Analyse de la variance avec R - exemple du cours

DU ETD / 2016

ChF - IUT Vannes / Dpt STID

2016

Problématique : Etude de la quantité d'alcool acquis en fonction du type de cidre

On veut comparer la quantité d'alcool acquis (en g/l) dans des cidres bruts, demi-secs et doux. Pour cela, on a analysé 8 cidres de chaque type et répertorié les quantités d'alcool acquis.

Démarche

1. Importation des données.
2. Mise en oeuvre de l'analyse de la variance avec R : fonctions **AovSum()**, **lm()** et **aov()**

- Représentation graphique (boites de dispersions)
- Proposer un modèle statistique permettant d'étudier le lien entre les variables pour répondre à la problématique
- Estimation des paramètres du modèle
- Indicateur de la qualité de l'ajustement

Mise en place de la session de travail

Sur l'espace de travail du DU sur l'ENT, récupérer le fichier *CIDRES.txt* et l'enregistrer dans le répertoire *H:/Mes Documents/modeleLineaire/data*

- Définir le répertoire de travail: *Session > Set Working Directory > Choose Directory* et sélectionner le répertoire *modeleLineaire*. La commande R est générée automatiquement :

```
setwd(dir = "C:/Businessdecision/Enseignement/STID - LP/Enseignement/ModeleLineaire-DU-ETD/")
```

Statistiques descriptives simples

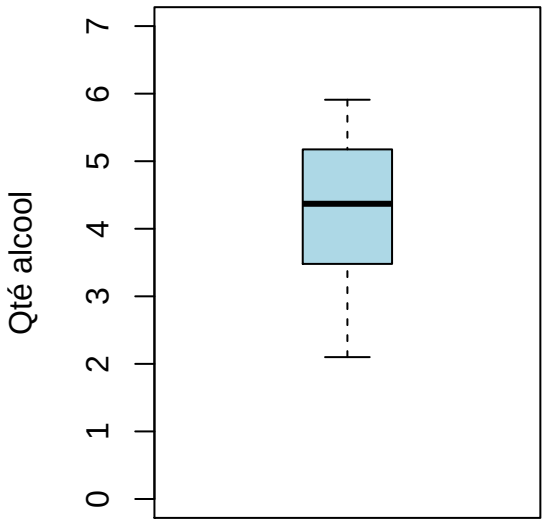
```
donnees=read.delim2("data/CIDRES.txt")
dim(donnees)
```

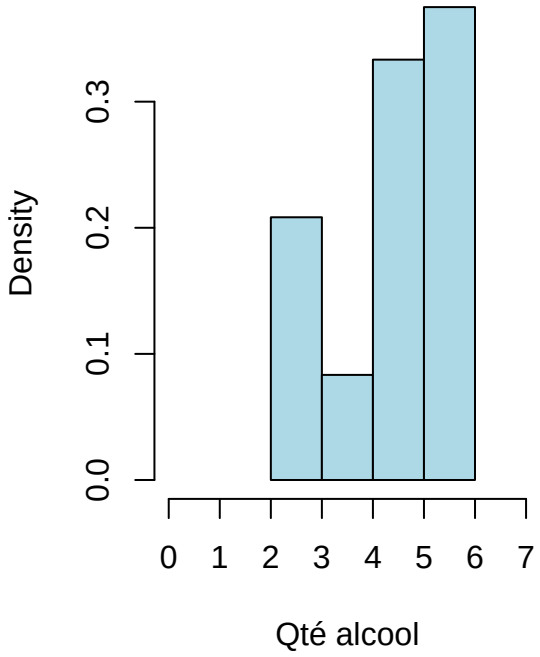
```
## [1] 24  2
```

```
summary(donnees)
```

```
##    qteAlcool       typeCidre
##  Min.   :2.100   Brut   :8
##  1st Qu.:3.650   DemiSec:8
##  Median :4.370   Doux   :8
##  Mean   :4.303
##  3rd Qu.:5.173
##  Max.   :5.910
```

```r
# Variable QteAlcool
boxplot(donnees$qteAlcool, col="lightblue", boxwex=0.5, ylim=c(0,7), pch="*", ylab="Qté alcool")
hist(donnees$qteAlcool, col="lightblue", breaks=5, xlim=c(0,7), xlab="Qté alcool", main="", freq=FALSE)
```

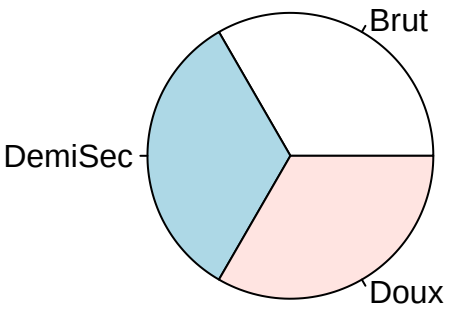





```r
# Variable TypeCidre
tab=table(donnees$typeCidre)
tab
```

```
##
##    Brut DemiSec    Doux
##       8       8       8
```

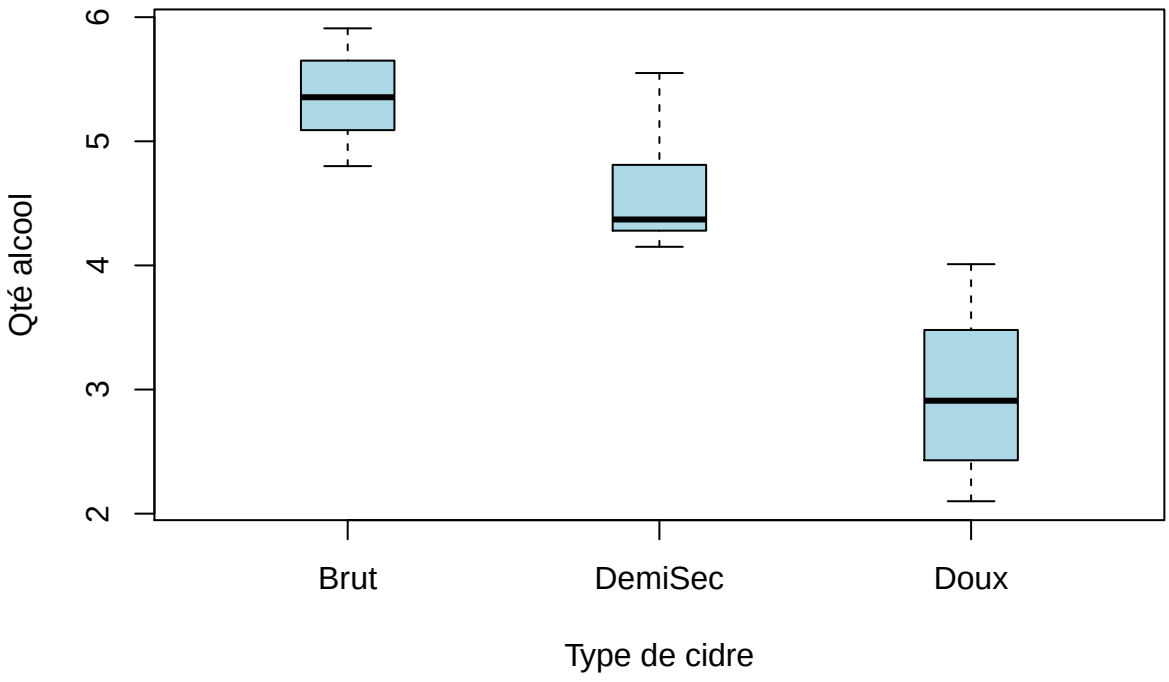
```r
barplot(height=tab, xlab="Type de cidre", ylab="Freq.")
pie(x = tab)
```



Représentation graphique

Représentez graphiquement les données à l'aide de boites de dispersion: la fonction R **boxplot** prend comme argument la variable à représenter (à expliquer), ici la quantité d'alcool, et différents paramètres graphiques permettant de personnaliser les boites de dispersions: entre autre, on peut utiliser ici *boxwex*: un coefficient appliqué à la largeur des boites (utile si on en représente plusieurs sur un même graphique), *col*: la couleur des boites, ou encore *pch*: la forme des points représentant les valeurs "atypiques" si il y en a, *xlab,ylab*: légendes des axes. Si, comme ici, on souhaite représenter notre variable à expliquer en fonction d'une variable qualitative, on spécifie à la fonction **boxplot** une formule, précisant la variable de groupe:

```
boxplot(formula=qteAlcool~typeCidre, data=donnees, boxwex=0.3, col="lightblue", pch=20, xlab="Type de c
```



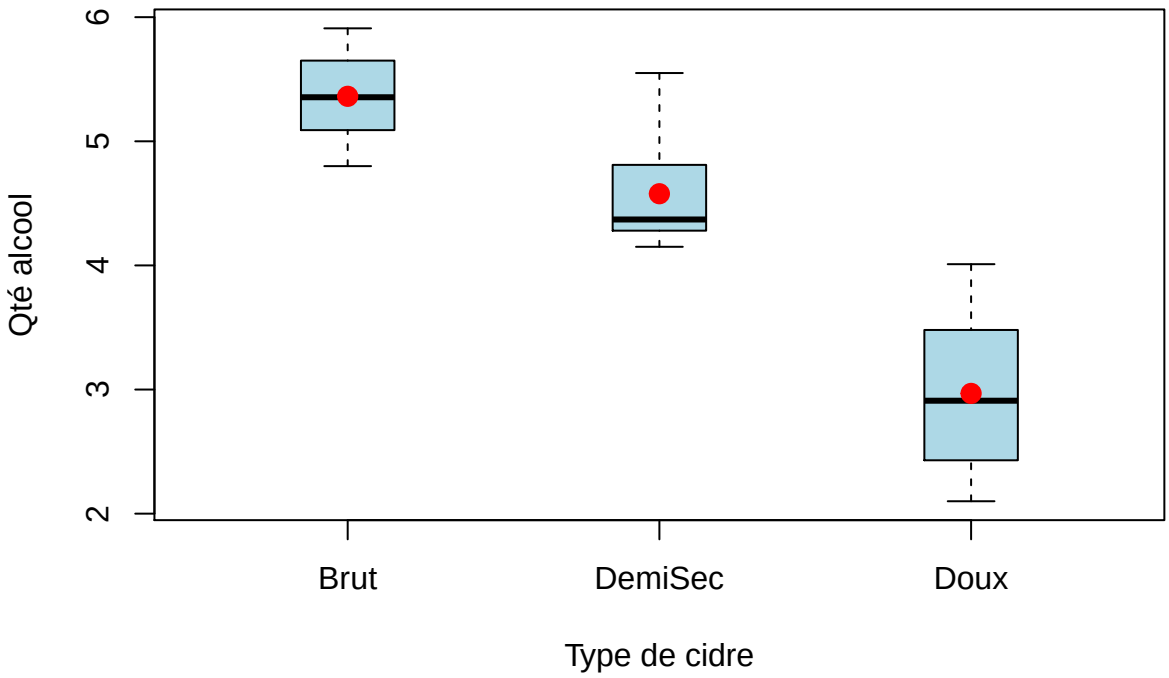
La fonction **by** permet de calculer, pour une variable quantitative, un indicateur (donné dans l'argument *FUN*) suivant les modalités d'une variable qualitative (donnée dans l'argument *INDICES*). On va l'utiliser ici pour calculer les moyennes de quantité d'alcool pour chaque type de cidre:

```
moy = by(donnees$qteAlcool,INDICES=donnees$typeCidre, FUN=mean )
moy
```

```
## donnees$typeCidre: Brut
## [1] 5.3625
## ------------------------------------------------------------
## donnees$typeCidre: DemiSec
## [1] 4.5775
## ------------------------------------------------------------
## donnees$typeCidre: Doux
## [1] 2.96875
```

On ajoute ensuite les points correspondants aux moyennes des 3 types de cidres sur le graphique précédent à l'aide de la fonction **points**. Cette fonction prend en argument deux vecteurs x et y contenant les coordonnées des points à ajouter sur le graphique, respectivement en abscisses (x) et en ordonnées (y), ainsi que des paramètres graphiques.

```
boxplot(formula=qteAlcool~typeCidre, data=donnees, boxwex=0.3, col="lightblue", pch=20, xlab="Type de c
points(x=1:3, y=moy, pch=20, col="red", cex=2)
```



Analyse de la variance

L'analyse de la variance permet d'étudier une variable quantitative en fonction d'une variable qualitative. Plusieurs fonction de R permettent de mener ce type d'analyse:

- Fonction **lm** (l'analyse de la variance fait partie de la classe de méthodes de modèle linéaire):

```
mod.lm=lm(formula=qteAlcool~typeCidre,data=donnees)
anova(mod.lm)
```

```
## Analysis of Variance Table
##
## Response: qteAlcool
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## typeCidre  2 23.8249 11.9125  41.619 4.944e-08 ***
## Residuals 21  6.0108  0.2862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod.lm)
```

```
##
## Call:
## lm(formula = qteAlcool ~ typeCidre, data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.8688 -0.3312 -0.1550  0.2525  1.0413
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.3625     0.1892  28.350  < 2e-16 ***
## typeCidreDemiSec  -0.7850     0.2675  -2.935  0.00792 **
## typeCidreDoux     -2.3938     0.2675  -8.949 1.31e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.535 on 21 degrees of freedom
## Multiple R-squared:  0.7985, Adjusted R-squared:  0.7794
## F-statistic: 41.62 on 2 and 21 DF,  p-value: 4.944e-08
```

- Fonction **AovSum** du package **FactoMineR** (à installer la 1ere fois) à charger à chaque session : utiliser l'utilitaire de R-studio et cocher le package dans la liste pour l'utiliser ou bien à l'aide de la commande *library*

```
library(FactoMineR)
```

Aide sur la fonction **AovSum**:

```
?AovSum
```

```
mod.aovSum=AovSum(formula=qteAlcool~typeCidre,data=donnees)
mod.aovSum
```

```
## Ftest
##                SS df      MS F value    Pr(>F)
## typeCidre 23.8249  2 11.9125  41.619 4.944e-08 ***
## Residuals  6.0108 21  0.2862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ttest
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.30292    0.10921 39.4015   <2e-16 ***
## typeCidre - Brut     1.05958    0.15444  6.8607   <2e-16 ***
## typeCidre - DemiSec  0.27458    0.15444  1.7779   0.0899 .
## typeCidre - Doux    -1.33417    0.15444 -8.6386   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fonction **aov**:

```
mod.aov = aov(formula=qteAlcool~typeCidre,data=donnees)
summary(mod.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## typeCidre    2 23.825  11.912   41.62 4.94e-08 ***
## Residuals   21  6.011   0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparaisons multiples

L'analyse de la variance permet de conclure qu'il existe au moins une différence significative entre les moyennes des I groupes (associés aux I modalités du facteurs), mais n'indique pas quelles sont les paires pour lesquelles ces différences de moyennes sont significatives. La réponse à cette question passe par la mises en oeuvre de tests de comparaisons multiples. Le test de Tukey est implémenté dans la fonction **TukeyHSD**. Cette fonction prend comme argument x le résultat de l'ANOVA obtenue avec la fonction **aov** et retourne les probabilités critiques ajustées:

```
res.Tukey = TukeyHSD( x = mod.aov, conf.level=0.95)
res.Tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = qteAlcool ~ typeCidre, data = donnees)
##
## $typeCidre
##                   diff       lwr       upr     p adj
## DemiSec-Brut -0.78500 -1.459256 -0.110744 0.0207321
## Doux-Brut    -2.39375 -3.068006 -1.719494 0.0000000
## Doux-DemiSec -1.60875 -2.283006 -0.934494 0.0000165
```

```
plot(res.Tukey)
```