Statistiques $\bowtie$ Forensiques

Quelques tests pour les Sciences Forensiques

Anne B. Dufour

11 juillet 2008

Ce document a pour vocation de rassembler les quelques tests statistiques de base trouvés dans la littérature scientifique ainsi que des tests plus spécifiques du domaine.

Table des matières

1 Comparaison de deux valeurs

1.1 Les valeurs sont deux proportions

On a suivi, sur une période de 20 ans, deux cohortes : 200 sujets fumeurs et 200 sujets non fumeurs. On a noté le nombre d'apparition de cancer dans chacune des cohortes : 40 chez les fumeurs ; 20 chez les non fumeurs. La différence d'apparition de cancer dans les deux cohortes est-elle significative ?

Soient k_1 et k_2 les fréquences absolues observées d'individus présentant un certain caractère A dans chacun des échantillons.
Soient n_1 et n_2 les effectifs observées dans chacun des échantillons.
La valeur de la statistique du test est :

$$z = \frac{f_1 - f_2}{\sqrt{(\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

où $\hat{p} = \frac{k_1+k_2}{n_1+n_2}$ et $\hat{q} = 1 - \hat{p}$.

```
tabac <- matrix(c(40, 160, 20, 180), 2)
colnames(tabac) = c("coh1", "coh2")
rownames(tabac) = c("fumeur", "non_fumeur")
tabac
```

```
           coh1 coh2
fumeur       40   20
non_fumeur  160  180
```

```
prop.test(c(40, 20), c(200, 200), correct = F)
```

```
        2-sample test for equality of proportions without continuity correction
data:  c(40, 20) out of c(200, 200)
X-squared = 7.8431, df = 1, p-value = 0.005101
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03070481 0.16929519
sample estimates:
prop 1 prop 2
   0.2    0.1
```

```
chisq.test(tabac, correct = F)
```

```
        Pearson's Chi-squared test
data:  tabac
X-squared = 7.8431, df = 1, p-value = 0.005101
```

1.2 Les valeurs sont deux moyennes et / ou deux variances

La feuille excel intitulée `CalculStatAuto AGL` présente, à partir de deux distributions X et Y les éléments nécessaires aux calculs suivants :

- ⋆ minimum, maximum, variance, écart-type, moyenne, coefficient de variation
- ⋆ les éléments pour réaliser le test de comparaison de deux variances observées (test de Fisher)
- ⋆ les éléments pour réaliser le test de comparaison de deux moyennes observées
  a) si les variances sont égales, par le test de Student
  b) si les variances sont inégales, par le test de Aspin-Welch

Valeur de la statistique de Fisher

$$F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}$$

Valeur de la statistique de Student

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\hat{\sigma}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

avec

$$\widehat{\sigma^2} = \frac{(n_1 - 1)\widehat{\sigma_1^2} + (n_2 - 1)\widehat{\sigma_2^2}}{n_1 + n_2 - 2}$$

Dans le cas où les variances sont inégales, le degré de liberté se calcule par la formule de Welch suivante :

$$\frac{(\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2})^2}{\frac{(\widehat{\sigma_1}/n_1)^2}{n_1 - 1} + \frac{(\widehat{\sigma_2}/n_2)^2}{n_2 - 1}}$$

La caféine agit sur les systèmes cardio-vasculaires et nerveux et sur divers processus métaboliques. Il est possible que les variations d'équilibre hormonal durant la grossesse modifient la manière dont la caféine est métabolisée en ralentissant ou en accélérant son élimination. Pour étudier l'existence de différences du métabolisme de la caféine chez la femme, on a administré une pilule de 250mg de caféine (environ trois tasses de café) à des femmes qui prenaient des contraceptifs oraux (oui pour CO) et des femmes qui n'en prenaient pas (non pour CO). Les femmes prenant des contraceptifs oraux ont un métabolisme semblable à celui des femmes enceintes. On a mesuré les concentrations de caféine et mesuré la vitesse avec laquelle la caféine disparaissait dans le sang (taux de demi-vie c'est-à-dire le temps que prend la concentration de caféine pour arriver à la moitié de sa valeur initiale, exprimé en heures).

```
conon <- c(3.47, 4.59, 4.72, 5.17, 5.3, 6.59, 7.01, 7.25, 7.28,
    7.3, 7.6, 8.16)
cooui <- c(5.54, 6.87, 7.26, 7.94, 7.98, 8.11, 12.04, 12.81, 13.04,
    14.28, 14.41, 15.47)
```

```
shapiro.test(conon)
```

```
        Shapiro-Wilk normality test
data:  conon
W = 0.9137, p-value = 0.2379
```

```
shapiro.test(cooui)
```

```
        Shapiro-Wilk normality test
data:  cooui
W = 0.8886, p-value = 0.1132
```

```
var.test(conon, cooui)
```

```
        F test to compare two variances
data:  conon and cooui
F = 0.1788, num df = 11, denom df = 11, p-value = 0.00815
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.05147036 0.62107148
sample estimates:
ratio of variances
         0.1787925
```

```
t.test(conon, cooui, var.equal = T)
```

```
        Two Sample t-test
data:  conon and cooui
t = -3.8896, df = 22, p-value = 0.0007892
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.555655 -1.996012
sample estimates:
mean of x mean of y
 6.203333 10.479167
```

```
t.test(conon, cooui, var.equal = F)
```

```
        Welch Two Sample t-test
data:  conon and cooui
t = -3.8896, df = 14.812, p-value = 0.001483
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.621547 -1.930119
sample estimates:
mean of x mean of y
 6.203333 10.479167
```

Connaissant X et Y deux distributions, on peut écrire la fonction donnant ce degré de liberté.

```
ddlw <- function(x, y) {
    num <- ((var(x)/length(x)) + (var(y)/length(y)))^2
    denom <- ((var(x)/length(x))^2/(length(x) - 1)) + ((var(y)/length(y))^2/(length(y) -
        1))
    res <- num/denom
    return(res)
}
```

Le grand avantage de R est de pouvoir recalculer et retrouver facilement les résultats d'un test statistique.

```
sum((conon - mean(conon))^2)
```

```
[1] 24.19487
```

```
sum((cooui - mean(cooui))^2)
```

```
[1] 135.3237
```

```
(sum((conon - mean(conon))^2) + sum((cooui - mean(cooui))^2))/(length(conon) +
    length(cooui) - 2)
```

```
[1] 7.250844
```

```
varcom <- (sum((conon - mean(conon))^2) + sum((cooui - mean(cooui))^2))/(length(conon) +
    length(cooui) - 2)
varcom * (1/length(cooui) + 1/length(conon))
```

```
[1] 1.208474
```

```
sqrt(varcom * (1/length(cooui) + 1/length(conon)))
```

```
[1] 1.099306
```

```
mean(conon) - mean(cooui)
```

```
[1] -4.275833
```

```
qt(0.025, 22)
```

```
[1] -2.073873
```

1.3 Exemple des Mégots

On a recherché la concentration d'ADN (en ng/μl) pour 24 mégots à l'aide de deux extracteurs différents : "extracteur auto CST" et "Qiagen DNA Mini manuel". Les quantités obtenues sont-elles identiques ?

```
megots <- read.csv("http://pbil.univ-lyon1.fr/R/donnees/megots.csv",
    header = TRUE, sep = "\t", dec = ",")
head(megots)
```

```
  AutoCST Qiagen
1   2.406  2.256
2   1.259  0.830
3   4.277  3.359
4   0.602  0.667
5   0.795  0.753
6   1.181  0.940
```

```
names(megots)
```

```
[1] "AutoCST" "Qiagen"
```

```
AutoCST <- megots$AutoCST
Qiagen <- megots$Qiagen
summary(AutoCST)
```

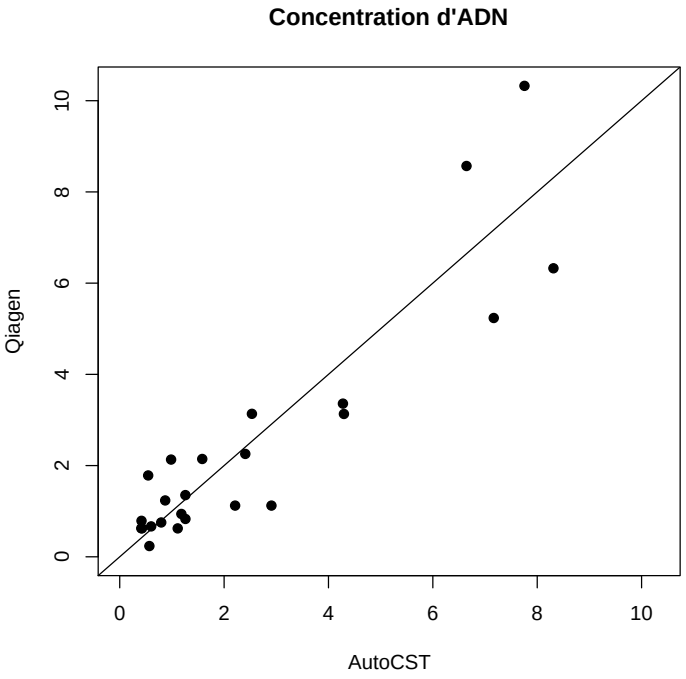
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.4130  0.7468  1.2580  2.5220  3.2500  8.3110
```

```
summary(Qiagen)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.236   0.780   1.294   2.472   3.133  10.330
```

Représentons graphiquement les résultats.

```
maxADN <- max(max(AutoCST), max(Qiagen))
plot(AutoCST, Qiagen, pch = 19, xlim = c(0, maxADN), ylim = c(0,
    maxADN), main = "Concentration d'ADN")
abline(0, 1)
```

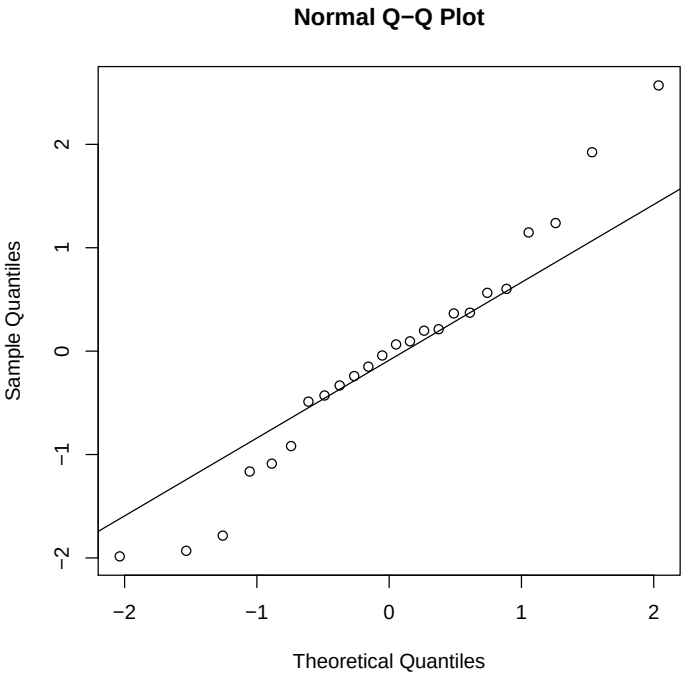



```
diffADN <- Qiagen - AutoCST
qqnorm(diffADN)
qqline(diffADN)
shapiro.test(diffADN)
```

```
        Shapiro-Wilk normality test
data:  diffADN
W = 0.9692, p-value = 0.6462
```

```
t.test(AutoCST, Qiagen, paired = T)
```

```
        Paired t-test
data:  AutoCST and Qiagen
t = 0.2178, df = 23, p-value = 0.8295
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4242127  0.5240460
sample estimates:
mean of the differences
             0.04991667
```



1.4 Comparaison sur 48 taches de sang issu du même tube sanguin

On a recherché la quantité d'ADN (en ng) pour 48 tâches de sang issu d'un même tube à essai. Les 48 tâches ont été séparées en deux groupes de 24 tâches et on a recherché la quantité d'ADN à l'aide de deux extracteurs différents : "extracteur auto CST" et "Qiagen DNA Mini manuel". Que peut-on dire de ces données ?

a) Quelle que soit l'analyse, il s'agit du même sang. On pourrait s'attendre à ce que tous les résultats soient identiques : toutes méthodes confondues, à l'intérieur de chaque méthode d'extraction

b) On compare ensuite les résultats par les deux méthodes d'extraction.

```
sangs <- read.csv("http://pbil.univ-lyon1.fr/R/donnees/sangs.csv",
    header = TRUE, sep = "\t", dec = ",")
head(sangs)
```

```
  X AutoCST Qiagen
1 NA   19.95  39.43
2 NA   14.58  19.94
3 NA   14.45  27.70
4 NA   22.70  36.50
5 NA   47.74  16.59
6 NA   19.59  32.94
```

```
names(sangs)
```

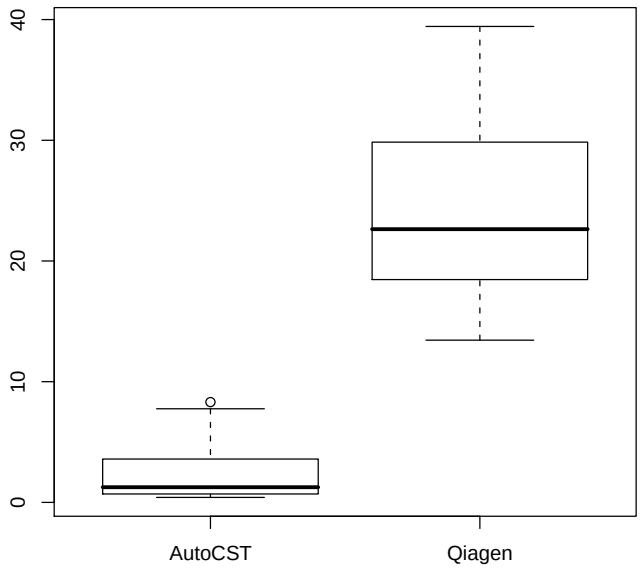
```
[1] "X"       "AutoCST" "Qiagen"
```

```
autoCST <- sangs$autoCST
Qiagen <- sangs$Qiagen
summary(AutoCST)
```

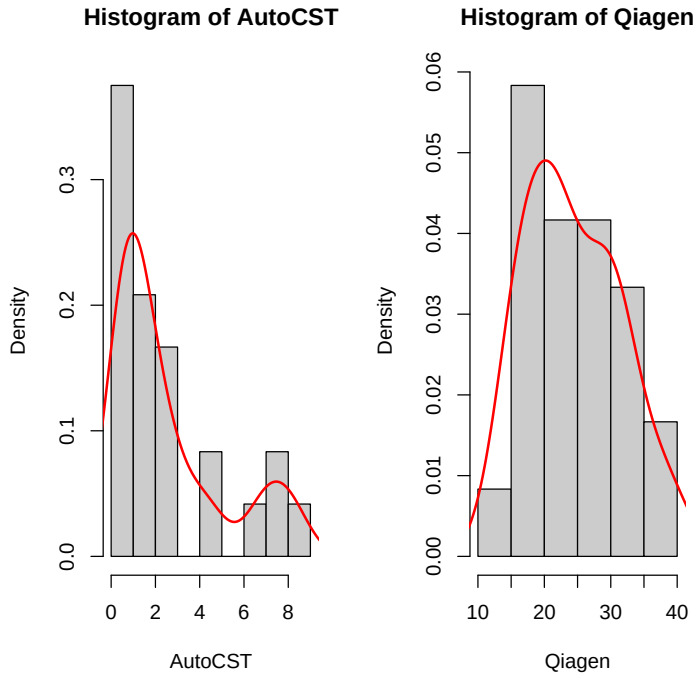
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4130  0.7468  1.2580  2.5220  3.2500  8.3110
summary(Qiagen)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.44   18.78   22.64   24.32   29.75   39.43
sd(AutoCST)
[1] 2.524578
sd(Qiagen)
[1] 7.073133
```

Représentation graphique

```
sangtube <- c(AutoCST, Qiagen)
choixmethod <- factor(rep(c(1, 2), c(24, 24)))
boxplot(sangtube ~ choixmethod, names = c("AutoCST", "Qiagen"))
dotchart(sangtube, group = choixmethod, pch = 19)
```




```
par(mfrow = c(1, 2))
hist(AutoCST, proba = T, col = grey(0.8))
lines(density(AutoCST), lwd = 2, col = "red")
hist(Qiagen, proba = T, col = grey(0.8))
lines(density(Qiagen), lwd = 2, col = "red")
```




```
 shapiro.test(AutoCST)
        Shapiro-Wilk normality test
data:  AutoCST
W = 0.7804, p-value = 0.0001435
 shapiro.test(Qiagen)
        Shapiro-Wilk normality test
data:  Qiagen
W = 0.9609, p-value = 0.4558
```

2 Comparaison de p valeurs

Chez le rat, on teste l'effet de l'ouabaïne sur la teneur en noradrénaline du myocarde. Les résultats sont dans le tableau ci-dessous. On note x la dose d'ouabaïne injectée et y la teneur en noradrénaline.

"http://pbil.univ-lyon1.fr/R/donnees/myo.txt.

2.1 Comparaison de p variances

La valeur de la statistique de Bartlett est :

$$B = (n-p)\ ln(\frac{\sum_{i=1}^{p}(n_i-1)\widehat{\sigma_i^2}}{n-p})\ -\ \sum_{i=1}^{p}(n_i-1)ln\widehat{\sigma_i^2}$$

```
 bartlett.test(myo$rep ~ dose.fac)
        Bartlett test of homogeneity of variances
data:  myo$rep by dose.fac
Bartlett's K-squared = 2.4182, df = 3, p-value = 0.4902
```

2.2 Comparaison de p moyennes

La valeur de la statistique du test est $F = \frac{CM_{inter}}{CM_{intra}}$.

```
 anova(lm(myo$rep ~ dose.fac))
Analysis of Variance Table
Response: myo$rep
          Df  Sum Sq Mean Sq F value  Pr(>F)
dose.fac   3 0.35203 0.11734  4.0797 0.01598 *
Residuals 28 0.80536 0.02876
---
Signif. codes:  0
```

3 Etudier les valeurs extrêmes

L'étude des valeurs extrêmes concerne une valeur au sein d'une distribution (`chisq.out.test`), une variance au sein d'un ensemble de variances (`cochran.test`) et une moyenne au sein d'un ensemble de moyenne (`grubbs.test`). Ces fonctions sont définies dans le paquet `outliers` [2].

```
 library(outliers)
```

3.1 Le test du Chi-Deux pour les valeurs extrêmes

La fonction réalise un test simple pour détecter la présence d'une valeur extrême. Le test est basé sur une distribution du chi-deux des carrés des différences entre les données et la moyenne de l'échantillon [1]. La variance de la population est supposée connue.
Les observations répliquées ci-dessous ont été obtenues.

```
 x <- c(4.85, 6.18, 6.28, 6.49, 6.69)
 x
[1] 4.85 6.18 6.28 6.49 6.69
```

Peut-on dire que la valeur 4.85 est une valeur extrême ?

```
 moy <- mean(x)
 moy
[1] 6.098
 xsort <- sort(x)
 xsort
[1] 4.85 6.18 6.28 6.49 6.69
 (xsort[5] - mean(x))^2/var(x)
[1] 0.6670803
 (xsort[1] - mean(x))^2/var(x)
[1] 2.964585
 valext <- max((xsort[5] - mean(x)), (mean(x) - xsort[1]))
 valtest <- (valext - mean(x))^2/var(x)
 pval <- 1 - pchisq(valtest, 1)
 chisq.out.test(x)

        chi-squared test for outlier

data:  x
X-squared = 2.9646, p-value = 0.0851
alternative hypothesis: lowest value 4.85 is an outlier
```

3.2 Le test de Dixon

Si on possède un ensemble de mesures répétées d'une quantité, une ou plusieurs valeurs peuvent différer grandement des autres valeurs. Le test de Dixon [1] permet de répondre au rejet ou non de ces valeurs.

La statistique du test se construit comme suit :

1. Les n valeurs sont classées par ordre croissant.

$$x_{(1)}\ x_{(2)}\ \cdots\ x_{(n)}$$

2. La statistique du test est le rapport entre la différence de la *valeur suspecte* à son plus proche voisin divisée par l'étendue de la distribution.

$$Q = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \text{ ou } Q = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$$

```
sort(x)
```
```
[1] 4.85 6.18 6.28 6.49 6.69
```
```
xsort <- sort(x)
(xsort[2] - xsort[1])/(max(x) - min(x))
```
```
[1] 0.7228261
```
```
val <- (xsort[2] - xsort[1])/(max(x) - min(x))
2 * pdixon(val, 5)
```
```
         1
0.04308173
```
```
qdixon(0.025, 5)
```
```
   1
0.71
```

```
dixon.test(x)
```
```
        Dixon test for outliers
data:  x
Q = 0.7228, p-value = 0.04308
alternative hypothesis: lowest value 4.85 is an outlier
```

3.3 Le test de Cochran

On dispose de 16 séries de 6 valeurs chacune contenant la "position" de l'allèle B1.

```
alleleB1 <- read.csv("http://pbil.univ-lyon1.fr/R/donnees/alleleB1.csv",
    header = TRUE, sep = "\t", dec = ",")
names(alleleB1)
```
```
[1] "Serie" "Val"
```

Le test de Cochran [3] compare la variance maximale de plusieurs groupes sur la somme des variances.

```
serie <- alleleB1$Serie
val <- alleleB1$Val
serie <- as.factor(serie)
summary(serie)
```
```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
 6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6
```

Il y a des séries avec valeurs manquantes.

```
allele <- na.omit(alleleB1)
serie <- allele$Serie
val <- allele$Val
serie <- as.factor(serie)
summary(serie)
```

```
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
6  6  6  6  6  6  6  6  6  6  6  6  5  5  6  6
```

Construction de la statistique de Cochran $C = \frac{\widehat{\sigma^2_{max}}}{\sum_{j=1}^{p} \widehat{\sigma^2_j}}$ où $\widehat{\sigma^2_{max}}$ est la variance la plus grande sur l'ensemble des p groupes.

```
tapply(val, serie, var)
```

```
           1            2            3            4            5            6
0.0024000000 0.0010966667 0.0029466667 0.0005766667 0.0010566667 0.0024300000
           7            8            9           10           11           12
0.0033466667 0.0017466667 0.0026966667 0.0018700000 0.0018800000 0.0009766667
          13           14           15           16
0.0010700000 0.0011500000 0.0031500000 0.0016800000
```

```
vartotal <- tapply(val, serie, var)
max(vartotal)/sum(vartotal)
```

```
[1] 0.1112835
```

```
cochran.test(val ~ serie)
```

```
        Cochran test for outlying variance
data:  val ~ serie
C = 0.1113, df = 5.875, k = 16.000, p-value = 1
alternative hypothesis: Group 7 has outlying variance
sample estimates:
           1            2            3            4            5            6
0.0024000000 0.0010966667 0.0029466667 0.0005766667 0.0010566667 0.0024300000
           7            8            9           10           11           12
0.0033466667 0.0017466667 0.0026966667 0.0018700000 0.0018800000 0.0009766667
          13           14           15           16
0.0010700000 0.0011500000 0.0031500000 0.0016800000
```

Pour retrouver les valeurs du test de Cochran, il faut retenir que :

1. les degrés de liberté représentent le nombre de valeurs par groupe c'est-à-dire n si les effectifs sont égaux et la moyenne des effectifs par groupe si non.
2. k représente le nombre de groupes

```
mean(summary(serie))
```

```
[1] 5.875
```

```
1 - pcochran(max(vartotal)/sum(vartotal), mean(summary(serie)),
    16)
```

```
[1] 1
```

3.4 Le test de Grubbs

L'objectif du test de Grubbs est de regarder si les moyennes calculées sur chaque groupe s'éloignent grandement ou non de la moyenne de l'ensemble. Il existe donc deux possibilités : comparaison de la plus grande moyenne à la moyenne de l'ensemble, comparaison de la plus petite moyenne à la moyenne de l'ensemble.

$$G = \frac{\bar{x} - \overline{x_{min}}}{s} \; ou \; G = \frac{\overline{x_{max}} - \bar{x}}{s}$$

```
tapply(val, serie, mean)
```

```
      1       2       3       4       5       6       7       8       9
111.2800 111.2717 111.2567 111.2583 111.2183 111.2450 111.2633 111.2533 111.2583
      10      11      12      13      14      15      16
111.2350 111.2600 111.2717 111.2480 111.2500 111.2850 111.2800
```

```
moytotal <- tapply(val, serie, mean)
grubbs.test(moytotal, opposite = T)
```

```
        Grubbs test for one outlier
data:  moytotal
G.15 = 1.5248, U = 0.8347, p-value = 0.9446
alternative hypothesis: highest value 111.285 is an outlier
```

```
grubbs.test(moytotal, opposite = F)
```

```
        Grubbs test for one outlier
data:  moytotal
G.5 = 2.2991, U = 0.6241, p-value = 0.09246
alternative hypothesis: lowest value 111.218333333333 is an outlier
```

Retrouvons ces valeurs :

```
(max(moytotal) - mean(moytotal))/sd(moytotal)
```

```
[1] 1.524752
```

```
moymax <- (max(moytotal) - mean(moytotal))/sd(moytotal)
1 - pgrubbs(moymax, 16)
```

```
[1] 0.9446248
```

```
moymin <- (min(moytotal) - mean(moytotal))/sd(moytotal)
1 - pgrubbs(moymin, 16)
```

```
[1] 0.09245908
```

4 Modèles linéaires et test de linéarité

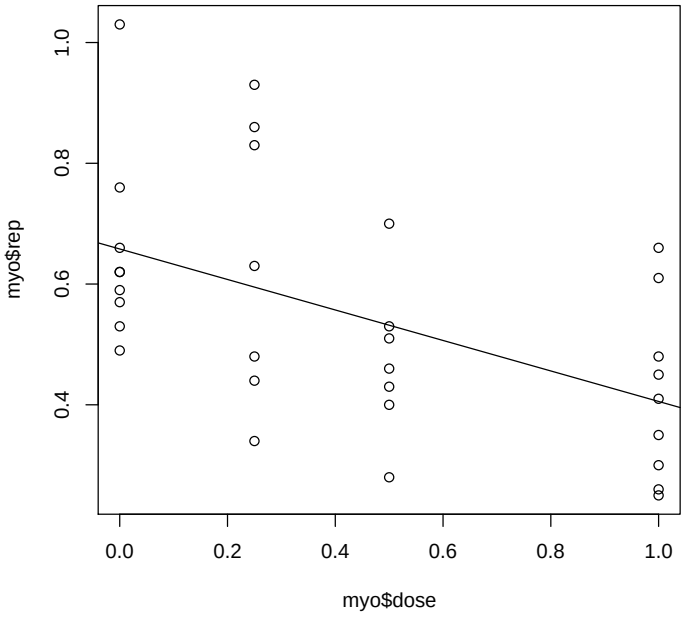
Reprenons l'exemple de l'effet de l'ouabaïne sur la teneur en noradrénaline du myocarde.

La variable `dose` peut être considérée soit comme une variable quantitative, soit comme une variable qualitative. Ceci est possible car elle est **contrôlée**.

4.1 La régression linéaire

Considérons la dose comme une variable quantitative et cherchons à expliquer la réponse par la dose dans un modèle simple : $\hat{y} = bx + a$.

```
plot(myo$dose, myo$rep)
abline(lm(rep ~ dose, data = myo))
```




```
lm1 <- lm(myo$rep ~ myo$dose)
anova(lm1)
```

```
Analysis of Variance Table
Response: myo$rep
          Df  Sum Sq Mean Sq F value   Pr(>F)
myo$dose   1 0.30875 0.30875  10.914 0.002475 **
Residuals 30 0.84864 0.02829
---
Signif. codes:  0
```

Une fois un modèle choisi, il faut toujours s'assurer de la normalité des résidus.

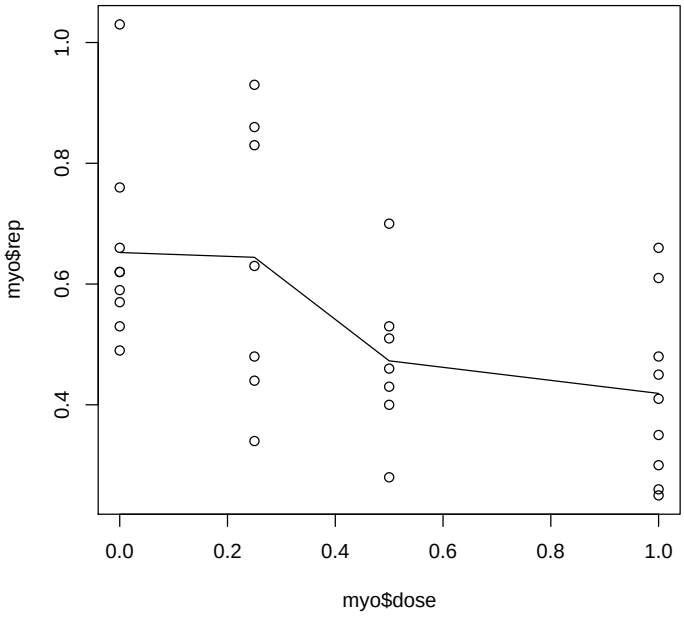
```
qqnorm(lm1$residuals)
qqline(lm1$residuals)
shapiro.test(lm1$residuals)
```

```
        Shapiro-Wilk normality test
data:  lm1$residuals
W = 0.9338, p-value = 0.04989
```

4.2 L'analyse de la variance à un facteur

Considérons la dose comme une variable qualitative et cherchons à expliquer la réponse par la dose dans le modèle suivant : $\hat{y} = \overline{y_{Classe}}$

```
plot(myo$dose, myo$rep)
dose.fac <- as.factor(myo$dose)
lines(myo$dose, predict(lm(myo$rep ~ dose.fac)))
```




```
 lm2 <- lm(myo$rep ~ dose.fac)
 anova(lm2)
Analysis of Variance Table
Response: myo$rep
           Df  Sum Sq Mean Sq F value  Pr(>F)
dose.fac    3 0.35203 0.11734  4.0797 0.01598 *
Residuals 28 0.80536 0.02876
---
Signif. codes:  0

 qqnorm(lm2$residuals)
 qqline(lm2$residuals)
 shapiro.test(lm2$residuals)
        Shapiro-Wilk normality test
data:  lm2$residuals
W = 0.9642, p-value = 0.3569
```

4.3 Le test de linéarité

Les deux modèles précédents sont emboités et donc comparables.

```
 anova(lm2, lm1)
Analysis of Variance Table
Model 1: myo$rep ~ dose.fac
Model 2: myo$rep ~ myo$dose
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     28  0.80536
2     30  0.84864 -2  -0.04328 0.7524 0.4805
 anova(lm1, lm2)
Analysis of Variance Table
Model 1: myo$rep ~ myo$dose
Model 2: myo$rep ~ dose.fac
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     30 0.84864
2     28 0.80536  2   0.04328 0.7524 0.4805
```

Références

[1] W.J. Dixon. Analysis of extreme values. *Ann. Math. Stat.*, 4 :488–506, 1950.

[2] Lukasz Komsta. *outliers : Tests for outliers*, 2007. R package version 0.13-2.

[3] G.W. Snedecor and W.G. Cochran. *Statistical Methods (seventh edition)*. Iowa State University Press, Ames, Iowa, United States of America, 1980.