Too many zeros? Not enough zeros? How to assess through inferential and graphical methods

Part II: A graphical tool for assessing the suitability of a count regression model

Jochen Einbeck[1] Paul Wilson[2]

[1]Durham University

[2]University of Wolverhampton

Wolverhampton, 9 May 2023

Problem

- 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?

Problem

- 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.

Problem

- 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 $y_1, \ldots, y_n$.
- Idea: check whether, for each count $k = 0, 1, 2, \ldots$, the number $N(k)$ is 'plausible' under the distribution $F(\hat{\mu}_i, \hat{\theta}_i)$.

Poisson-Binomial distribution

- The random variable $N(k)$ follows a Poisson–Binomial distribution with parameters $p_1(k), \ldots, p_n(k)$, where

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

  is the probability of observing the count k under covariate x_i and model F (Chen and Liu, 1997).
- The $p_i(k)$ can be estimated by $\hat{p}_i(k) = P(k|\hat{\mu}_i, \hat{\theta}_i)$ from the fitted model.

Poisson-Binomial distribution

- The random variable $N(k)$ follows a Poisson–Binomial distribution with parameters $p_1(k), \ldots, p_n(k)$, where

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

  is the probability of observing the count k under covariate x_i and model F (Chen and Liu, 1997).

- The $p_i(k)$ can be estimated by $\hat{p}_i(k) = P(k|\hat{\mu}_i, \hat{\theta}_i)$ from the fitted model.
  - For instance, in the special case that $F(\mu_i, \theta_i)$ corresponds to Pois(μ_i), one has $\hat{p}_i(k) = \exp(-\hat{\mu}_i)\hat{\mu}_i^k/k!$.
  - This scenario was discussed in the previous talk with focus on the case $k = 0$.
  - This talk generalizes those ideas to general k and F and proposes a generic diagrammatic tool.

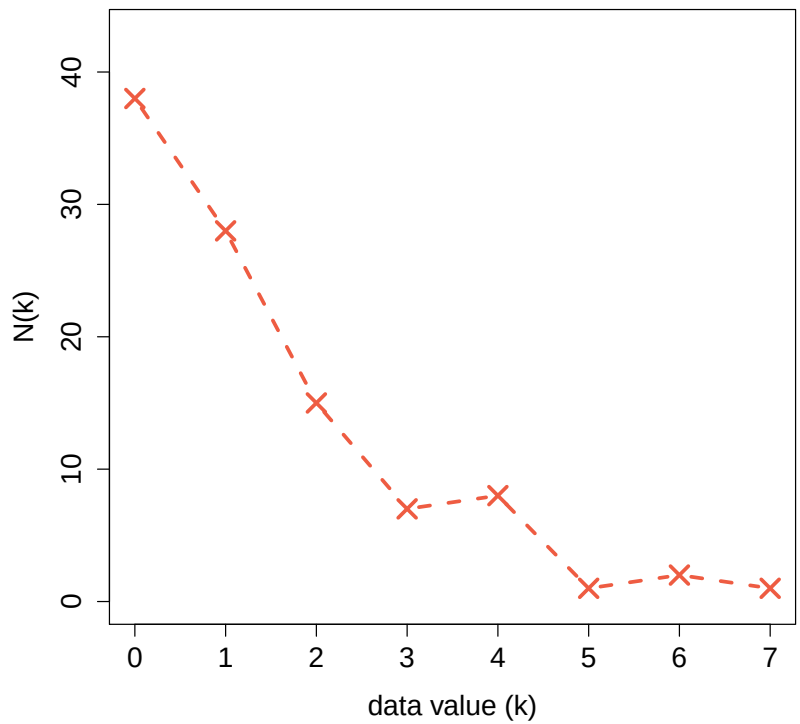
Plausibility intervals for N(k)

- 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 using the R package `poibin` (Hong, 2013).
- 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.

Example: simulated data

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

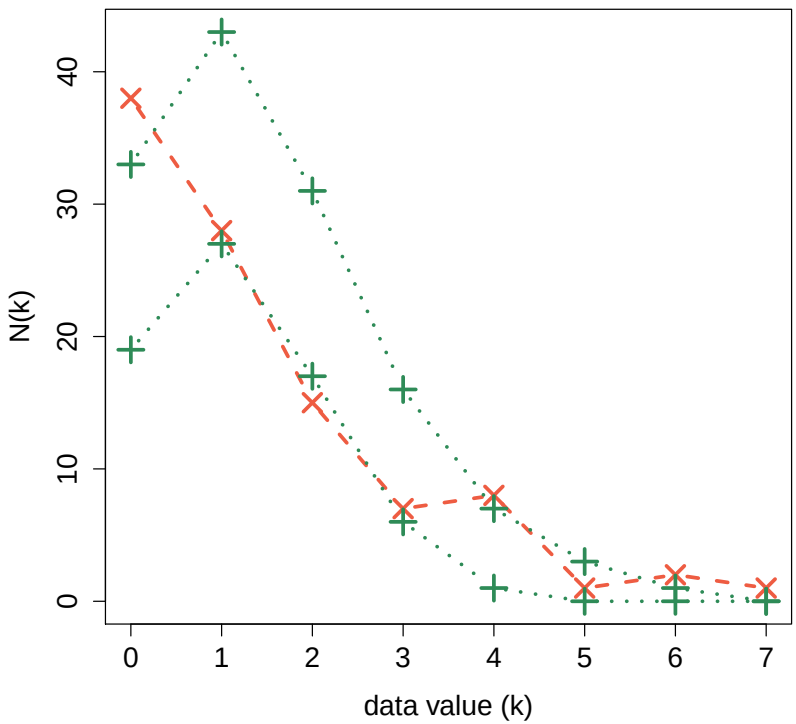
k	$N(k)$
0	38
1	28
2	15
3	7
4	8
5	1
6	2
7	1



Example: simulated data

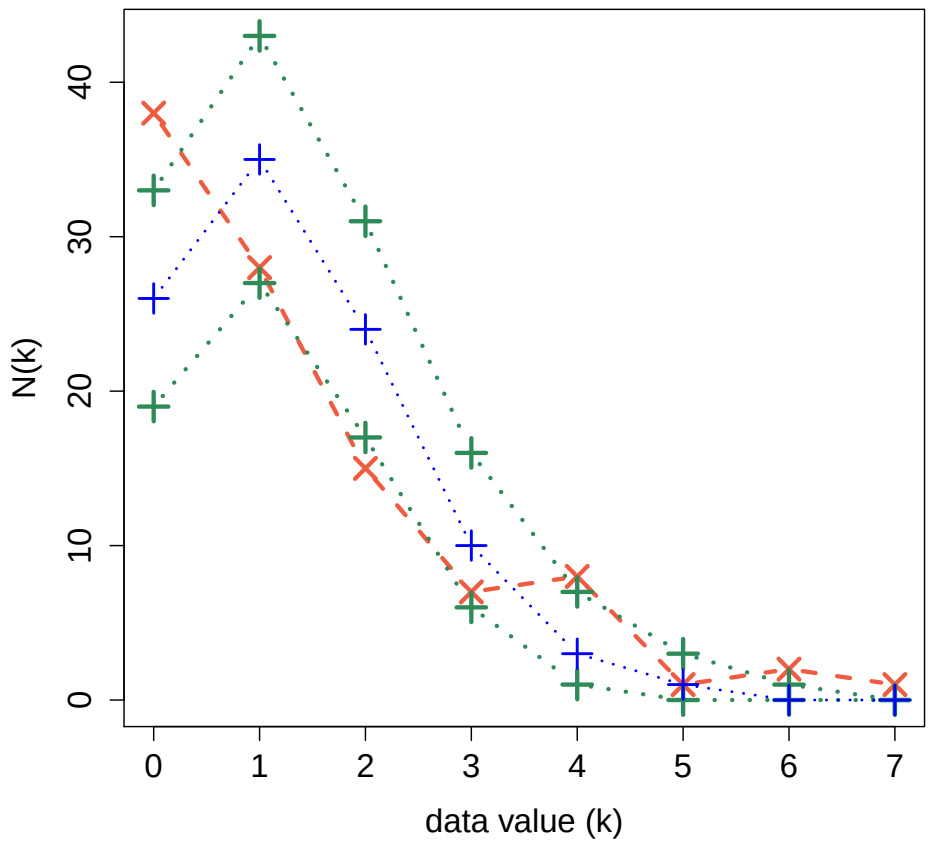
- Consider $F(\mu) \sim \text{Pois}(\mu)$ with $\hat{\mu} = \bar{y}$, so $\hat{p}(k) = e^{-\bar{y}} \frac{\bar{y}^k}{k!}$.

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 $N(k)$.



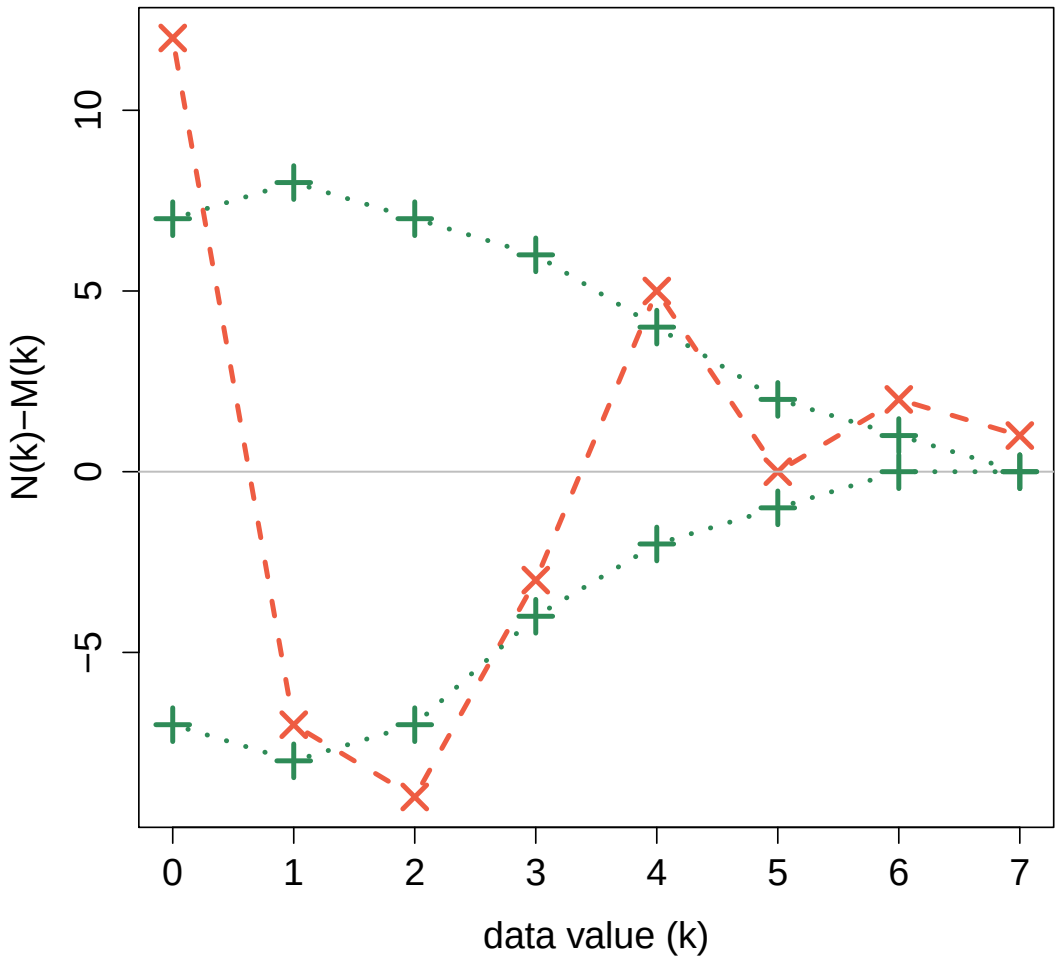
Median-adjustment

- 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 $N(k)$.

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

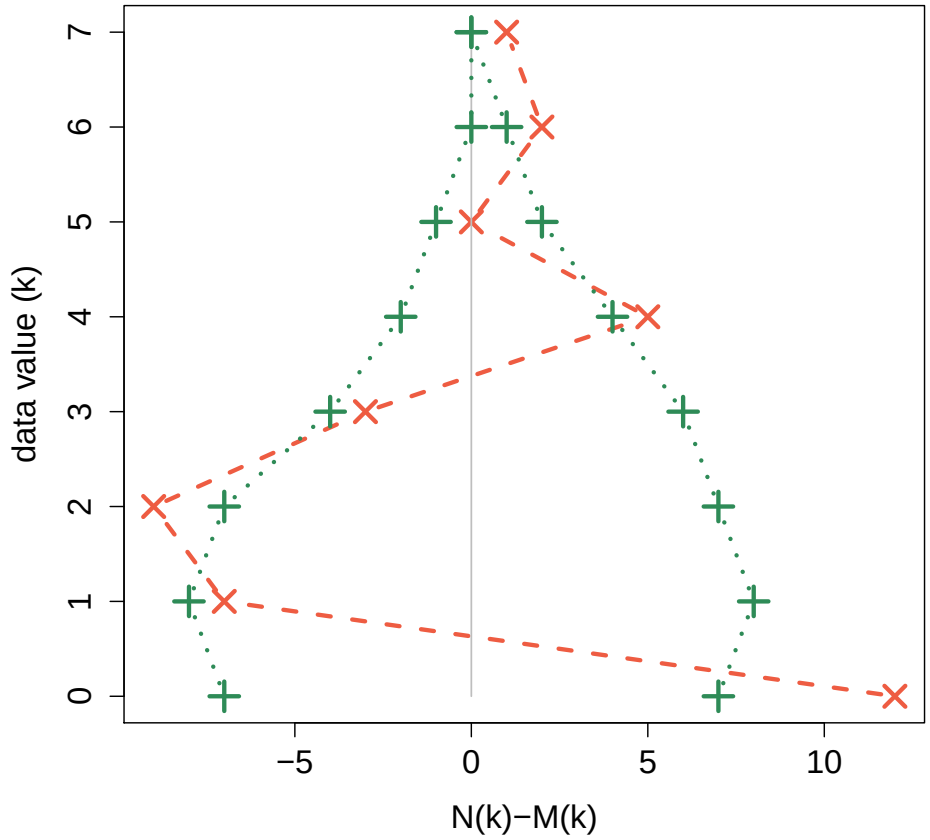
Median–adjusted bounds

- Diagnostic plot for the accuracy of the Poisson assumption.



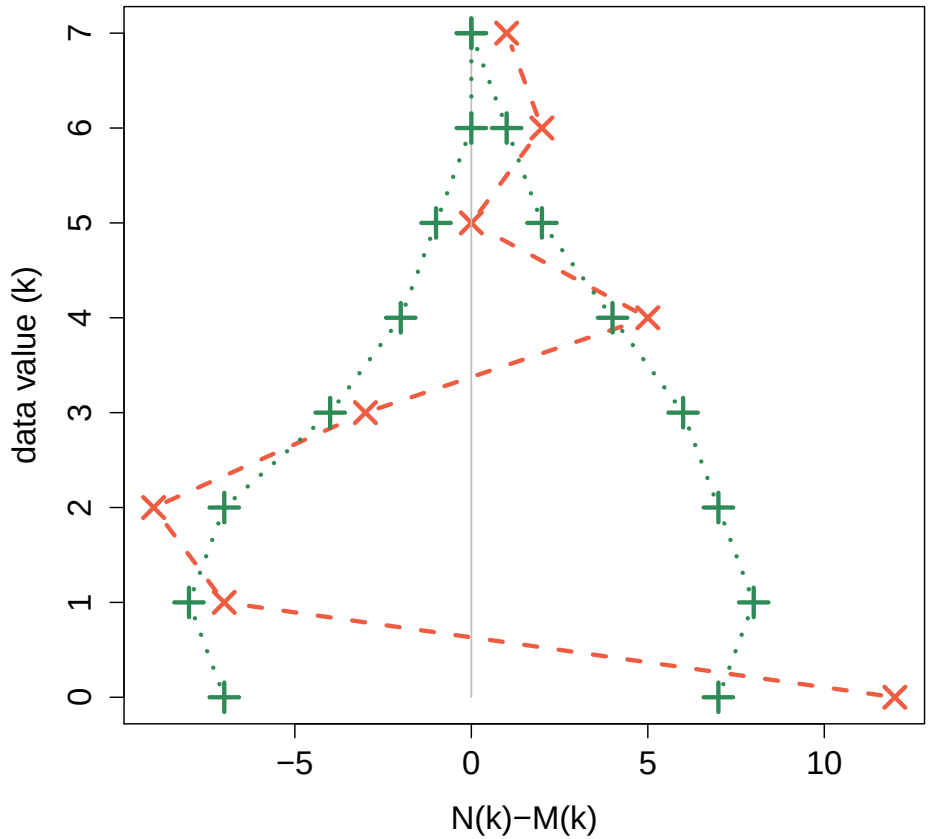
Median–adjusted bounds: Variant

- Exchange horizontal and vertical axis:



Median–adjusted bounds: Variant

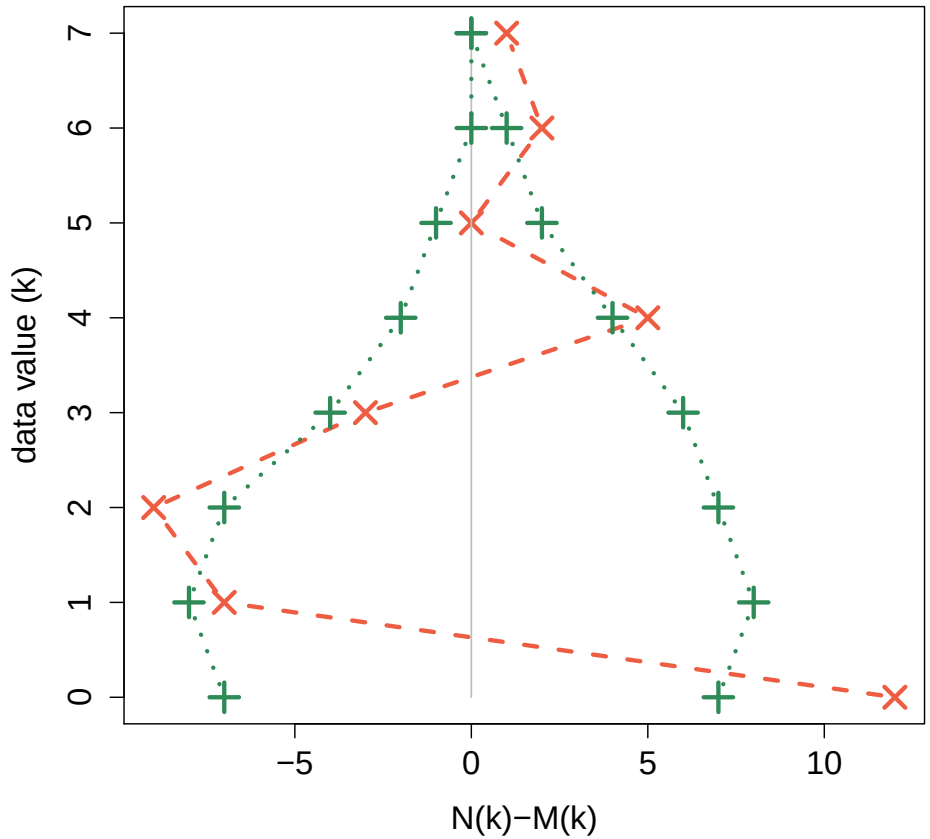
- Exchange horizontal and vertical axis:



- ‘Christmas tree diagram’/ ‘Quantile band plot’ (Wilson & Einbeck, 2021)

Median–adjusted bounds: Variant

- Exchange horizontal and vertical axis:

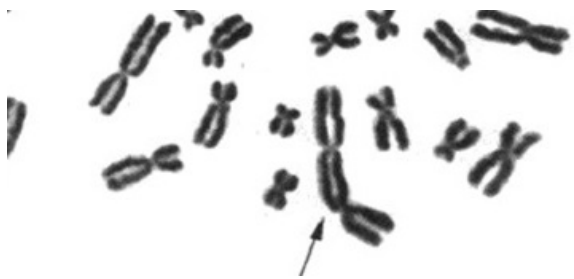


- 'Christmas tree diagram'/ 'Quantile band plot' (Wilson & Einbeck, 2021)
- Adequate models have the 'decoration' inside the tree.

Example: Biodosimetry 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
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



Example: Biodosimetry 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.

	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	$n = 7200$

Modelling of biodosimetry data

- These are $n = 7200$ observations of the type (dose_i, y_i), with y_i being a count in $0, \ldots, 8$.
- X–rays are sparsely ionizing — the literature suggests a quadratic dose model in this case.

Modelling of biodosimetry data

- These are $n = 7200$ observations of the type (dose_i, y_i), with y_i being a count in $0, \ldots, 8$.
- X–rays are sparsely ionizing — the literature suggests a quadratic dose model in this case.
- Link function:
  - Cytogeneticists prefer identity link.
  - Being among Statisticians (?), I will use the log link.
- Response (count) distribution:
  - It is widely accepted that the number of dicentrics in irradiated blood samples is Poisson distributed.
  - However, under partial body exposure, we would expect a deviation from this assumption...

Modelling of biodosimetry data

- These are $n = 7200$ observations of the type (dose_i, y_i), with y_i being a count in $0, \ldots, 8$.
- X–rays are sparsely ionizing — the literature suggests a quadratic dose model in this case.
- Link function:
  - Cytogeneticists prefer identity link.
  - Being among Statisticians (?), I will use the log link.
- Response (count) distribution:
  - It is widely accepted that the number of dicentrics in irradiated blood samples is Poisson distributed.
  - However, under partial body exposure, we would expect a deviation from this assumption...

Modelling of biodosimetry data

- These are $n = 7200$ observations of the type (dose_i, y_i), with y_i being a count in $0, \ldots, 8$.
- X–rays are sparsely ionizing — the literature suggests a quadratic dose model in this case.
- Link function:
  - Cytogeneticists prefer identity link.
  - Being among Statisticians (?), I will use the log link.
- Response (count) distribution:
  - It is widely accepted that the number of dicentrics in irradiated blood samples is Poisson distributed.
  - However, under partial body exposure, we would expect a deviation from this assumption...
- Consider the initial model $y_i|\text{dose}_i \approx \text{Pois}(\mu_i)$ with

$$\mu_i \equiv E(y_i|\text{dose}_i) = \exp\left(\beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{dose}_i^2\right)$$

Diagnostics for Biodosimetry data

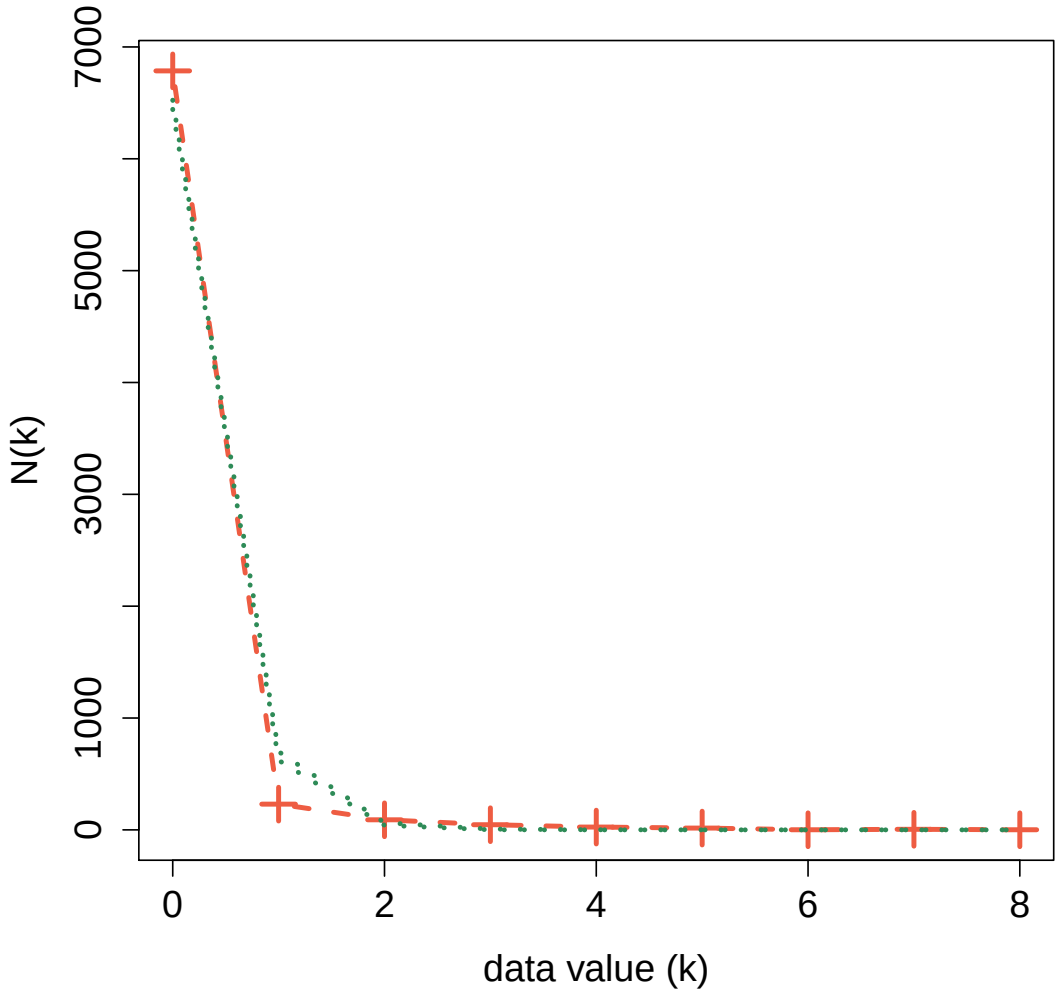
Do the same as before. That is,

- 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

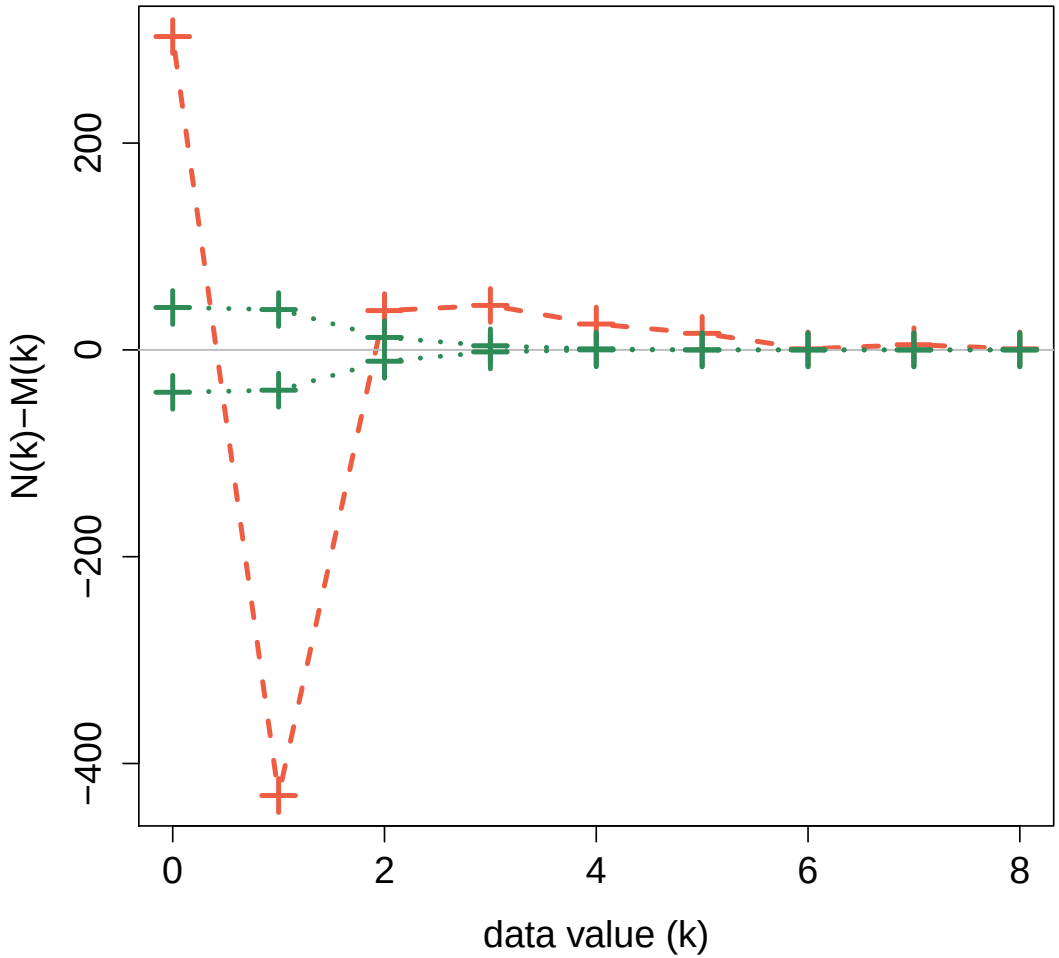
- ...without median– adjustment:



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

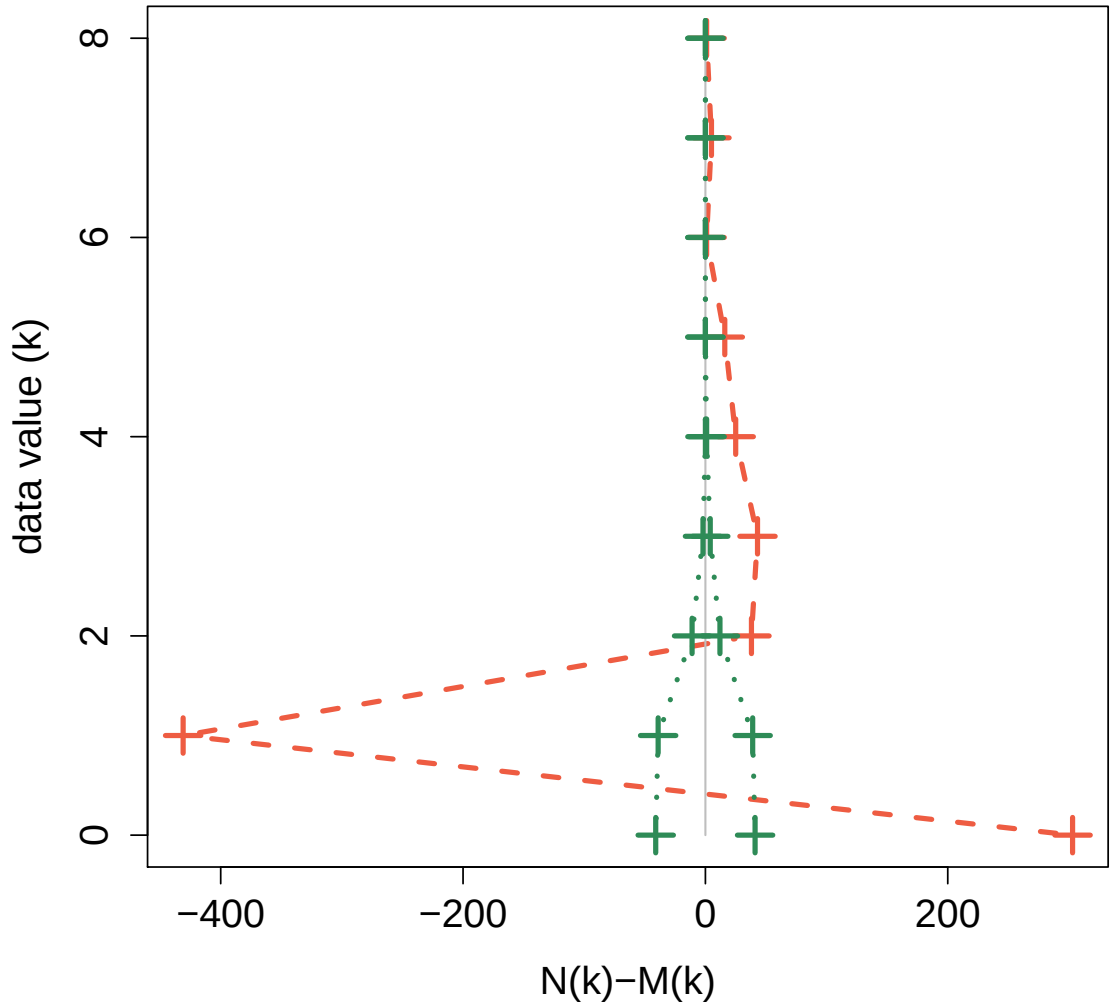
Diagnostics for biodosimetry data

- ...with median– adjustment:

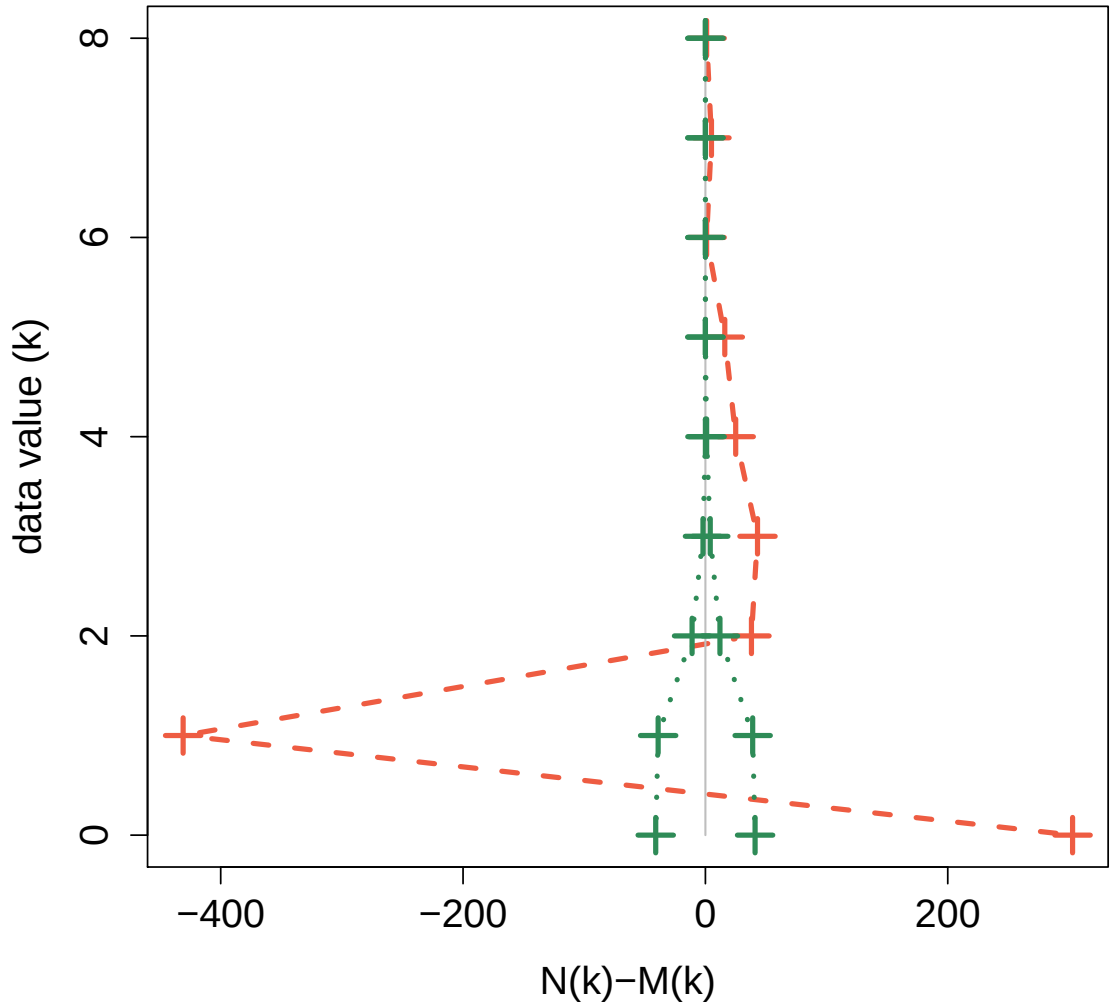


- much better!

Christmas tree diagram: Poisson hypothesis

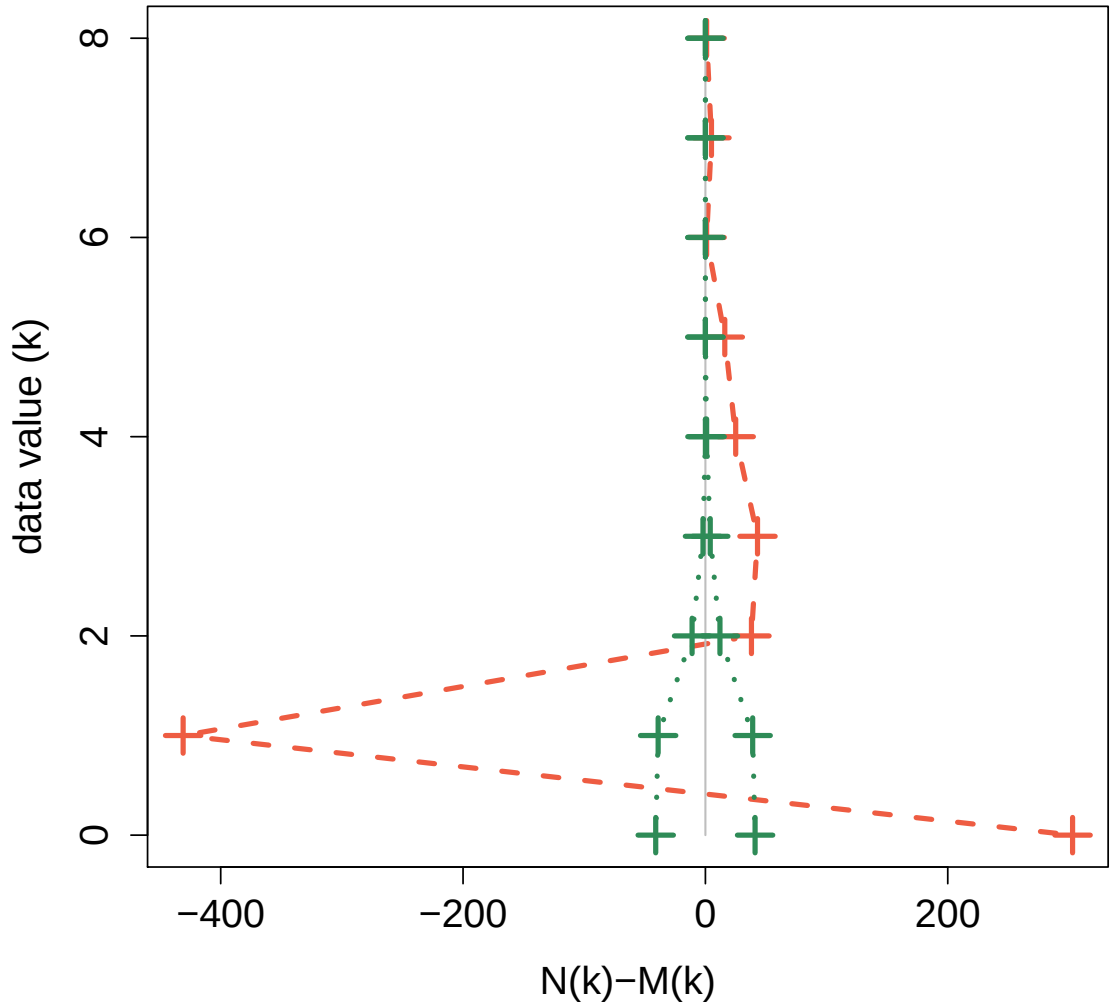


Christmas tree diagram: Poisson hypothesis



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

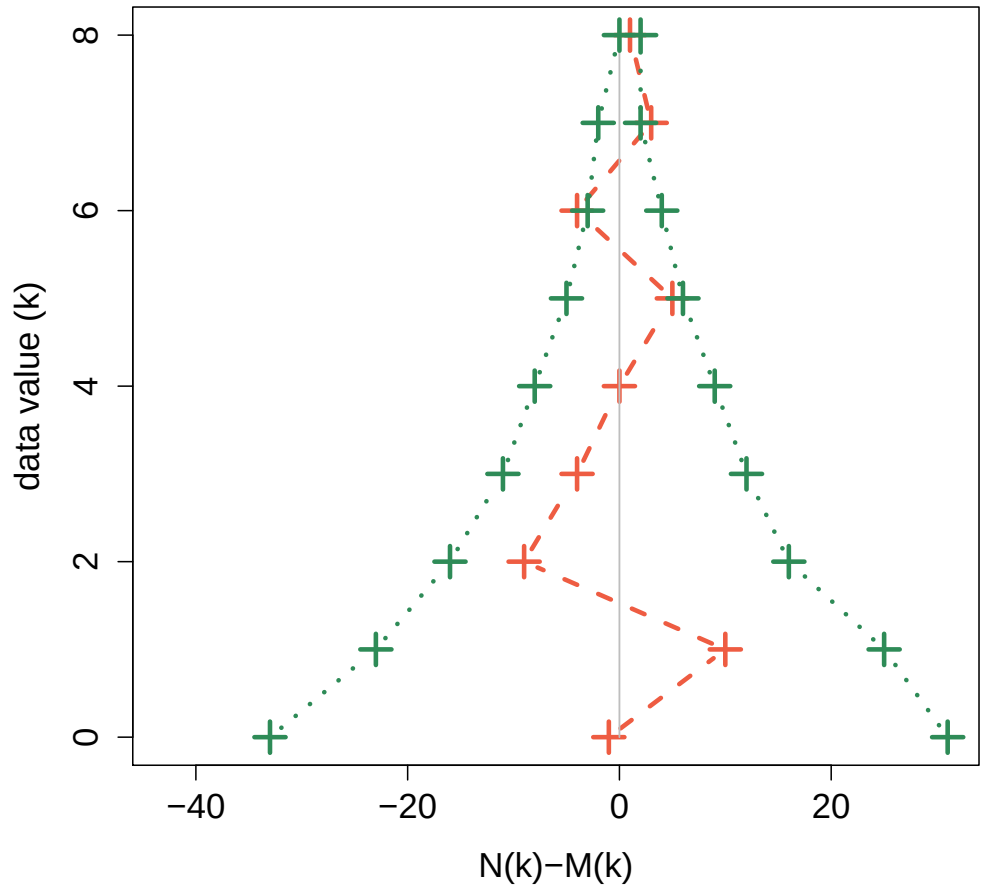
Christmas tree diagram: Poisson hypothesis



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

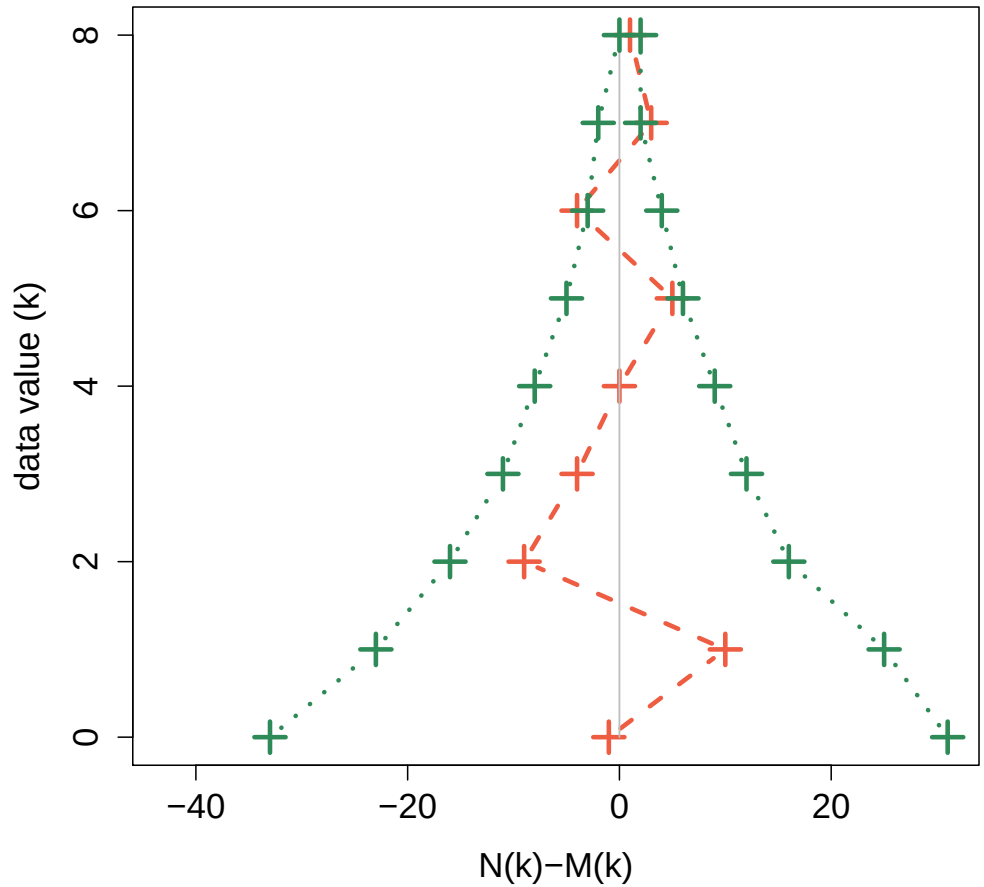
Christmas tree diagram: ZIP hypothesis

- 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.



Christmas tree diagram: ZIP hypothesis

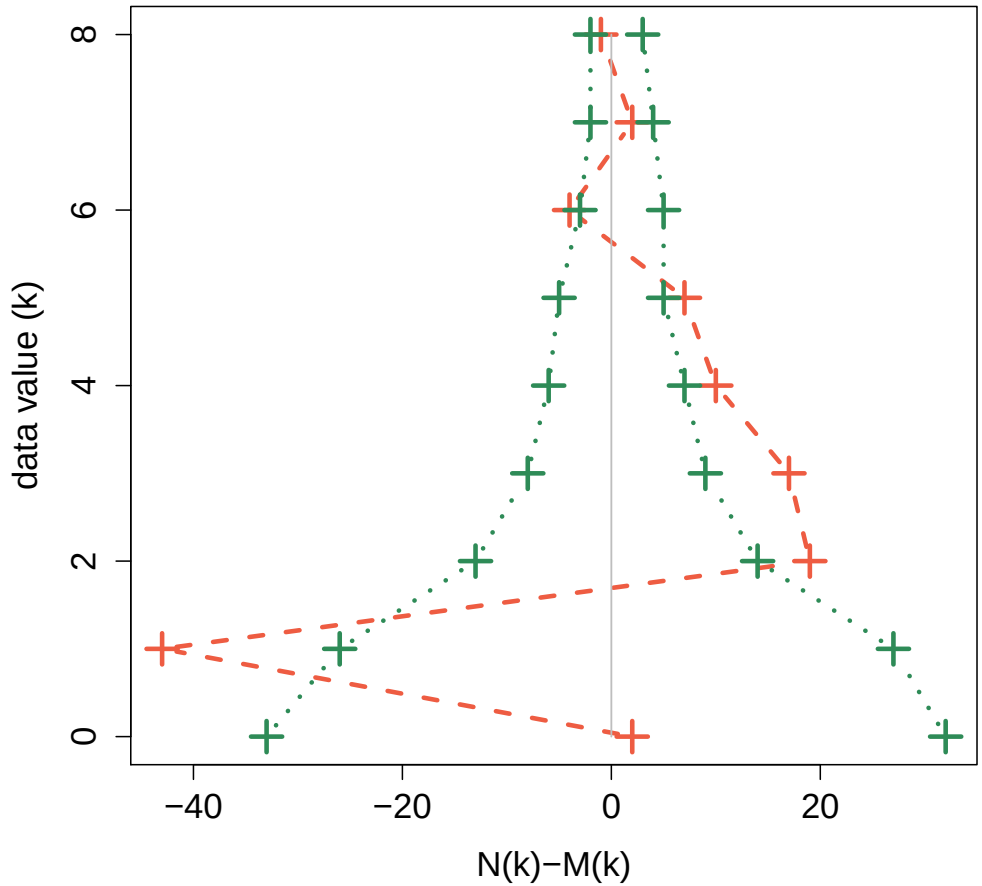
- Do all the same as before, but now compute $\hat{\mu}_i$, $\hat{\theta}_i$, and $\hat{p}_i(k)$, using the zero–inflated Poisson (ZIP) model as the hypothesized model.



- indicates a good fit.

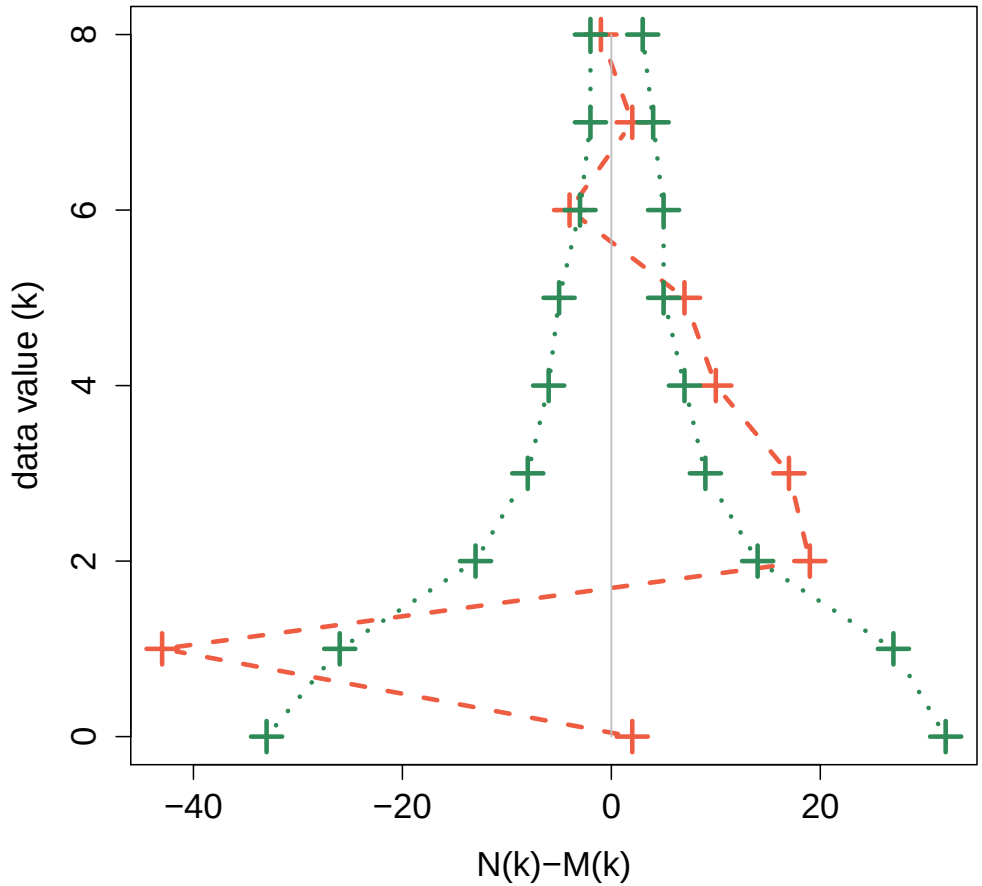
Christmas tree diagram: NB hypothesis

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



Christmas tree diagram: NB hypothesis

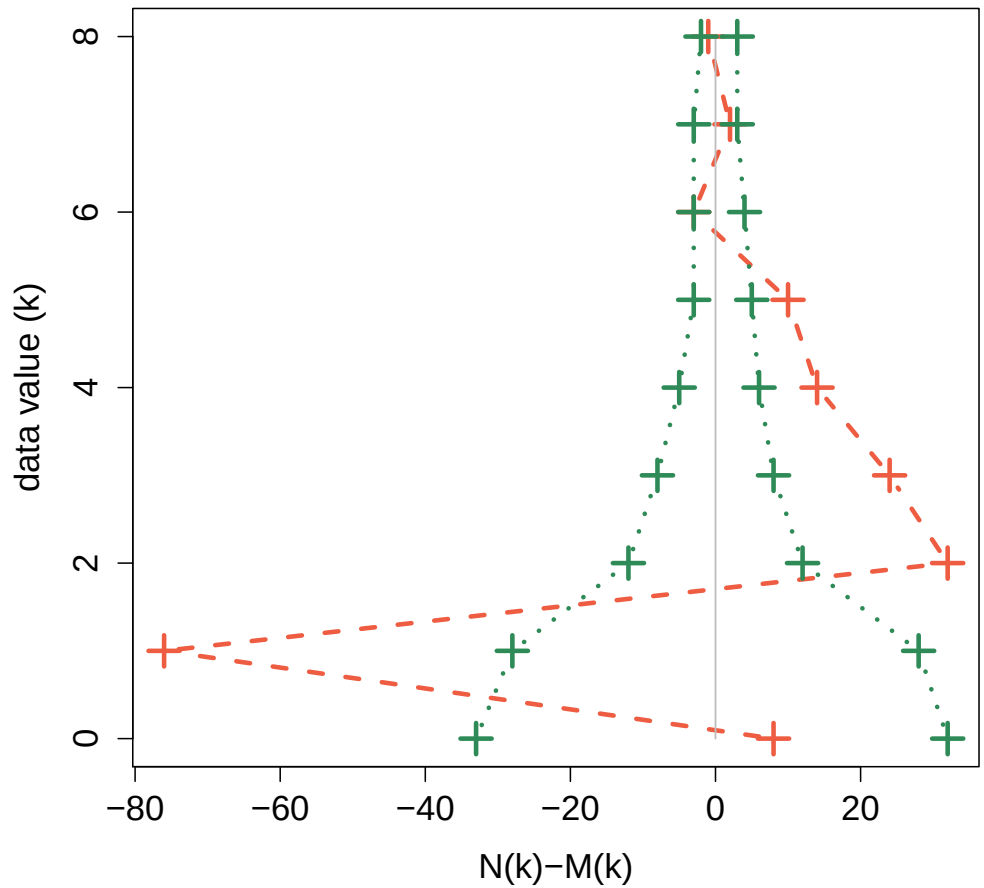
- Repeat the procedure using the negative Binomial model as the hypothesized model.



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

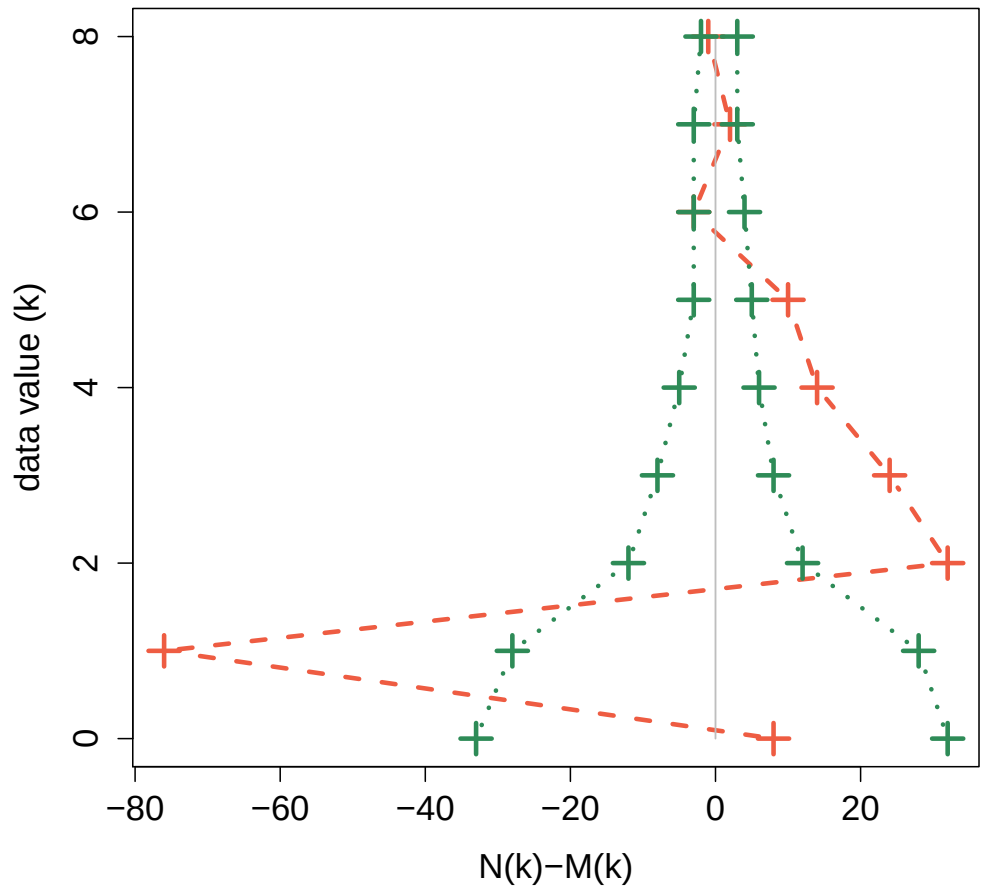
Christmas tree diagram: PIG hypothesis

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



Christmas tree diagram: PIG hypothesis

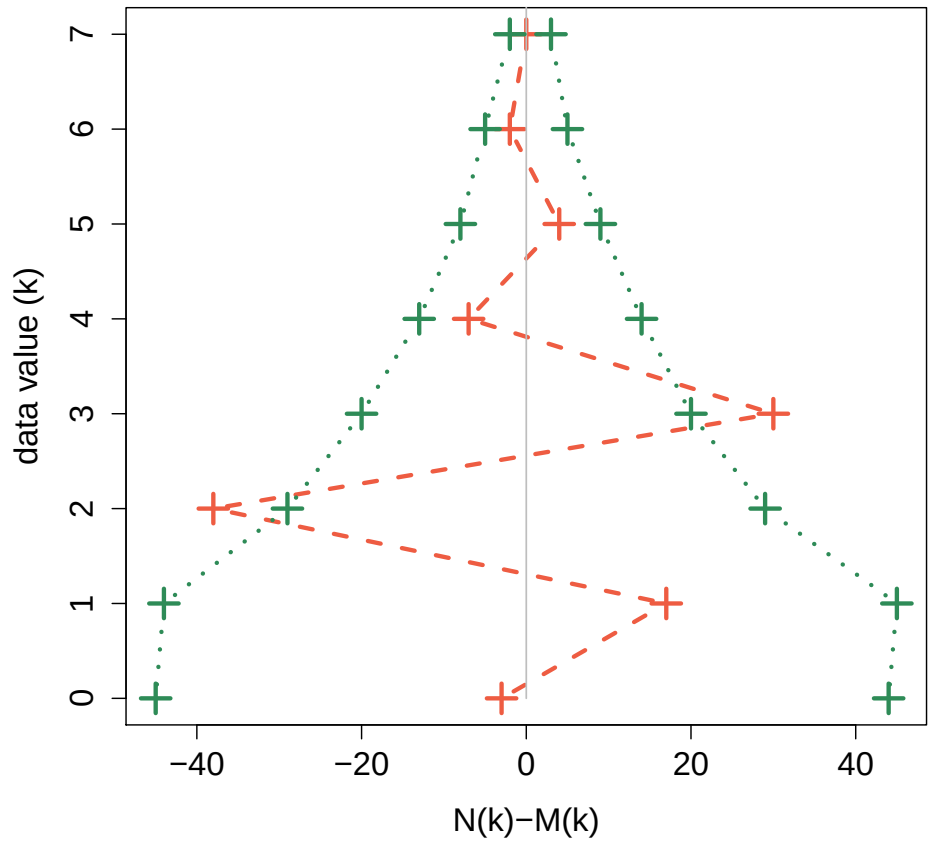
- Repeat the procedure using the Poisson inverse Gaussian (PIG) model as the hypothesized model.



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

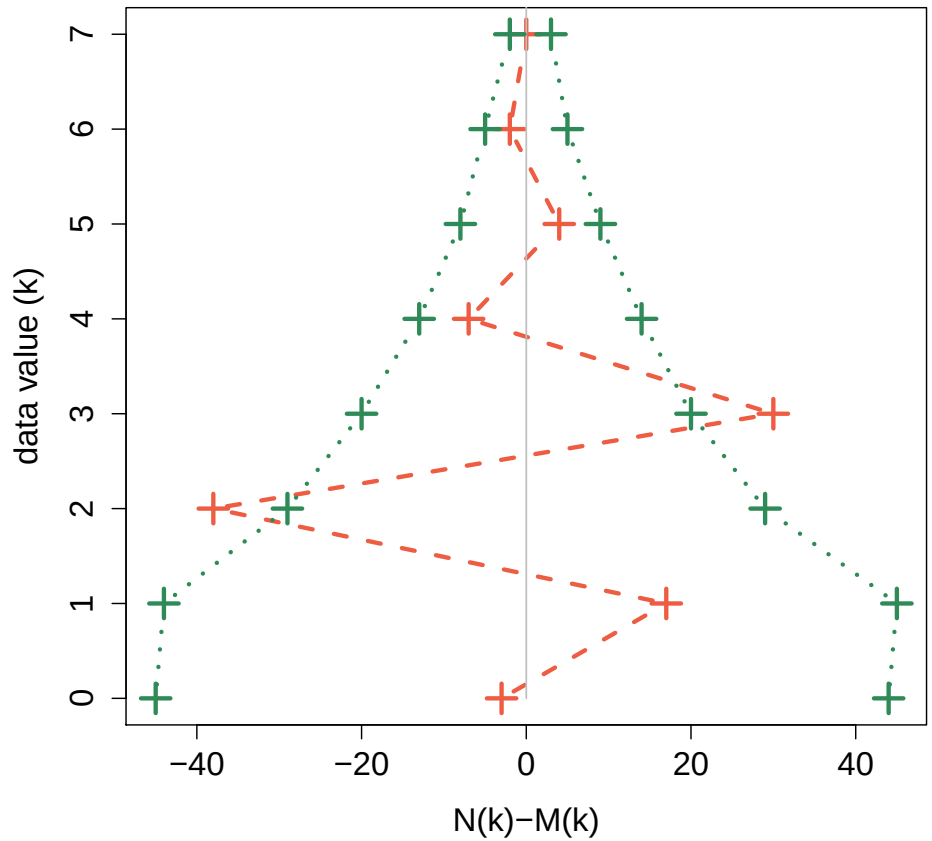
Alternative data set: Whole body exposure

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



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 tree covers ten possible counts, at a significance level of 0.1 one would expect one piece of decoration to fall outside the tree purely by chance.

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 tree covers ten possible counts, at a significance level of 0.1 one would expect one piece of decoration to fall outside the tree purely by chance.
- One could adjust this through a Bonferroni correction etc.
- However, we do believe that the corresponding inflated boundaries would be rather meaningless.

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 tree covers ten possible counts, at a significance level of 0.1 one would expect one piece of decoration to fall outside the tree purely by chance.
- One could adjust this through a Bonferroni correction etc.
- However, we do believe that the corresponding inflated boundaries would be rather meaningless.
- 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 $\chi^2_{1,0.95} = 3.84$:

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

- Confirms that Poisson is adequate for whole body exposure but inadequate for partial body exposure (Oliveira et al, 2016).

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 $\chi^2_{1,0.95} = 3.84$:

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

- Confirms that Poisson is adequate for whole body exposure but inadequate for partial body exposure (Oliveira et al, 2016).
- ...but the score test does not tell us whether it's the zero's which cause the problem, nor whether the data are zero–inflated or –deflated!

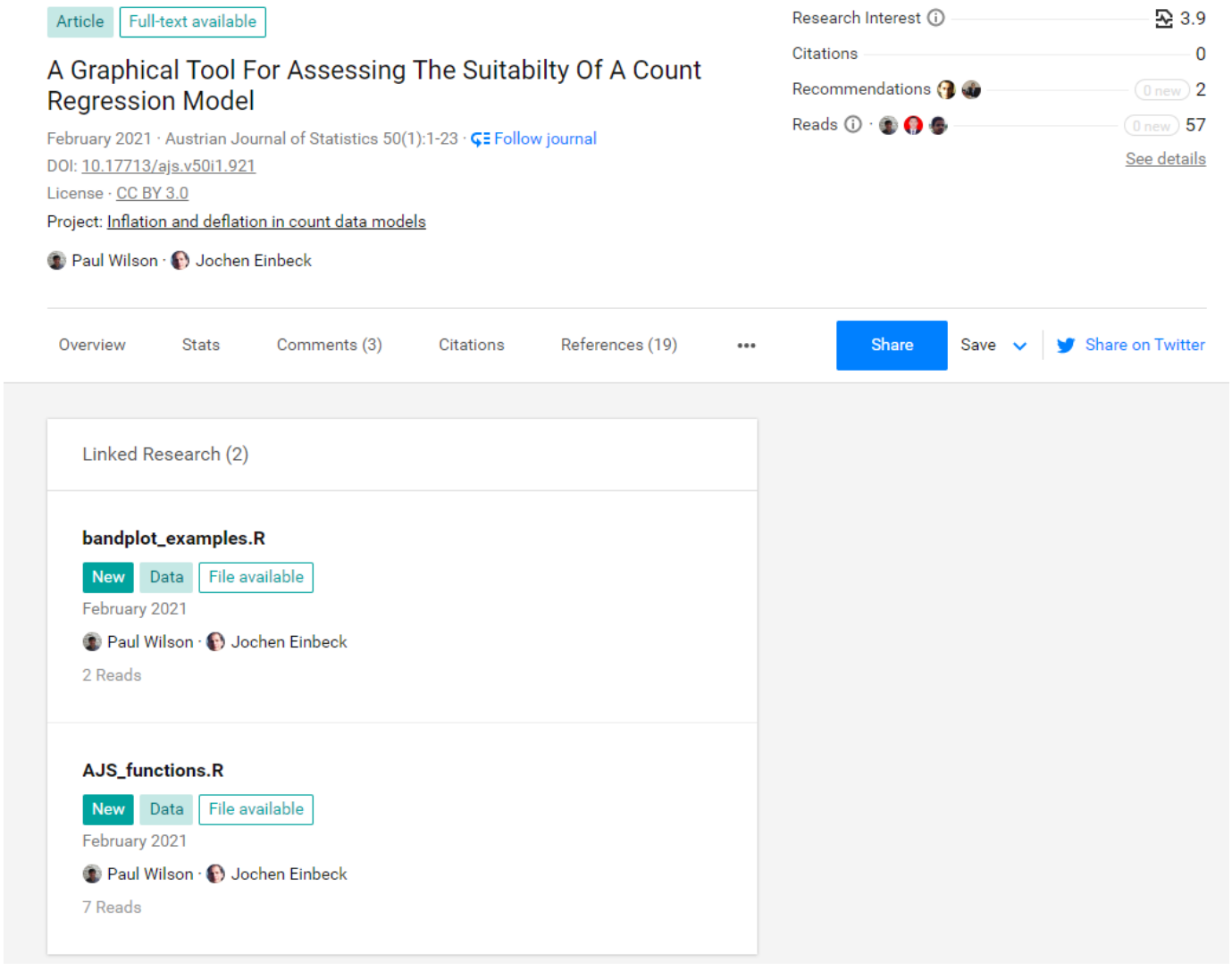
Conclusion

- We have provided a simple diagrammatic tool to assess the adequacy of any given count data model.
- For each count k, bounds are constructed as quantiles of the Poisson-Binomial distribution.
- How exactly to compute the quantiles? Traditional quantiles, as produced by `poibin`, can behave infavorably for discrete distributions; we therefore advocate the use of 'mid-quantiles' (Wilson & Einbeck, 2021).
- Estimation of model parameters when the model is inadequate can possibly be tricky!

Conclusion

- We have provided a simple diagrammatic tool to assess the adequacy of any given count data model.
- For each count k, bounds are constructed as quantiles of the Poisson-Binomial distribution.
- How exactly to compute the quantiles? Traditional quantiles, as produced by `poibin`, can behave infavorably for discrete distributions; we therefore advocate the use of 'mid-quantiles' (Wilson & Einbeck, 2021).
- Estimation of model parameters when the model is inadequate can possibly be tricky!
  - For the work carried out in this talk, all parameters have been estimated under the hypothesized model.
  - In the special case of $F \sim$ Pois and $k = 0$, an improved mean estimator $\hat{\mu}_i$ has been proposed in the previous talk.
  - More work required for the more general case of an arbitrary count/distribution.

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

References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Hong, Y. (2013). poibin: The Poisson Binomial Distribution. R package version 1.2. https://CRAN.R-project.org/package=poibin

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

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