distplot {vcd}

http://www.inside-r.org/packages/cran/vcd/docs/distplot

Diagnostic Distribution Plots

Package: vcd
Version: 1.3-1

Description

Diagnostic distribution plots: poissonness, binomialness and negative binomialness plots.

Usage

```
distplot(x, type = c("poisson", "binomial", "nbinomial"),
  size = NULL, lambda = NULL, legend = TRUE, xlim = NULL, ylim = NULL,
  conf_int = TRUE, conf_level = 0.95, main = NULL,
  xlab = "Number of occurrences", ylab = "Distribution metameter",
  gp = gpar(cex = 0.5), name = "distplot", newpage = TRUE, pop = TRUE, ...)
```

Arguments

`x`
: either a vector of counts, a 1-way table of frequencies of counts or a data frame or matrix with frequencies in the first column and the corresponding counts in the second column.

`type`
: a character string indicating the distribution.

`size`
: the size argument for the binomial and negative binomial distribution. If set to `NULL` and `type` is `"binomial"`, then `size` is taken to be the maximum count. If set to `NULL` and `type` is `"nbinomial"`, then `size` is estimated from the data.

`lambda`
: parameter of the poisson distribution. If type is `"poisson"` and `lambda` is specified a leveled poissonness plot is produced.

`legend`
: logical. Should a legend be plotted?

`xlim`
: limits for the x axis.

`ylim`
: limits for the y axis.

`conf_int`
: logical. Should confidence intervals be plotted?

`conf_level`
: confidence level for confidence intervals.

`main`
: a title for the plot.

`xlab`
: a label for the x axis.

`ylab`
: a label for the y axis.

`gp`
: a `"gpar"` object controlling the grid graphical parameters of the points.

`name`
: name of the plotting viewport.

`newpage`
: logical. Should `grid.newpage` be called before plotting?

`pop`
: logical. Should the viewport created be popped?

`...`
: further arguments passed to `grid.points`.

Details

`distplot` plots the number of occurrences (counts) against the distribution metameter of the specified distribution. If the distribution fits the data, the plot should show a straight line. See Friendly (2000) for details.

References

D. C. Hoaglin (1980), A poissonness plot, *The American Statistican*, **34**, 146--149.

D. C. Hoaglin \& J. W. Tukey (1985), Checking the shape of discrete distributions. In D. C. Hoaglin, F. Mosteller, J. W. Tukey (eds.), *Exploring Data Tables, Trends and Shapes*, chapter 9. John Wiley \& Sons, New York.

M. Friendly (2000), *Visualizing Categorical Data*. SAS Institute, Cary, NC.

Examples

```
## Real data examples:
data("HorseKicks")
data("Federalist")
distplot(HorseKicks, type = "poisson")
distplot(Federalist, type = "poisson")
```

Author(s)

Achim Zeileis Achim.Zeileis@R-project.org

<u>Poissonness Plot</u>

Graphical techniques can help to appreciate how closely a sample of discrete data follows a member of one of the common families of discrete distributions, such as the Poisson, Binomial, Negative Binomial.

Poisson distribution is often used as a model for the occurrence of locally rare events.

Examples: # NDAs approved/year, #Supreme Court vacancies/term, # ships arriving in a harbor/day.

In queueing theory: assume Poisson arrivals and exponential waiting times

$$P_\lambda(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

lambda = average # occurrences/unit time

k=0,1,,2, . . .

For a sample of size N,

$$\hat{\lambda} = \bar{x} = \frac{\sum k\, n_k}{N}$$

traditional estimator,
also the Maximum Likelihood Estimator

A single discrepant cell count, n_k, can affect $\bar{x}$ so would like to be able to use a resistant estimator.

Plot Construction

Assume, for some fixed λ, each observed frequency, n_k equals the expected frequency, $m_k = Np_k$. Then, setting $n_k = Np_k = Ne^{-\lambda}\lambda^k/k!$, and taking logs of both sides gives

$$\log(n_k) = \log N - \lambda + k \log \lambda - \log k! \ .$$

This can be rearranged to a linear equation in k,

$$\phi(n_k) \equiv \log\left(\frac{k!\, n_k}{N}\right) = -\lambda + (\log \lambda)\, k \ . \tag{3.13}$$

The left side of Eqn. (3.13) is called the ***count metameter***, and denoted $\phi(n_k)$. Hence, plotting $\phi(n_k)$ against k should give a straight line of the form $\phi(n_k) = a + bk$ with

- slope = $\log \lambda$
- intercept = $-\lambda$

when the observed frequencies follow a Poisson distribution. If the points in this plot are close enough to a straight line, then an estimate of λ may be obtained from the slope b of the line, $\hat{\lambda} = e^b$ should be reasonably close in value to the MLE of λ, $\hat{\lambda} = \bar{x}$. In this case, we might as well use the MLE as our estimate.

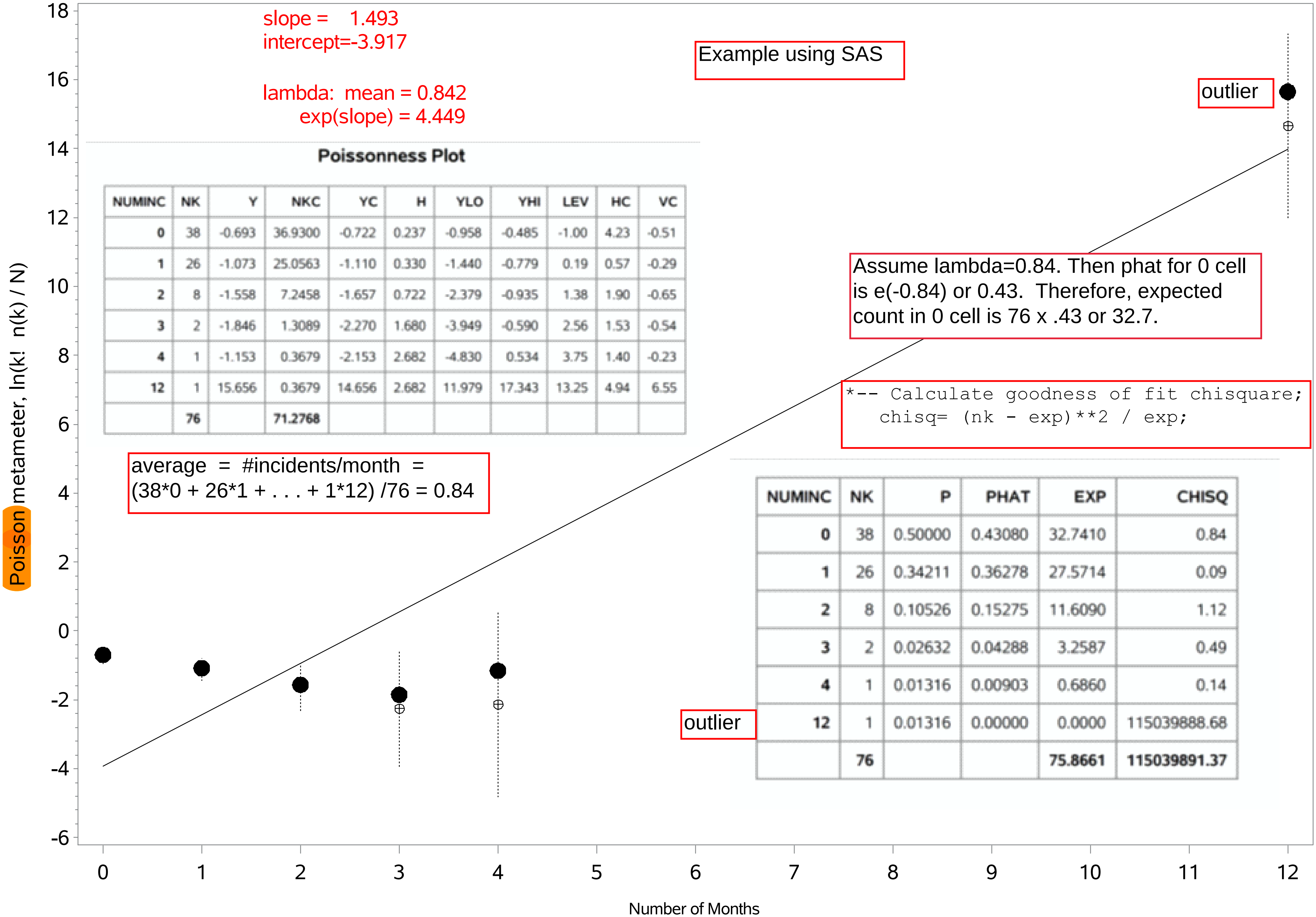
References:
Hoaglin (1980), "A Poissonness Plot", The American Statistician, 34, pp. 146-149.
Hoaglin and Tukey (1985), "Checking The Shape of Discrete Distributions". In Hoaglin, Mosteller, and Tukey, editors, "Exploring Data Tables, Trends, and Shapes", chapter 9, John Wiley and Sons, New York.
Friendly (2000), "Visualizing Categorical Data", SAS Publishing, Cary, NC, pp. 49-56.

Example (Poissonness Plot)

Incidents of international terrorism/month in US during 1968-1974

# Incidents	Number of Months
0	38
1	26
2	8
3	2
4	1
12	1
	N=76 months were studied
note: the 12 incidents were later discovered to be related.	
Independence of events is assumed by the Poisson distribution	

Poissonness Plot



Poissonness Plot

NUMINC	NK	Y	NKC	YC	H	YLO	YHI	LEV	HC	VC
0	38	-0.693	36.9300	-0.722	0.237	-0.958	-0.485	-1.00	4.23	-0.51
1	26	-1.073	25.0563	-1.110	0.330	-1.440	-0.779	0.19	0.57	-0.29
2	8	-1.558	7.2458	-1.657	0.722	-2.379	-0.935	1.38	1.90	-0.65
3	2	-1.846	1.3089	-2.270	1.680	-3.949	-0.590	2.56	1.53	-0.54
4	1	-1.153	0.3679	-2.153	2.682	-4.830	0.534	3.75	1.40	-0.23
12	1	15.656	0.3679	14.656	2.682	11.979	17.343	13.25	4.94	6.55
	76		71.2768							

NUMINC	NK	P	PHAT	EXP	CHISQ
0	38	0.50000	0.43080	32.7410	0.84
1	26	0.34211	0.36278	27.5714	0.09
2	8	0.10526	0.15275	11.6090	1.12
3	2	0.02632	0.04288	3.2587	0.49
4	1	0.01316	0.00903	0.6860	0.14
12	1	0.01316	0.00000	0.0000	115039888.68
	76			75.8661	115039891.37

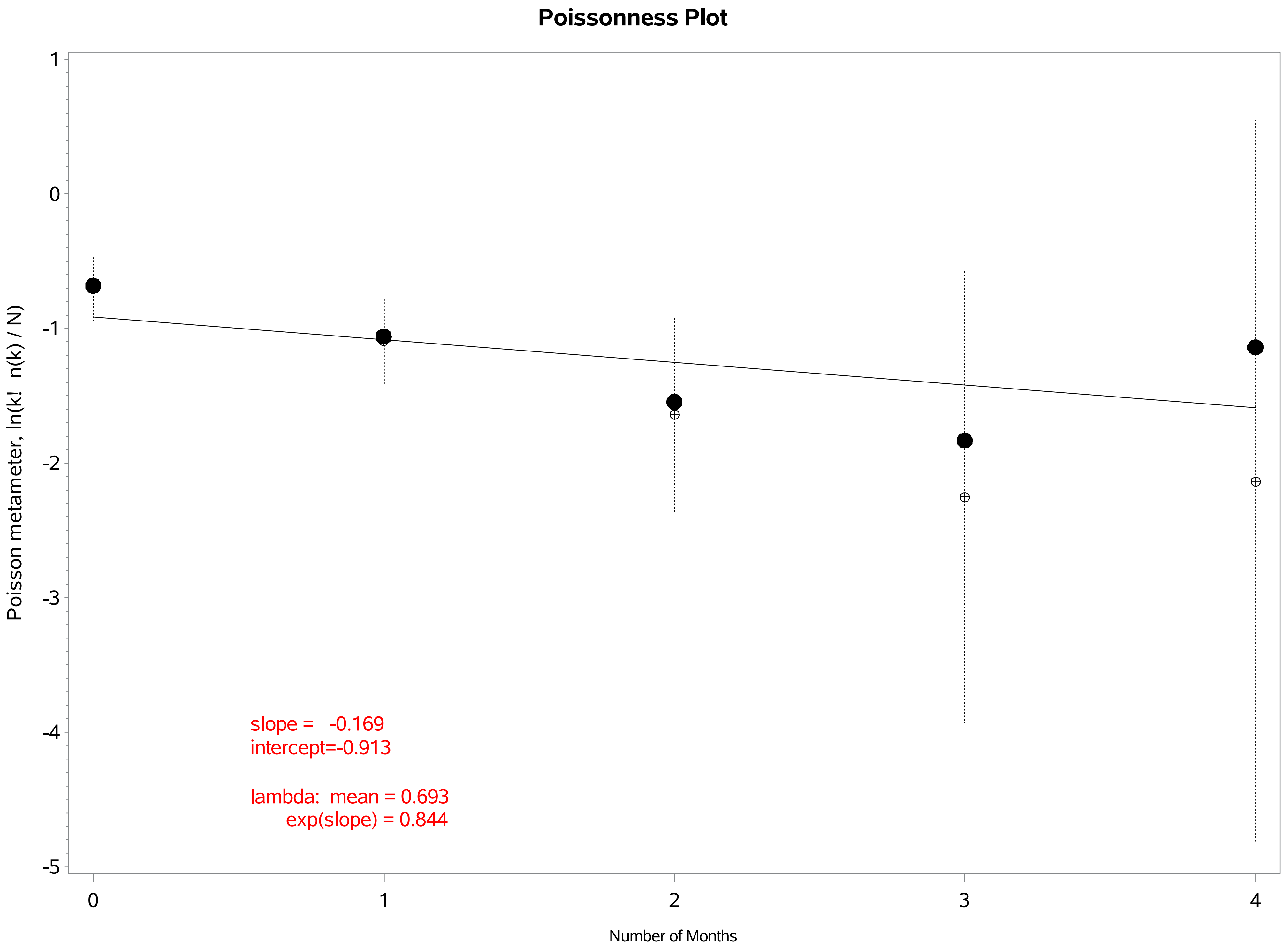
Poissonness Plot

NUMINC	NK	Y	NKC	YC	H	YLO	YHI	LEV	HC	VC
0	38	-0.680	36.9247	-0.709	0.235	-0.944	-0.474	-1.00	4.26	0.07
1	26	-1.059	25.0527	-1.097	0.329	-1.426	-0.767	0.44	1.34	-0.11
2	8	-1.545	7.2447	-1.644	0.721	-2.365	-0.923	1.88	2.61	-0.30
3	2	-1.833	1.3087	-2.257	1.679	-3.936	-0.577	3.33	1.98	-0.28
4	1	-1.139	0.3679	-2.139	2.682	-4.816	0.547	4.77	1.78	0.01
	75		**70.8985**							

calculations redone with outlier eliminated

NUMINC	NK	P	PHAT	EXP	CHISQ
0	38	0.50667	0.49991	37.4930	0.00686
1	26	0.34667	0.34660	25.9952	0.00000
2	8	0.10667	0.12016	9.0117	0.11357
3	2	0.02667	0.02777	2.0827	0.00328
4	1	0.01333	0.00481	0.3610	1.13108
	75			**74.9435**	**1.25479**

average = #incidents/month =
(38*0 + 26*1 + . . . + 1*4) /75 = 0.69


Poissonness Plot
Poisson metameter, ln(k! n(k) / N)
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
Number of Months
slope = -0.169
intercept=-0.913
lambda: mean = 0.693
exp(slope) = 0.844

Poissonness Plot

DEATHS	NK	Y	NKC	YC	H	YLO	YHI	LEV	HC	VC
0	109	-0.607	107.894	-0.617	0.130	-0.748	-0.487	-1.00	7.66	0.05
1	65	-1.124	64.070	-1.138	0.207	-1.345	-0.931	0.64	3.09	-0.16
2	22	-1.514	21.242	-1.549	0.417	-1.966	-1.132	2.28	5.47	0.12
3	3	-2.408	2.318	-2.666	1.318	-3.984	-1.348	3.92	2.97	-0.43
4	1	-2.120	0.368	-3.120	2.691	-5.797	-0.415	5.56	2.06	-0.20
	200		**195.892**							

Classic data set: number of deaths due to horsekicks recorded by Prussian army during Crimean War. Ten army corps were observed over 20 years giving a total of 200 observations.

DEATHS	NK	P	PHAT	EXP	CHISQ
0	109	0.545	0.54335	108.670	0.00100
1	65	0.325	0.33144	66.289	0.02506
2	22	0.110	0.10109	20.218	0.15705
3	3	0.015	0.02056	4.111	0.30025
4	1	0.005	0.00313	0.627	0.22201
	200			**199.915**	**0.70537**

xbar=0.61

(109x0 + 65x1 + 22x2 + 3x3 + 1x4)/200

= 122/200 = 0.61

see HorseKicks.R

Poissonness Plot