DEA Analyse et Modélisation des Systèmes Biologiques
Introduction au logiciel R - 2001/2002

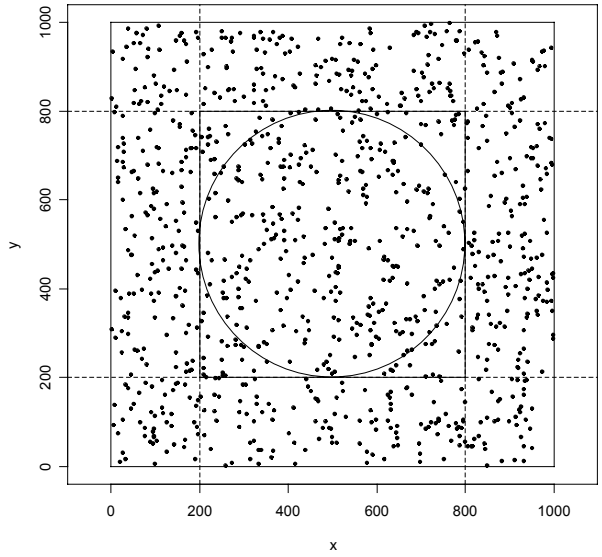
Récupérer les fichiers exofunc.txt, geyser.txt, exoxy.txt et exochats.txt à l'endroit habituel.

On n'est pas obligé de répondre à toutes les questions pour obtenir un très bon résultat.

1. Exécuter la fonction funcexam contenue dans le fichier exofunc.txt. Que fait-elle ?

2. Le fichier exoxy.txt contient les coordonnées de 1000 points dans le carré [0,1000]x[0,1000]. Les abscisses et les ordonnées sont deux échantillons aléatoires simples d'une loi uniforme.

```
> w_read.table("exoxy.txt",h=T)
> plot(w,pch=20,asp=1)
> rect(0,0,1000,1000)
> abline(h=c(200,800),v=c(200,800),lty=2)
> rect(200,200,800,800)
> symbols(500,500,circle=300,add=T,inc=F)
```



Estimer le nombre de points qui sont à l'intérieur du carré [200,800]x[200,800].

3. Estimer le nombre de points qui sont à l'intérieur du cercle.

4. Compter le nombre de points qui sont effectivement à l'intérieur du carré [200,800]x[200,800].

5. Compter le nombre de points qui sont effectivement à l'intérieur du cercle.

6. Quand on lance les deux ordres :

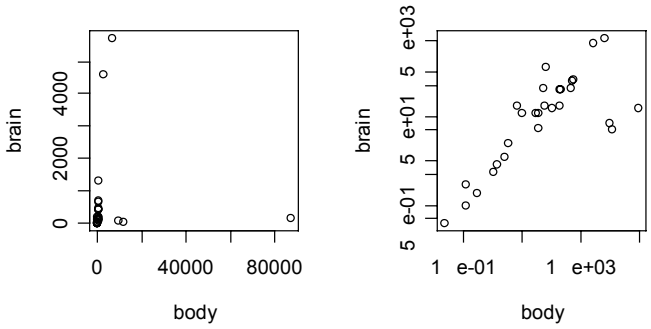
```
> plot(0,0)
> legend(locator(1),"coucou")
```

que faut il faire pour reprendre la main ?

7. Dans la librairie Mass, le jeu de données Animals donne le poids moyen du corps (kg) et le poids moyen du cerveau (g) pour 28 espèces terrestres (Rousseeuw (P.J.) & Leroy (A.M.) (1987) Robust Regression and Outlier Detection. Wiley, p. 57).

```
> library(MASS)
```

```
> data(Animals)
> par(mfrow=c(1,2))
> plot.default(Animals)
> plot.default(Animals,xxxxxxxxxxx)
```



Quelle est la chaîne de caractères qui a été utilisée pour avoir le second graphe en double échelle logarithmique ?

8. Le fichier "Class.txt" contient l'information :

```
age  →  fat   →   sex¶
23.00 → 9.50  →   m¶
23.00 → 27.90  →  f¶
27.00 → 7.80  →   m¶
...¶
58.00 → 33.00  →  f¶
58.00 → 33.80  →  f¶
60.00 → 41.10  →  f¶
61.00 → 34.50  →  f¶
```

La petite flêche indique une tabulation. Lequel des ordres suivants faut-il utiliser pour en faire un data.frame dans R ?

```
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)
```

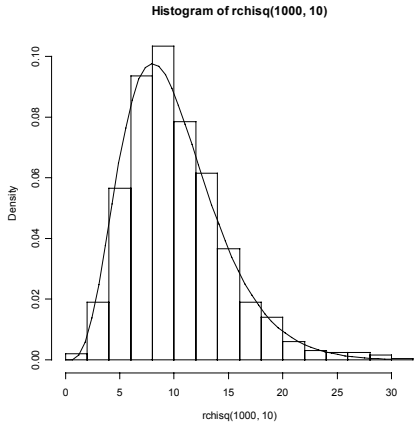
9. Quelle est la probabilité pour qu'une variable aléatoire qui suit une loi normale N(0,1) soit supérieure à 6 ?

10. Si $X_1, X_2, ..., X_p$ sont indépendantes et suivent des lois normales N(0,1), quelle est la variance de la variable $X_1^2 + X_2^2 + ... + X_p^2$? (N.B. on peut utiliser la documentation de dchisq).

11. Quelle est la probabilité pour qu'une variable aléatoire qui suit une loi exponentielle de moyenne 2 soit inférieure ou égale à 10 ?

12. Compléter la partie cachée pour obtenir le graphe ci-contre ?

```
> hist(rchisq(1000,10),pro=T,nclass=20)
> x0_seq(0,30,le=50)
> lines(xxxxxxxxxxxxxxxxxxxxx)
```



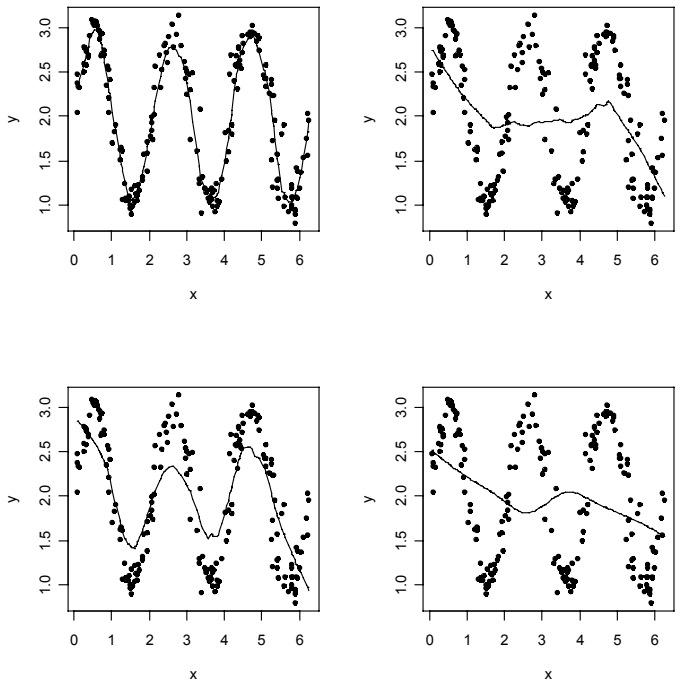
13. Pourquoi l'ordre qui suit génère t'il une erreur ?

```
hist(w,nclass=10,p=T)
Error in hist.default(w, nclass = 10, p = T) :
        argument 3 matches multiple formal arguments
```

14. Le *Old Faithful geyser* du *Yellowstone National Park* (Wyoming) a été observé pendant 15 jours. On a mesuré (Azzalini, A. & Bowman, A.W. (1994) A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society*, C : 39, 357-366. *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.) le temps écoulé entre le début de deux éruptions successives (en minutes). Les données sont dans le fichier geyser.txt. Caractériser la distribution de cette variable.

15. Quelle est le premier quartile, la moyenne, la médiane et le troisième quartile de la suite de nombres 1, 3, 5, 7, ..., 99 ?

16. Pour obtenir la figure ci-dessous, on a utilisé :




```
> par(mfrow=c(2,2))
> plot(x,y,pch=20)
> lines(lowess(x,y,XXX))
> plot(x,y,pch=20)
> lines(lowess(x,y,XXX))
> plot(x,y,pch=20)
> lines(lowess(x,y,XXX))
> plot(x,y,pch=20)
> lines(lowess(x,y,XXX))
```

Le paramètre caché a pris exactement une fois et une seule chacune des valeurs 0.1, 0.3, 0.5 et 0.7. Sachant qu'on n'a modifié aucun des paramètres graphiques autre que mfrow, indiquer clairement dans quel ordre ces valeurs ont été utilisées.

17. Que pensez-vous de l'édition qui suit ?

```
> sample(12)
 [1]  8  3 10  4 11  5  7  6 12  8  1  2
```

18. On a utilisé :

```
> x
  [1] "a" "d" "c" "e" "b" "c" "b" "c" "b" "c" "b" "b" "c" "a" "c" "a" "d" "a" "b"
 [20] "c" "b" "b" "c" "b" "a" "c" "e" "b" "d" "b" "c" "c" "d" "b" "b" "a" "b" "d"
 [39] "b" "d" "a" "b" "a" "d" "c" "b" "a" "e" "a" "c" "c" "e" "c" "d" "c" "b" "b"
 [58] "c" "d" "d" "a" "e" "b" "b" "b" "a" "b" "c" "d" "c" "a" "e" "b" "b" "d" "d"
 [77] "b" "b" "b" "d"
```

```
> table(x)
x
 a  b  c  d  e
13 28 19 14  6
> length(unique(x))
[1] xxxx
```

Quelle est la valeur obtenue ?

19. Un test de Wilcoxon bilatéral peut-il être significatif au seuil de 5% si on ne possède que 4 individus dans chacun des groupes ?

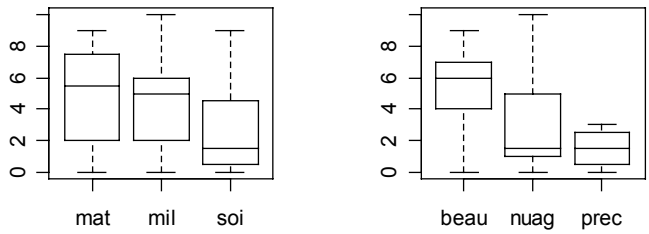
20. Dans les données originales de Mendel (Mendel, G. (1956) Mathematics of heredity. In : The world of mathematics. Part V. Newman, J.R. (Ed.) Tempus Books of Microsoft Press. 923-934.), on trouve :

Expt. 2. Colour of albumen. 258 plants yielded 8023 seeds, 6022 yellow and 2001 green; their ratio, therefore is as 3.01 to 1 ...

Peut-on affirmer que les données sont trop proches du modèle (3/4 1/4) pour être honnêtes ?

21. Dans l'île de Kerguelen, un transect a été parcouru 31 fois et lors de chaque parcours on a compté le nombre de chats *Felis catus* L. rencontrés (extrait des données de D. Pontier). Les données sont dans le fichier exochats.txt. Peut-on admettre que les effectifs de chats sont homogènes (dans un sens à préciser) ?

22. On connait, en outre, pour chaque sortie, l'heure du comptage (mat pour le début de la journée, mil pour milieu de la journée, soi pour la fin de la journée) et l'indication des conditions météorologiques (beau pour beau temps, nuag pour temps nuageux et prec pour précipitations). Comment ce graphique a-t-il été obtenu ?



23. Acceptez-vous l'hypothèse "l'heure de la mesure et la condition météorologique sont indépendantes" ?

24. Que pouvez vous dire de l'objet créé par l'ordre `w_split(chats,heure:temps)`

25. Le nombre de chats rencontrés dépend il des facteurs contrôlés ?

Nom :
Prénom:

DEA Analyse et Modélisation des Systèmes Biologiques / R - 2001/2002

Merci d'utiliser l'espace imparti pour vos réponses.

1. Exécuter la fonction funcexam contenue dans le fichier exofunc.txt. Que fait-elle ?

2. Estimer le nombre de points à l'intérieur du carré.

3. Estimer le nombre de points à l'intérieur du cercle.

4. Compter le nombre de points à l'intérieur du carré.

5. Compter le nombre de points à l'intérieur du cercle.

6. Que faut il faire pour reprendre la main ?

7. Dans la librairie Mass, le jeu de données Animals ...

8. Lequel des ordres suivants faut-il utiliser pour en faire un data.frame dans R ?

9. Quelle est la probabilité pour qu'une variable N(0,1) soit supérieure à 6 ?

10. Quelle est la variance de la variable $X_1^2 + X_2^2 + ... + X_p^2$?

11. Quelle est la probabilité pour qu'une variable aléatoire qui suit une loi exponentielle de moyenne 2 soit inférieure ou égale à 10 ?

12. Compléter la partie cachée pour obtenir le graphe ci-dessous ?

13. Pourquoi l'ordre qui suit génère une erreur ?

14. Caractériser la distribution de cette variable.

15. Quelle est le premier quartile, la moyenne, la médiane et le troisième quartile de la suite de nombres 1, 3, 5, 7, ..., 99 ?

16. Indiquer clairement dans quel ordre ces valeurs ont été utilisées.

17. Que pensez-vous de l'édition qui suit ?

18. Quelle est la valeur obtenue ?

19. Un test de Wilcoxon bilatéral peut-il être significatif au seuil de 5% si on ne possède que 4 individus dans chacun des groupes ?

20. Peut-on affirmer que les données sont trop proches du modèle (3/4 1/4) pour être honnêtes ?

21. Peut-on admettre que ces données sont homogènes (dans un sens à préciser) ?

22. Comment ce graphique a-t-il été obtenu ?

23. Acceptez-vous l'hypothèse "l'heure et la condition météorologique sont indépendantes" ?

24. Que pouvez vous dire de l'objet créé par l'ordre `w_split(chats,heure:temps)`

25. Le nombre de chats rencontrés dépend il des facteurs contrôlés ?

DEA Analyse et Modélisation des Systèmes Biologiques
Introduction au logiciel R - 2001/2002

Solution

1. Exécuter la fonction funcexam contenue dans le fichier exofunc.txt. Que fait-elle ?

```
Elle fait un camembert en 20 couleurs avec le titre "BRAVO : vous avez correctement
chargé la fonction"
```

2. Estimer le nombre de points à l'intérieur du carré.

```
1000*0.6*0.6 = 360
```

3. Estimer le nombre de points à l'intérieur du cercle.

```
> pi*300*300/1000
[1] 282.7
```

4. Compter le nombre de points à l'intérieur du carré ?

```
> sum( (w$x>200) & (w$x<800) & (w$y>200) & (w$y<800) )
[1] 339
```

5. Compter le nombre de points à l'intérieur du cercle ?

```
> sum( sqrt( (w$x-500)^2 + (w$y-500)^2 ) <300)
[1] 268
```

6. Que faut il faire pour reprendre la main ?

```
Il faut cliquer dans la fenêtre graphique, ce qui provoque l'édition de la chaîne de
caractères.
```

7. Dans la librairie Mass, le jeu de données Animals ...

```
> plot.default(Animals,xxxxxxxxxxx)

> ?plot.default
> plot(Animals,log="xy")
```

8. Lequel des ordres suivants faut-il utiliser pour en faire un data.frame dans R ?

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

C'est le second.
```

9. Quelle est la probabilité pour qu'une variable N(0,1) soit supérieure à 6 ?

```
> 1-pnorm(6)
[1] 9.866e-10
```

10. Quelle est la variance de la variable $X_1^2 + X_2^2 + ... + X_p^2$?

```
La variable suit une loi Chi2 à p degrés de liberté et sa variance est 2p. Voir :

? dchisq
The chi-squared distribution with `df'= n degrees of freedom has density

        f_n(x) = 1 / (2^(n/2) Gamma(n/2))  x^(n/2-1) e^(-x/2)
```

```
for x > 0.  The mean and variance are n and 2n.
```

11. Quelle est la probabilité pour qu'une variable aléatoire qui suit une loi exponentielle de moyenne 2 soit inférieure ou égale à 10 ?

```
> ?pexp # (i.e., mean `1/rate')
> pexp(10,0.5)
[1] 0.9933
```

12. Compléter la partie cachée pour obtenir le graphe ci-dessous ?

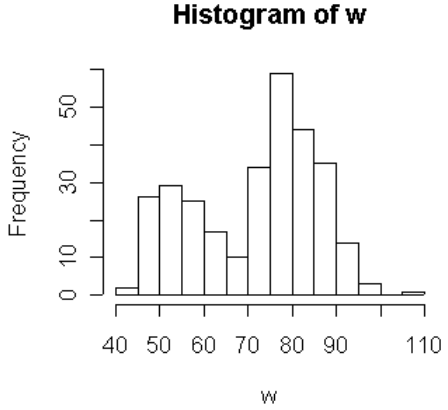
```
> lines(xxxxxxxxxxxxxxxxxxxxx)

> lines(x0,dchisq(x0,10))
```

13. Pourquoi l'ordre qui suit génère une erreur ?

```
hist(w,nclass=10,p=T)
Error in hist.default(w, nclass = 10, p = T) :
        argument 3 matches multiple formal arguments

Parce que deux paramètres ont l'abbréviation p utilisée :

hist.default(x, breaks, freq = NULL, probability = !freq,
          include.lowest = TRUE,
          right = TRUE, col = NULL, border = par("fg"),
          main = paste("Histogram of" , xname),
          xlim = range(breaks), ylim = NULL,
          xlab = xname, ylab,
          axes = TRUE, plot = TRUE, labels = FALSE,
          nclass = NULL, ...)
```

14. Caractériser la distribution de cette variable.

```
> w_read.table("geyser.txt")$V1
> hist(w,nclass=10)
C'est un mélange de deux distributions relativement symétriques unimodales autour de
55 et 80 minutes.
```



15. Quelle est le premier quartile, la moyenne, la médiane et le troisième quartile de la suite de nombres 1, 3, 5, 7, ..., 99 ?

```
> summary(seq(1,99,by=2))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    25.5    50.0    50.0    74.5    99.0
```

16. indiquer clairement dans quel ordre ces valeurs ont été utilisées.

```
Réponse 0.1, 0.5, 0.3, 0.7 (paramètre de lissage et remplissage 1-1, 1-2, 2-1 et 2-2)
```

17. Que pensez-vous de l'édition qui suit ?

```
Elle a été modifiée car 8 apparaît 2 fois et 9 n'apparaît pas. On attend une
permutation des 12 premiers entiers.
```

18. Quelle est la valeur obtenue ?

```
> length(unique(x))
[1] XXXX

La réponse est 5 (le nombre de valeurs différentes)
```

19. Un test de Wilcoxon bilatéral peut-il être significatif au seuil de 5% si on ne possède que 4 individus dans chacun des groupes ?

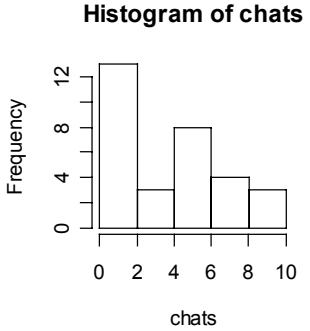
```
Oui, mathématiquement on a deux chances sur 70 de trouver les deux groupes séparés:
> 2/70
[1] 0.02857

oui, expérimentalement on a ce résultat :

> wilcox.test(1:4,5:8)

        Wilcoxon rank sum test

data:  1:4 and 5:8
W = 0, p-value = 0.02857
alternative hypothesis: true mu is not equal to 0
```

20. Peut-on affirmer que les données sont trop proches du modèle (3/4 1/4) pour être honnêtes ?

```
> chisq.test(c(6022,2001),p=c(3/4,1/4))

        Chi-squared test for given probabilities

data:  c(6022, 2001)
X-squared = 0.015, df = 1, p-value = 0.9025
Il n'y a rien d'anormal à ce résultat. Il n'est pas trop élevé (ajustement trop bon).
```

21. Peut-on admettre que ces données sont homogènes (dans un sens à préciser) ?

```
> exochats_read.table("exochats.txt",h=T)
> chats_exochats$chats
> heure_exochats$heure
> temps_exochats$temps

> chats
 [1]  9  7  6  8  5  1  0  3  6  0  7  5  6  7  2  5  1  6 10  2  1  2  3  4  2
[26]  9  0  5  1  1  0
```




```
> hist(chats)

La distribution est manifestement bimodale.


> chisq.test(chats)

        Chi-squared test for given probabilities

data:  chats
X-squared = 69, df = 30, p-value = 6.6e-05

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(chats)

On ne peut admettre une distribution aléatoire des chats entre les sorties

> mean(chats)
[1] 4
> var(chats)
[1] 9.2
```

```
Ce n'est pas un échantillon d'une loi de Poisson.
```

22. Comment ce graphique a-t-il été obtenu ?

```
> par(mfrow=c(1,2))
> plot(heure,chats)
> plot(temps,chats)
```

23. Acceptez-vous l'hypothèse "l'heure de la mesure et la condition météorologique sont indépendantes" ?

```
> table(exo$heure,exo$temps)

     beau nuag prec
  mat    5    3    0
  mil    9    3    3
  soi    3    4    1

> chisq.test(table(exo$heure,exo$temps),sim=T,B=10000)

        Pearson's Chi-squared test with simulated p-value (based on 10000
        replicates)

data:  table(exo$heure, exo$temps)
X-squared = 3.763, df = NA, p-value = 0.4462

On n'accepte jamais une hypothèse nulle mais on a aucune raison de la rejeter.
```

24. Que pouvez vous dire de l'objet créé par l'ordre `w_split(chats,heure:temps)`

```
> w
$"mat:beau"
[1] 9 7 6 8 5

...

$"soi:prec"
[1] 0

C'est une liste triant les valeurs des observations par combinaison des facteurs
contrôlés
```

25. Le nombre de chats rencontrés dépend il des facteurs contrôlés ?

```
Il dépend surtout du temps avec une inter-action. On peut admettre deux réponses.

> anova(lm(chats~heure*temps))
Analysis of Variance Table

Response: chats
           Df Sum Sq Mean Sq F value Pr(>F)
heure       2   19.2     9.6    1.52  0.240
temps       2   52.6    26.3    4.16  0.029 *
heure:temps 3   58.8    19.6    3.10  0.047 *
Residuals  23  145.4     6.3
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

> kruskal.test(chats,temps)

        Kruskal-Wallis rank sum test

data:  chats and temps
Kruskal-Wallis chi-squared = 7.079, df = 2, p-value = 0.02903

> anova(glm(chats~heure*temps,family=poisson),test="Chi")
Analysis of Deviance Table

Model: poisson, link: log

Response: chats
```

```
Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                           30       82.1
heure        2      5.1        28       77.0       0.1
temps        2     14.8        26       62.2   0.00060
heure:temps  3     18.1        23       44.1   0.00042

> unlist(lapply(w,mean))
mat:beau mat:nuag mil:beau mil:nuag mil:prec soi:beau soi:nuag soi:prec
   7.000    1.333    4.333    6.000    2.000    5.000    1.750    0.000
> unlist(lapply(w,length))
mat:beau mat:nuag mil:beau mil:nuag mil:prec soi:beau soi:nuag soi:prec
       5        3        9        3        3        3        4        1
Voir les summary des modèles. Quand il fait beau, les meilleurs moments sont le matin
et le soir. Quand il ne fait pas beau, c'est le milieu de la journée. L'heure joue un
rôle conditionellement au temps qu'il fait.
```