A graphical tool for assessing the suitability of count regression models, with applications in biological dosimetry

Jochen Einbeck (joint work with Paul Wilson)

Birkbeck, 8th December 2021

Motivating context: Radiation biodosimetry

- Radiation incident leading to (potentially) exposed individuals
- Contracted radiation dose can be estimated retrospectively by exploiting the radiation-induced change in certain biomarkers
- 'Gold–standard': Dicentric chromosomes (resulting from unsuccessful DNA-damage response)

Dicentric chromosomes: Laboratory data

Frequency of dicentric chromosomes in human lymphocytes after *in vitro* exposure to doses between 1 and 5Gy of 200kV X–rays. The irradiated blood was mixed with non-irradiated blood in a proportion 1:3 in order to mirror a partial body exposure scenario.

	Frequency of counts									
dose	0	1	2	3	4	5	6	7	8	# cells
1	2713	78	8	0	1	0	0	0	0	2800
2	1302	71	22	5	0	0	0	0	0	1400
3	1116	46	28	7	2	1	0	0	0	1200
4	929	18	14	22	13	2	0	1	1	1000
5	726	17	18	12	9	13	1	4	0	800

Clearly, many 0's! But too many for Poisson–model?

General setup: Count data models

- Given: univariate count data $y_1, \ldots, y_n$.
- Is it plausible to assume that $y_1, \ldots, y_n$ are generated from a given (hypothesized) count distribution F?
- Specifically, denote $F = F(\mu_i, \theta_i)$, with both $\mu_i = E(Y_i|x_i)$ and θ_i (possibly) depending on covariates x_i.
- Assume that a routine to obtain estimates $\hat{\mu}_i = \hat{E}(Y_i|x_i)$ and $\hat{\theta}_i$ is readily available.
- Denote $N(k)$, for $k = 0, 1, 2, \ldots$, the number of observed counts k in $y_1, \ldots, y_n$.

General setup: Count data models

Frequency of dicentric chromosomes in human lymphocytes after *in vitro* exposure to doses between 1 and 5Gy of 200kV X–rays. The irradiated blood was mixed with non-irradiated blood in a proportion 1:3 in order to mirror a partial body exposure scenario.

	k									
x	0	1	2	3	4	5	6	7	8	# cells
1	2713	78	8	0	1	0	0	0	0	2800
2	1302	71	22	5	0	0	0	0	0	1400
3	1116	46	28	7	2	1	0	0	0	1200
4	929	18	14	22	13	2	0	1	1	1000
5	726	17	18	12	9	13	1	4	0	800
$N(k)$	6786	230	90	46	25	16	1	5	1	$n = 7200$

Aim of this talk

We will develop a graphical tool which helps to decide whether, for each count $k = 0, 1, 2, \ldots$, the number $N(k)$ is 'plausible' under the distribution F.

Distribution of N(k)

What is the distribution of the number of counts, $N(k)$, when $y_i \sim F(\mu_i, \theta_i)$? Denoting the probability of observing the count k under covariate x_i and model F as

$$p_i(k) = P(k|\mu_i, \theta_i),$$

it is clear that $N(k)$ is just the sum of Bernoulli r.v.'s with success probability $p_1(k), \ldots, p_n(k)$.

Consider firstly the case without covariates. Then $\mu_1 = \ldots = \mu_n \equiv \mu$, $\theta_1 = \ldots = \theta_n \equiv \theta$, and hence

$$p_1(k) = \ldots = p_n(k) \equiv p(k)$$

so that clearly

$$N(k) \sim \text{Bin}(n, p(k))$$

Distribution of N(k)

In the situation with covariates, the distribution of $N(k)$ is a bit more complicated, and is known as the Poisson–Binomial distribution

$$P(N(k) = \ell) = \left\{ \prod_{i=1}^{n} (1 - p_i(k)) \right\} \sum_{i_1 < \cdots < i_\ell} w_{i_1} \cdots w_{i_\ell} \tag{1}$$

with parameters $p_1(k), \ldots, p_n(k)$.
Here, $w_i \equiv w_i(k) = \frac{p_i(k)}{1-p_i(k)}$, $i = 1, 2, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, i_2, \ldots, i_\ell$ from $\{1, 2, \ldots, n\}$ (Chen and Liu, 1997).

- R implementation available in R package `poibin` (Hong, 2013).
- Note this is different (and unrelated) to the compound Poisson Binomial distribution.

Example: Poisson–Binomial distribution

Nine urns are filled with black balls and white balls. Urn 1 contains 10% white balls, urn 2 contains 20% etc. A ball is drawn from each urn.

What is a 90% prediction interval for the number of white balls drawn?

If 8 white balls were drawn, is this consistent with the percentages stated above?

```
probs <- c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9)
qpoibin(c(0.05,0.95), pp=probs)
[1] 2 7
1-(ppoibin(7, pp=probs))
[1] 0.00736272
```

Estimating parameters

The Poisson–Binomial distribution of the counts $N(k)$ depends on the parameters $p_i(k) = P(k|\mu_i, \theta_i), i = 1, \ldots, n$.

These parameters are unknown and have to be estimated from the data.

Candidate estimate: $\hat{p}_i(k) = P(k|\hat{\mu}_i, \hat{\theta}_i)$, where $\hat{\mu}_i$ and $\hat{\theta}_i$ come from the fitted count data model F in question.

- For instance, in the special case that $F(\mu_i, \theta_i)$ corresponds to Pois(μ_i), one has $\hat{p}_i(k) = \exp(-\hat{\mu}_i)\hat{\mu}_i^k/k!$.
- Clearly, this raises the question on the accuracy of $\hat{\mu}_i$ when the model F is wrong. Put aside for now.

Plausibility intervals for $N(k)$

- Knowing the distribution of $N(k)$, one can derive intervals of plausible values of $N(k)$ by considering appropriate quantiles from this distribution.

- For fixed k, appropriate lower and upper quantiles, say $q_{\alpha/2}(k)$ and $q_{1-\alpha/2}(k)$ of the Poisson-Binomial distribution can be computed[1]; e.g. using the R package `poibin`.

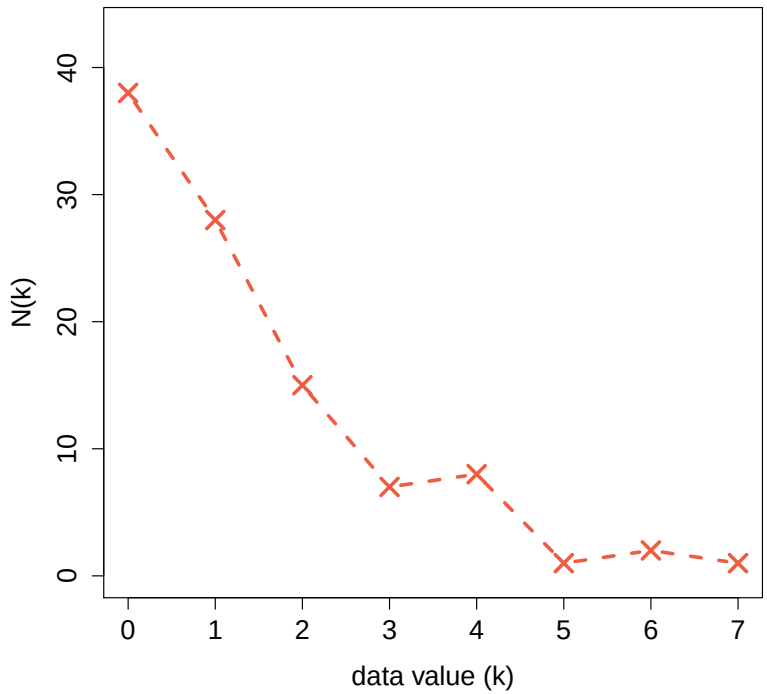
- Do this for a range of values of k, and plot intervals $(q_{\alpha/2}(k), q_{1-\alpha/2}(k))$ alongside observed values $N(k)$ as a function of k.

[1] alternative quantiles can be used, such as 'mid-quantiles'

Example: simulated data

$n = 100$ observations $y_1, \ldots, y_n$ simulated from a Zero-inflated Poisson (ZIP) distribution with Poisson parameter $\lambda = 1.5$ and zero-inflation parameter $p = 0.2$.

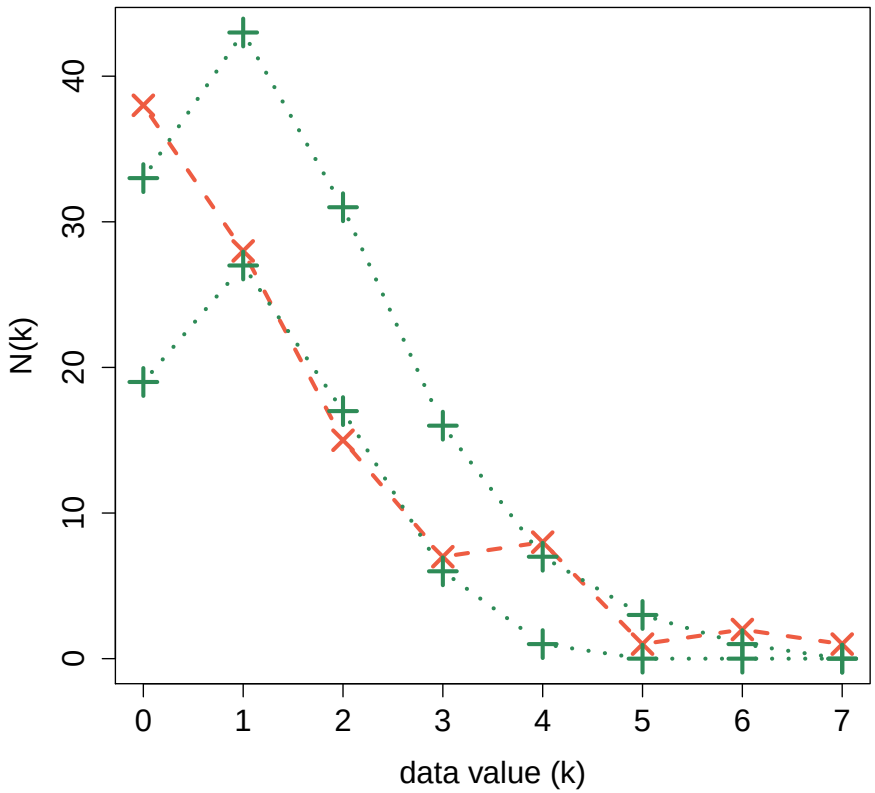
k	$N(k)$
0	38
1	28
2	15
3	7
4	8
5	1
6	2
7	1



Example: simulated data

Consider $F(\mu) \sim \text{Pois}(\mu)$ with $\hat{\mu} = \bar{y}$, so $\hat{p}(k) = e^{-\bar{y}} \frac{\bar{y}^k}{k!}$.

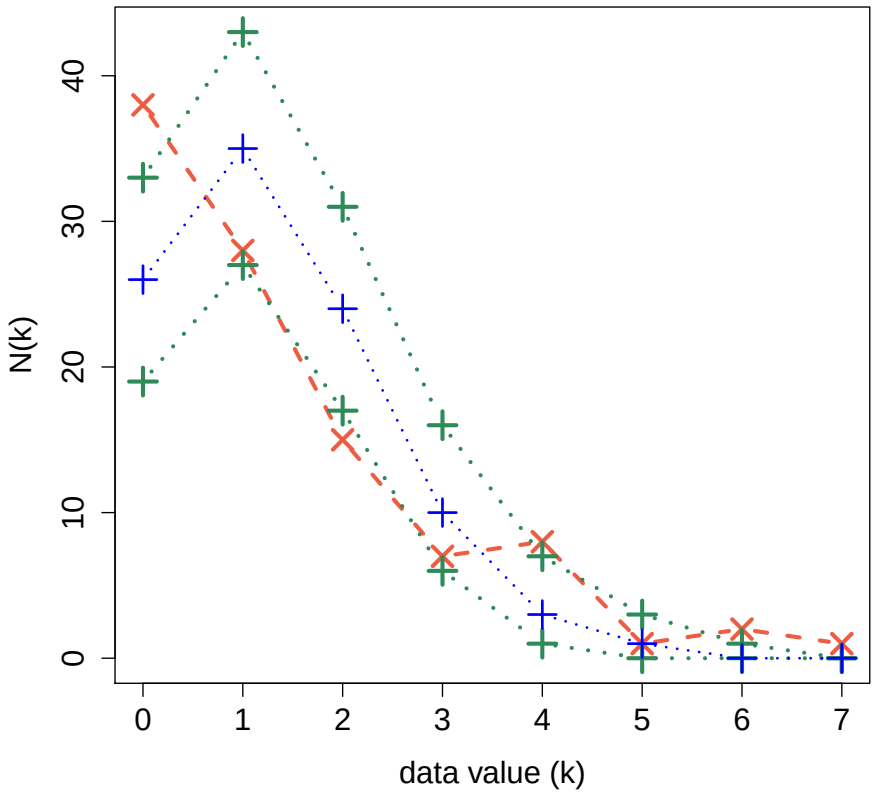
k	$N(k)$	$q_{0.05}(k)$	$q_{0.95}(k)$
0	38	19	33
1	28	27	43
2	15	17	31
3	7	6	16
4	8	1	7
5	1	0	3
6	2	0	1
7	1	0	0



Median-adjustment

The previous graph can be difficult to read if the sample size is large, and so the bounds get very tight.

We therefore adjust it by subtracting the medians $M(k) = \text{med}(N(k))$ from all values, where the median is taken wrt to the Poisson-Binomial distribution of $N(k)$.

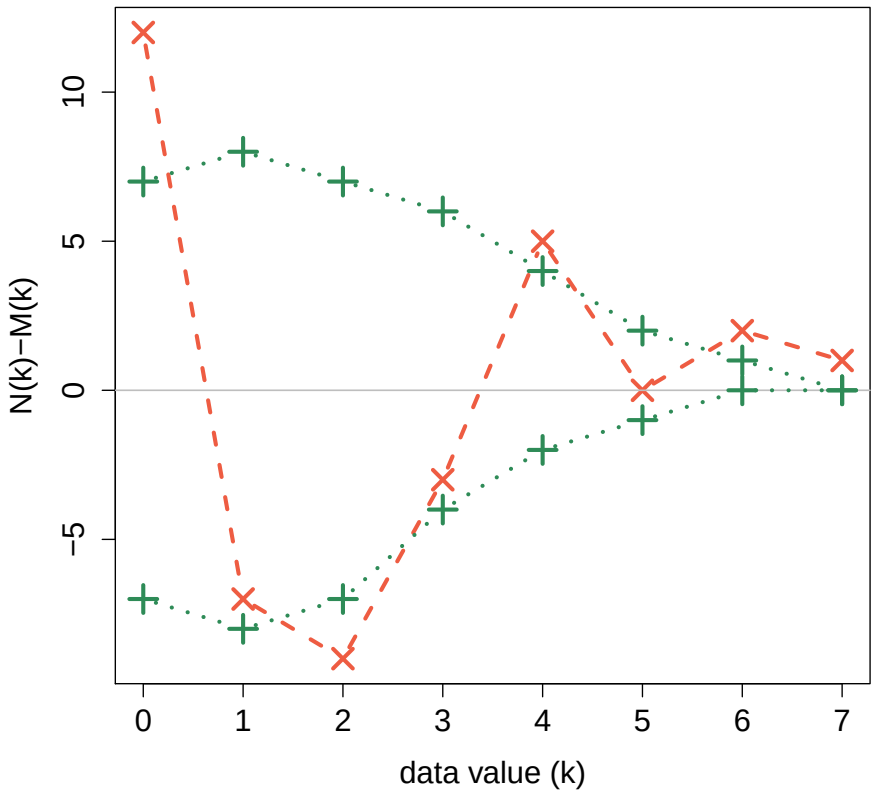


Median-adjustment

k	$N(k)$	$M(k)$	$N(k)$–$M(k)$	$q_{0.05}(k)$–$M(k)$	$q_{0.95}(k)$– $M(k)$
0	38	26	12	-7	7
1	28	35	-7	-8	8
2	15	24	-9	-7	7
3	7	10	-3	-4	6
4	8	3	5	-2	4
5	1	1	0	-1	2
6	2	0	2	0	1
7	1	0	1	0	0

Median-adjusted bounds

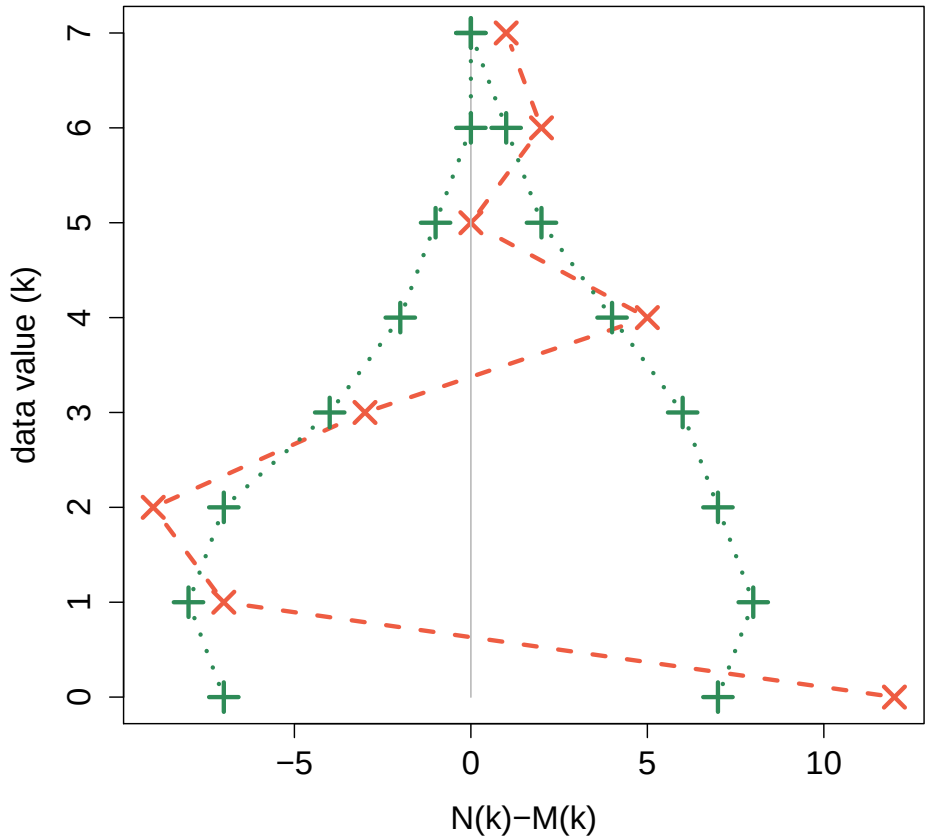
Diagnostic plot for the accuracy of the Poisson assumption.



Median-adjusted bounds: Vertical version

Exchange horizontal and vertical axis:

'Quantile band plot'



Return to biodosimetry data

Recall: These are data which resemble 'partial body exposure'.
Hence, we would expect inflation of zero's in the response.

	Frequency of counts								
dose	0	1	2	3	4	5	6	7	8
1	2713	78	8	0	1	0	0	0	0
2	1302	71	22	5	0	0	0	0	0
3	1116	46	28	7	2	1	0	0	0
4	929	18	14	22	13	2	0	1	1
5	726	17	18	12	9	13	1	4	0

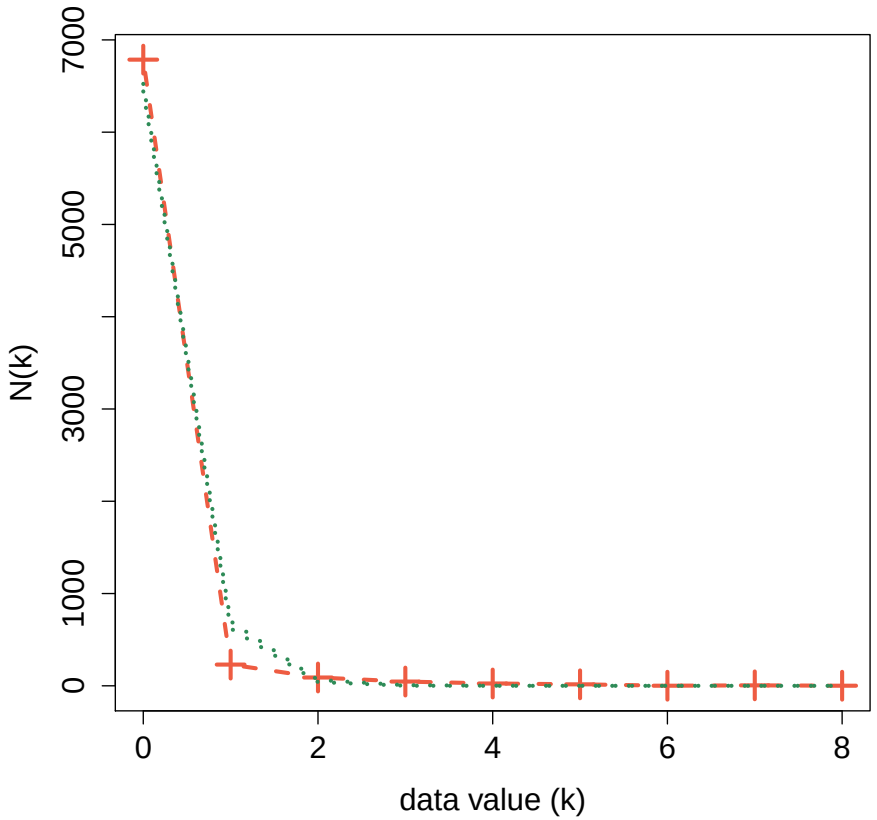
Let's check: Are these more zero's than one would reasonably expect under the Poisson assumption?

Do the same as before. That is,

- estimate
  $\hat{\mu}_i = \exp\{\hat{\beta}_0 + \hat{\beta}_1 \text{dose}_i + \hat{\beta}_2 \text{dose}_i^2\}$;
- build
  $\hat{p}_i(k) = \exp\{-\hat{\mu}_i\}\hat{\mu}_i^k / k!$;
- Use Poisson–Binomial distribution with parameters $\hat{p}_i(k)$.

k	$N(k)$	$q_{0.05}(k)$	$q_{0.95}(k)$
0	6786	6442	6524
1	230	622	700
2	90	41	64
3	46	1	7
4	25	0	1
5	16	0	0
6	1	0	0
7	5	0	0
8	1	0	0

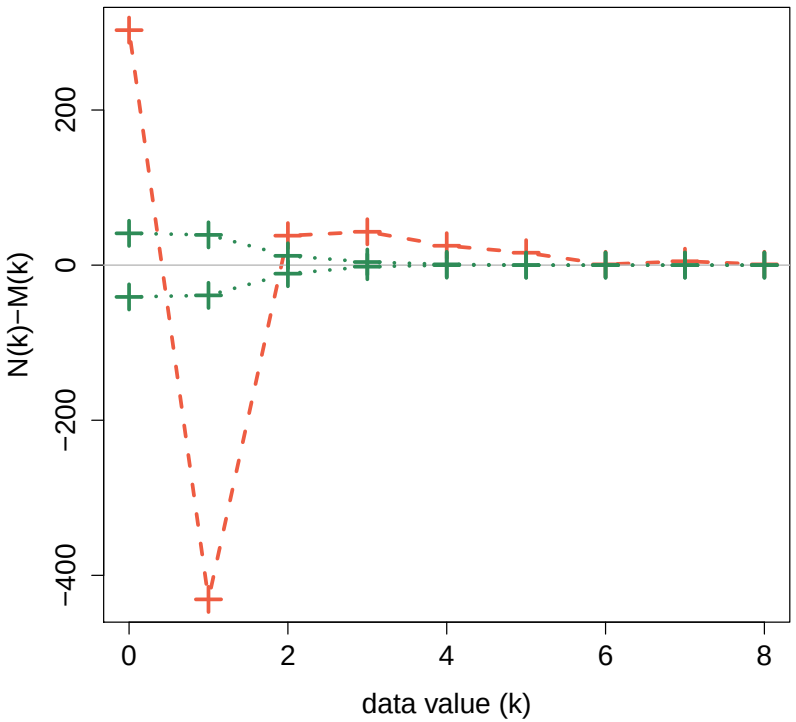
Diagnostics for biodosimetry data

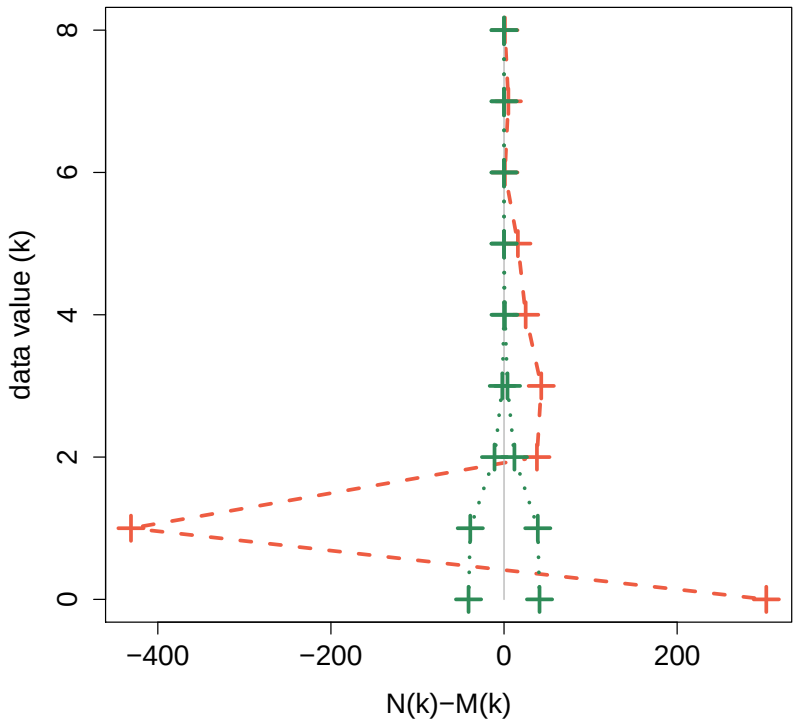


... does not look very useful since boundaries are very close.

Diagnostics for biodosimetry data

... so apply median–adjustment and rotate:

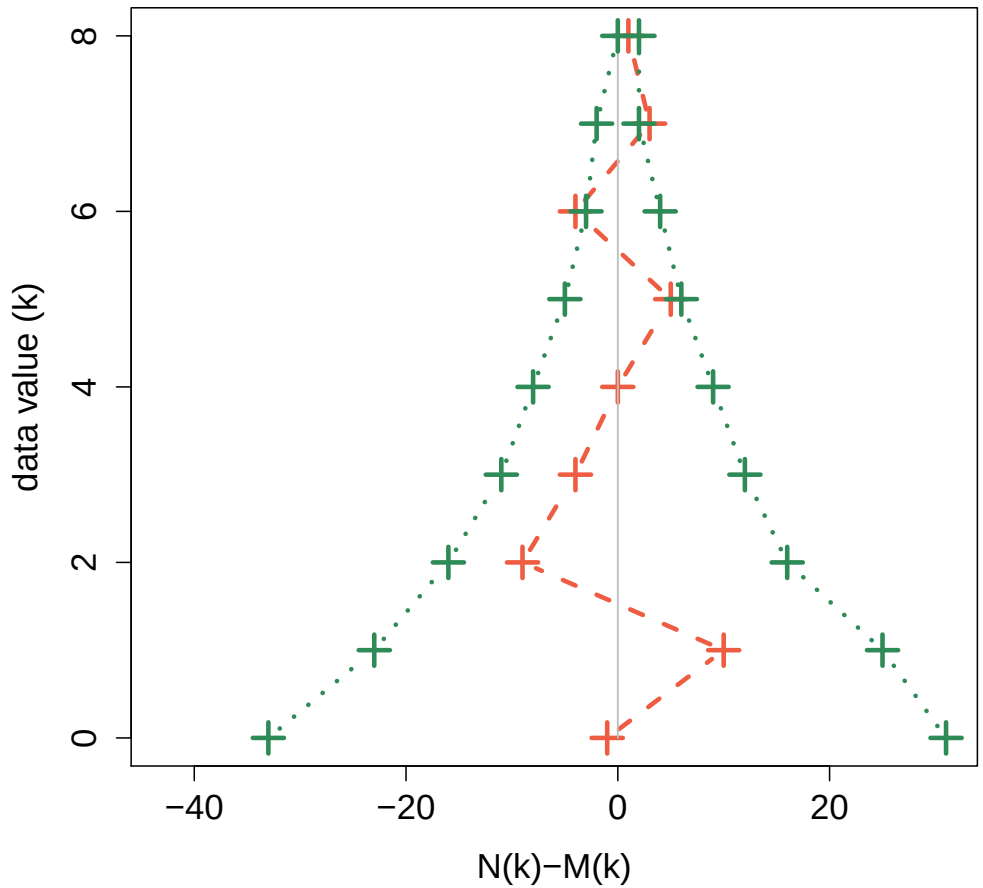




We clearly observe zero-inflation (and associated 1-deflation).

Quantile band plot: ZIP hypothesis

Do all the same as before, but now compute $\hat{\mu}_i$, $\hat{\theta}_i$, and $\hat{p}_i(k)$, using the zero-inflated Poisson (ZIP) model as the hypothesized model.

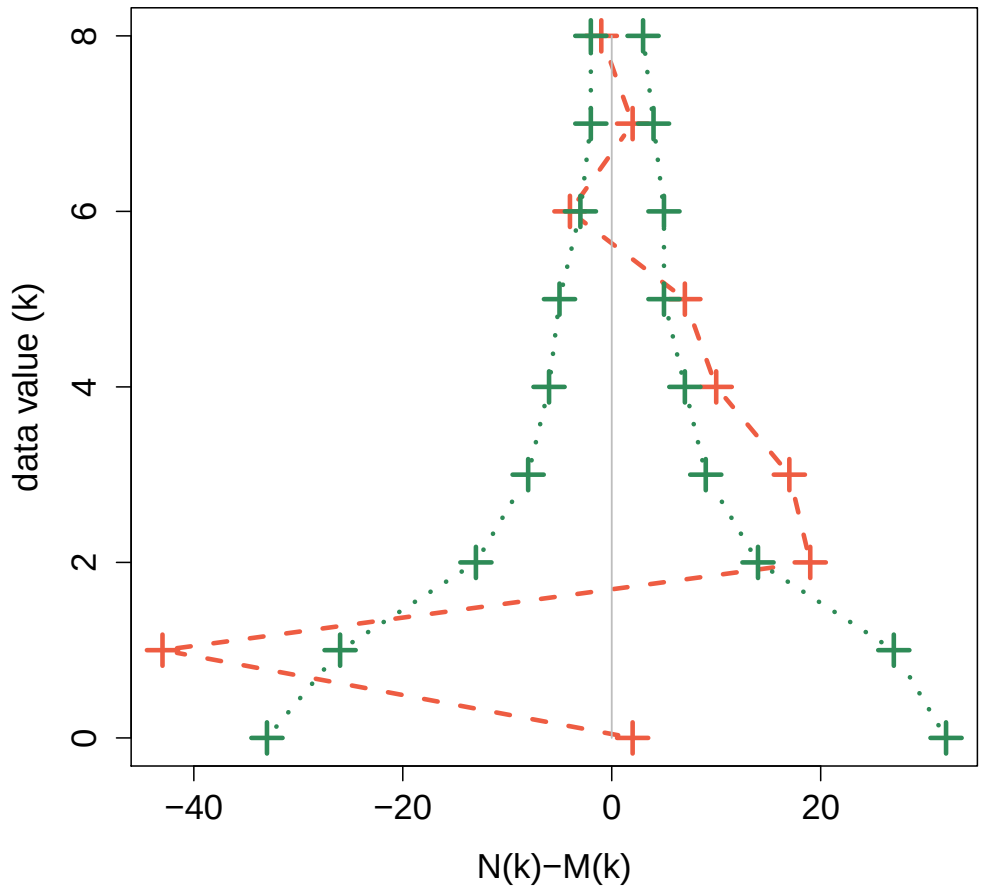


...indicates a good fit.

Quantile band plot: NB hypothesis

Repeat the procedure using the negative Binomial model as the hypothesized model.

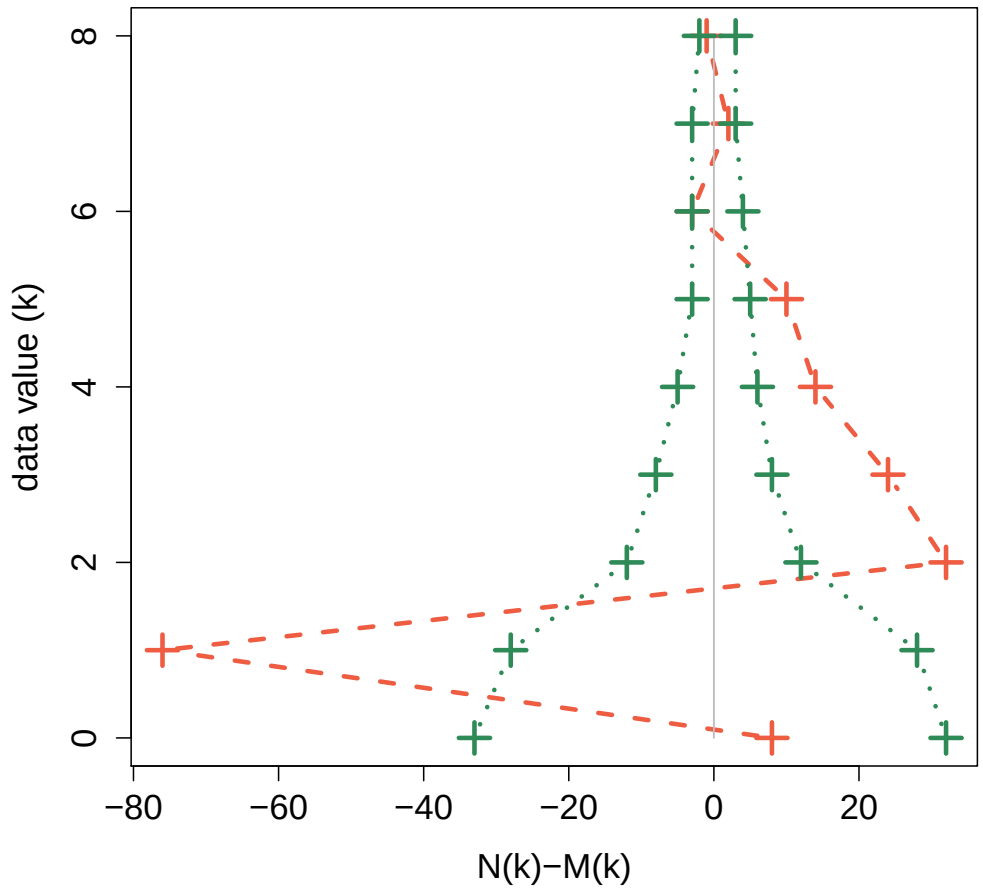
... indicates that the NB model does not capture the data well.



Quantile band plot: PIG hypothesis

Repeat the procedure using the Poisson inverse Gaussian (PIG) model as the hypothesized model.

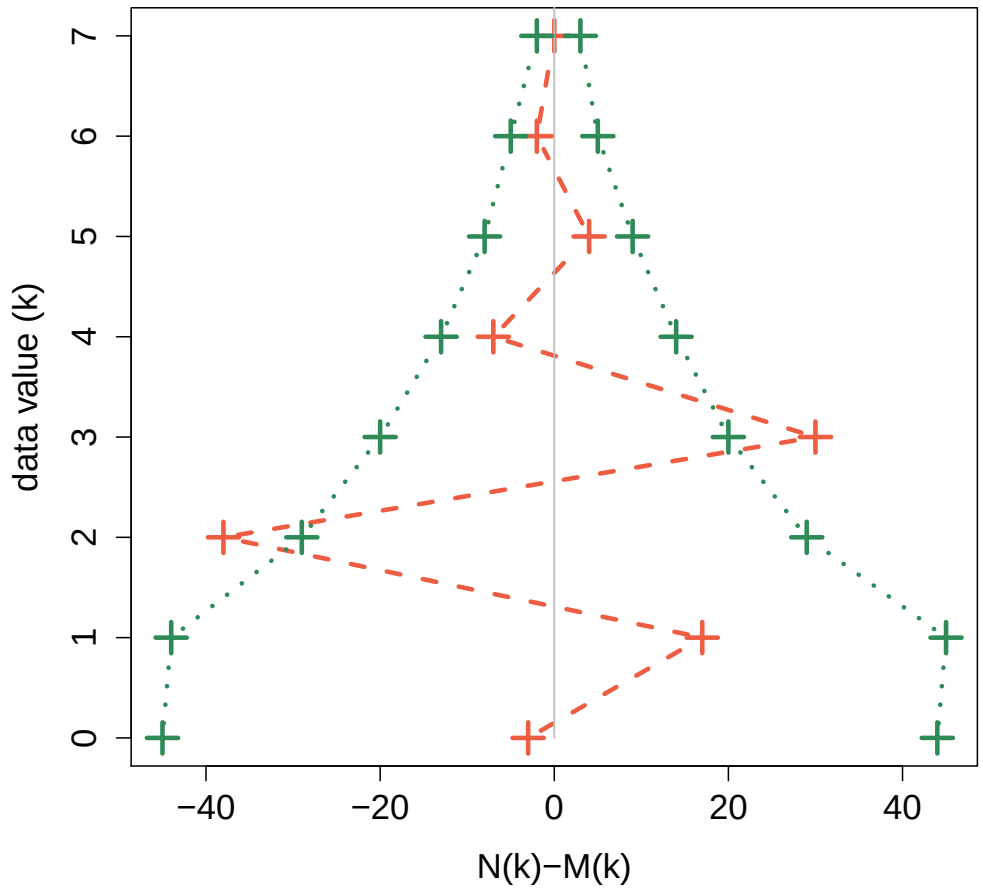
... the PIG model does not capture the data well either.



Alternative data set: Whole body exposure

Counts of dicentric chromosomes in 4400 blood cells after *in vitro* 'whole body' exposure with 200kV X-rays from 0 to 4.5Gy.

... indicates that Poisson model is fairly reasonable.



Multiple testing ?

- If considered as a series of statistical tests over counts $k = 0, 1, 2, ...$, one can argue that multiple testing issues arise.
- For instance, if the diagram covers ten possible counts, at a significance level of 0.1 one would expect the countline to fall beyond the quantile bounds once, purely by chance.
- One could adjust this through a Bonferroni correction etc (which leads, in our view, to meaningless boundaries).
- Hence, we do not make such a correction, but explicitly do not advocate this procedure as a testing procedure.
- It should rather be seen as a diagnostic device, similar as a residual plot or a QQ-plot.

Comparison with score tests

- Alternatively, one can carry out traditional score tests.
- For instance, consider H_0: Poisson versus H_1: ZIP or H_1: NB.
- Score test statistic $T = S^T J^{-1} S$, where S and J are the score function and Fisher Information matrix (resp.) evaluated under the Poisson model. Asymptotically, $T \sim \chi^2(1)$.
- Resulting values of T, to be compared with $\chi^2_{1,0.95} = 3.84$ (Oliveira et al, 2016):

Test	Body exposure	
	Partial	Whole
Pois/ZIP	1996.30	1.00
Pois/NB	6009.35	0.90

Comparison with score tests

- Confirms that Poisson is adequate for whole body exposure but inadequate for partial body exposure.

- ...but the score test does not tells us whether it's at all the zero's which cause the problem, nor whether the data are zero–inflated or –deflated!

Back to an initial question...

The procedure needs to 'know', or estimate, $p_i(k) = P(k|\mu_i, \theta_i)$ and hence the distributional parameters μ_i and θ_i.

We do not consider this estimation step as part of the methodology for the quantile band plot. The plot assesses the plausibility of the distributional assumption F, for given $F = F(\hat{\mu}_i, \hat{\theta}_i)$ as supplied by the user.

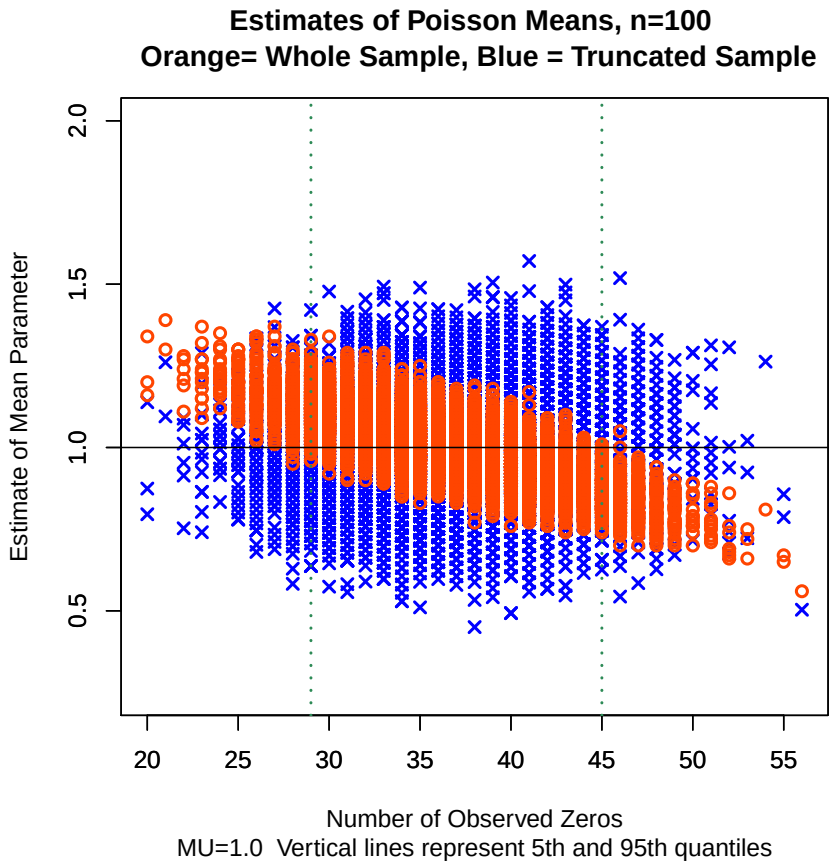
It is still worth asking: How reliable is the estimation of these parameters if the distributional 'hypothesis' is wrong?

Special case: $F = \text{Pois}(\lambda)$

How much is the estimate of λ affected by the number of zero's?

Experiment: Sample repeatedly 100 observations from Pois(1). Estimate the Poisson mean, for each run, by

- the usual whole sample mean, $\bar{y}$,
- the zero-truncated Poisson (ZTP) model based on the positive response data (Dietz & Böhning, 2000).



The mean λ of the Poisson distribution and the mean ζ of the ZTP distribution are theoretically related as

$$\zeta = \frac{\lambda e^{\lambda}}{e^{\lambda} - 1} \equiv g(\lambda)$$

The inverse function, $\lambda = g^{-1}(\zeta)$, does not have an analytical closed form expression, but can be approximated by

$$\lambda_{ZTP} \approx \frac{\zeta[1 - \exp(-h(\zeta))]^2 - [h(\zeta)]^2 \exp(-h(\zeta))}{1 - [h(\zeta) + 1] \exp(-h(\zeta))}$$

where $h(\zeta) = \zeta[1 - \exp(\frac{1}{\zeta} - \zeta)]$ (Ridout and Demétrio, 1992).

So, what to do?

Since the whole sample mean is biased under zero-modification, but the ZTP estimator has high variance, Wilson & Einbeck (2019) suggest a 'hybrid' (weighted) mean estimate, $\hat{\lambda}_h = 2/3 \times \bar{y} + 1/3 \times \hat{\lambda}_{ZTP}$.

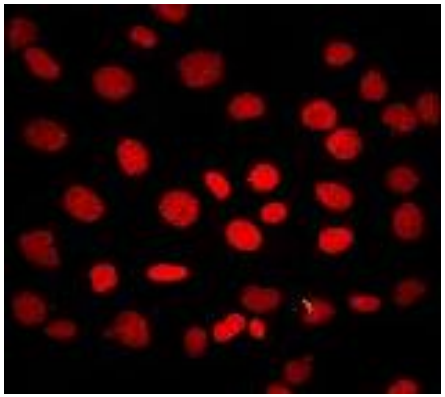
Simulation studies show that this will not harm the mean estimate if there is no zero-modification (inflation/deflation) present.

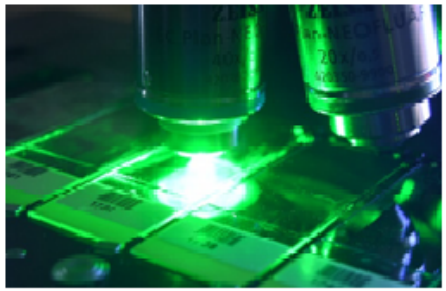
No such adjustments required for higher counts, or two-parameter distributions.

[Alternative approach for finding $P(N(0) = n_0)$ without requiring estimation of λ, based on an occupancy argument, and by conditioning on the total $n\bar{y}$: Fernandéz-Fontelo et al (2018).]

A protein-based radiation biomarker: γ-H2AX

- Following radiation-induced double strand breaks, the H2AX histone reacts with phosphorylation, in this state then referred to as γ-H2AX.
- The resulting foci can be counted manually or in a semi-automated way, using immunofluorescence microscopy.
- Typically, one examines a sample of 500-2000 (blood) cells on a given 'slide' and then records the number of foci per cell on each slide.





A protein-based radiation biomarker: γ-H2AX

Data from BfS, Germany, one slide with foci counts from 2006 cells, irradiated at 0.5Gy:

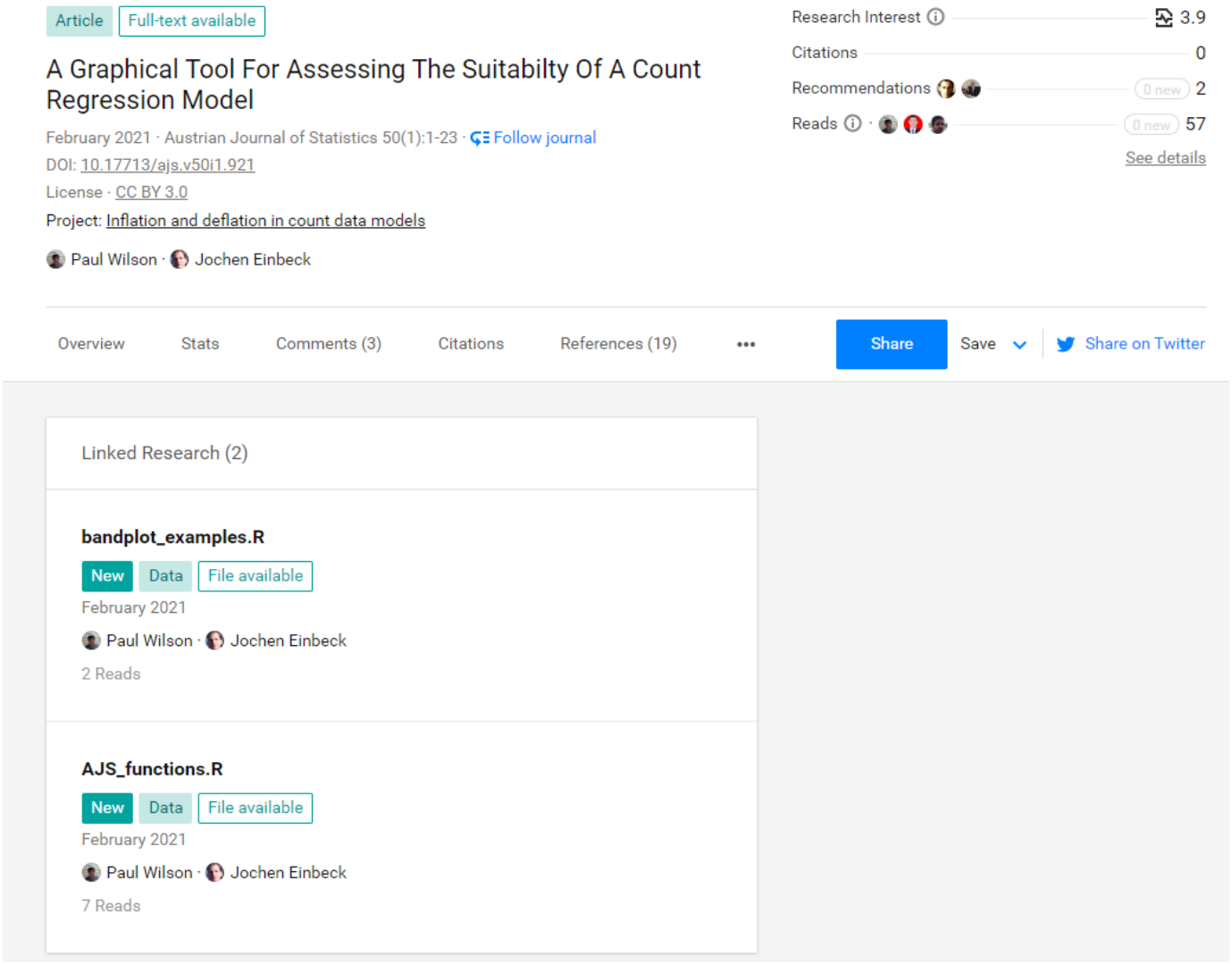
```
x <- read.table("h2ax-counts05Gy.dat")$x
table(x)
```

```
## x
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14
## 340 538 455 306 181  71  46  26  16   7   9   4   2   2   3
```

```
sum(table(x))
```

```
## [1] 2006
```

Find our code on ResearchGate...



Article · Full-text available

A Graphical Tool For Assessing The Suitabilty Of A Count Regression Model

February 2021 · Austrian Journal of Statistics 50(1):1-23 · Follow journal
DOI: 10.17713/ajs.v50i1.921
License · CC BY 3.0
Project: Inflation and deflation in count data models

Paul Wilson · Jochen Einbeck

Research Interest 3.9
Citations 0
Recommendations 0 new 2
Reads · 0 new 57
See details

Overview Stats Comments (3) Citations References (19) ••• Share Save Share on Twitter

Linked Research (2)

bandplot_examples.R

New Data File available
February 2021
Paul Wilson · Jochen Einbeck
2 Reads

AJS_functions.R

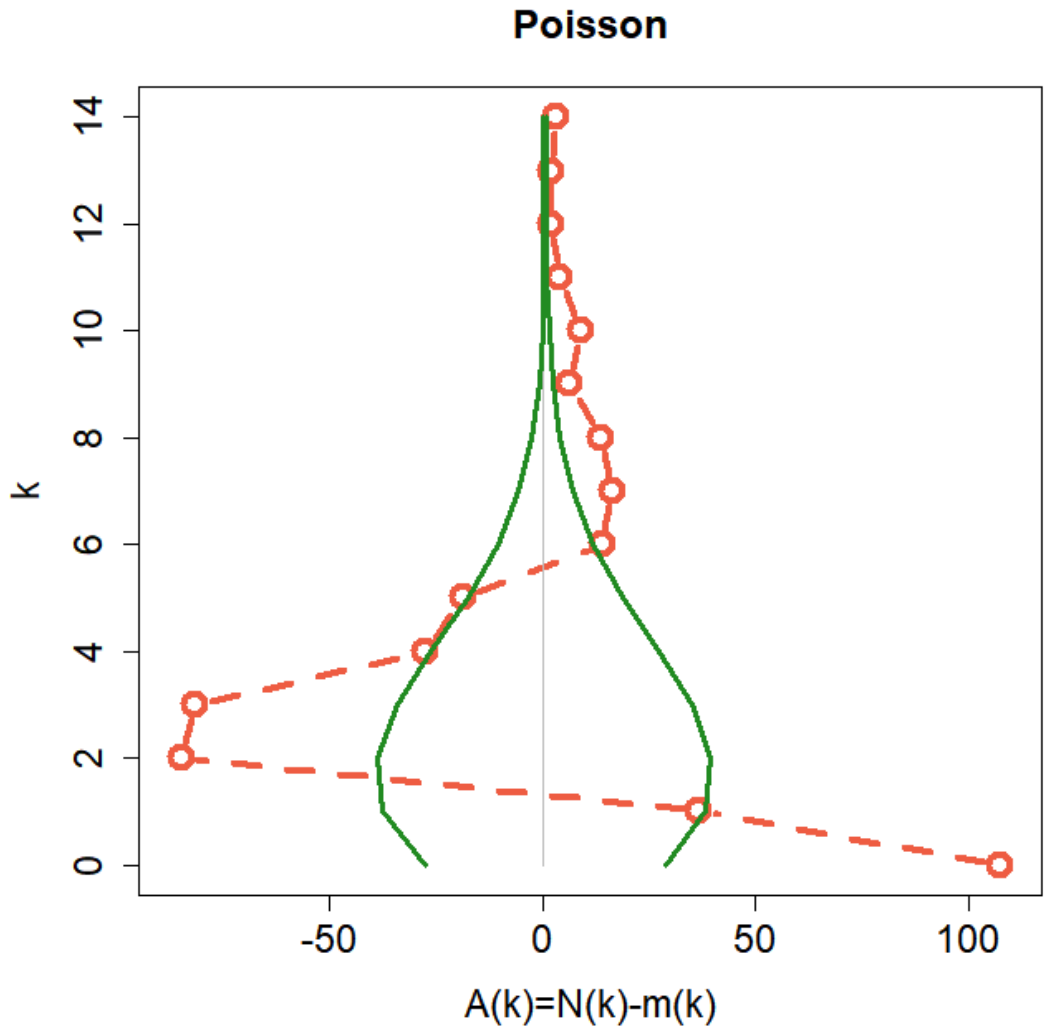
New Data File available
February 2021
Paul Wilson · Jochen Einbeck
7 Reads

Quantile band plot using Poisson mean

```
m  <-  mean(x)
m

## [1] 2.153539
```

```
quantbandplot(x, m)
```



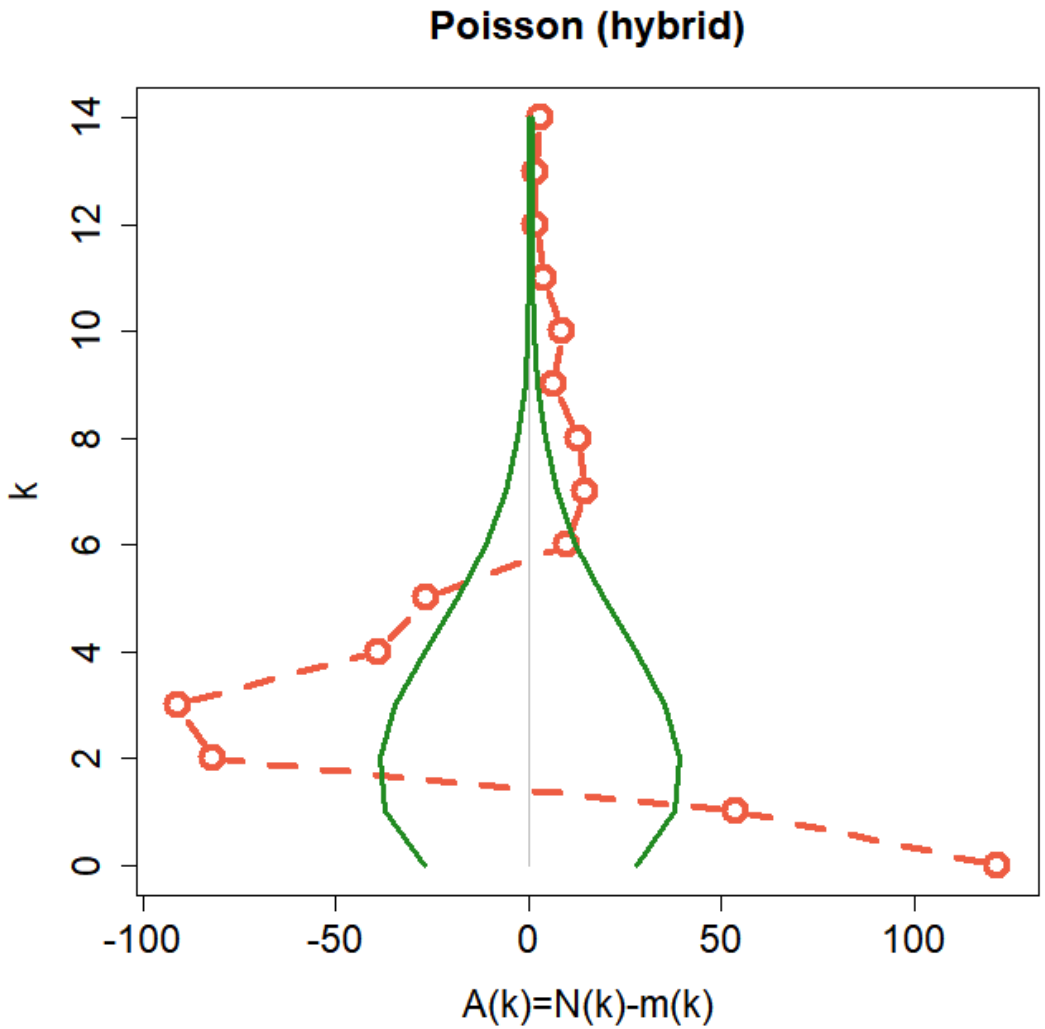
Hybrid mean estimator

```
h2<- function(t){
  h1<-function(t)  t*(1-exp(1/t-t))
  h2a<-function(t)  t*(1-exp(-h1(t)))^2-(h1(t))^2*exp(-h1(t))
  h2b<-function(t)  1-(h1(t)+1)*(exp(-h1(t)))
  return(h2a(t)/h2b(t))
}
mt <-  h2(mean(x[x>0]))  # mean estimate from ZPT, see slide 32
mh <-  2/3*m + 1/3*mt
mh    #$


## [1] 2.217165
```

```
quantbandplot(x, mh, type="vertical", main= "Poisson (hybrid)")
```

Hybrid mean estimator



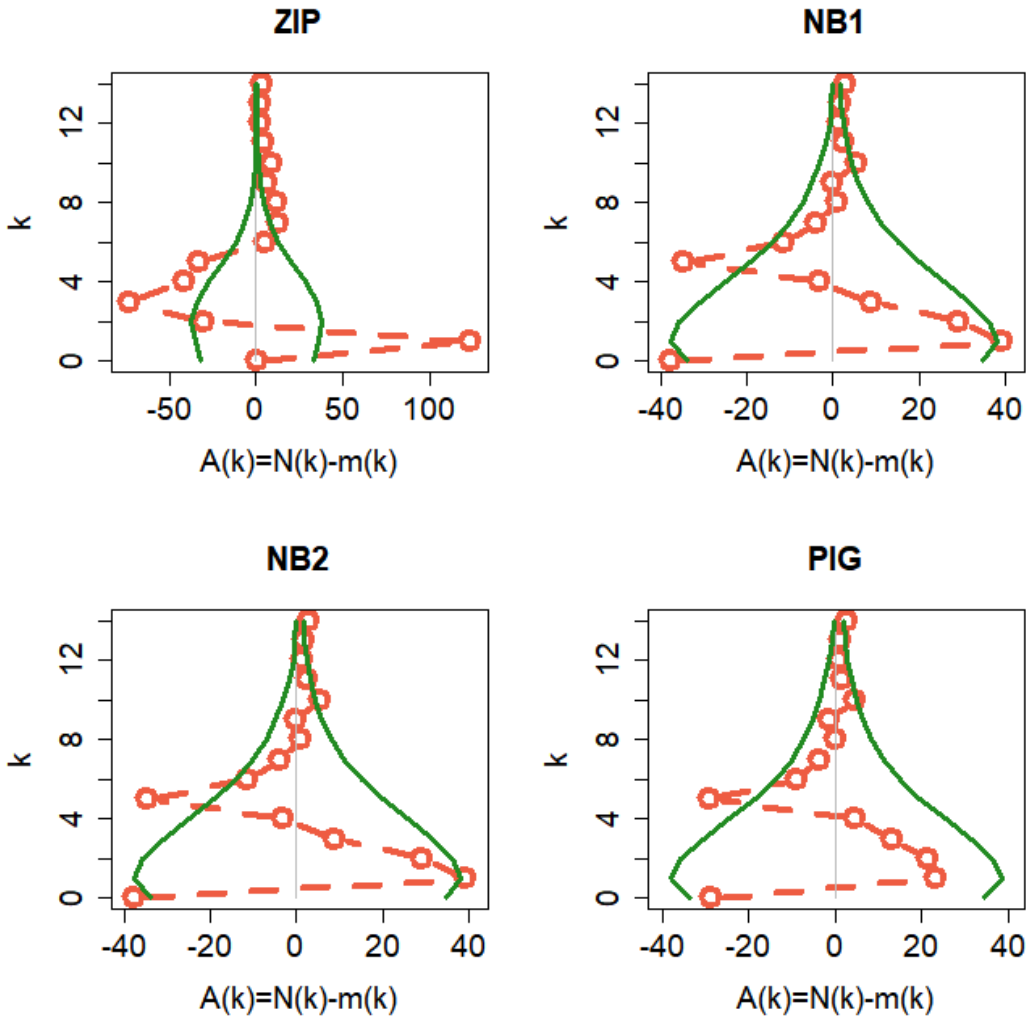
ZIP, NB, and PIG distributions

```
require(gamlss)

x.zip <- gamlss(x~1, sigma.fo=~1, family="ZIP")
x.nb1 <- gamlss(x~1, sigma.fo=~1, family="NBII")
x.nb2 <- gamlss(x~1, sigma.fo=~1, family="NBI")
x.pig <- gamlss(x~1, sigma.fo=~1, family="PIG")

par(mfrow=c(2,2))
quantbandplot(x, muhat= x.zip$mu.fv, disphat=x.zip$sigma.fv,
  dist="zip", main="ZIP")
quantbandplot(x, muhat= x.nb1$mu.fv, disphat=x.nb1$sigma.fv,
  dist="NBI", main="NB1")
quantbandplot(x, muhat= x.nb2$mu.fv, disphat=x.nb2$sigma.fv,
  dist="NBII", main="NB2")
quantbandplot(x, muhat= x.pig$mu.fv, disphat=x.pig$sigma.fv,
  dist="pig", main="PIG")
```

ZIP, NB, and PIG distributions



Conclusion

We have presented an efficient diagrammatic tool to detect the presence of deflation or inflation of any digit in count data (regression) models.

For each count k, bounds are constructed as quantiles of the Poisson-Binomial distribution.

Question not discussed today: How exactly to compute such quantiles? Traditional quantiles, as produced by `poibin`, can behave infavorably for discrete distributions; we therefore advocate, and use in `quantbandplot`, 'mid-quantiles' (Wilson & Einbeck, 2021).

References

Dietz E & Böhning D (2000). On estimation of the Poisson parameter in zero–modified Poisson models. *Computational Statistics & Data Analysis* **34**, 547–548.

Fernández-Fontelo A et al (2018). An exact goodness-of-fit test based on the occupancy problems to study zero-inflation in biological dosimetry data. *Radiation Protection Dosimetry* **179**, 317–326.

Oliveira M et al (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. Biometrical Journal **58**, 259-279.

Ridout M & Demétrio C (1992). Generalized linear models for positive count data. *Revista de Matemática e Estatística* **10**, 139–148.

References

Wilson P & Einbeck J (2019). A new and intuitive test for zero modification. *Statistical Modelling* **19**, 341–361.

Wilson P & Einbeck J (2021). A graphical tool for assessing the suitability of a count regression model. *Austrian Journal of Statistics* **50**, 1–23.