Sujet d'examen corrigé

D. Chessel

LICENCE BO - UE BMS - 02/2001 (2 HEURES)

Une feuille de réponse est jointe a l'énoncé. Répondre aux questions *strictement* dans la place impartie et *justifier* vos réponses par un argument qui vous paraît approprié. Une réponse non justifiée n'est pas prise en compte. *Les réponses sont en italique.*

Question 1. Quel est le résultat du dernier tirage du loto en mars 2000 ?

Question qui avait été annoncée en cours !

29/03	13a	03086	36-27-47-35-39-44-33	27-35-36-39-44-47-33
29/03	13b	03087	32-12-01-17-26-43-37	01-12-17-26-32-43-37

Question 2. Dans une grande population, on a déterminé le sexe de 100 individus et trouvé 40 males et 60 femelles. La valeur 0.5 appartient-elle à l'intervalle de confiance de la fréquence des mâles au seuil de 90% ?

NON. On doit faire un test contre l'hypothèse p = 0.5. La variable nombre de mâles dans l'échantillon suit une loi binomiale de paramètres 100 et 0.5, donc approximativement une loi normale de moyenne 50 et de variance 25. Avec la probabilité de 0.9 elle est comprise entre $50-1.64\times5$ *et* $50+1.64\times5$. *L'observation 40 est dans la zone de rejet du test et 0.5 n'appartient pas à l'intervalle de confiance.*

Je veux tester par un test de Wilcoxon l'hypothèse nulle « Les deux échantillons sont extraits de la même population » contre l'hypothèse alternative « La moyenne de l'échantillon 1 est différente de celle de l'échantillon 2 ». Je dispose de 4 observations pour chacun des échantillons. Je ne dispose pas de la table de la distribution de Wilcoxon.

Question 3. Vrai ou Faux ? Quel que soit le résultat expérimental le test ne sera pas significatif au seuil de 5%.

C'est vrai. Si l'échantillon 1 contient les rangs 1-2-3-4, la probabilité pour que la somme des rangs soit inférieure ou égale à l'observation vaut $\frac{1}{\binom{8}{4}}=\frac{4\times3\times2\times1}{8\times7\times6\times5}=\frac{1}{70}=0.02857$. *Le test bilatéral associé aura donc un seuil supérieur à 5%.*

Le problème est posé par Nathalie Mugnier dans son rapport de statistique "Les naissances au Québec" qui indique ce site :

Données statistiques > Démographie > Les naissances et la fécondité

Naissances et taux de fécondité selon l'âge de la mère, Québec, 1995-1999

```
> naiss
   age  n95  n96  n97  n98  n99
1   12    0    1    1    1    0
2   13    6    1    2    6    1
3   14   28   44   32   27   24
4   15  125  119  121   96   92
5   16  295  305  314  291  239
…
11  22 3167 3360 3166 3040 2947
12  23 3875 3661 3578 3663 3451
13  24 4729 4459 4142 3860 3932
14  25 5312 4968 4651 4496 4299
15  26 5693 5805 5196 4760 4715
…
21  32 5291 5091 4885 4309 4055
22  33 4305 4455 4280 3997 3678
23  34 3654 3625 3590 3329 3339
24  35 2905 2898 2827 2787 2763
25  36 2194 2297 2233 2064 2336
…
31  42  170  217  197  206  191
32  43   90   91  117  122  110
33  44   49   43   44   53   49
34  45   22   25   20   26   27
35  46    5    7    7    3   10
36  47    4    4    4    3    4
37  48    2    6    0    0    1
38  49    0    1    0    1    0
39  50    1    0    0    0    0
```

Cette statistique donne l'âge de la mère à la naissance des enfants nés de 1995 à 1999 au Québec.

```
> n95_rep(naiss$age,naiss$n95)
> n96_rep(naiss$age,naiss$n96)
> n97_rep(naiss$age,naiss$n97)
> n98_rep(naiss$age,naiss$n98)
> n99_rep(naiss$age,naiss$n99)
> mean(n95)
[1] 28.22
> mean(n99)
[1] 28.39
> length(n95)
[1] 87258
> length(n99)
[1] 73580
> sum(naiss$n95)
[1] *****
> t.test(n95,n99)
```

```
        Welch Two Sample t-test

data:  n95 and n99
t = -6.514, df = 153713, p-value = 7.335e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2228 -0.1197
sample estimates:
mean of x mean of y
    28.22     28.39
```

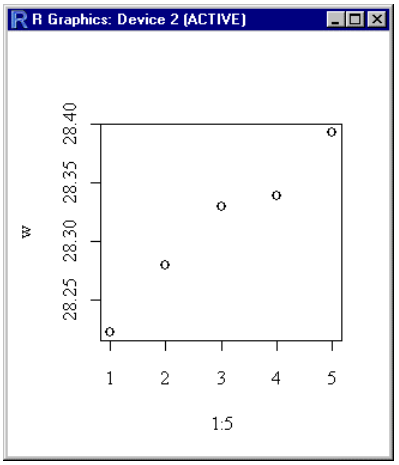
Question 4. Quel est la valeur cachée par les caractères ********* ?

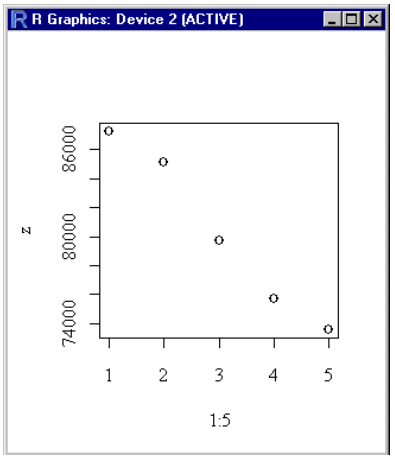
La fonction rep donne une statistique de 87285 valeurs, la somme des nombres de répétitions.

Question 5. Quelle information est-elle apportée par le test ?

On compare les deux échantillons de la variable âge de la mère à la naissance pour les deux années considérées. Il est certain que cet âge a augmenté en moyenne de 0.17 année.

```
> w_rep(0,5)
> w[1]_mean(n95)
> w[2]_mean(n96)
> w[3]_mean(n97)
> w[4]_mean(n98)
> w[5]_mean(n99)
> z_apply(naiss[,2:6],2,sum)
```

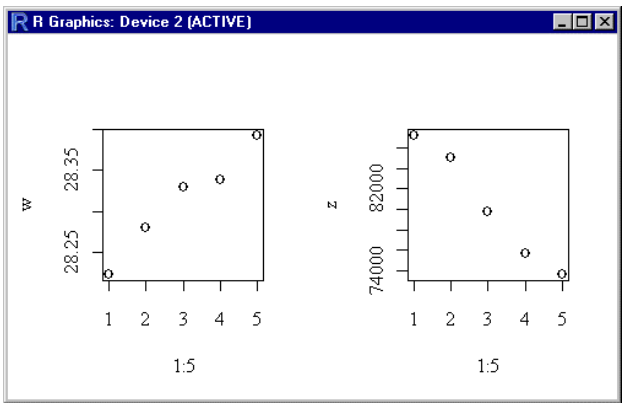




Question 6. Quelles sont les commandes qui ont donné ces figures ?

```
> plot(1:5,w)
> plot(1:5,z)
```

Question 7. Quelles sont les commandes qui donnent cette figure ?




```
> par(mfrow=c(1,2))
> plot(1:5,w)
> plot(1:5,z)
```

Question 8. Que pensez-vous de l'assertion "il y a moins de deux chances sur 100 de trouver 5 valeurs rangées dans l'ordre" ?

Elle est justifiée dans l'espace de probabilités des 5 ! permutations de 5 valeurs. Il y a 1 chance sur 120 de les trouver dans l'ordre croissant et 1 chance sur 120 de les trouver dans l'ordre décriossant, donc moins de 2 chances sur 100 de les trouver rangées dans l'ordre.

Question 9. Qu'avez vous appris des indications qui précèdent ?

Au Québec, de 1995 à 1999, le nombre des naissances décroit régulièrement et l'âge de la mère croit régulièrement.

Le problème est posé par Pierre Pichard dans son rapport de statistique "Analyse statistique des temps d'attente des éruptions du volcan Aso"

Paul J. Buklarewicz
P.O. Box 854
Volcano, Hawaii 96785
(808) 967-8294 (phone)

Les années d'éruption du volcan Aso (Japon), entre 1250 et 1950 sont :

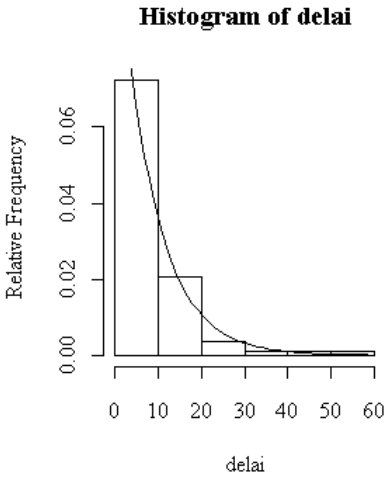
```
> aso
```

```
 [1] 1265 1269 1270 1272 1273 1274 1281 1286 1305 1324 1331 1335 1340 1346 1369
[16] 1375 1376 1377 1387 1388 1434 1438 1473 1485 1505 1506 1522 1533 1542 1558
[31] 1562 1563 1564 1576 1582 1583 1584 1587 1598 1611 1612 1613 1620 1631 1637
[46] 1649 1668 1675 1683 1691 1708 1709 1765 1772 1780 1804 1806 1814 1815 1826
[61] 1827 1828 1829 1830 1854 1872 1874 1884 1894 1897 1906 1916 1920 1927 1928
[76] 1929 1931 1932 1933 1934 1935 1938 1949 1950
```

Source : Wickman, F.E. (1966) Respose-period patterns of volcanoes. Arkiv för Mineralogi och Geologi : Bd. 4, Häfte 4, 291-350.

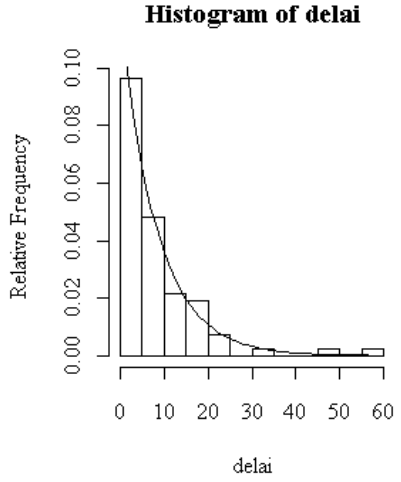
```
> delai_diff(aso)
> hist(delai,proba=T)
> lines (0:60,dexp(0:60,1/mean(delai))))
Error: syntax error
> lines (0:60),dexp(0:60,1/mean(delai)))
Error: syntax error
> lines (0:60),dexp(0:60,1/mean(delai))
Error: syntax error
```

Question 10. Quel est la commande correcte pour obtenir la figure suivante ?




```
lines (0:60,dexp(0:60,1/mean(delai)))
```

Question 11. Quel paramètre a t'il été modifié pour obtenir la figure suivante ?



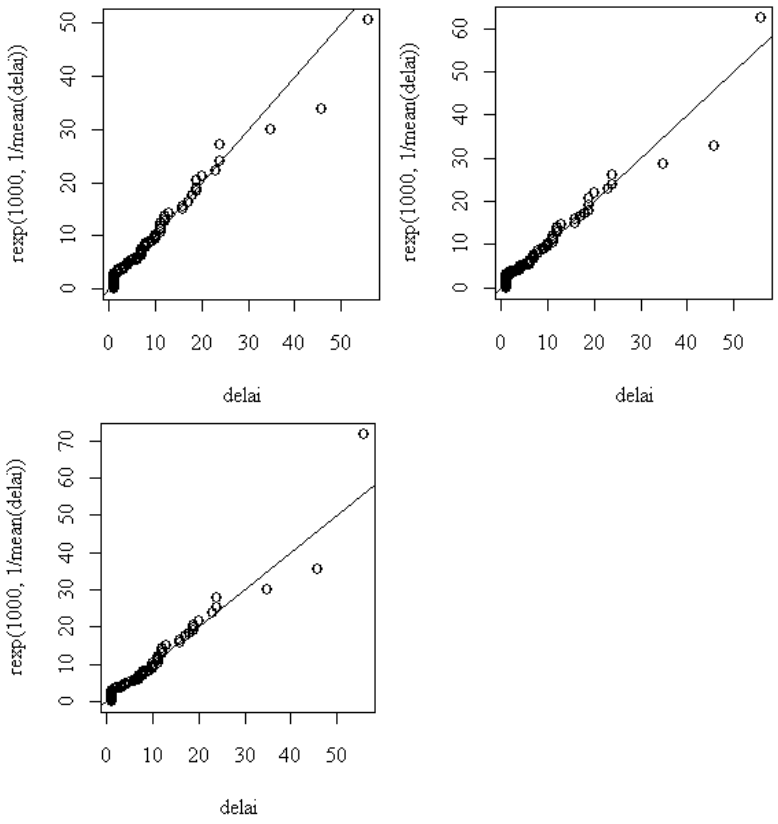
Le nombre de class (nclass)

```
> ks.test(delai, "pexp",1/mean(delai))

        One-sample Kolmogorov-Smirnov test

data:  delai
D = 0.1991, p-value = 0.002768
alternative hypothesis: two.sided
```

Question 12. Quelle est votre conclusion ?

On a fait un test d'ajustement sur la fonction de répartition d'une loi exponentielle avec estimation au maximum de vraisemblance du paramètre. L'ajustement graphique sur la densité est bon mais le test rejette l'hypothèse nulle au risque de 2 pour mille. On ne peut pas accepter l'hypothèse.

```
> qqplot(delai,rexp(1000, 1/mean(delai)))
> abline(0,1)
```



Question 13. Pourquoi n'obtient-on pas toujours le même résultat ?

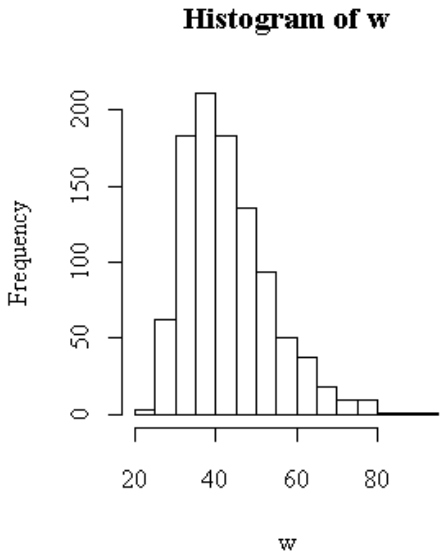
Parce qu'on utilise rexp qui donne un échantillon aléatoire simple d'une densité exponentielle de paramètre donné. Deux exécutions successives ne donnent pas le même résultat.

```
> f1_function() {return(max(rexp(100,1/mean(delai))))}
```

Question 14. Que fait cette fonction ?

Elle renvoie la plus grande valeur d'un échantillon aléatoire simple d'une loi exponentielle de paramètre donné (estimation à partir de l'échantillon observé).

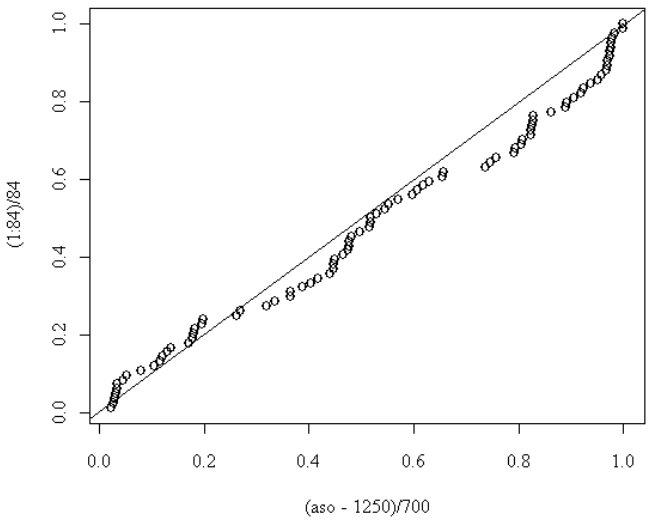
```
> w_rep(0,1000)
> for (i in 1:1000) w[i]_f1()
> hist(w)
> max(delai)
[1] 56
> sum(w>max(delai))
[1] 120
```



Question 15. Le temps de repos du volcan entre 1709 et 1765 vous semble t'il incompatible avec l'hypothèse d'un processus aléatoire ?

Non, on a fait 1000 tirages de la même loi et obtenu la distribution d'échantillonnage approchée (méthode de Monte-Carlo) de la plus grande valeur. L'observation est dépassée dans 12 % des cas, ce qui n'est pas extraordinaire.

```
> plot((aso-1250)/700,(1:84)/84)
> abline(0,1)
> ks.test(aso,"punif",1250,1950)

        One-sample Kolmogorov-Smirnov test

data:  aso
D = 0.1367, p-value = 0.08674
alternative hypothesis: two.sided
```



Question 16. L'intensité de l'activité du volcan vous semble t'elle avoir évoluer pendant la période de l'étude ?

Non. On a fait un test d'ajustement sur la variable date avec une loi uniforme sur l'intervalle de l'étude. Le test mesure comment la fonction de répartition empirique s'éloigne de la fonction de répartition théorique. Dans 8 % des cas, sous l'hypothèse nulle on a un écart au moins aussi grand et je considère que le test n'est pas significatif.

```
> hist(aso,nclass=70,plot=F)
$breaks
 [1] 1260 1270 1280 1290 1300 1310 1320 1330 1340 1350 1360 1370 1380 1390 1400
[16] 1410 1420 1430 1440 1450 1460 1470 1480 1490 1500 1510 1520 1530 1540 1550
[31] 1560 1570 1580 1590 1600 1610 1620 1630 1640 1650 1660 1670 1680 1690 1700
[46] 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800 1810 1820 1830 1840 1850
[61] 1860 1870 1880 1890 1900 1910 1920 1930 1940 1950

$counts
 [1] 3 3 2 0 1 0 1 3 1 0 1 3 2 0 0 0 0 2 0 0 0 1 1 0 2 0 1 1 1 1 3 1 4 1 0 4 0 2
[39] 1 0 1 1 1 1 2 0 0 0 0 0 1 2 0 0 2 2 5 0 0 1 0 2 1 2 1 2 3 6 2
> p10_c(0,hist(aso,nclass=70,plot=F)$counts)
> p10
 [1] 0 3 3 2 0 1 0 1 3 1 0 1 3 2 0 0 0 0 2 0 0 0 1 1 0 2 0 1 1 1 1 3 1 4 1 0 4 0
[39] 2 1 0 1 1 1 1 2 0 0 0 0 0 1 2 0 0 2 2 5 0 0 1 0 2 1 2 1 2 3 6 2
```

Question 17. Pourquoi rajouter un 0 devant ?

Le logiciel a défini 69 classes en éliminant la borne 1250. On doit donc rajouter l'intervalle 1250/1260 dendant lequel il y a eu 0 irruption. On aura ainsi le nombre des éruuptions par classes de 10 ans.

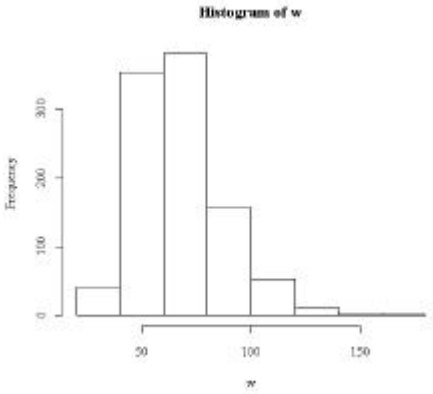
```
> table(p10)
p10
 0  1  2  3  4  5  6
26 21 13  6  2  1  1
> sum(p10)
[1] 84
> mean(p10)
[1] 1.2
```

```
> var(p10)
[1] 1.728
> 84*dpois(0:4,mean(p10))
[1] 25.300 30.360 18.216  7.286  2.186
> 84-sum(84*dpois(0:4,mean(p10)))
[1] 0.6506
> obs_c(26,21,13,6,4)
> 2.186+0.6506
[1] 2.837
> the_c(25.3,30.36,18.22,7.286,2.837)
> obs
[1] 26 21 13  6  4
> the
[1] 25.300 30.360 18.220  7.286  2.837
> sum( ((obs-the)^2)/the)
[1] 5.104
> pchisq(5.104,3)
[1] 0.8357
```

Question 18. Commenter ce qui précède.

On a compté puis tabulé le nombre d'éruptions par classes de 10 ans. Le test est le Khi2 d'ajustement à une loi de Poisson. Le Khi2 observé est dépassé dans 16% sous l'hypothèse nulle qui ne peut être rejetée. On n'a pas de raison de rejeter l'hypothèse d'une loi de Poisson.

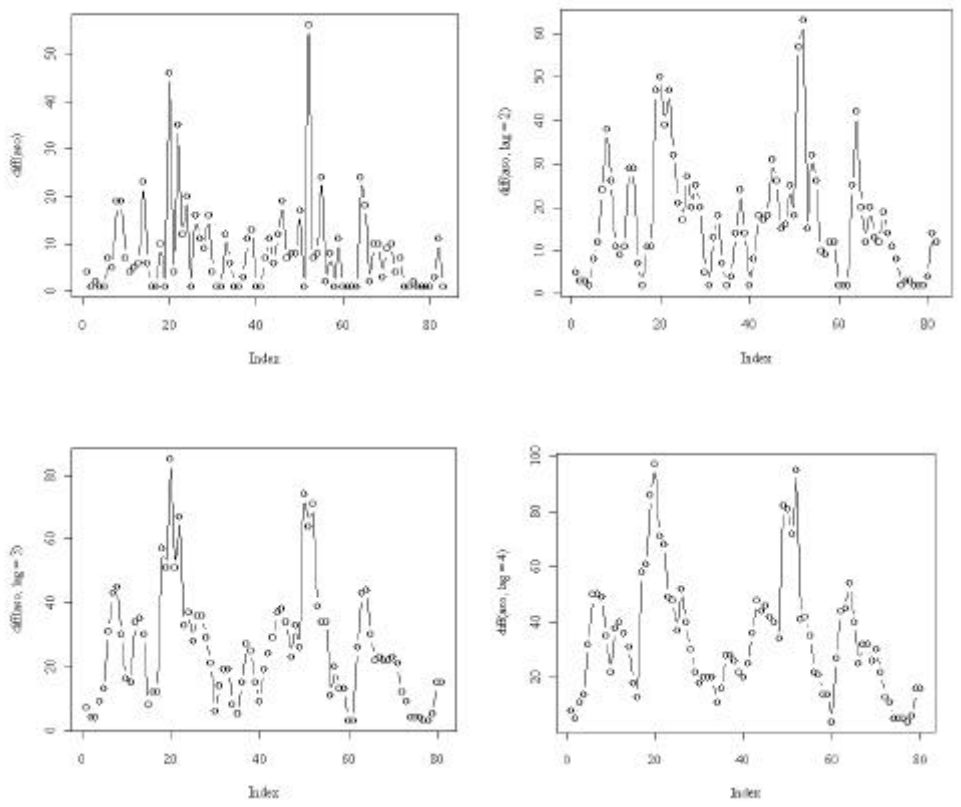
```
>  f2_function() {return(var(rexp(100,1/mean(delai))))}
> for (i in 1:1000) w[i]_f2()
> hist(w)
> sum(w>94.8)
[1] 100
```



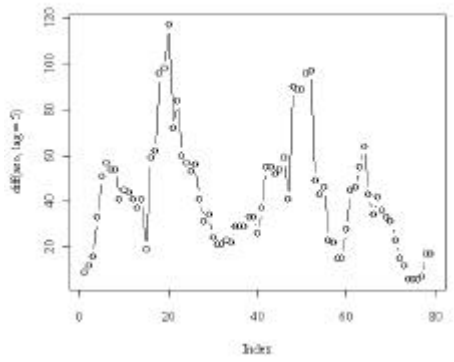
Question 19. La variance du temps d'attente de la prochaine irruption du volcan vous semble t'elle incompatible avec l'hypothèse d'un processus aléatoire ?

Non, on a refait un test de Monte-Carlo sur la variance de l'échantillon qui est dépassé dans 10 % des simulations. On ne peut rejeter l'hypothèse nulle.

```
> plot(diff(aso),type="b")
> plot(diff(aso,lag=2),type="b")
> plot(diff(aso,lag=3),type="b")
> plot(diff(aso,lag=4),type="b")
> plot(diff(aso,lag=5),type="b")
```



```
> plot(diff(asosim,lag=5),type="b")
```



Question 20. Donner une phrase pour commenter l'état actuel de l'analyse.

Aucun des tests effectués n'indique un écart franc au modèle aléatoire à l'exception d'un seul. La question est de savoir en quoi on s'écarte du modèle de l'hypothèse avec la fonction de répartition du temps d'attente. Ni l'intervalle maximum ni la variance ne sont en cause. Les dénombrements sont poissoniens. L'intensité est constante. Le temps d'attente aux évènements suivants au pas 2,3, … semble suggérer une structure interne au processus avec quatre phases. Pierre Pichard s'est posé une question difficile.