DEA Analyse et Modélisation des Systèmes Biologiques

Introduction au logiciel S-PLUS© - 1998/1999

Contrôle sur machine (avec solution)

D. Chessel & J. Thioulouse

1. Questions

Q 1 Que vous inspire l'affichage suivant ?

```
> 2*rnorm(1)
[1] -6.4829
```

Q 2 Quel est le résultat ?

```
> print(pi,digits=12)
```

Q 3 L'ordre ci-dessous peut-il renvoyer la valeur 0 ?

```
> sum(rbinom(900,1,0.01))-rbinom(1,900,0.01)
```

Q 4 A quelle condition le résultat suivant est-il possible ?

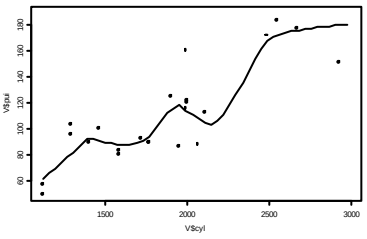
```
> sort(sample(c("a","b",12.7,"c",n0,"aaa","bb"),6,replace=F))
[1] "10"  "a"   "aaa" "b"   "bb"  "c"
```

Q 5 Quelle valeur a t'elle été affectée à n0 ?

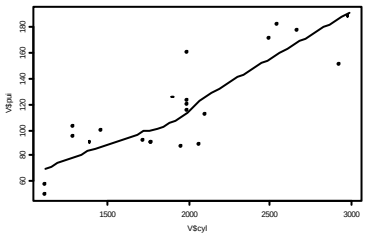
```
> 3*length(cumsum(sample(n0)))-5
[1] 7
```

Q 6 A-t-on nécessairement `0<=a<b<=1` pour obtenir ce qui suit ?

```
> scatter.smooth(V$cyl,V$pui,span=a) # résultat A
> scatter.smooth(V$cyl,V$pui,span=b) # résultat B
```



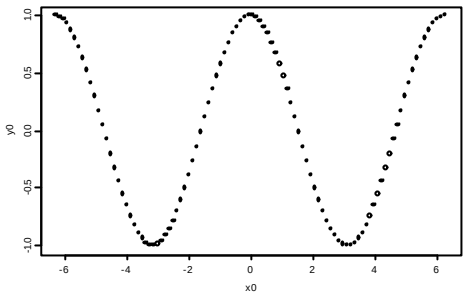
A



B

Q7 Quelle chaîne de caractères a-t-elle été utilisée pour **xxxxx** ?

```
> x0<-seq(-2*pi,2*pi,le=100)
> y0<-XXXXX(x0)
> plot(x0,y0)
```



Q8 A quel résultat vous attendez vous ?

```
> quantile(x,probs=seq(0,1,0.1))
   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
    1  1.9  2.8  3.7  4.6  5.5  6.4  7.3  8.2  9.1   10
> (quantile(x,probs=0.73))<8.2
```

Sur 38 points d'écoute ornithologique on a extrait les données sur deux espèces, la pie et la corneille. Les résultats sont :

	Corneille	
	absente	présente
Pie absente	13	8
Pie présente	6	11

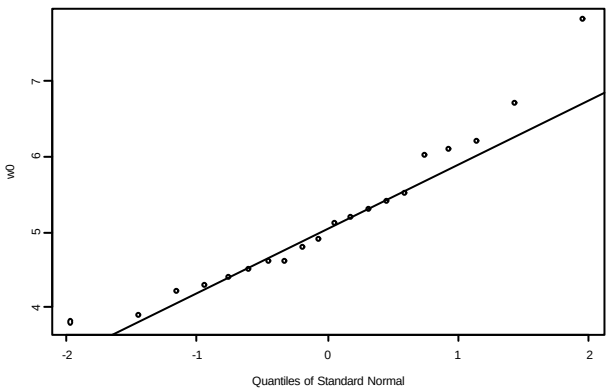
Q9 Quelle est la valeur du Chi2 de cette table de contingence et sa signification ?

Lors d'une étude de croissance, la hauteur de 20 plantules a été mesurée (en cm) :

4.6	7.8	4.8	6	5.4	3.9	4.5	6.2	4.9	5.3	6.7	4.2	3.8	5.1	6.1	5.2	4.3	4.6	4.4	5.5

Q10 Peut-on considérer que cette variable suit une loi normale ? (Note : voir `ks.gof`)

Q11 Quels ordres ont-ils donné ce graphique ?



Dans une conversation autour d'une méthode sur l'analyse des niches écologiques M. Montadert propose un tableau très caractéristique du problème des courbes de réponse aux variables environnementales. Dans n = 102 stations on connaît la présence ou l'absence (0/1) de la gélinotte (geli) et un descriptif de la végétation avec (entre autres) les variables

abondance du Framboisier (fra), abondance du Noisetier (noi) et abondance du Sorbier (sor). fra, noi et sor sont les variables explicatives. geli (binaire) est la variable à expliquer.

Pour contrôler l'implantation des données, utiliser le tableau 1. Le fichier G.txt est disponible dans le répertoire indiqué oralement et sera lu par un ordre du type :

```
> G<-read.table("--- G.txt",h=T,sep=",")
> G
    geli  fra  noi  sor
  1    0 0.75 3.00 0.75
  2    0 1.00 2.75 3.00
  3    0 1.00 2.75 1.25
...
100    1 1.00 1.50 2.00
101    1 0.50 0.75 2.00
102    1 0.50 0.75 3.25
```

	fra	noi	sor
1	1.25	3	0.25
2	1.25	3	3
3	0.75	2.5	2.75
4	1	1.5	4
5	0.25	2.5	2.5
6	0.75	1.75	2
7	1.25	1	3.25
8	1	2.5	2.5
9	1.25	2.25	3.75
10	0.5	2	3.5
11	0	2	1.25
12	0.5	3	2.75
13	0.5	1.5	2
14	1	2.25	2.75
15	0.5	3	2.25
16	0	2.33	2.33
17	0.75	1.25	1.25
18	0.5	0.75	1.25
19	1	1.5	2
20	0.5	0.75	2
21	0.5	0.75	3.25

	fra	noi	sor		fra	noi	sor		fra	noi	sor
1	0.75	3	0.75	28	0	1	1.5	55	1	1.25	2.5
2	1	2.75	3	29	0	0.25	1	56	0.25	0.75	1.5
3	1	2.75	1.25	30	0	1.67	1.67	57	0.5	1.5	2.75
4	1	3	3.25	31	0	1.25	1.25	58	0	2	2.75
5	0	1.25	0.25	32	0.25	3	2.75	59	0	1.5	2
6	1	1.25	1.75	33	0.25	0.25	0.75	60	0	0	0.25
7	0.25	1.75	3	34	0.5	0.75	0.75	61	0.25	0.25	0.75
8	0	1	0.75	35	0.75	0	0.75	62	0.25	0	1.5
9	1	1	0.5	36	1	0	1.25	63	0.25	0.5	1.75
10	1	3	1	37	1	1.25	3.25	64	0.5	0.75	1.75
11	0.5	2.25	2.25	38	0.25	2.25	2	65	0.75	1	1.25
12	0.75	0.75	2.75	39	0	1	1.5	66	1	1	1.5
13	1	0.5	0.5	40	0	2.25	2.25	67	0.5	0	1
14	0	1.75	1.5	41	0	0.75	1.5	68	0.5	1	1.5
15	0.5	1.25	1.75	42	0	0.75	2	69	0.5	0.25	1.75
16	0.5	1.25	2.25	43	0.25	1	1	70	0.25	0.5	1.25
17	0.75	1.75	3.75	44	0	1.33	2.33	71	0	1	2.75
18	1	1.5	3	45	0	1.25	2.5	72	0.25	0.75	1.75
19	1	1.75	3	46	0.25	0	1.25	73	0.25	0.5	1.5
20	1.25	2	2.5	47	0	0.25	3.25	74	1	1	0.75
21	0.25	0	1.25	48	1.5	2	2.75	75	1	0.5	2
22	1	2.25	2.5	49	0.5	0.75	3.25	76	0.75	0.75	3.25
23	0.25	0	0.75	50	1	1	2	77	0.25	0.75	2.25
24	0	2.75	3	51	0.5	0	1	78	0.5	1.25	3
25	0.25	2	2.5	52	0.75	0.25	0.25	79	0	0.25	0.75
26	0.25	1.25	2.75	53	0.75	0.25	1.5	80	0.75	2.75	1.5
27	0.25	0.5	1.5	54	1	1	1.75	81	0.67	3	1.33

Tableau 1 : A gauche 21 stations avec Gelinotte, à droite 81 stations sans Gelinotte.

Q12 Donner les moyennes des variables fra, noi et sor.

Q13 La corrélation entre l'abondance du Sorbier et celle du Noisetier est-elle significativement non nulle ?

Q14 Quelle est la première valeur propre de l'analyse en composantes principales normée du tableau des trois variables d'habitat ?

On appelle tot la variable qui attribue à une station la somme des trois explicatives :

```
> tot<-G$fra+G$noi+G$sor
```

La corrélation entre la première coordonnée de cette ACP et la variable tot montre que la somme simple suffit largement à synthétiser les trois explicatives.

Q15 Quelle est cette corrélation ?.

Q16 Quelle est la fréquence de la présence de la Gélinotte pour les stations où tot est inférieure (respectivement supérieure ou égale) à 4 ?

Q17 La régression logistique montre-t-elle que la probabilité de rencontrer la Gélinotte dépend de chacune des explicatives ?

On a alors deux possibilités :

```
> tot<-G$fra+G$noi+G$sor
glm(G$geli ~ tot, family = binomial)
glm(G$geli ~ G$fra+G$noi+G$sor, family = binomial)
```

Q18 Avez-vous une préférence ?

La suite s'en tiendra au modèle

```
> glm0 <- glm(G$geli ~ tot, family = binomial)
```

La probabilité estimée de rencontrer la gélinotte pour une station où la note cumulée des trois explicatives vaut 0 est donnée par :

```
> predict(glm0,newdata=data.frame(tot=0),type="response")
       1
 0.01312
```

Les paramètres du modèle sont :

```
> glm0$coefficients
 (Intercept)    tot
       -4.32 0.6913
```

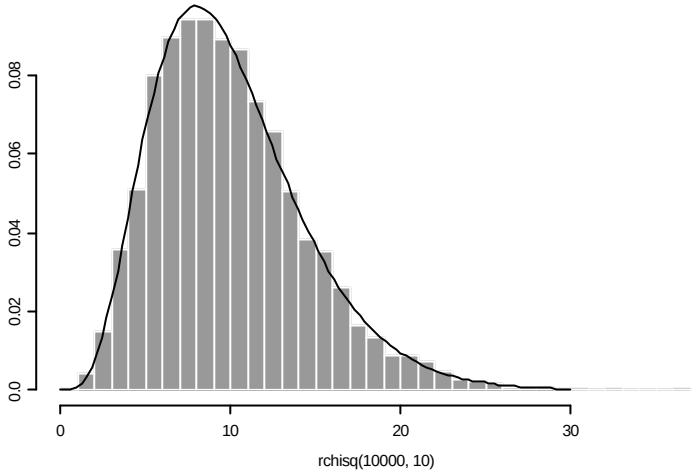
Q19 Quelle relation existe-t-il entre les deux ?

Q20 Reporter sur l'emplacement prévu à cet effet la courbe de réponse de la Gélinotte à la variable tot.

Pour les amateurs :

Q21 Donner le F de l'analyse de variance de la variable tot pour le facteur présence-absence de la gélinotte.

Q22 Quelle est la loi de probabilité représentées ci-dessous ?



Q23 Comment a été tracée la courbe ?

2. Solutions

Q 1

Cette valeur est hautement improbable puisque deux fois une loi normale est normale et n'est inférieure à -4 que dans 2,5 % des cas.

Q 2

```
[1] 3.14159265359
```

Q 3

OUI, car c'est une valeur prise par la différence entre deux variables binomiales de paramètres n = 900 et p = 1/100, sensiblement deux variables poissoniennes de paramètres m = 9.

Q 4

A condition d'avoir affecter à n0 la valeur 10 ou "10". Sample extrait au hasard 6 des 7 valeurs et sort les trie par ordre alphabétique.

Q 5

> n0<-4. sample(4) renvoie une permutation sur 1,...,4, cumsum est un vecteur de 4 valeurs, length() est sa longueur 4, 3*4-5 = 12 - 5 = 7

Q 6

OUI. Utiliser > ?scatter.smooth. Donne span smoothing parameter, valeur par défaut 2/3. B est plus lisse que A.

Q 7

Il faut mettre cos pour **XXXXX**. Trace la fonction cosinus sur -2*pi, +2*pi.

Q 8

On attend **T** pour true. C'est vrai le quantile 0.73 est plus petit que le quantile 0.80.

Q 9

Le Chi2 vaut 1.703

```
> w0<-matrix(c(13,6,8,11),nrow=2)
> w0
     [,1] [,2]
[1,]   13    8
[2,]    6   11
> chisq.test(w0)

   Pearson's chi-square test with Yates' continuity correction

data:  w0
X-square = 1.703, df = 1, p-value = 0.1919
```

Q 10

OUI, le test de normalité sur la fonction de répartition n'est pas significatif.

```
> w0
 [1] 4.6 7.8 4.8 6.0 5.4 3.9 4.5 6.2 4.9 5.3 6.7 4.2 3.8 5.1 6.1 5.2
[17] 4.3 4.6 4.4 5.5
```

DESCRIPTION

Performs a one or two sample Kolmogorov-Smirnov test, which tests the relationship between two distributions.

USAGE

ks.gof(x, y = NULL, alternative = "two.sided", distribution = "normal", ...)

```
> ks.gof(w0)

   One sample Kolmogorov-Smirnov Test of Composite Normality

data:  w0
ks = 0.1193, p-value = 0.5
alternative hypothesis:
  True cdf is not the normal distn. with estimated parameters
sample estimates:
 mean of x standard deviation of x
     5.165                 1.004

Warning messages:
  The Dallal-Wilkinson approximation, used to calculate
     the p-value in testing composite normality,
     is most accurate for p-values <= 0.10 .
   The calculated p-value is 0.679 and so is set to 0.5 . in: dall.w
ilk(test, nx)
```

Q 11

C'est un graphe Quantile-Quantile (QQ-plot). Il faut faire :

```
> qqnorm(w0)
> qqline(w0)
```

Q 12

Les moyennes valent 0.5286, 1.329 et 1.933.

```
> apply(G[,2:4],2,mean)
    fra   noi   sor
 0.5286 1.329 1.933
On peut faire aussi mean(G$fra) ou mean(G[,2]), ...
```

Q 13

OUI, la corrélation est hautement significative :

```
> cor.test(G$noi,G$sor)

   Pearson's product-moment correlation

data:  G$noi and G$sor
t = 4.216, df = 100, p-value = 0.0001
alternative hypothesis: true coef is not equal to 0
sample estimates:
    cor
 0.3885
```

Q 14

La première valeur propre est le carré de la première valeur singulière. Elle vaut 1.580

```
> pca0<-princomp(G[,2:4],cor=T)
> pca0$sdev*pca0$sdev
 Comp. 1 Comp. 2 Comp. 3
    1.58  0.8165  0.6038
```

Q 15

La correlation vaut 0.9878. Faire la somme ou prendre la combinaison linéaire de variance maximale donne sensiblement la même chose.

```
> cor(pca0$scores[,1],tot)
[1] 0.9878
```

Q 16

La fréquence vaut 7.5% quand tot<4 et 35% quand tot est >=4.

```
> mean(G$geli[tot<4])
[1] 0.07547
> mean(G$geli[tot>=4])
[1] 0.3469
```

Q 17

Les régressions logistiques sont très significatives avec p = 2/100, p = 3/10000 et p = 6/1000.

```
> anova(glm(geli ~ fra, family = binomial, data = G), test = "Chi")
Analysis of Deviance Table

Binomial model

Response: geli
```

```
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev Pr(Chi)
NULL                   101      103.7
 fra  1    5.566       100       98.2 0.01831
> anova(glm(geli ~ noi, family = binomial, data = G), test = "Chi")
Analysis of Deviance Table

Binomial model

Response: geli

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev   Pr(Chi)
NULL                   101      103.7
 noi  1    12.87       100       90.9 0.0003337
> anova(glm(geli ~ sor, family = binomial, data = G), test = "Chi")
Analysis of Deviance Table

Binomial model

Response: geli

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev  Pr(Chi)
NULL                   101      103.7
 sor  1    7.518       100       96.2 0.006109
```

Q 18

La première est préférable. La vraisemblance des deux modèles est sensiblement la même mais le premier utilise 2 paramètres au lieu de 4.

```
> glm(G$geli ~ tot, family = binomial)
Call:
glm(formula = G$geli ~ tot, family = binomial)

Coefficients:
 (Intercept)    tot
       -4.32 0.6913

Degrees of Freedom: 102 Total; 100 Residual
Residual Deviance: 86.63
> glm(G$geli ~ G$fra+G$noi+G$sor, family = binomial)
Call:
glm(formula = G$geli ~ G$fra + G$noi + G$sor, family = binomial)

Coefficients:
 (Intercept)  G$fra  G$noi  G$sor
      -4.229 0.8399 0.8342 0.5012

Degrees of Freedom: 102 Total; 98 Residual
Residual Deviance: 86.07

> anova(glm(G$geli ~ tot, family = binomial, data = G), test = "Chi")
Analysis of Deviance Table

Binomial model

Response: G$geli

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev    Pr(Chi)
```

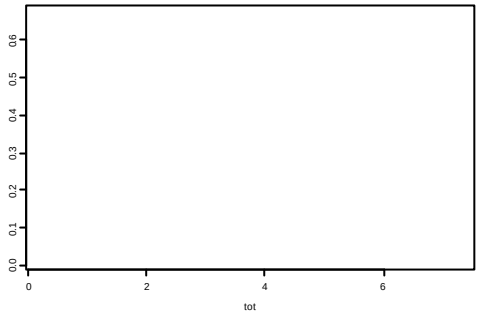
```
NULL                   101      103.7
 tot  1    17.09       100       86.6 0.00003564
```

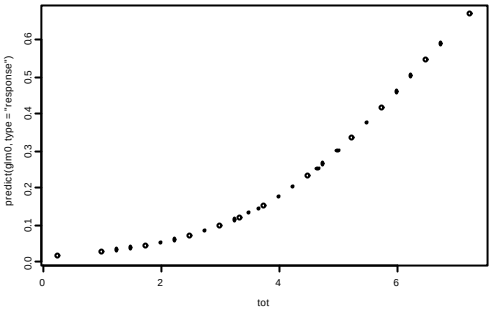
Q 19

Utiliser la fonction de lien pour vérifier :

```
> 1/(1+exp(4.32))
[1] 0.01313
```

Q 20





Q 21

La valeur du F est 18.45

```
> anova(lm(tot~as.factor(G$geli)),test="F")
Analysis of Variance Table

Response: tot

Terms added sequentially (first to last)
                  Df Sum of Sq Mean Sq F Value      Pr(F)
as.factor(G$geli)   1      43.9   43.89   18.45 0.00004049
        Residuals 100     237.9    2.38
```

Q 22

C'est une loi Chi2 à 10 degrés de liberté.

Q 23

```
> x0<-seq(0,30,le=100)
> lines(x0,dchisq(x0,df=10))
```