# TD2

ING1 EC551 Statistiques descriptives

Galharret Jean-Michel
département MSC
https://galharret.github.io/WEBSITE/cours_ONIRIS.html

## Exercice 1 : Lien entre deux variables quantitatives

1. Calculer la matrice de corrélation entre les différents types d'acide gras contenus dans les huiles d'olives (fonction ***cor***).

```
cor(h_olive[,3:10])
```

2. A l'aide du package ***corrplot*** que vous chargerez au préalable, représenter la matrice de corrélation sous la forme ci-dessous (fonction ***corrplot***). Quelles corrélations vous semblent intéressantes ?

***Remarque*** La matrice de corrélation précédente n'est pas facile à lire donc on va utiliser deux libraries de R qui vont permettre d'améliorer cette lecture.

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
M <- cor(h_olive[,3:10])
corrplot(M,type="upper")
```

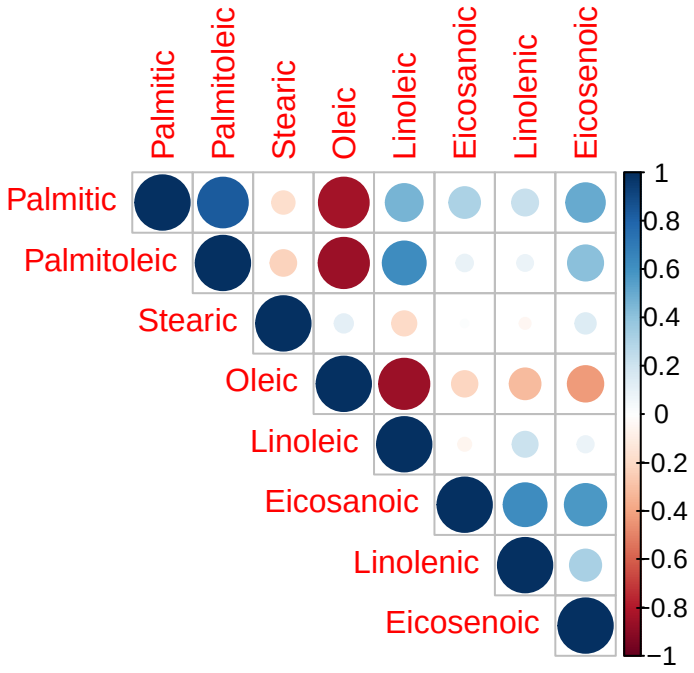



Figure 1: Matrice de corrélation

```
corrplot(M, method = "number", type="upper")
```

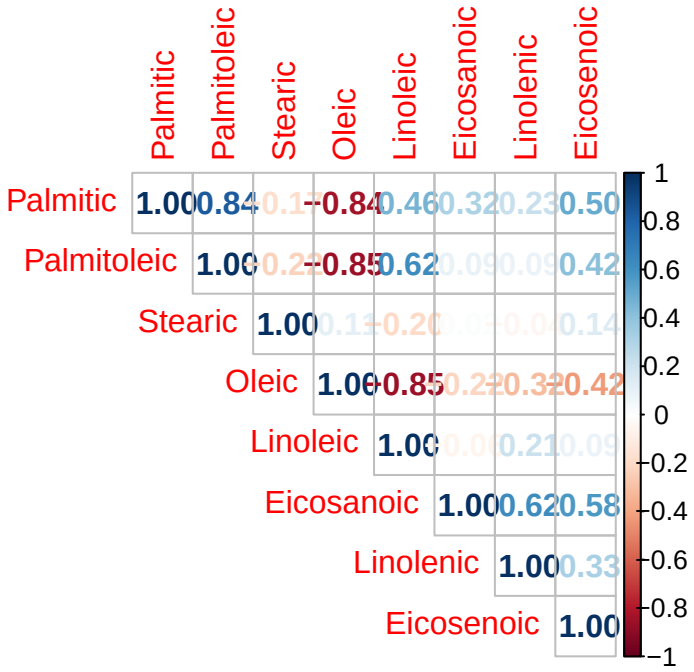



Figure 2: Matrice de corrélation

La présence de l'acide Palmitic est fortement lié à la présence de l'acide Palmitoleic, l'acide Linolenic. A l'inverse la présence de l'acide Palmitic est associée à l'absence de l'acide Oleic.

3. A l'aide du package ***psych*** que vous chargerez au préalable, reproduire le graphique ci-dessous (fonction ***panels.pairs***). Ecrire le lien qui existe entre les variables considérées.

```
library(psych)
pairs.panels(h_olive[,c(3,4,6)])
```

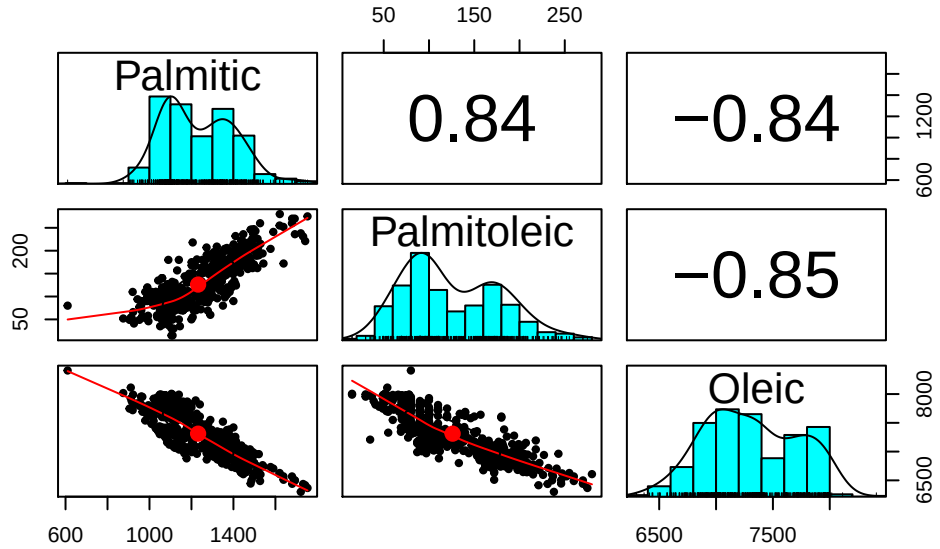



Figure 3: Graphe des corrélations

```
pairs.panels(h_olive[,c(3,4,6)],smooth = "",lm=T)
```

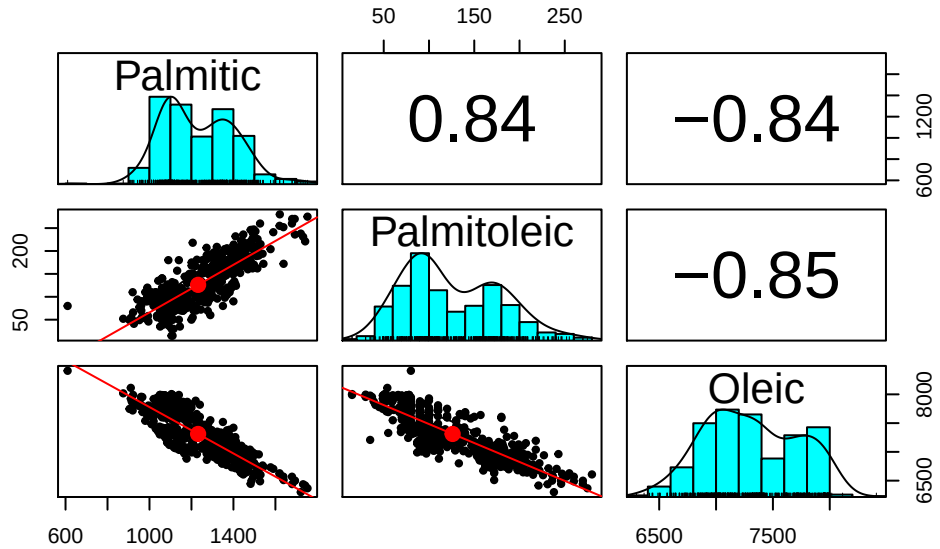


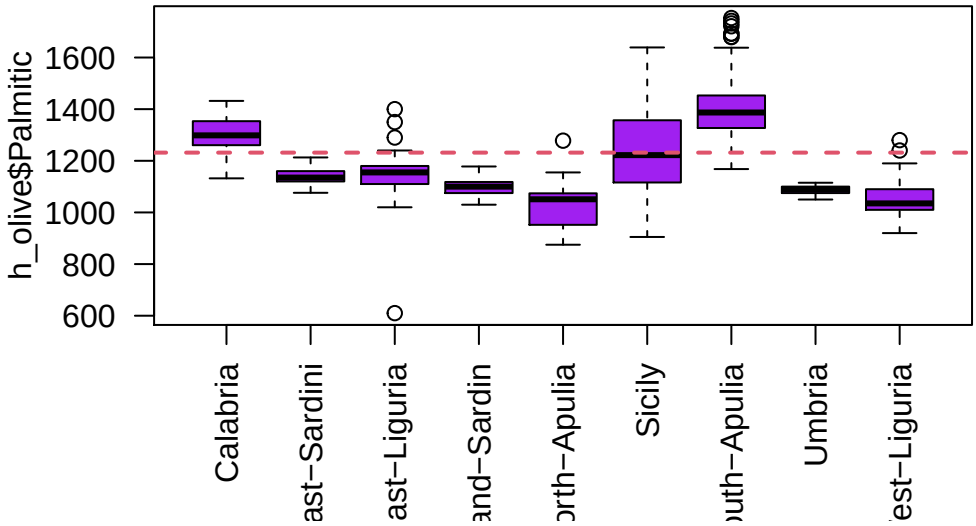

Figure 4: Graphe des corrélations

On a un lien qui est fort entre ces trois variables positif entre Palmitic et Palmitoleic, négatif entre Palmitic et Oleic.

# Exercice 2 : Lien entre 1 variable qualitative et une variable quantitative

On cherche le lien entre la région et la présence de l'acide gras Palmitic :

```
boxplot(h_olive$Palmitic~h_olive$Region,las=2,xlab="",col="purple")
abline(h=mean(h_olive$Palmitic),col=2,lty=2,lwd=2)
```




***Remarque*** on peut considérer que le lien va exister puisqu'on voit clairement sur la boite à moustache que dans certaines régions la présence d'acide palmitique est bien au dessus de la moyenne globale alors que pour d'autres régions elle est bien en dessous.

1. Définissez sur R la fonction VAR telle que $Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$.

```
VAR<-function(x){return(mean((x-mean(x))^2))}
```

Attention : comme on l'a déjà vu précédemment la fonction ***var*** définie dans R calcule $\frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - \bar{x})^2$, c'est à dire une estimation de la variance de la population.

2. Calculer les moyennes et les variances de l'acide palmitique contenue dans les huiles en fonction de la région considérée (fonction ***by***). Les résultats seront stockées dans deux vecteurs $Moy\_Region$ et $Var\_Region$ (pour transformer une liste en vecteur on utilisera la fonction ***as.vector***).

```
moy_Region<-as.vector(by(h_olive$Palmitic,h_olive$Region,mean))
var_Region<-as.vector(by(h_olive$Palmitic,h_olive$Region,VAR))
```

3. En déduire quelle région a en moyenne le plus d'acide palmitique et dans quelle région les huiles sont le plus dispersées en termes d'acide palmitique.

```
max(moy_Region)
```

```
[1] 1395.67
```

```
max(var_Region)
```

```
[1] 31722.73
```

4. En utilisant le cours (diapo 74) calculer les variances inter et intra de l'acide palmique selon la région considérée.

```
N<-dim(h_olive)[1]
n_k<-as.vector(table(h_olive$Region))
m<-mean(h_olive$Palmitic)
V_inter<-sum(n_k*(moy_Region-m)^2)/N
V_intra<-sum(n_k*var_Region)/N
V_inter/(V_inter+V_intra)
```

```
[1] 0.7005781
```

5. Quel est l'acide gras dont la présence est le plus lié à la région considérée ? On pourra utiliser la question précédente.

```
p<-c()
for(i in 3:10){
  moy_region<-as.vector(by(h_olive[,i],h_olive$Region,mean))
  var_region<-as.vector(by(h_olive[,i],h_olive$Region,VAR))
  n_region<-table(h_olive$Region)
  moy<-mean(h_olive[,i])
  N<-length(h_olive[,i])
  V_inter<-sum(n_region*(moy_region-moy)^2)/N
  V_intra<-sum(n_region*var_region)/N
  p<-c(p,V_inter/VAR(h_olive[,i]))
}
names(p)<-colnames(h_olive)[3:10]
which.max(p)
```

```
Eicosenoic
         8
```

# Exercice 3 : Lien entre deux variables qualitatives

Téléchargez et ouvrez sur Connect le fichier ***mais.txt*** :

1. Etablir la table de contingence des variables enracinements et couleur de l'épi. On ajoutera la somme marginale des colonnes

```
Tab<-table(mais$Enracinement,mais$Couleur)
Tab2<-cbind(Tab,margin.table(Tab,margin=2))
colnames(Tab2)[4]<-"Somme"
```

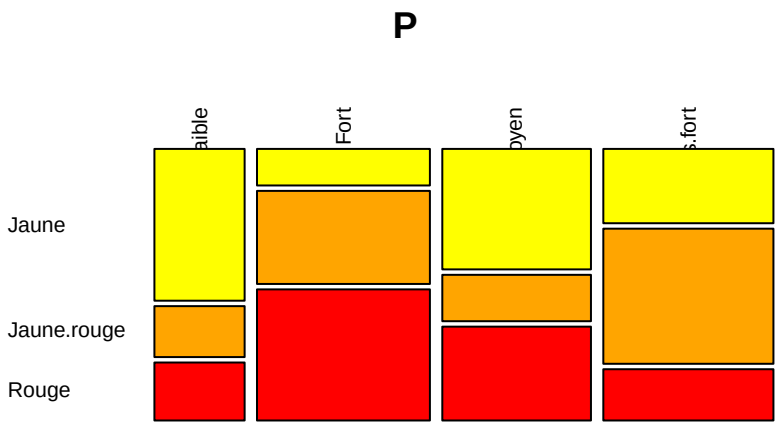
On obtient la table :

```
kable(Tab2,format = "latex",booktabs = TRUE)
```

| | Jaune | Jaune.rouge | Rouge | Somme |
|---|---|---|---|---|
| Faible | 13 | 2 | 3 | 48 |
| Fort | 6 | 7 | 13 | 22 |
| Moyen | 17 | 3 | 8 | 29 |
| Tres.fort | 12 | 10 | 5 | 48 |

Etablir les profils de l'enracinement selon la couleur (fonction ***proportions***).

```
P<-proportions(Tab,margin=2)
mosaicplot(P,col=c("yellow","orange","red"),las=2)
```




Les profils d'enracinement selon la couleur du maïs sont donnés par le graphe précédent, celui-ci

illustre le potentiel lien entre ces variables. Attention il n'est pas facile à interpréter !

2. Pour calculer la valeur du $\chi^2$ on va utiliser la fonction ***chisq.test***. A partir de l'aide déterminer la valeur correspondante.

On peut faire le calcul à partir de la formule du cours en utilisant les valeurs observées et les valeurs théoriques :

```
res<-chisq.test(Tab)
```

```
Warning in chisq.test(Tab): Chi-squared approximation may be incorrect
```

```
res$observed
```

```
           Jaune Jaune.rouge Rouge
Faible        13           2     3
Fort           6           7    13
Moyen         17           3     8
Tres.fort     12          10     5
```

```
res$expected
```

```
              Jaune Jaune.rouge    Rouge
Faible     8.727273    4.000000 5.272727
Fort      12.606061    5.777778 7.616162
Moyen     13.575758    6.222222 8.202020
Tres.fort 13.090909    6.000000 7.909091
```

```
C2<-sum((res$observed-res$expected)^2/res$expected)
```

Ou bien on peut directement utiliser la valeur qui est stockée dans statistique :

```
C2<-chisq.test(P)$statistic
```

```
Warning in chisq.test(P): Chi-squared approximation may be incorrect
```

3. En déduire $V^2$ (diapo 82).

```
N<-dim(mais)[1]
V2<-C2/(N*min(dim(Tab)[1]-1,dim(Tab)[2]-1))
V2
```

```
  X-squared
0.002772053
```

Ici le lien entre les deux variables qualitatives est très faible, le fait d'être enraciné est très peu lié à la couleur du mais.