DEA Analyse et Modélisation des Systèmes Biologiques

Introduction au logiciel S-PLUS© - 1999/2000

Contrôle sur machine (avec solution)

D. Chessel & J. Thioulouse

1. Questions

1.1. Généralités

Question 11 : Quelle est la probabilité pour qu'une variable aléatoire prenne la valeur 15 si elle suit une loi binomiale de paramètre $n = 27$ et $p = 0.45$?

Question 12 : Quelle est la probabilité pour qu'une variable aléatoire dépasse la valeur 5.26 si elle suit une loi Khi2 à 3 degrés de liberté ?

Question 13 : Quelle est la valeur x telle que $P(X \leq x) = 0.00001$ si X suit une loi de Student à à 13 degrés de liberté ?

Question 14 : Vous désirez refaire un graphique que vous avez déjà tracé avec la fonction `plot` lors d'une séance précédente dans S-Plus. Quel ordre utiliserez vous pour afficher vos précédentes utilisations de la fonction `plot` ?

Question 15 : A quoi servent les deux commandes UNIX suivantes :

```
setenv S_WORK /mnt/users/dea/dupont/Splus
setenv S_CLHISTFILE /mnt/users/dea/dupont/Splus/.Splus_history
```

Question 16 : on veut lire le fichier ci-dessous pour en faire un data.frame :

```
age fat      sex
23.00        9.50  m
23.00        27.90 f
27.00        7.80  m
27.00        17.80 m
39.00        31.40 f
41.00        25.90 f
45.00        27.40 m
49.00        25.20 f
50.00        31.10 f
53.00        34.70 f
53.00        42.00 f
```

```
54.00        29.10 f
56.00        32.50 f
57.00        30.30 f
58.00        33.00 f
58.00        33.80 f
60.00        41.10 f
61.00        34.50 f
```

Lequel des ordres suivants faut-il utiliser ?

```
ordre 1 : class<-read.table(file="Class.txt")
ordre 2 : class<-read.table("Class.txt",header=T)
ordre 3 : class<-read.table(file="Class.txt",header=T,",",1)
ordre 4 : class<-read.table("Class.txt",header=T,1)
```

Question 17 : On veut créer une fonction S-Plus qui s'exécute automatiquement au lancement de S-Plus. Quel nom faut-il lui donner ?

Question 18 : A quoi servent les fonctions source et dput ?

1.2. Statistiques non paramétriques

Question 21 : La variable mesurée est la longueur de la mâchoire inférieure de 10 chacals mâles et 10 chacals femelles (*Canis Aureus*) conservées au British Museum [1]. La variable mesurée différe-t-elle entre les deux sexes chez cette espèce ?

```
males      120  107  110  116  114  111  113  117  114  112
femelles   110  111  107  108  110  105  107  106  111  111
```

Question 22 : La variable mesurée est le temps de survie de patients atteints d'un cancer et traités avec un médicament donné [2]. Cette variable dépend du type de cancer.

```
Estomac  124  42   25    45  412    51      1112  46    103   876  146   340   396
Poumon  1235  24 1581  1166   40  727      3808 791   1804  3460  719
```

Le test t (t.test) donne un rejet de l'égalité des moyennes à $p = 0.005$ et le test de Wilcoxon donne un rejet à $p = 0.02$. Pourquoi ? Lequel des deux résultats vous semble le plus solide ?

Question 23 : Les 142 catastrophes aériennes [3] survenues entre le 01/01/1972 et le 31/12/1975 sont énumérées dans la chronique cata. La date de chaque accident est donnée en nombre de jours entre 0 et 1461. Ces données sont dans le fichier "cata.txt" à l'endroit habituel.

```
> cata
  [1]    7   17   21   21   26   34   36   42   63   74   79  104  107  109  111
 [16]  126  129  139  142  150  166  167  170  176  181  181  181  211  211  224
 [31]  227  229  240  245  254  257  268  276  276  287  295  301  304  309  323
 [46]  333  338  343  343  355  358  364  388  395  416  418  418  428  430  444
 [61]  466  491  505  515  515  517  518  537  547  547  558  570  570  578  591
 [76]  600  605  607  617  620  637  640  652  662  672  673  687  708  716  721
 [91]  722  732  740  741  748  757  761  784  793  803  805  825  843  848  853
[106]  888  909  948  954  955  957  981  982  985  989 1033 1055 1066 1069 1087
[121] 1093 1094 1112 1126 1130 1154 1167 1209 1271 1308 1308 1311 1327 1338 1340
[136] 1363 1366 1369 1392 1399 1418 1422
```

On suppose qu'un accident arbitraire survient à un quelconque moment dans l'intervalle
$[a = 0, b = 1461]$.

D'après le théorème central limite, la moyenne des dates d'observations :

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

suit, pour un tel processus poissonien, une loi normale de moyenne $E(Y) = \frac{b-a}{2}$ et de variance $V(Y) = \frac{(b-a)^2}{12n}$. Quelle hypothèse a votre préférence entre :

1 - « La fréquence des catastrophes augmente pendant cette période »
2 - « La fréquence des catastrophes est constante pendant cette période »
3 - « La fréquence des catastrophes diminue pendant cette période » ?

1.3. Modèle linéaire

Question 31 : Quelle est l'erreur intoduite dans la reproduction du listing ci-dessous :

```
> x_seq(from=1, to=20, by=1)
> y_5.75-2.365*x+rnorm(20,sd=0.2)
> summary(lm(y~x))

Call: lm(formula = y ~ x)
Residuals:
     Min      1Q   Median    3Q    Max
 -0.3587 -0.1174 -0.02288 0.104 0.5179

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept)    5.6895    0.0961    59.2284    0.0000
          x   -2.3573    0.0080  -293.9633    0.0000

Residual standard error: 20.68 on 18 degrees of freedom
Multiple R-Squared: 0.9998
F-statistic: 86410 on 1 and 18 degrees of freedom, the p-value is 0

Correlation of Coefficients:
  (Intercept)
x -0.8765
```

Question 32 : Les données viennent d'une recherche sur une nouvelle méthode [4] de mesure de la composition du corps et donne le pourcentage de graisse (fat), l'âge (age) et le sexe (sex) de 18 personnes adultes en bonne santé agées de 23 à 61 ans. Comment l'âge et le pourcentage de graisse sont-ils reliés ? Cette relation est elle différente chez les hommes (m) et chez les femmes (f).

```
> human
                  age  fat sex     age  fat sex
                1  23  9.5   m    11  53 42.0   f
                2  23 27.9   f    12  54 29.1   f
                3  27  7.8   m    13  56 32.5   f
                4  27 17.8   m    14  57 30.3   f
```

```
 5  39 31.4   f    15  58 33.0   f
 6  41 25.9   f    16  58 33.8   f
 7  45 27.4   m    17  60 41.1   f
 8  49 25.2   f    18  61 34.5   f
 9  50 31.1   f
10  53 34.7   f
```

Les données sont dans le fichier "human.txt" à l'endroit habituel .

Question 33 :

```
> aa
 [1] "coefficients"  "residuals"     "fitted.values" "effects"
 [5] "R"             "####"          "assign"        "df.residual"
 [9] "contrasts"     "terms"         "####"
```

Quels sont les noms qui manquent ?

1.4. Modèle linéaire généralisé

Importer le data.frame fum depuis le fichier "fum.txt" du répertoire habituel [5].

```
> fum
    POP DEATHS AGE SMOKE          POP DEATHS AGE SMOKE
 1  656     18  40    NO   19 4531    149  40    Y2
 2  359     22  45    NO   20 3030    169  45    Y2
 3  249     19  50    NO   21 2267    193  50    Y2
 4  632     55  55    NO   22 4682    576  55    Y2
 5 1067    117  60    NO   23 6052   1001  60    Y2
 6  897    170  65    NO   24 3880    901  65    Y2
 7  668    179  70    NO   25 2033    613  70    Y2
 8  361    120  75    NO   26  871    337  75    Y2
 9  274    120  80    NO   27  345    189  80    Y2
10  145      2  40    Y1   28 3410    124  40    Y3
11  104      4  45    Y1   29 2239    140  45    Y3
12   98      3  50    Y1   30 1851    187  50    Y3
13  372     38  55    Y1   31 3270    514  55    Y3
14  846    113  60    Y1   32 3791    778  60    Y3
15  949    173  65    Y1   33 2421    689  65    Y3
16  824    212  70    Y1   34 1195    432  70    Y3
17  667    243  75    Y1   35  436    214  75    Y3
18  537    253  80    Y1   36  113     63  80    Y3
```

Sur chaque ligne on trouve une population d'hommes (au sens statistique) d'une classe d'âge donnée et ayant le même comportement par rapport au tabac. POP est l'effectif du groupe, DEATHS est le nombre de morts dans le groupe, AGE désigne la classe d'âge (on a retenu la borne inférieure de l'intervalle) et SMOKE décrit le comportement (NO pour non fumeur, Y1 pour fumeur de cigares et pipes seulement, Y2 pour fumeur de cigarettes et autres, Y3 pour fumeur de cigarettes seulement. On rappelle que `glm ( (n succès, n échecs)~paramètres, family = binomial)` permet de travailler sur ce type de données (exemples fiche 4 p. 24).

Question 41 : Qu'obtient-on avec l'ordre suivant ?

```
> tapply(fum$DEATHS,as.factor(fum$AGE),sum)
```

Question 42 : Donner la fréquence des décès par classe d'âge, tout type de comportements confondus.

Question 43 : Quel glm redonne comme prédiction ce modèle simple.

Question 44 :

```
> glm1
Call:
glm(formula = cbind(DEATHS, POP - DEATHS) ~ AGE + SMOKE, family=binomial, data = fum)

Coefficients:
 (Intercept)        AGE     SMOKE1     SMOKE2    SMOKE3
    -6.74105 0.08421371 0.02368168 0.09153101 0.1100668

Degrees of Freedom: 36 Total; 31 Residual
Residual Deviance: 39.44169



> contrasts(fum$SMOKE)
   [,1] [,2] [,3]
NO   -1   -1   -1
Y1    1   -1   -1
Y2    0    2   -1
Y3    0    0    3
```

Indiquer comment on retrouve, avec l'information qui précède, la probabilité de décès pour le groupe "Y2" de la classe d'âge 70 ans.

Question 45 : Quel ordre utiliser pour obtenir ce qui suit :

```
Analysis of Deviance Table

Binomial model

Response: cbind(DEATHS, POP - DEATHS)

Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev Pr(Chi)
 NULL                    35   4917.031
  AGE  1 4643.757        34    273.274       0
SMOKE  3  233.833        31     39.442       0
```

Question 46 : Quel ordre utiliser pour obtenir ce qui suit :

```
Call: glm(formula = cbind(DEATHS, POP - DEATHS) ~ AGE + SMOKE, family = binomial,
data
    = Fum)
Deviance Residuals:
       Min         1Q      Median        3Q      Max
 -2.663295 -0.7209114 -0.03990509 0.6282081 2.053219

Coefficients:
                  Value  Std. Error     t value
(Intercept) -6.74105029 0.086955256 -77.5232063
        AGE  0.08421371 0.001339912  62.8501659
     SMOKE1  0.02368168 0.027308065   0.8672047
     SMOKE2  0.09153101 0.011190883   8.1790701
     SMOKE3  0.11006680 0.007273717  15.1321260
```

Question 47 : Résumer l'information acquise grâce au modèle.

2. Solutions

2.1. Généralités

Question 11

```
> dbinom(15, 27, 0.45)
[1] 0.08369243
```

Question 12

```
> 1 - pchisq(5.26, 3)
[1] 0.1537191
```

Question 13

```
> qt(1e-005, 13)
[1] -6.501145
```

Question 14

```
> history("plot")
```

Question 15

La première fixe le dossier de travail et la seconde fixe le fichier de l'historique des commandes.

Question 16

Le bon ordre est le second car il y a des noms de variables, des séparateurs non « , » et pas de noms d'individus.

Question 17

La fonction doit s'appeler « `.First` ».

Question 18

Les fonctions servent à lire et écrire des fonctions S-Plus dans des fichiers textes.

2.2. Statistique non paramétrique

Question 21 : La variable mesurée diffère t'elle entre les deux sexes chez cette espèce ?
Oui : le test de Wilcoxon donne une p-value de 5 pour mille

```
> males <- c(120, 107, 110, 116, 114, 111, 113, 117, 114, 112)
> femelles <- c(110, 111, 107, 108, 110, 105, 107, 106, 111, 111)
> males
 [1] 120 107 110 116 114 111 113 117 114 112
> femelles
 [1] 110 111 107 108 110 105 107 106 111 111
> wilcox.test(males, femelles)
Warning messages:
```

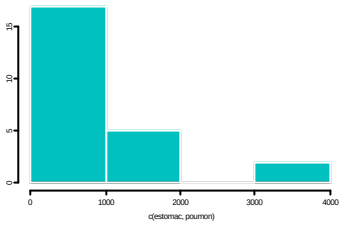
```
  Warning in wil.rank.sum(x, y, alternative, exact, correct): cannot compute exact p-
value with ties

   Wilcoxon rank-sum test

data:  males and femelles
rank-sum normal statistic with correction Z = 2.8171, p-value = 0.0048
alternative hypothesis: true mu is not equal to 0
```

Question 22 : Pourquoi ? Lequel des deux résultats vous semble le plus solide ?

```
estomac_c(124,42,25,45,412,51,1112,46,103,876,146,340,396)
poumon_c(1235,24,1581,1166,40,727,,3808,791,1804,3460,719)
hist(c(estomac,poumon))
```



La normalité des données est grossièrement fausse. Le test paramétrique est invalide. Le test non paramétrique est largement plus solide.

Question 23

```
> z0 <- mean(cata)
> z0 <- z0 - 1461/2
> z0 <- z0/sqrt((1461 * 1461)/12/142)
> z0
[1] -3.579939
> 1 - pnorm(z0)
[1] 0.9998282
```

La date moyenne est très significativement inférieure à la moyenne attendue. Il y a plus de catastrophes au début. Je décide pour «La fréquence des catastrophes diminue pendant cette période ».

2.3. Modèle linéaire

Question 31 : Quelle est l'erreur intoduite ?

```
> y_5.75-2.365*x+rnorm(20,sd=0.2)
```

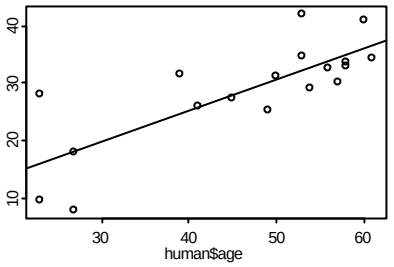
On fabrique un modèle linéaire dont l'écart-type résiduel vaut 0.2

```
Residual standard error: 20.68 on 18 degrees of freedom
```

Cet écart-type résiduel est estimé à 20.68. C'est cette valeur qui a été grossièrement modifiée.

Question 32 : Comment l'âge et le pourcentage de graisse sont-ils reliés ?

```
> plot(human$age,human$fat)
```



L'adiposité croit avec l'âge. Un modèle linéaire est acceptable.

```
> lm0_lm(fat~age*sex,data=human)
> anova(lm0)
Analysis of Variance Table

Response: fat

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
      age  1  891.8737 891.8737 44.27361 0.00001089
      sex  1  168.7867 168.7867  8.37876 0.01176673
  age:sex  1   78.8532  78.8532  3.91436 0.06789632
Residuals 14  282.0243  20.1446
```

Cette relation est elle différente chez les hommes (m) et chez les femmes (f) ?

Oui, on peut refuser au niveau 1% l'égalité des ordonnées à l'origine. Vu le nombre de mesures, on peut même refuser au risque de 6% l'égalité des pentes.

Question 33 :

```
> names(lm0)
 [1] "coefficients"  "residuals"     "fitted.values" "effects"
 [5] "R"             "rank"          "assign"        "df.residual"
 [9] "contrasts"     "terms"         "call"
```

Les noms qui manquent sont rank et call.

2.4. Modèle linéaire généralisé

Question 41 : Qu'obtient-on avec l'ordre suivant ?

```
> tapply(fum$DEATHS, as.factor(fum$AGE), sum)
  40  45  50   55   60   65   70  75  80
 293 335 402 1183 2009 1933 1436 914 625
```

On obtient le nombre de morts par classes d'âge.

Question 42 : Donner la fréquence des décès par classe d'âge, tout type de comportements confondus.

```
> a0 <- tapply(fum$DEATHS, as.factor(fum$AGE), sum)
> a1 <- tapply(fum$POP, as.factor(fum$AGE), sum)
> a1
   40   45   50   55    60   65   70   75   80
 8742 5732 4465 8956 11756 8147 4720 2335 1269
> print(a0/a1, digits = 3)
     40     45   50    55    60    65    70    75    80
 0.0335 0.0584 0.09 0.132 0.171 0.237 0.304 0.391 0.493
```

Question 43 : Quel glm redonne comme prédiction ce modèle simple ?

```
> glm0 <- glm(cbind(fum$DEATHS, fum$POP - fum$DEATHS) ~ as.factor(AGE), data = fum,
   family = binomial)
> predict0 <- predict(glm0, type = "response")
> print(predict0[1:9], digits = 3)
      1      2    3     4     5     6     7     8     9
 0.0335 0.0584 0.09 0.132 0.171 0.237 0.304 0.391 0.493
```

La réponse est `glm(cbind(fum$DEATHS, fum$POP - fum$DEATHS) ~ as.factor(AGE), data = fum, family = binomial)`

Question 44 :

```
> glm1
Call:
glm(formula = cbind(DEATHS, POP - DEATHS) ~ AGE + SMOKE, family = binomial, data
    = fum)

Coefficients:
 (Intercept)        AGE     SMOKE1     SMOKE2    SMOKE3
    -6.74105 0.08421371 0.02368168 0.09153101 0.1100668

Degrees of Freedom: 36 Total; 31 Residual
Residual Deviance: 39.44169
> contrasts(fum$SMOKE)
   [,1] [,2] [,3]
NO   -1   -1   -1
Y1    1   -1   -1
Y2    0    2   -1
Y3    0    0    3
```

On calcule la prédiction sur le lien.

```
> -6.74105 +0.08421371*70+2*0.09153101 -0.1100668
[1] -0.7730951
```

On inverse par $x = \log\left(\frac{p}{1-p}\right) \Leftrightarrow p = \frac{\exp(x)}{1+\exp(x)}$

```
> exp(-0.7730951)/(1+exp(-0.7730951))
[1] 0.31581
> predict(glm1)[fum$AGE==70&fum$SMOKE=="Y2"]
        25
 -0.7730952
> predict(glm1,type="response")[fum$AGE==70&fum$SMOKE=="Y2"]
       25
 0.3158099
```

Question 45 : Quel ordre utiliser pour obtenir ce qui suit :
`anova(glm1, test="Chisq")` *donne le résultat*

Question 46 : Quel ordre utiliser pour obtenir ce qui suit :
`summary(glm1)` *donne le résultat*

Question 47 : Résumer l'information acquise grâce au modèle.

La probabilité de décès est une fonction croissante de l'âge. Le comportement face au tabac dans cette étude est un facteur significatif. Le coefficient du premier contraste (test Y1 contre NO) n'est pas significatif. Par contre Y2 contre les deux premiers et Y3 contre les trois premiers est très significatif. On peut dire «Pipe+Cigare = Témoins, Cigarette + Autres > Témoins, Cigarette seule > Cigarette + Autres > Pipe + Cigare = Témoins).

```
                Value  Std. Error     t value
(Intercept) -6.74105029 0.086955256 -77.5232063
        AGE  0.08421371 0.001339912  62.8501659
     SMOKE1  0.02368168 0.027308065   0.8672047
     SMOKE2  0.09153101 0.011190883   8.1790701
     SMOKE3  0.11006680 0.007273717  15.1321260
```

Il n'y a pas d'interaction.

```
> glm2 <- glm(cbind(DEATHS, POP - DEATHS) ~ AGE * SMOKE, family = binomial, data =
   fum)
> anova(glm2, test = "Chisq")
Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev   Pr(Chi)
     NULL                    35   4917.031
      AGE  1 4643.757        34    273.274 0.0000000
    SMOKE  3  233.833        31     39.442 0.0000000
AGE:SMOKE  3    4.393        28     35.049 0.2220546
```

3. Références

1 Manly, B.F.J. (1991) Randomization and Monte Carlo methods in biology. Chapman & Hall, London. 1-281.

2 Cameron, E. & Pauling, L. (1978) Supplemental ascorbate in the supportive treatment of cancer: re-evaluation of prolongation of survival times in terminal human cancer. *Proceeding of the National Academy of Sciences of the USA* : 75, 4538-4542.

3 Eddy P., Potter E. & Page B. (1976). *Destination désastre*. Grasset, Paris. pp. 330-332

4 Mazess R.B., Peppler W.W. & Gibbons M. (1984) Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition* : 40, 834-839. *In* Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994*) A handbook of small data sets.* Chapman & Hall, London. 1-458.

[5] Best E.W.R. & Walker C.B. (1964) A canadian study of smoking and health. *Canadian Journal of Public Health*, 55, 1. *In* Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994*) A handbook of small data sets.* Chapman & Hall, London. 1-458.