7.1 Principe des tests de permutation

Les tests de permutation sont une bonne alternative aux tests de rang par leur robustesse et en se basant sur une distribution empirique et non théorique. On va permuter les observations et calculer la statistique de test correspondante. On répète les étapes précédentes autant de fois que nécessaire. On va donc calculer toutes les valeurs possibles de la statistique.

7.2 Applications sous R

Un package couramment utilisé est le package *coin*.

```
library(coin)
```

Exemple 1

```
oneway_test(Petal.Length~Species, data=datairis,
       alternative="less", paired=FALSE, var.equal=TRUE,
       distribution = approximate(nresample = 10000))
```

```
##
```

```
##  Approximative Two-Sample Fisher-Pitman Permutation Test
##
## data:  Petal.Length by Species (versicolor, virginica)
## Z = -7.8248, p-value < 1e-04
## alternative hypothesis: true mu is less than 0
```

Exemple 2

```
wilcox_test(Ozone~Month, data = datawilcoxon,
            distribution = approximate(10000))
```

```
##
##  Approximative Wilcoxon-Mann-Whitney Test
##
## data:  Ozone by Month (May, September)
## Z = -1.4844, p-value = 0.1428
## alternative hypothesis: true mu is not equal to 0
```

Exemple 3

```
kruskal_test(Ozone~Month, data=datakruskal,
             distribution = approximate(nresample = 10000))
```

```
##
##  Approximative Kruskal-Wallis Test
##
## data:  Ozone by Month (May, Jun, Jul, Aug, Sep)
## chi-squared = 26.309, p-value < 1e-04
```

Chapitre 8

Tester l'association entre deux variables quantitatives : les tests de corrélation

8.1 Le test de Pearson

On dispose de deux variables quantitatives continues.

Les conditions d'application de ce test sont les suivantes :

- Chacune des variables suit une loi normale.
- Relation linéaire entre les deux variables (tracer un nuage de points)
- Absence d'outliers

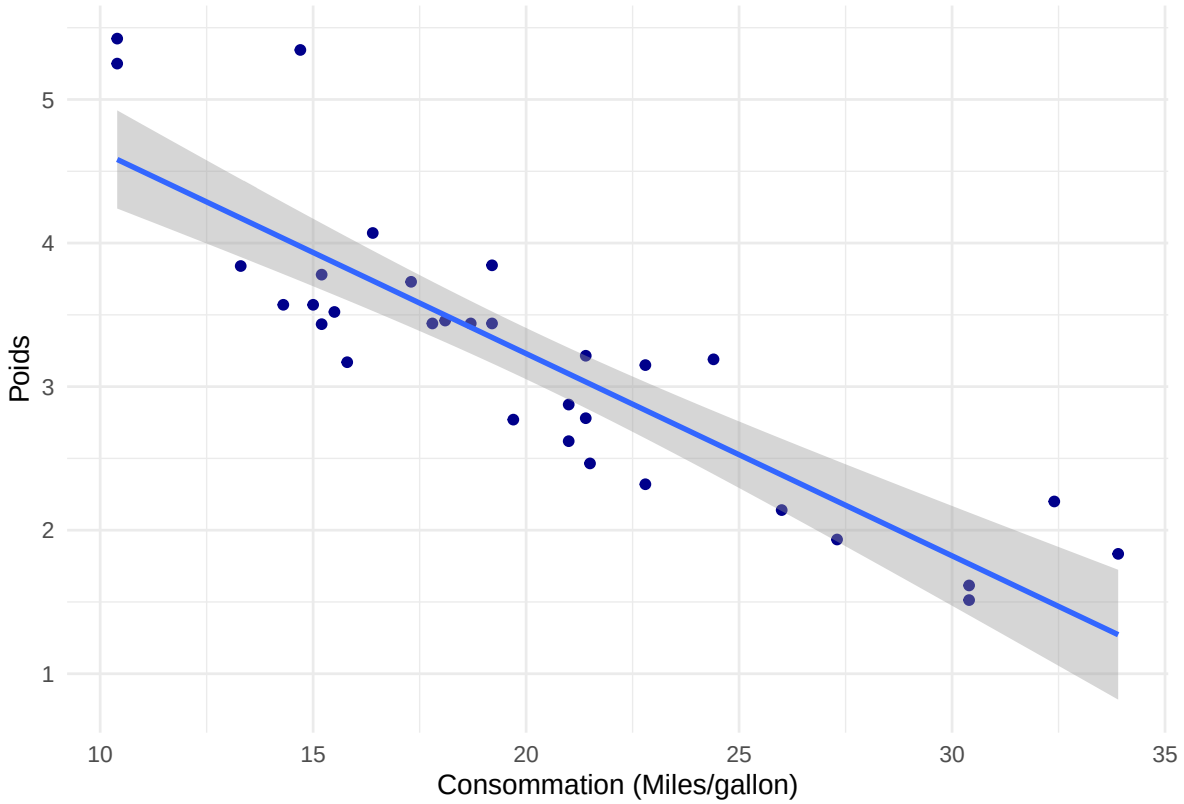
Les hypothèses d'application de ce test sont les suivantes :

- H_0 : Il n'existe pas de relation linéaire entre les deux variables
- H_1 : Il existe une relation linéaire entre les deux variables

Prenons le jeu de données *mtcars*. On veut tester si il existe une relation linéaire entre la consommation et le poids du véhicule.

```
ggplot(mtcars, aes(x = mpg, y = wt)) +
  geom_point(colour = "darkblue") +
  xlab("Consommation (Miles/gallon)") +
  ylab("Poids") +
  geom_smooth(method="lm") +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

```
shapiro.test(mtcars$wt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$wt
## W = 0.94326, p-value = 0.09265
```

Les conditions d'application du test semblent respectées.

```
cor.test(mtcars$mpg, mtcars$wt, method="pearson",
         alternative = "two.sided")
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$mpg and mtcars$wt
## t = -9.559, df = 30, p-value = 1.294e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9338264 -0.7440872
## sample estimates:
##        cor
## -0.8676594
```

On conclut qu'il existe une corrélation entre la consommation et le poids du véhicule.

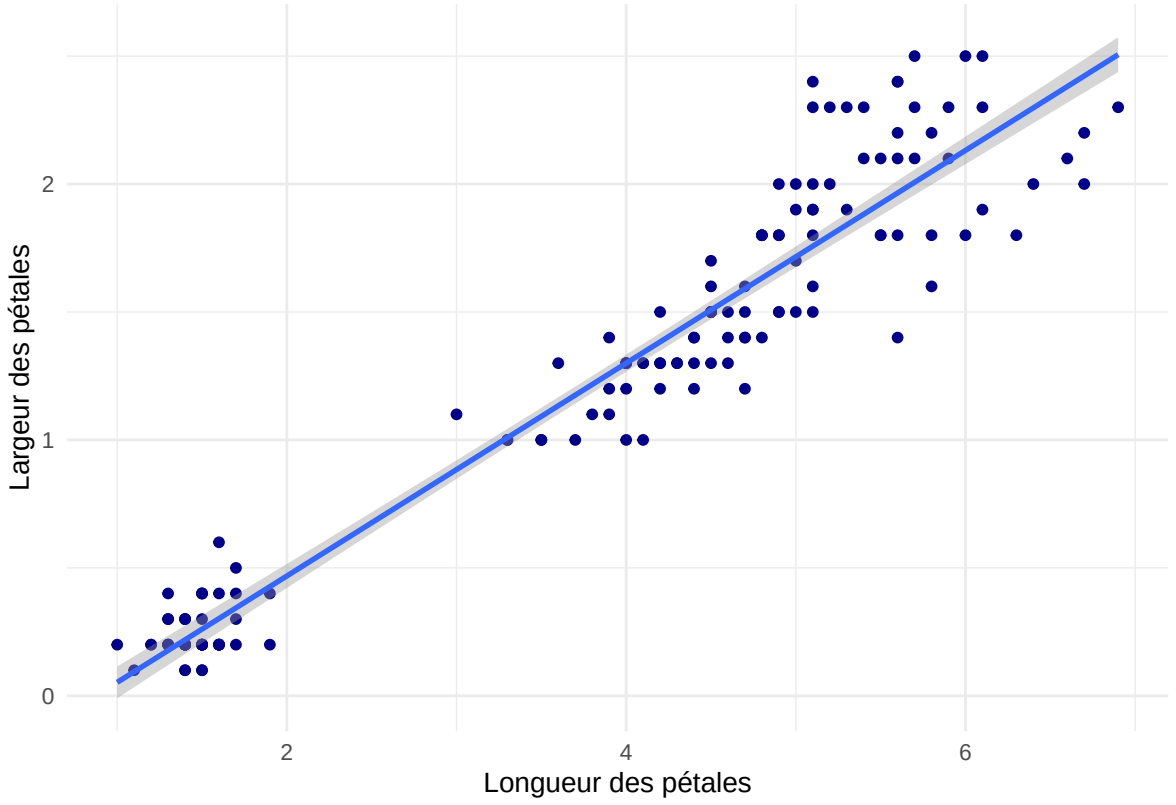
8.2 Tests non paramétriques : le test de Kendall et le test de Spearman

Ces tests sont basés sur les rangs et s'appliquent lorsque l'hypothèse de normalité n'est pas respectée.

Reprenons le jeu de données iris. Existe-t-il une corrélation entre la longueur des pétales et la largeur des pétales ?

```
ggplot(iris, aes(x = Petal.Length, y = Petal.Width)) +
  geom_point(colour = "darkblue") +
  xlab("Longueur des pétales") +
  ylab("Largeur des pétales") +
  geom_smooth(method="lm") +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
shapiro.test(iris$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

```
shapiro.test(iris$Petal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Width
## W = 0.90183, p-value = 1.68e-08
```

Appliquons un test non paramétrique.

Effectuons un test de Kendall :

```
cor.test(iris$Petal.Length, iris$Petal.Width,
         method="kendall", alternative = "two.sided")
```

```
##
##  Kendall's rank correlation tau
##
## data:  iris$Petal.Length and iris$Petal.Width
## z = 13.968, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.8068907
```

Effectuons un test de Spearman :

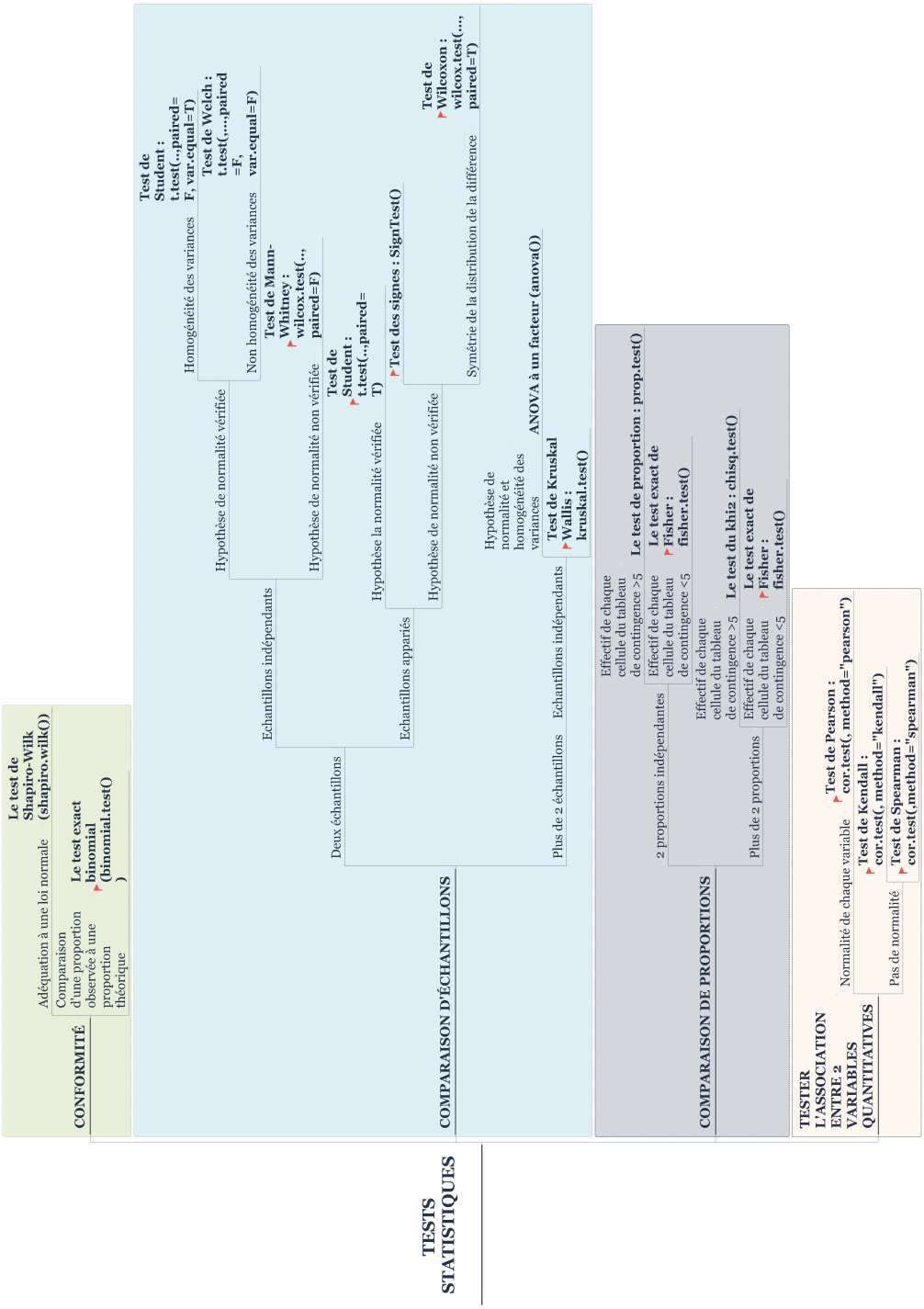
```
cor.test(iris$Petal.Length, iris$Petal.Width,
         method="spearman", alternative = "two.sided")
```

```
## Warning in cor.test.default(iris$Petal.Length, iris$Petal.Width, method =
## "spearman", : Impossible de calculer la p-value exacte avec des ex-aequos

##
##  Spearman's rank correlation rho
##
## data:  iris$Petal.Length and iris$Petal.Width
## S = 35061, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9376668
```

On conclut qu'il existe une corrélation entre la longueur des pétales et la largeur des pétales.

Synthèse


TESTS STATISTIQUES
CONFORMITÉ
Adéquation à une loi normale
Le test de Shapiro-Wilk (shapiro.wilk())
Comparaison d'une proportion observée à une proportion théorique
Le test exact binomial (binomial.test())
COMPARAISON D'ÉCHANTILLONS
Deux échantillons
Echantillons indépendants
Hypothèse de normalité vérifiée
Homogénéité des variances
Test de Student : t.test(...paired= F, var.equal=T)
Non homogénéité des variances
Test de Welch : t.test(,...,paired =F, var.equal=F)
Hypothèse de normalité non vérifiée
Test de Mann-Whitney : wilcox.test(.., paired=F)
Echantillons appariés
Hypothèse la normalité vérifiée
Test de Student : t.test(...paired= T)
Hypothèse de normalité non vérifiée
Test des signes : SignTest()
Symétrie de la distribution de la différence
Test de Wilcoxon : wilcox.test(..., paired=T)
Plus de 2 échantillons
Echantillons indépendants
Hypothèse de normalité et homogénéité des variances
ANOVA à un facteur (anova())
Hypothèse de normalité non vérifiée
Test de Kruskal Wallis : kruskal.test()
COMPARAISON DE PROPORTIONS
2 proportions indépendantes
Effectif de chaque cellule du tableau de contingence >5
Le test de proportion : prop.test()
Effectif de chaque cellule du tableau de contingence <5
Le test exact de Fisher : fisher.test()
Plus de 2 proportions
Effectif de chaque cellule du tableau de contingence >5
Le test du khi2 : chisq.test()
Effectif de chaque cellule du tableau de contingence <5
Le test exact de Fisher : fisher.test()
TESTER L'ASSOCIATION ENTRE 2 VARIABLES QUANTITATIVES
Normalité de chaque variable
Test de Pearson : cor.test(, method= "pearson")
Pas de normalité
Test de Kendall : cor.test(, method="kendall")
Test de Spearman : cor.test(,method="spearman")

Références

Anderson, Edgar. 1935. “The Irises of the Gaspe Peninsula.” *Bulletin of the American Iris Society* 59 : 2–5.

Andri et mult. al., Signorell. 2020. *DescTools : Tools for Descriptive Statistics*. https://cran.r-project.org/package=DescTools.

Ben-Shachar, Mattan S., Dominique Makowski, and Daniel Lüdecke. 2020. “Compute and Interpret Indices of Effect Size.” *CRAN*. https://github.com/easystats/effectsize.

Champely, Stephane. 2020. *Pwr : Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Elsevier Science.

Dobson, A. J. 1983. *Introduction to Statistical Modelling*. Science Paperbacks. Chapman ; Hall.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA : Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Frédéric Bertrand, Myriam Maumy-Bertrand. 2010. *Initiation à La Statistique Avec R : Cours, Exemples, Exercices Et Problèmes Corrigés*. Dunod.

Hand, D. J., F. Daly, K. McConway, D. Lunn, and E. Ostrowski. 1993. *A Handbook of Small Data Sets*. Chapman & Hall Statistics Texts. Taylor & Francis.

Henderson, Harold V., and Paul F. Velleman. 1981. “Building Multiple Regression Models Interactively.” *Biometrics* 37 (2) : 391–411.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. “Simultaneous Inference in General Parametric Models.” *Biometrical Journal* 50 (3) : 346–63.

Hothorn, Torsten, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. 2008. “Implementing a Class of Permutation Tests : The coin Package.” *Journal of Statistical Software* 28 (8) : 1–23. https://doi.org/10.18637/jss.v028.i08.

Kassambara, Alboukadel. 2020. *Ggpubr : ’Ggplot2’ Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Millot, Gaël. 2011. *Comprendre Et Réaliser Des Tests Statistiques à l’aide de r, Manuel de*

Biostatistique, 2eme Édition. Editions De Boeck.

Patil, Indrajeet. 2018. “ggstatsplot : ’Ggplot2’ Based Plots with Statistical Details.” *CRAN*. https://doi.org/10.5281/zenodo.2074621.

Pohlert, Thorsten. 2020. *PMCMRplus : Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended*. https://CRAN.R-project.org/package=PMCMRplus.

Saporta, Gilbert. 2006. *Probabilités, Analyse Des Données Et Statistique*. Editions Technip.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York : Springer. http://www.stats.ox.ac.uk/pub/MASS4/.

Verzani, John. 2005. *Using R for Introductory Statistics*. Chapman & Hall/CRC.

Wickham, Hadley. 2007. “Reshaping Data with the Reshape Package.” *Journal of Statistical Software* 21 (12). http://www.jstatsoft.org/v21/i12/paper.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43) : 1686. https://doi.org/10.21105/joss.01686.

Zar, Jerrold H. 1984. *Biostatistical Analysis*. Prentice Hall ; 2nd edition.