AMT Homework assignment 1

Inference for mixed populations

22nd February 2006

Soutrik BANERJEE

Introduction

The data considered here were obtained from a double-blind, randomised, parallel group, multicentre study in which a placebo treatment was compared with a new anti-epileptic drug (AED), in combination with one or two other AEDs. The randomisation of epilepsy patients took place after a 12-week baseline period that served as a stabilisation period for the use of AEDs, and during which the total number of seizures were counted. There were a total of 45 patients, who were assigned the placebo treatment and 44 patients, who were given the new (active) treatment, out of a total of 89 patients. Patients were followed-up for a period of 16 weeks ; some patients were even followed-up for 27 weeks, however in the given dataset, the counts of seizures for 16 weeks are only presented.

In the given data set, the response variable, Y, was the total number (count) of seizures during the 16-week period. In addition, the patients' ID, a treatment indicator (a binary variable), and a baseline rate (a continuous variable) of the average number of epileptic seizures per week were also provided.

Distribution of the number of seizures in the given sample

The histogram of the Y variable (Figure 1) shows a very positively skewed distribution. A lognormal distribution (or approximation, if we consider the data to be continuous) was fitted to the sample distribution of the Y variable, which was found to be satisfactory ($P > .25$). The plot of the logarithm of the Y variable is also shown in Figure 2, which resembles a Normal distribution. Often medical data follow a highly positively skewed distribution with the standard deviation proportionately increasing with the mean. Under these circumstances a logarithmic transformation almost always solves the problem (*cf.* JM Bland). In the following sections, however an alternative approach is used, where Poisson distribution or its mixture is used to fit the Y variable taking into account its discrete (count) nature.

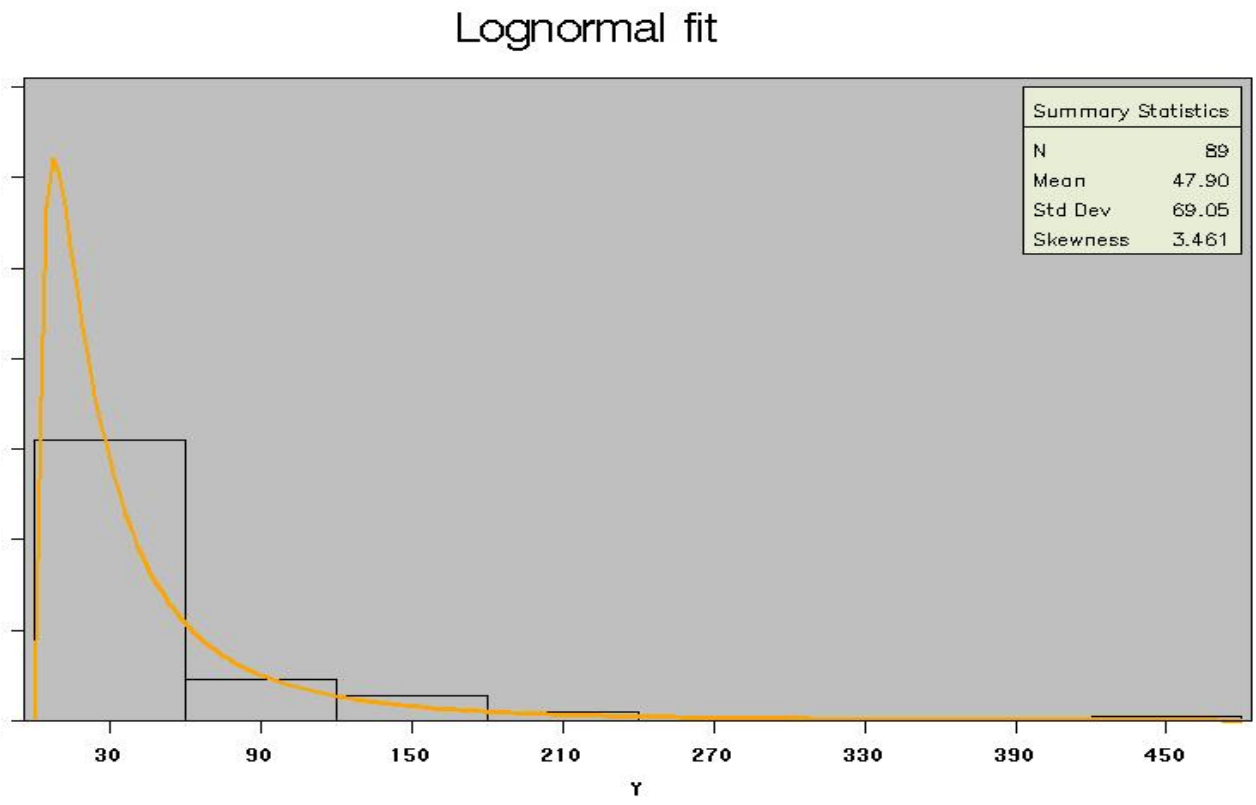

Figure 1.

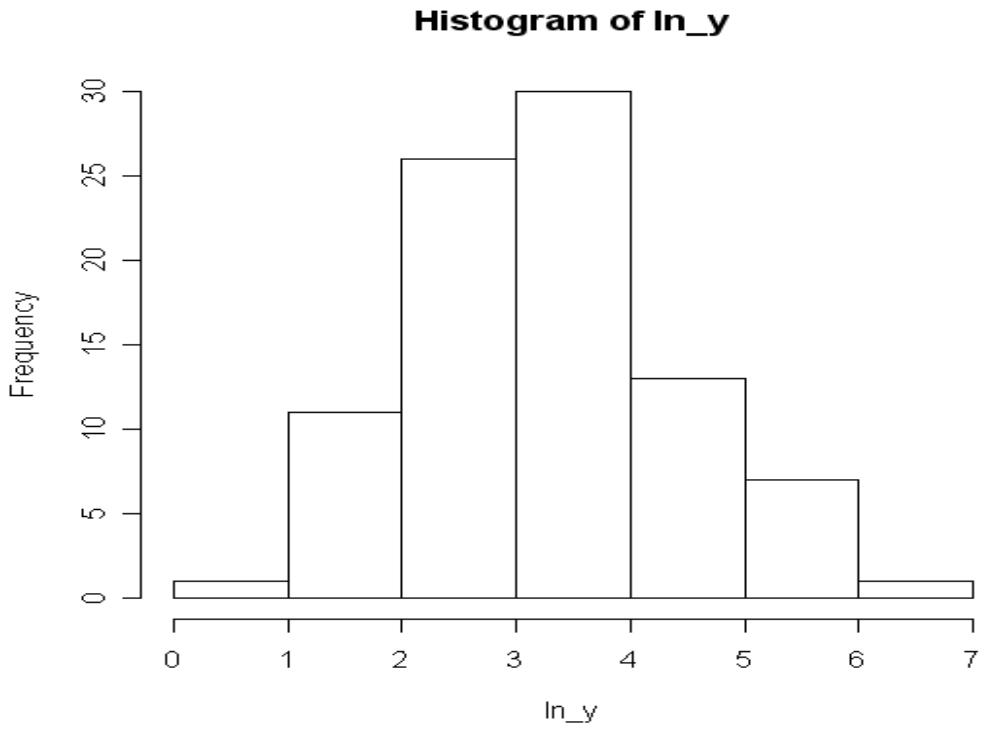

Figure 2.

Mean and variance of the response variable

The mean number of seizures was found to be 47.9 and the variance to be 4768.43. As the given response variable is a count data (non-negative integer values), Poisson modelling was considered in this case. It can be seen that there is gross overdispersion from the discrepancy between the mean and variance of the sample response variable. A possible cause of overdispersion is heterogeneity in the sample. In order to explore this phenomenon, analysis of a finite mixture model was considered using the software C.A.MAN© developed by Böhning and Schlattmann. In Figure 3, the distribution of the Y variable is plotted along with a fitted single Poisson distribution (line) with mean = 47.9, showing a poor fit.

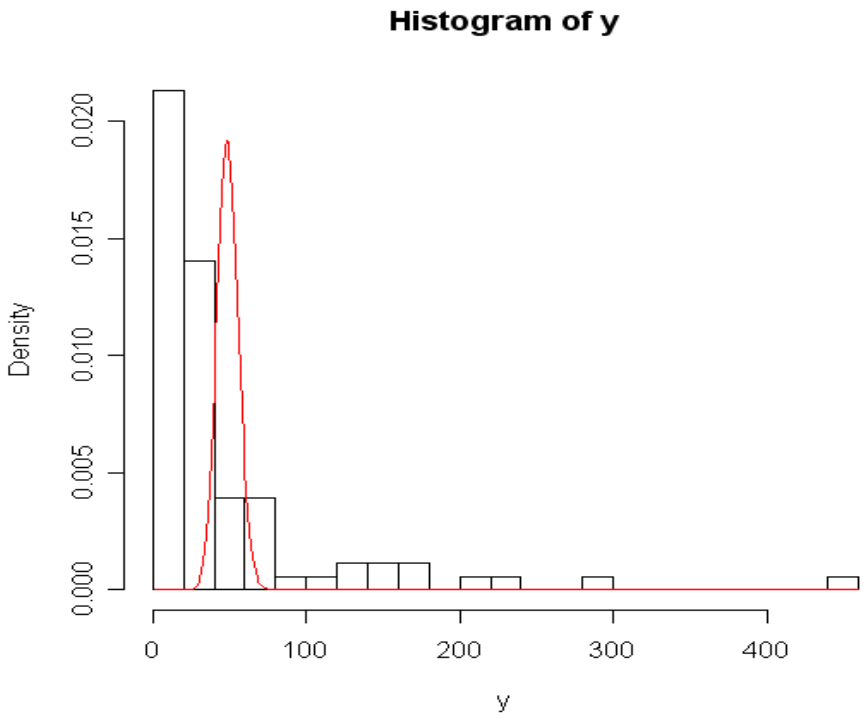

Figure 3.

Method

The dataset contains 56 different values of the Y variable (the number of seizures) out of a total 89 patients. The minimum value was 1 and maximum value was 459. The first question was to determine the number of support points *g* to fit a Poisson mixture model. The method was NPMLE with parameter grid size as default, VEM algorithm, full NR (Newton-Raphson) method, accuracy level of 0.00001 (the smallest), with number of maximum iterations equal to 30,000 was chosen at the outset. It returned 7 support points with two support points getting 0 weights. The programme then prompted for a possible refinement of the number of parameters and of the values of parameter estimates using the EM algorithm, which yielded 5 support points. The accuracy level was changed to 0.01 (the largest) and the same procedure was performed.

Following, it was tried to reduce the number of support points (using fixed support size) from 5 to 3 (keeping the accuracy level the smallest each time).

It was also tried to see if there were any difference in the number of parameters and parameter estimates by starting with a high parameter grid value, which was 50 in this case. The maximum NP (non-parametric) log-likelihood values with the number of estimates are given in the Table 1 for the all the steps.

Using the special option (option 10) in C.A.MAN, a classification rule was made to divide the patients in different clusters.

Finally, an exploration of a possible association between the clusters and known covariates was done.

It was not possible to plot the graphics associated with C.A.MAN results owing to the fact that the graphics (option 8) resulted in stopping of the program. Hence, each time the graphics (option 8) was used, the program had to be restarted. In addition, the current work is done in Open Office software, hence an elegant presentation of mathematical expressions could not be done in this report. The softwares used were SAS, C.A.MAN, and R.

Results

Specifications	***No. of support points***	***Maximum NP Log-likelihood values***
Parameter grid size = default, accuracy 0.00001	5	- 431.80580
Parameter grid size = default, accuracy 0.01	5	- 431.76440
Fixed support size, accuracy 0.00001	5	- 426.65630
Fixed support size, accuracy 0.00001	4	**- 400.98470**
Fixed support size, accuracy 0.00001	3	- 344.63060*
Parameter grid size = 50, accuracy 0.00001	12	- 412.78620

Table 1 showing the number of support points and maximum NP log-likelihood values for different specifications. One can see that the best maximum NP log-likelihood value was obtained when the number of fixed support points were 4. One can also see that for the number of support points equal to 12 by use of a larger parameter grid size (= 50) in the beginning, there was no substantial improvement in the maximum NP log-likelihood value. By changing the accuracy level from the smallest to the largest, there was very little change in the maximum NP log-likelihood value, neither was a change in the number of support points noted. The asterisk (*) denotes that for the fixed support size of 3, the results are unacceptable due to the fact that maximum directional gradient function was less than unity. It is also worthwhile to mention here that the maximum NP log-likelihood values and parameter estimates were slightly different (for a given fixed support size and accuracy level), if the starting values of the parameter estimates were entered different in C.A.MAN. Therefore, the starting values for the parameter estimates are also included in the appendix, when fixed support size procedure was done.

Estimated parameter weights (π_j)	***Estimated mean of no. of seizures (λ_j)***
0.3668	10.18
0.3512	27.28
0.1584	62.50
0.1236	145.86

Table 2 shows the different estimated probabilities with estimated mean no. of seizures for each sub-population in the finite Poisson mixture model with 4 support points.

Components	***Observed patients (out of 89)***	***Observed proportions***	***Estimated prior proportions (π_j)***
1	33	0.37	0.3668
2	31	0.35	0.3512
3	14	0.16	0.1584
4	11	0.12	0.1236

Table 3 shows the observed no. of patients classified into each component group. It can be seen that the observed proportions are almost equal to the estimated prior proportions. A detailed, case-by-case, classification can be seen in the appendix.

In the next step, the variance of the mixture was calculated. Note that the sample variance was 4768.43. The variance of the mixture is given by :

$$\text{Var}(Y) = \Sigma\, \pi_j\lambda_j^2 - (\Sigma\, \pi_j\lambda_j)^2 + \Sigma\, \pi_j\lambda_j = 1888.23$$

One can notice that the overdispersion is not completely corrected by the 4-component Poisson mixture model.

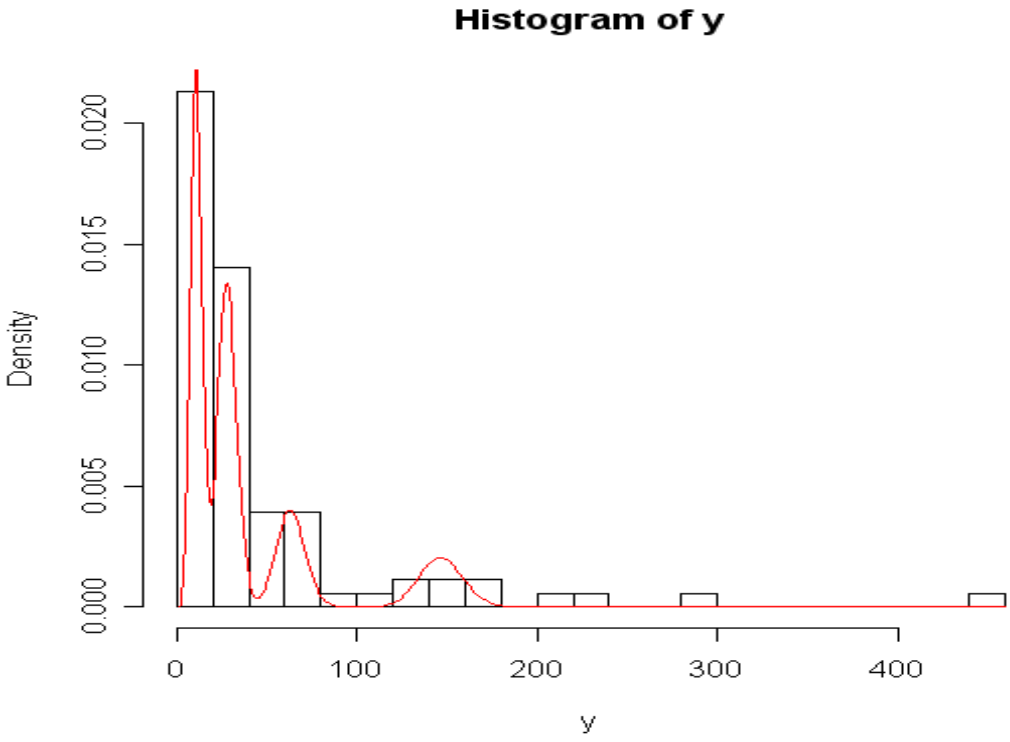

Figure 4. Figure showing a better fit with a 4-component Poisson mixture model, although some of the data on the extreme right of the graph are not "captured" (precisely 4 patients). This may explain the failure to correct the overdispersion. The latter might be improved by adding more components, however a parsimonious structure of 4-component Poisson mixture model would be lost. Besides, it is well known that patients, who are prone to have repeated seizures may have a variable number of seizures, which does not necessarily signify that they belong to a different group by aetiology or by risk of seizures.

Discussion

Primary epileptic seizures are classified into :

a) Partial seizures :-
   i) Simple (consciousness not impaired)
   ii) Complex (consciousness impaired)
b) Secondary Generalised (Jacksonian) from partial seizures
c) Generalised :-
   i) Absenses (petit mal)
   ii) Tonic-clonic (grand mal) – the commonest of all seizures
   iii) Myoclonic jerk – sudden violent jerk of the body or a part of it (e.g., disobedient leg)
   iv) Atonic (flaccid type)

v) Akinetic (motionless)

Each type has its own characteristic recurrence pattern, EEG, prognosis, treatment, etc. In the present dataset, information on the treatment as well as the causes of epileptic seizures were not provided. On the basis of the number of seizures, one may empirically classify the patients into one of the plausible 4 components (*cf.* Table 3). It may be interesting in future to study the role of different predictive factors that might distinguish the above 4 groups.

In an attempt to find a potential association between these 4 clusters and the 2 given covariates, *viz.*, the treatment indicator and the baseline rate of average number of epileptic seizures per week. Using proc GLM (or ANOVA) in SAS, the baseline rate was found to be significant ($P < .0001$), but the treatment indicator was 'borderline' significant ($P = .0563$) with overall significance at .05 level ($P < .0001$). This implies that the level of baseline seizure rate (prior odds) was predictive of the outcome of the number of seizures during the 16-week period (posterior odds).

In addition, by using proc GENMOD on the *count* Y variable (using log link and Poisson distribution *without* Pearson's scaling), both the treatment indicator and baseline seizure rate were found to be significant ($P < .0001$ for both), but only the baseline seizure rate was significant ($P < .0001$) if one uses the Pearson's scaling to account, in part, for the overdispersion factor, thus widening the standard errors and hence providing a better fit. (The treatment indicator was not significant at .05 level ($P < .0802$) in the latter case where Pearson's scaling was used.) This shows that the results obtained by treating the Y variable as count or as a class variable of 4 components tally with each other concerning its association with the covariates.

The results obtained in the analysis of the Poisson mixture model appears promising and simple to interpret, since only 4 clusters were identified. As the causes of epileptic fits can be several, this classification may only, in part, explain the prognostic nature or potential risks on the part of the patient. However, the rôle of a full diagnostic evaluation of the causes of epilepsy cannot be overemphasised.

References :

[1] Bland JM. An Introduction to Medical Statistics (1987), 3rd Ed., Enlish Language Book Society

[2] C.A.MAN 2.0 (Computer Assisted Mixture Analysis) (1997). http://www.medizin.fu-berlin.de/sozmed/caman.html

[3] Molenberghs G and Verbeke G (2005). *Models for Discrete Longitudinal Data*. New York : Springer-Verlag

[4] *Oxford Handbook of Clinical Medicine* (reprint 2002), 5th Ed., Oxford University Press, USA.

[5] Verbeke G and Molenberghs G (2005). *Advanced Modelling Techniques* (Biostatistics course notes, UHasselt).

Appendix

SAS codes :

```
data Amt2;
set Amt;
ln_y = log(y);
run;quit;

proc univariate data = Amt2 normal;
var y ln_y;
qqplot;
run;quit;

title 'Lognormal fit';
ods select Lognormal.ParameterEstimates Lognormal.GoodnessOfFit MyPlot;
proc univariate data = Amt2;
    var y;
    histogram / lognormal (w = 3 theta = est color = orange)
            cframe = ligr
            vaxis = axis1
            name = 'MyPlot';
    inset n mean (5.3) std = 'Std Dev' (5.3) skewness (5.3) /
    pos = ne header = 'Summary Statistics' cfill = ywh;
    axis1 label = (a = 90 r = 0);
run;
```

Parameters for Lognormal Distribution

Parameter	Symbol	Estimate
Threshold	Theta	0.004601
Scale	Zeta	3.246507
Shape	Sigma	1.096923
Mean		46.90969
Std Dev		71.61117

Goodness-of-Fit Tests for Lognormal Distribution

Test	---Statistic----		-----p Value-----
Kolmogorov-Smirnov	D	0.06653647	Pr > D >0.250
Cramer-von Mises	W-Sq	0.06266986	Pr > W-Sq >0.250
Anderson-Darling	A-Sq	0.39870825	Pr > A-Sq >0.250

```
data Amt3;
set Amt;
if Y < 18 then Y1 = 0;
if Y => 18 & Y < 45 then Y1 = 1;
if Y => 45 & Y < 110 then Y1 = 2;
if Y => 110 then Y1 = 3;
```

```
run;quit;

proc glm data = Amt3;
class Y1 trt;
model Y1 = trt bserate / solution;
run;quit;

proc glm data = Amt3;
class Y1 trt;
model Y1 = bserate / solution;
run;quit;

proc genmod data = Amt3;
class trt;
model Y = trt bserate / dist = poi link = log;
run;quit;

proc genmod data = Amt3;
class trt;
model Y = trt bserate / dist = poi link = log scale = pearson;
run;quit;
```

C.A.MAN outputs :

```
The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:DEFAULT
DISTRIBUTION:POISSON
ALGORITHM:     VEM
STEP-LENGTH: FULL NR
ACCURACY:   .000010
NUMBER OF ITERATIONS:30000

step       43       max. dir. derivative   1.000007
The NPMLE consists of  5 support points
Result after combining equal estimates:

weight:      .3668 parameter:     10.176220
weight:      .3512 parameter:     27.281510
weight:      .1584 parameter:     62.501390
weight:      .0786 parameter:    145.846200
weight:      .0450 parameter:    238.946800

Log-Likelihood at iterate :     - 431.80580
```

```
The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:DEFAULT
DISTRIBUTION:POISSON
```

ALGORITHM: VEM
STEP-LENGTH: FULL NR
ACCURACY: **.010000**
NUMBER OF ITERATIONS:30000

step 21 max. dir. derivative 1.008191
The NPMLE consists of 5 support points
Result after combining equal estimates:

weight: .3765 parameter: 10.359420
weight: .3432 parameter: 27.644990
weight: .1566 parameter: 62.717090
weight: .0786 parameter: 145.846200
weight: .0450 parameter: 238.946800

Log-Likelihood at iterate : - 431.76440

Minimum of your data is : 1.000000
Maximum of your data is : 459.000000
Please enter number of support points : **5**

Starting values for fixed support size:
10 .2
30 .2
50 .2
100 .2
150 .2

The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:**ENTERED**
DISTRIBUTION:POISSON
ALGORITHM: **FIXED**
STEP-LENGTH: NONE
ACCURACY: **.000001**
NUMBER OF ITERATIONS:30000

step 56 max. dir. derivative 1.000001
The NPMLE consists of 5 support points
Result after combining equal estimates:

weight: .3539 parameter: 9.937642
weight: .3483 parameter: 26.326710
weight: .1253 parameter: 54.117100
weight: .0601 parameter: 82.873210
weight: .1124 parameter: 167.604600

Log-Likelihood at iterate : - 426.65630

Minimum of your data is : 1.000000
Maximum of your data is : 459.000000

Please enter number of support points : **4**

Starting values for fixed support size:
10 .25
30 .25
50 .25
100 .25

The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:**ENTERED**
DISTRIBUTION:POISSON
ALGORITHM: **FIXED**
STEP-LENGTH: NONE
ACCURACY: **.000001**
NUMBER OF ITERATIONS:30000

step 39 max. dir. derivative 1.000001
The NPMLE consists of 4 support points
Result after combining equal estimates:

weight: .3668 parameter: 10.176050
weight: .3512 parameter: 27.281190
weight: .1584 parameter: 62.501200
weight: .1236 parameter: 145.856800

Log-Likelihood at iterate : **- 400.98470**

Minimum of your data is : 1.000000
Maximum of your data is : 459.000000
Please enter number of support points : **3**

Starting values for fixed support size:
10 .333
30 .333
100 .333

The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:**ENTERED**
DISTRIBUTION:POISSON
ALGORITHM: **FIXED**
STEP-LENGTH: NONE
ACCURACY: **.000001**
NUMBER OF ITERATIONS:30000

step 5 max. dir. derivative **.992873**
The NPMLE consists of 3 support points
Result after combining equal estimates:

weight: .3943 parameter: 10.689330
weight: .3376 parameter: 28.787010
weight: .2681 parameter: 67.535980

Log-Likelihood at iterate : - 344.63060

The program will use the following options
to compute NPMLE:
DATA-FILE:epimix.dat
PARAMETER-GRID:**COMPUTED**
DISTRIBUTION:POISSON
ALGORITHM: VEM
STEP-LENGTH: FULL NR
ACCURACY: **.000001**
NUMBER OF ITERATIONS:30000

step 30001 max. dir. derivative 1.000002
The NPMLE consists of 12 support points
Result after combining equal estimates:

weight: .0069 parameter: 1.112560
weight: .1581 parameter: 6.376639
weight: .2283 parameter: 14.284870
weight: .2064 parameter: 24.158250
weight: .1370 parameter: 35.876100
weight: .0795 parameter: 57.951490
weight: .0601 parameter: 73.676870
weight: .0330 parameter: 121.922900
weight: .0455 parameter: 163.072600
weight: .0227 parameter: 214.530000
weight: .0112 parameter: 286.983200
weight: .0112 parameter: 459.000000

Log-Likelihood at iterate : - 412.78620

---

datum 1.000000 belongs to cluster 1
datum 4.000000 belongs to cluster 1
datum 5.000000 belongs to cluster 1
datum 6.000000 belongs to cluster 1
datum 7.000000 belongs to cluster 1
datum 8.000000 belongs to cluster 1
datum 9.000000 belongs to cluster 1
datum 10.000000 belongs to cluster 1
datum 11.000000 belongs to cluster 1
datum 12.000000 belongs to cluster 1
datum 13.000000 belongs to cluster 1
datum 14.000000 belongs to cluster 1
datum 15.000000 belongs to cluster 1
datum 16.000000 belongs to cluster 1
datum 17.000000 belongs to cluster 1
datum 18.000000 belongs to cluster 2
datum 19.000000 belongs to cluster 2

datum	21.000000 belongs to cluster	2
datum	22.000000 belongs to cluster	2
datum	23.000000 belongs to cluster	2
datum	24.000000 belongs to cluster	2
datum	25.000000 belongs to cluster	2
datum	27.000000 belongs to cluster	2
datum	29.000000 belongs to cluster	2
datum	30.000000 belongs to cluster	2
datum	32.000000 belongs to cluster	2
datum	33.000000 belongs to cluster	2
datum	34.000000 belongs to cluster	2
datum	35.000000 belongs to cluster	2
datum	37.000000 belongs to cluster	2
datum	39.000000 belongs to cluster	2
datum	41.000000 belongs to cluster	2
datum	45.000000 belongs to cluster	3
datum	46.000000 belongs to cluster	3
datum	50.000000 belongs to cluster	3
datum	51.000000 belongs to cluster	3
datum	59.000000 belongs to cluster	3
datum	60.000000 belongs to cluster	3
datum	61.000000 belongs to cluster	3
datum	62.000000 belongs to cluster	3
datum	64.000000 belongs to cluster	3
datum	68.000000 belongs to cluster	3
datum	72.000000 belongs to cluster	3
datum	78.000000 belongs to cluster	3
datum	84.000000 belongs to cluster	3
datum	110.000000 belongs to cluster	4
datum	125.000000 belongs to cluster	4
datum	131.000000 belongs to cluster	4
datum	155.000000 belongs to cluster	4
datum	159.000000 belongs to cluster	4
datum	165.000000 belongs to cluster	4
datum	176.000000 belongs to cluster	4
datum	209.000000 belongs to cluster	4
datum	221.000000 belongs to cluster	4
datum	287.000000 belongs to cluster	4
datum	459.000000 belongs to cluster	4

R codes :

```
# dividing by 3 is a scaling parameter to fit the figure; single Poisson mean

hist(y, nclass=25, prob=T, col=0)
lines((dpois(0:459, 47.9)/3), type="l", col=2)

# dividing by 2 is a scaling parameter to fit the figure; 4-component Poisson mixture

hist(y, nclass=25, prob=T, col=0)
lines((0.3668*dpois(0:459, 10.18)/2)
```

```
+(0.3512*dpois(0:459, 27.28)/2)
+(0.1584*dpois(0:459, 62.5)/2)
+(0.1236*dpois(0:459, 145.86)/2), type="l", col=2)
```