Procédure analyse

Florian Fort

02/04/2020

## I - Qu’est ce que je veux faire ?

1 - Une variable qualitative (un trait trouvé dans la biblio) est-t-elle différente entre plusieurs groupes qualitatifs (ex: stades de succession) ?

**Xhi deux**

2 - Une variable quantitative(un trait mesuré) est-t-elle différente entre plusieurs groupes qualitatifs (ex: stades de succession) ?

**ANOVA (ANalyse Of VAriance)**

3 - Trouver une relation entre 2 variables quantitatives (2 traits)

**Test de correlation**

## - II - Préparation du fichier

Faire un fichier simple, propre : pas de colonnes fusionnées pour un titre, pas d’acent, les décimales doivent être marquée par des virgules ou des points mais pas les deux.

Le fichier doit être formaté en fonction de ce que je veux tester (ex: fichier de moyenne si test se fait sur moyenne)

Le fichier = 1 feuille excel (pas un classeur)

L’enregistrer au format .csv (séparateur “;”)

Fichier<-**read.table**("data\_quanti.csv", header=TRUE, dec= ",", sep =";")  
Fichier

chi<-**chisq.test**(Tableau)  
chi

**barplot**(Tableau)

m<-**aov**(Hauteur**~**Stade,data=Fichier)  
**summary**(m)

**par**(mfrow=**c**(2,2)) *# affiche une fenêtre graphique 2x2*  
**plot**(m) *# affiche les résidus du modèle*

**shapiro.test**(m**$**residuals)

**bartlett.test** (m**$**residuals **~** Fichier**$**Stade)

**library**(agricolae)  
hsd<-**HSD.test**(m,"Stade")  
hsd

### Ouvrir le fichier :

Spécifiez votre répertoire courant

**setwd**("~/Enseignement 2019-2020/L3-1A/UED/Stats")

Fichier = Nom que je donne au fichier (mettre ce qu’on veut) Ce qui est entre “…” = chemin d’accès au fichier. Il se termine par le nom sous lequel j’ai enregistré mon fichier.txt (si enregistré en txt) header= TRUE signifie que mes colonnes ont des titres (si ce n’est pas le cas écrire FALSE) dec = " , " signifie que mes décimales sont séparées par des virgules sep = “;” signifie que mes colonnes sont séparées par un point virgule (format d’enregistrement)

### Vérifier que le fichier soit OK

Fichier<-**read.table**("data\_quanti.csv", header=TRUE, dec= ",", sep =";")  
Fichier

## Stade Genre\_Espèces Hauteur Encombrement Volume Longueur  
## 1 I Bromus\_madritensis 22.1250 266.00000 4.228000e+03 16.11250  
## 2 I Calendula\_arvensis 13.4500 44.68750 4.117992e+02 4.38750  
## 3 I Crepis\_sancta 14.9125 56.59375 5.661217e+02 5.36250  
## 4 I Diplotaxis\_euricoides 34.8750 373.50000 9.926750e+03 6.30000  
## 5 I Euphorbia\_helioscopia 10.4875 2.88750 2.081417e+01 1.38250  
## 6 I Geranium\_rotundifolium 4.1125 112.12500 3.194833e+02 6.27500  
## 7 I Lamium\_amplexicaule 11.3125 54.25000 4.063333e+02 3.50625  
## 8 I Lepidium\_draba 38.2500 168.12500 4.403917e+03 5.08125  
## 9 I Malva\_sylvesteris 10.7500 878.00000 7.266833e+03 9.86250  
## 10 I Medicago\_arabica 7.2875 30.30125 1.526712e+02 6.38750  
## 11 I Senecio\_vulgaris 15.5625 8.81250 9.679167e+01 5.55000  
## 12 I Veronica\_cymbalaria 7.8125 188.68750 1.053917e+03 1.59125  
## 13 II Bellis\_sylvestris 5.7500 142.40625 5.447292e+02 7.06250  
## 14 II Bituminaria\_bituminosa 28.6250 276.75000 5.847917e+03 2.06250  
## 15 II Brachypodium\_ramosum 38.3750 270.37500 6.892083e+03 10.95000  
## 16 II Bromus\_erectus 21.7500 215.25000 3.204542e+03 16.33750  
## 17 II Dactylis\_glomerata 19.8125 887.37500 1.182458e+04 15.68750  
## 18 II Gallium\_aparine 20.8750 720.50000 1.108333e+04 3.03750  
## 19 II Plantago\_lanceoloata 12.5625 245.43750 2.045625e+03 10.26250  
## 20 II Rumex\_scutatus 14.7500 244.25000 2.539667e+03 10.40000  
## 21 II Sanguisorba\_minor 12.2500 285.00000 2.372167e+03 8.80000  
## 22 II Sedum\_album 7.6250 198.56250 1.063042e+03 1.16250  
## 23 II Thymus\_vulgaris 14.2500 492.75000 4.848833e+03 0.44875  
## 24 II Vicia\_sativa 34.5000 105.00000 3.644375e+03 3.80000  
## 25 III Arum\_italicum 32.1250 489.37500 1.588963e+04 23.40000  
## 26 III Cistus\_monspeliensis 100.2500 4382.00000 4.764359e+05 3.28500  
## 27 III Coronilla\_glauca 123.3750 10980.50000 1.492558e+06 1.80625  
## 28 III Hedera\_helix 247.0000 22442.87500 8.869940e+06 12.87500  
## 29 III Laurus\_nobilis 232.8750 28391.25000 1.250905e+07 11.20000  
## 30 III Osyris\_alba 78.1250 3452.50000 3.952900e+05 3.13750  
## 31 III Phillyrea media\_Phillyrea 273.7500 51130.62500 1.455111e+07 3.87500  
## 32 III Quercus\_ilex 566.8750 495575.50000 4.040774e+08 5.77500  
## 33 III Rhammus\_alaternus 267.6000 35314.51250 1.164752e+07 5.01250  
## 34 III Ruscus\_aculeatus 69.5250 826.31625 7.454494e+04 2.90625  
## 35 III Smilax\_aspera 157.5000 2020.25000 3.364508e+05 7.51250  
## 36 III Viburnum\_tinus 183.8750 13035.25000 2.656196e+06 10.30000  
## Surface SLA LDMC Alloc\_Tige Alloc\_Repro  
## 1 7.70625000 28.587158 223.32897 0.39771521 0.149545065  
## 2 2.88000000 26.201456 111.03218 0.28766373 0.394360618  
## 3 2.96750000 35.316224 91.65116 0.02964477 0.528516814  
## 4 9.62250000 19.999908 148.60860 0.51548295 0.283563699  
## 5 0.51204167 25.567276 238.94114 0.46002090 0.188297872  
## 6 3.97500000 27.452790 161.65464 0.02580912 0.000000000  
## 7 2.58520833 33.085978 177.14006 0.22272958 0.270994998  
## 8 9.74250000 22.733948 170.85738 0.38626288 0.138906785  
## 9 17.19500000 21.664274 149.21568 0.20061126 0.000000000  
## 10 3.72375000 28.578954 139.02748 0.13327779 0.056344779  
## 11 5.63625000 22.381331 98.18843 0.41886193 0.199607330  
## 12 0.91791667 32.714926 99.43898 0.27326467 0.067175363  
## 13 5.81000000 19.248098 139.27560 0.10173858 0.000000000  
## 14 2.50000000 22.126193 181.62839 0.52591445 0.000000000  
## 15 1.36833333 16.477819 396.03906 0.70854981 0.000000000  
## 16 3.60250000 14.936133 312.76835 0.51775431 0.000000000  
## 17 5.07750000 21.637529 286.96491 0.41649500 0.000000000  
## 18 0.71367857 26.344744 152.07596 0.30955736 0.000000000  
## 19 5.39875000 13.975637 176.16663 0.18139912 0.030382139  
## 20 19.66875000 20.559907 130.93019 0.06739046 0.000000000  
## 21 4.44125000 14.117634 257.41003 0.15983221 0.000000000  
## 22 0.89781250 20.453078 114.53708 0.42396538 0.000000000  
## 23 0.03032983 9.296342 348.75162 0.75713443 0.000000000  
## 24 2.84000000 30.387553 170.47474 0.27719245 0.000000000  
## 25 50.21375000 24.310487 94.20965 0.00000000 0.000000000  
## 26 1.30175000 6.629085 361.83938 0.36367652 0.000000000  
## 27 1.22687500 27.697096 177.16228 0.59249078 0.157966276  
## 28 26.97500000 10.334522 366.12651 0.56751927 0.000000000  
## 29 24.95125000 8.316270 509.18125 0.40352787 0.000000000  
## 30 0.99393750 8.754543 290.43422 0.66519492 0.000000000  
## 31 4.94166667 7.055050 488.21883 0.59958460 0.041268163  
## 32 10.68625000 6.474570 534.09164 0.46939373 0.000000000  
## 33 8.55875000 9.669953 414.81957 0.41327991 0.020050537  
## 34 2.32562500 8.856395 395.88840 0.74579957 0.005605078  
## 35 22.51375000 11.016393 365.93711 0.53298490 0.000000000  
## 36 32.01125000 9.067163 446.70736 0.45261687 0.009031196

### si le fichier est gros, préférez les commandes :

**str**(Fichier)

## 'data.frame': 36 obs. of 11 variables:  
## $ Stade : Factor w/ 3 levels "I","II","III": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Genre\_Espèces: Factor w/ 36 levels "Arum\_italicum",..: 6 7 10 12 13 15 17 19 20 21 ...  
## $ Hauteur : num 22.1 13.4 14.9 34.9 10.5 ...  
## $ Encombrement : num 266 44.69 56.59 373.5 2.89 ...  
## $ Volume : num 4228 411.8 566.1 9926.8 20.8 ...  
## $ Longueur : num 16.11 4.39 5.36 6.3 1.38 ...  
## $ Surface : num 7.706 2.88 2.967 9.623 0.512 ...  
## $ SLA : num 28.6 26.2 35.3 20 25.6 ...  
## $ LDMC : num 223.3 111 91.7 148.6 238.9 ...  
## $ Alloc\_Tige : num 0.3977 0.2877 0.0296 0.5155 0.46 ...  
## $ Alloc\_Repro : num 0.15 0.394 0.529 0.284 0.188 ...

**names** (Fichier)

## [1] "Stade" "Genre\_Espèces" "Hauteur" "Encombrement"   
## [5] "Volume" "Longueur" "Surface" "SLA"   
## [9] "LDMC" "Alloc\_Tige" "Alloc\_Repro"

## - III - Tests

### 1 CHI DEUX

#### Les hypothèses

H0 : Il y a indépendance entre les deux variables H1 : Les deux variables ne sont pas indépendantes

Le tableau permettant l’analyse est une matrice d’effectif distribués dans les différentes cases à partir de leurs valeurs pour les deux variables.

#### Conditions validité pour effectuer le test du Chi2

Condition n°1 : L’effectif total du tableau de contingence (N..) doit être supérieur ou égal à 20 Condition n°2 : L’effectif marginal du tableau de contingence (Ni. ou N.j) doit toujours être supérieur ou égal à 5 (somme ligne ou colonne). Condition n°3 : L’effectif théorique (N\*ij) des cases du tableau de contingence doit être supérieur à 3

#### Application

Pour cette analyse il vous faut rentrer vos données dans une table que vous saisissez.

Ici le tableau est une matrice avec colonnes = stades 1 2 3 et lignes = trait qualit (ex : les differents modes de dispersion). Dans le tableau, on trouve donc par ex le nombre de sp anémochores du stade 1, 2 et 3 (en ligne 1, colonnes 1, 2 et 3 respectivement), en ligne 2 le nb d’sp zoochores pour le stade 1 , 2 et 3, etc.

Tableau = nom de mon tableau; nrow = nb lignes et ncol= nb colonnes, byrow= j’ai écrit la liste des valeurs par ligne (pas par colonne)

Tableau<-**matrix**(**c**(5,30,3, 11,12,13), nrow = 2, ncol=3, byrow=TRUE,  
 dimnames = **list**(**c**("anémochorie", "zoochorie"),  
 **c**("stade1", "stade2", "stade3")))

Mais vous pouvez utiliser la fonction table à partir de deux varaibles qualitatives

Tab<-**table**(Y,Y1)  
Tab

## Y1  
## Y D E  
## A 10 0  
## B 5 5  
## C 0 10

Ensuite vous utilisez la fonction chisq.test pour faire le test Vous pouvez aussi afficher les valeurs attendues pour tester les conditions de validité

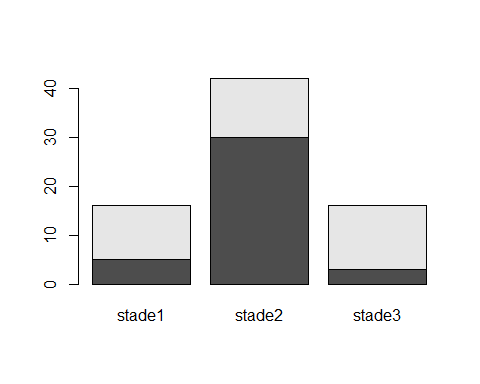
chi<-**chisq.test**(Tableau)  
chi

##   
## Pearson's Chi-squared test  
##   
## data: Tableau  
## X-squared = 16.172, df = 2, p-value = 0.0003078

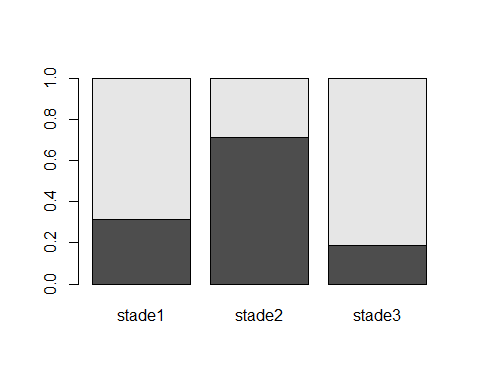
chi**$**expected

## stade1 stade2 stade3  
## anémochorie 8.216216 21.56757 8.216216  
## zoochorie 7.783784 20.43243 7.783784

En terme de représentation on optera pour des diagrammes en bare avec les valeurs brutes ou avec les proportion d’espèces dans chaque stade.



tab<-**prop.table**(Tableau,2)  
**barplot**(tab)



### 2 ANOVA

H0 : Il n’y a pas de différence de moyenne entre les groupes H1 : Au moins deux groupes ont des moyennes différentes

#### Il faut d’abord faire le test avant de vérifier les conditions d’application

#### La variable stade doit être déclarée comme facteur si cela n’est pas compris automatiquement par R.

FStade<-**as.factor**(Fichier**$**Stade)

On utilise la fonction aov(Y~X) pour réaliser une ANOVA

m<-**aov**(Hauteur**~**Stade,data=Fichier)  
**summary**(m)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Stade 2 250191 125095 18.11 4.9e-06 \*\*\*  
## Residuals 33 227880 6905   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Si p-value (ici, Pr(<F)) < 0.05, alors l’hypothèse nulle (= les moyennes sont identiques entre les stades) est rejettée. Au moins, un stade de succession possède une moyenne différente des autres stades. F value = statistique du test (plus F est grand, plus les stades sont différents entre eux)

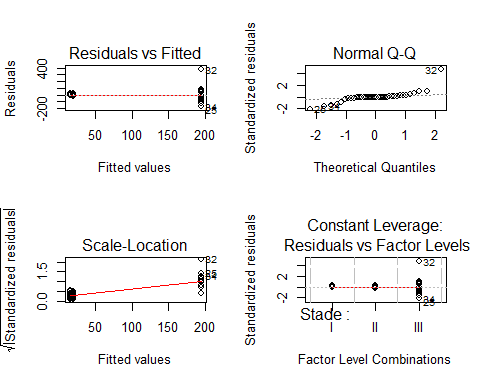
#### Conditions de validité à vérifier (pour chaque traits) :

Afficher les valeurs des résidus du modèle

m<-**aov**(Hauteur**~**Stade,data=Fichier)  
**summary**(m)

**par**(mfrow=**c**(2,2)) *# affiche une fenêtre graphique 2x2*  
**plot**(m) *# affiche les résidus du modèle*

**shapiro.test**(m**$**residuals)



##### Normalité des résidus du modèle

H0 : les résidus suivent une distribution normale H1 : les résidus ne suivent pas une distribution normale

**shapiro.test**(m**$**residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m$residuals  
## W = 0.68679, p-value = 1.766e-07

Si la p-value est > à 0.05, on ne peut pas rejeter l’hypothèse nulle. Les résidus du modèle sont donc considérées comme suivant une loi normale.

##### Vérification homoscédasticité (égalité des variances) :

H0 : les résidus du modèle ont la même variance pour les différents stades H1 : les résidus du modèles ont des variances différentes pour au moins deux stades

m<-**aov**(Hauteur**~**Stade,data=Fichier)  
**summary**(m)

**par**(mfrow=**c**(2,2)) *# affiche une fenêtre graphique 2x2*  
**plot**(m) *# affiche les résidus du modèle*

**shapiro.test**(m**$**residuals)

**bartlett.test** (m**$**residuals **~** Fichier**$**Stade)

##   
## Bartlett test of homogeneity of variances  
##   
## data: m$residuals by Fichier$Stade  
## Bartlett's K-squared = 76.118, df = 2, p-value < 2.2e-16

Si la p-value est > à 0.05, on ne peut pas rejeter l’hypothèse nulle on considère que les variances des résidus sont égales.

##### Si ces deux conditions sont vérifiées (résidus suivent une loi normale et égalité des variances)

alors je peux faire un test dit paramétrique, sinon il faut faire un test non paramétrique ou transformer les données

##### Si les conditions d’application sont remplies on peut faire un test post-hoc qui ermet de connaitre quels stades sont différents les uns des autres

Ici je vous propose un test post-hoc de Tukey qui va définir des groupes ayant des moyennes homogènes

**library**(agricolae)  
hsd<-**HSD.test**(m,"Stade")  
hsd

## $statistics  
## MSerror Df Mean CV MSD  
## 6905.453 33 76.52604 108.5892 83.24509  
##   
## $parameters  
## test name.t ntr StudentizedRange alpha  
## Tukey Stade 3 3.470189 0.05  
##   
## $means  
## Hauteur std r Min Max Q25 Q50 Q75  
## I 15.91146 10.70601 12 4.1125 38.250 9.81875 12.38125 17.20312  
## II 19.26042 10.23269 12 5.7500 38.375 12.48438 17.28125 23.46875  
## III 194.40625 143.16784 12 32.1250 566.875 94.71875 170.68750 252.15000  
##   
## $comparison  
## NULL  
##   
## $groups  
## Hauteur groups  
## III 194.40625 a  
## II 19.26042 b  
## I 15.91146 b  
##   
## attr(,"class")  
## [1] "group"

#### Dans le cas de la hauteur présenté ici on peut faire une transformation log() car on a beaucoup de petites valeurs et et quelques grandes

m<-**aov**(**log**(Hauteur)**~**Stade,data=Fichier)  
**summary**(m)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Stade 2 43.55 21.777 49.02 1.31e-10 \*\*\*  
## Residuals 33 14.66 0.444   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**shapiro.test**(m**$**residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m$residuals  
## W = 0.99165, p-value = 0.9935

**bartlett.test** (m**$**residuals **~** Fichier**$**Stade)

##   
## Bartlett test of homogeneity of variances  
##   
## data: m$residuals by Fichier$Stade  
## Bartlett's K-squared = 0.95952, df = 2, p-value = 0.6189

hsd<-**HSD.test**(m,"Stade")  
hsd

## $statistics  
## MSerror Df Mean CV MSD  
## 0.4442336 33 3.472931 19.19153 0.6676798  
##   
## $parameters  
## test name.t ntr StudentizedRange alpha  
## Tukey Stade 3 3.470189 0.05  
##   
## $means  
## log(Hauteur) std r Min Max Q25 Q50 Q75  
## I 2.578798 0.6390190 12 1.414031 3.644144 2.276569 2.512444 2.832825  
## II 2.817667 0.5742074 12 1.749200 3.647406 2.524419 2.838778 3.148280  
## III 5.022329 0.7711299 12 3.469635 6.340139 4.545328 5.136841 5.529415  
##   
## $comparison  
## NULL  
##   
## $groups  
## log(Hauteur) groups  
## III 5.022329 a  
## II 2.817667 b  
## I 2.578798 b  
##   
## attr(,"class")  
## [1] "group"

#### Si même après transformation (par un log par exemple)les conditions d’applications ne sont pas remplies on doit faire un test non paramétrique

### Test non parametriques

Les tests non paramétriques pour remplacer les ANOVA ne pas sur les mêmes critères que l’ANOVA et ne sont donc pas des équivalents.

#### Kruskall-Wallis :

H0 : Il n’y a pas de différence entre les distributions des groupes H1 : Au moins deux groupes ont des distributions différentes

**kruskal.test**(Fichier**$**Hauteur**~**Fichier**$**Stade)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Fichier$Hauteur by Fichier$Stade  
## Kruskal-Wallis chi-squared = 22.456, df = 2, p-value = 1.329e-05

Ensuite on peut faire une comparaison des échantillons deux a deux qui donnes les différences entre les pares d’échantillons (p-values < 0.05 les échantillons sont différents)

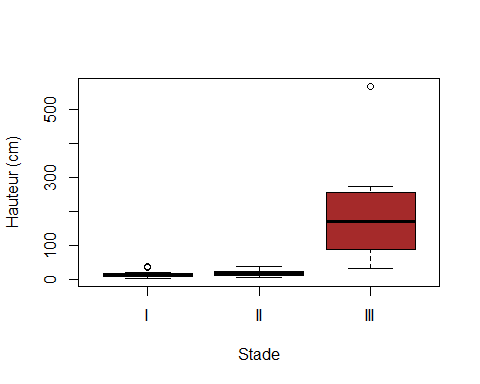
**pairwise.wilcox.test**(Fichier**$**Hauteur,Fichier**$**Stade)

##   
## Pairwise comparisons using Wilcoxon rank sum test   
##   
## data: Fichier$Hauteur and Fichier$Stade   
##   
## I II   
## II 0.38 -   
## III 8.9e-06 8.9e-06  
##   
## P value adjustment method: holm

#### Représentation graphique

Pour ce type de donnée on optera pour des boites à moustache

**boxplot**(Hauteur**~**Stade, data= Fichier, ylab="Hauteur (cm)",col=**c**("lightgreen","darkgreen","brown"))



### 3 CORRELATION

#### Vérification normalité des données

A faire pour chaque trait Si la p-value est > à 0.05, on ne peut pas rejeter l’hypothèse nulle selon laquelle les données sont issues d’une population normalement distribuée. Les données sont donc considérées comme suivant une loi normale. Un test paramétrique peut être utilisé.

**shapiro.test** (Y)

##   
## Shapiro-Wilk normality test  
##   
## data: Y  
## W = 0.97744, p-value = 0.949

**shapiro.test**(Y1)

##   
## Shapiro-Wilk normality test  
##   
## data: Y1  
## W = 0.86245, p-value = 0.02618

#### Test paramétrique

A faire pour chaque “paire” de traits si p-value < 0.05, alors l’hypothèse nulle (=absence de corrélation) est rejetée au risque de 5%; on considère que les deux traits sont corrélés.

**cor.test**(Y,Y1, method="pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Y and Y1  
## t = 1.5946, df = 13, p-value = 0.1348  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.1359638 0.7593909  
## sample estimates:  
## cor   
## 0.4044701

cor = coefficient r de correlation

#### Test non paramétrique

Idem que pour le test paramétrique mais remplacer dans les lignes de commandes par method=“spearman”.

**cor.test**(Fichier**$**Hauteur,Fichier**$**Encombrement, method="spearman")

##   
## Spearman's rank correlation rho  
##   
## data: Fichier$Hauteur and Fichier$Encombrement  
## S = 1750, p-value = 4.007e-07  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.7747748

#### Représentation

Ici j’ai fait le choix de transformer en log les variables pour avoir une représentation plus claire et une relation linéaire

**plot**(**log**(Fichier**$**Hauteur)**~log**(Fichier**$**Encombrement),xlab="Encombrement (log, cm2)",ylab="Hauteur (cm)",col=Fichier**$**Stade,pch=19)

