Semaine 6 - Tests statistiques classiques

Semaine 6

Table des matières

Exercice 3 - Données anatomiques de chats domestiques

On étudie un jeu de données présentant le poids du cœur et du corps de chats mâles et femelles utilisés dans des expériences sur la digitaline. Tous les chats étaient adultes et pesaient plus de 2 kg. (Référence : R. A. Fisher (1947), *The analysis of covariance method for the relation between a part and the whole*, Biometrics, 3, 65–68.)

Les variables sont :

- Sexe (Sex) : Facteur indiquant le sexe du chat ("F" pour femelle, "M" pour mâle).
- Poids corporel (Bwt) : Poids du corps en kg.

- Poids du cœur (Hwt) : Poids du cœur en g.

```
library(MASS)
data(cats)
```

```
attach(cats)

# Nombre de femelles
nF<-sum(Sex=="F")
nF
```

```
## [1] 47
```

```
# Nombre de mâles
nM<-sum(Sex=="M")
nM
```

```
## [1] 97
```

```
# Poids du corps des femelles
BwtF<-Bwt[Sex=="F"]
mean(BwtF)
```

```
## [1] 2.359574
```

```
# Poids du corps des mâles
BwtM<-Bwt[Sex=="M"]
mean(BwtM)
```

```
## [1] 2.9
```

```
# Poids du coeur des femelles
HwtF<-Hwt[Sex=="F"]
mean(HwtF)
```

```
## [1] 9.202128
```

```r
# Poids du coeur des mâles
HwtM<-Hwt[Sex=="M"]
mean(HwtM)
```

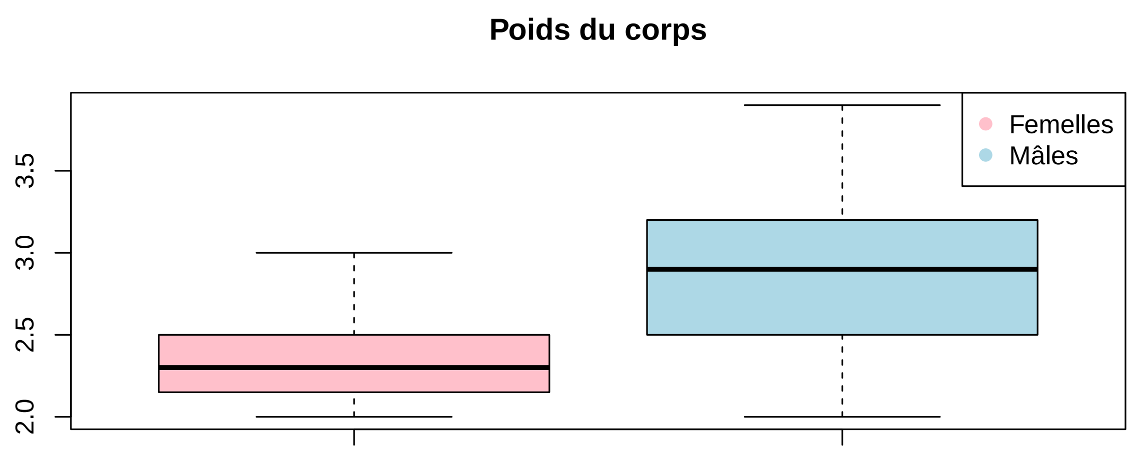
```
## [1] 11.32268
```

Question 1

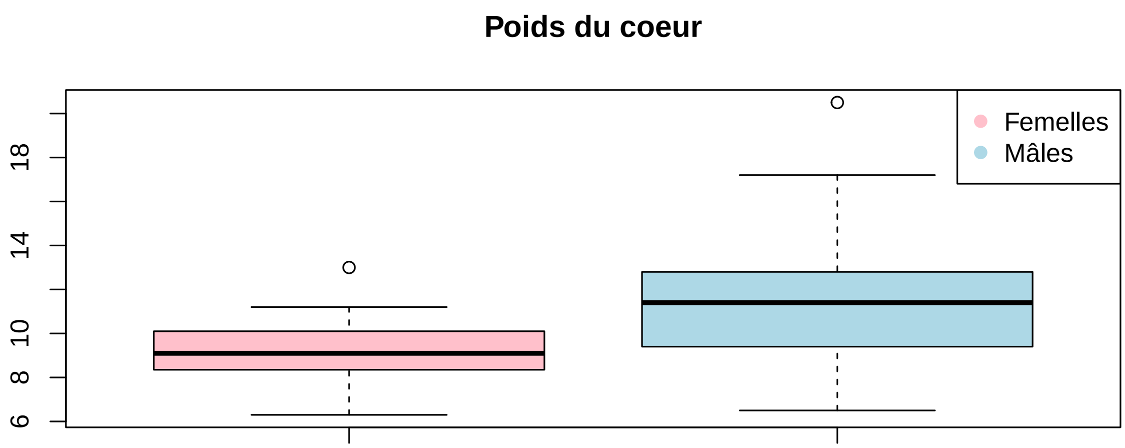
A l'aide de boîtes à moustaches, comparer graphiquement les données relatives au poids du corps chez les mâles et les femelles. Evaluer la différence des poids moyens entre les deux sexes. Résumer et commenter le résultat.

Solution.

```r
# Poids du corps
boxplot(BwtF,BwtM,col = c("pink", "lightblue"),main="Poids du corps")
legend(x = "topright", col = c("pink","lightblue"), pch =19, legend = c("Femelles","Mâles"))
```




```r
# Autre possibilité
# boxplot(Bwt ~ Sex, data = cats, col = c("pink", "lightblue"),main="Poids du corps")
# legend(x = "topright", col = c("pink","lightblue"), pch =19, legend = c("Femelles","Mâles"))

# Poids du coeur
boxplot(HwtF,HwtM,col = c("pink", "lightblue"),main="Poids du coeur")
legend(x = "topright", col = c("pink","lightblue"), pch =19, legend = c("Femelles","Mâles"))
```



Les boxplots montrent une nette différence de distribution des poids entre les mâles (plus lourds) et les femelles, plus prononcée sur le poids du corps que sur le poids du coeur.

La différence des poids moyens se calcule de la manière suivante

```
# calcule la difference inter-sexe
# on peut utiliser with(cats, )
# with() sert à rendre le code plus court et lisible lorsqu'on travaille avec un data frame.
with(cats, mean(BwtF) - mean(BwtM))
```

```
## [1] -0.5404255
```

```
with(cats, mean(HwtF) - mean(HwtM) )
```

```
## [1] -2.120553
```

```
# Autre possibilité :
cat("Différence des poids moyens du corps = ", mean(BwtF)-mean(BwtM),"\n")
```

```
## Différence des poids moyens du corps =  -0.5404255
```

```
cat("Différence des poids moyens du coeur = ", mean(HwtF)-mean(HwtM),"\n")
```

```
## Différence des poids moyens du coeur =  -2.120553
```

Question 2

Les données concernant le poids des mâles peuvent elles être considérées comme étant échantillonnées selon la loi normale ? Même question pour les femelles. Tester graphiquement à l'aide de la fonction `qqnorm`, puis à l'aide `shapiro.test`.

Solution.

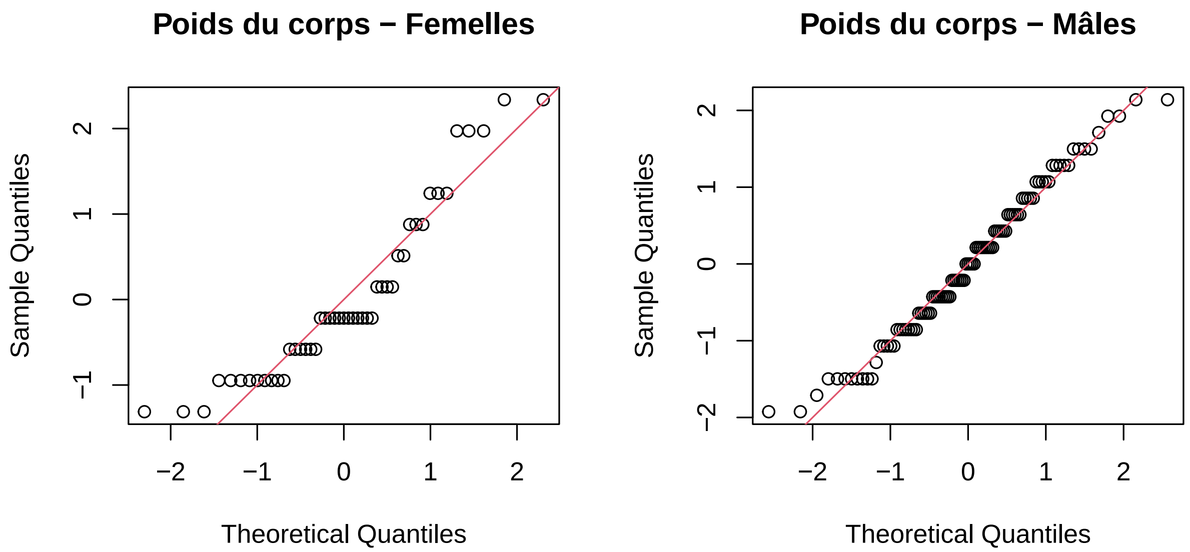
```
## deux graphiques males et femelles
par(mfrow = c(1,2))

# Poids du corps
qqnorm(scale(BwtF),main="Poids du corps - Femelles") ; abline(0, 1, col = 2)
cat("Test de normalité - Poids du corps - Femelles :","\n")
```

```
## Test de normalité - Poids du corps - Femelles :
```

```r
shapiro.test(BwtF)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BwtF
## W = 0.89096, p-value = 0.0003754
```

```r
qqnorm(scale(BwtM),main="Poids du corps - Mâles") ; abline(0, 1, col = 2)
```




```r
cat("Test de normalité - Poids du corps - Mâles :","\n")
```

```
## Test de normalité - Poids du corps - Mâles :
```

```r
shapiro.test(BwtM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BwtM
## W = 0.97883, p-value = 0.119
```

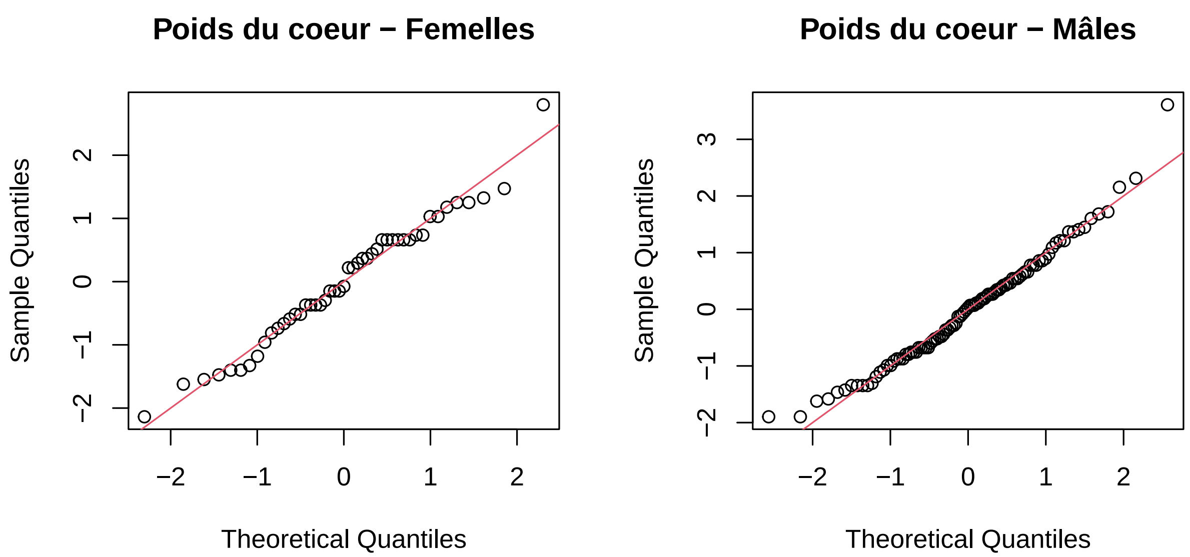
```r
# Poids du coeur

qqnorm(scale(HwtF),main="Poids du coeur - Femelles") ; abline(0, 1, col = 2)
cat("Test de normalité - Poids du coeur - Femelles :","\n")
```

```
## Test de normalité - Poids du coeur - Femelles :
```

```r
shapiro.test(HwtF)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  HwtF
## W = 0.9788, p-value = 0.5435
```

```r
qqnorm(scale(HwtM),main="Poids du coeur - Mâles") ; abline(0, 1, col = 2)
```




```r
cat("Test de normalité - Poids du coeur - Mâles :","\n")
```

```
## Test de normalité - Poids du coeur - Mâles :
```

```r
shapiro.test(HwtM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  HwtM
## W = 0.97727, p-value = 0.09039
```

Pour le poids du corps des femelles, la faible p-valeur du test de Shapiro et la forme du graphe de probabilités amènent à conclure que la loi normale n'est pas un modèle approprié. Pour les autres données, la loi normale convient.

Faute de modèle alternatif pour le moment, on admettra quand même l'hypothèse de loi normale pour le poids du corps des femelles.

Question 3

On suppose les données issues de la loi normale. Proposer un test de significativité pour la difference moyenne de poids entre les deux sexes et calculer une p-valeur pour ce test.

Solution. On applique la formule de calcul de la statistique de test t et on obtient une p-valeur. L'écart-type (pooled) est donné par

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

et la statistique de Student est

$$t = 7.330667$$

```r
# on calcule tout ca en R
# la formule de la statistique de test

# Poids des corps

# calcul de la variance pooled : cf resumé des tests
sp2 <- ((nF-1)*var(BwtF)+(nM-1)*var(BwtM))/(nF+nM-2)

# calcul de la t.stat : cf resumé des tests
t.stat <- (mean(BwtM)-mean(BwtF))/sqrt(sp2*(1/nM+1/nF))
t.stat
```

```
## [1] 7.330667
```

La valeur p est égale à

```r
# on utilise la fonction de repartition de la loi t
2*(1-pt(abs(t.stat),df=nF+nM-2))
```

```
## [1] 1.590461e-11
```

```r
# ou bien
2*pt(abs(t.stat),df=nF+nM-2,lower=FALSE)
```

```
## [1] 1.590445e-11
```

Que conclut-on ? La p-valeur est très faible, donc on rejette l'hypothèse d'égalité des moyennes.

Question 4

Calculer une p-valeur pour ce test à l'aide de la commande `t.test`. Que remarquez vous ? Expliquez brièvement les différences avec le test précédent (hypothèse d'égalité des variances).

Solution.

```r
# Poids des corps
t.test(BwtF,BwtM)
```

```
##
##  Welch Two Sample t-test
##
## data:  BwtF and BwtM
## t = -8.7095, df = 136.84, p-value = 8.831e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6631268 -0.4177242
## sample estimates:
## mean of x mean of y
##  2.359574  2.900000
```

La p-valeur n'est pas exactement la même que celle calculée précédemment (mais la conclusion, rejeter l'égalité des moyennes, est la même). La différence vient de l'égalité des variances. t.test teste l'égalité des variances. Si elle n'est pas rejetée, t.test fait le test de Student classique et on doit trouver la même p-valeur que celle vue plus haut. Si elle est rejetée, t.test fait le test de Welch, donc c'est normal de ne pas trouver la même chose.

```r
# Test d'égalité des variances.
library(car)
```

```
## Warning: le package 'car' a été compilé avec la version R 4.1.3
```

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

## Warning: le package 'carData' a été compilé avec la version R 4.1.3
```

```
group<-factor(c(rep(1,nF),rep(2,nM)))
leveneTest(c(BwtF,BwtM) ~ group)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  19.431 2.044e-05 ***
##       142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'égalité des variances est bien rejetée, donc t.test a fait le test de Welch. Mais on peut imposer à t.test de faire le test de Student classique avec l'option var.equal=TRUE.

```
t.test(BwtF,BwtM,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  BwtF and BwtM
## t = -7.3307, df = 142, p-value = 1.59e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6861584 -0.3946927
## sample estimates:
## mean of x mean of y
##  2.359574  2.900000
```

On retrouve bien la statistique de test et la p-valeur calculées plus haut.

Même chose avec les poids des coeurs.

```
# Poids des coeurs

# calcul de la variance pooled : cf resumé des tests
sp2 <- ((nF-1)*var(HwtF)+(nM-1)*var(HwtM))/(nF+nM-2)

# calcul de la t.stat : cf resumé des tests
t.stat <- (mean(HwtM)-mean(HwtF))/sqrt(sp2*(1/nM+1/nF))
t.stat
```

```
## [1] 5.353924
```

```
# p-valeur
2*(1-pt(abs(t.stat),df=nF+nM-2))
```

```
## [1] 3.379786e-07
```

```
t.test(HwtF,HwtM)
```

```
##
##  Welch Two Sample t-test
##
## data:  HwtF and HwtM
## t = -6.5179, df = 140.61, p-value = 1.186e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -2.763753 -1.477352
## sample estimates:
## mean of x mean of y
##  9.202128 11.322680
```

```
t.test(HwtF,HwtM,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  HwtF and HwtM
## t = -5.3539, df = 142, p-value = 3.38e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.903517 -1.337588
## sample estimates:
## mean of x mean of y
##  9.202128 11.322680
```

La p-valeur est encore très faible, on rejette l'hypothèse d'égalité des moyennes.

Remarque : plutôt que de tester si les moyennes dont différentes, on aurait envie de tester si le poids moyen des hommes est supérieur à celui des femmes. On peut le faire directement avec t.test en fixant alternative à "less".

```
# Poids des corps.
t.test(BwtF,BwtM,alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  BwtF and BwtM
## t = -8.7095, df = 136.84, p-value = 4.416e-15
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.4376663
## sample estimates:
## mean of x mean of y
##  2.359574  2.900000
```

```
# Poids des coeurs
t.test(HwtF,HwtM,alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  HwtF and HwtM
## t = -6.5179, df = 140.61, p-value = 5.93e-10
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.581858
## sample estimates:
## mean of x mean of y
##  9.202128 11.322680
```

On conclut clairement que les poids moyens des corps et coeurs des mâles sont supérieurs à ceux des femelles.