

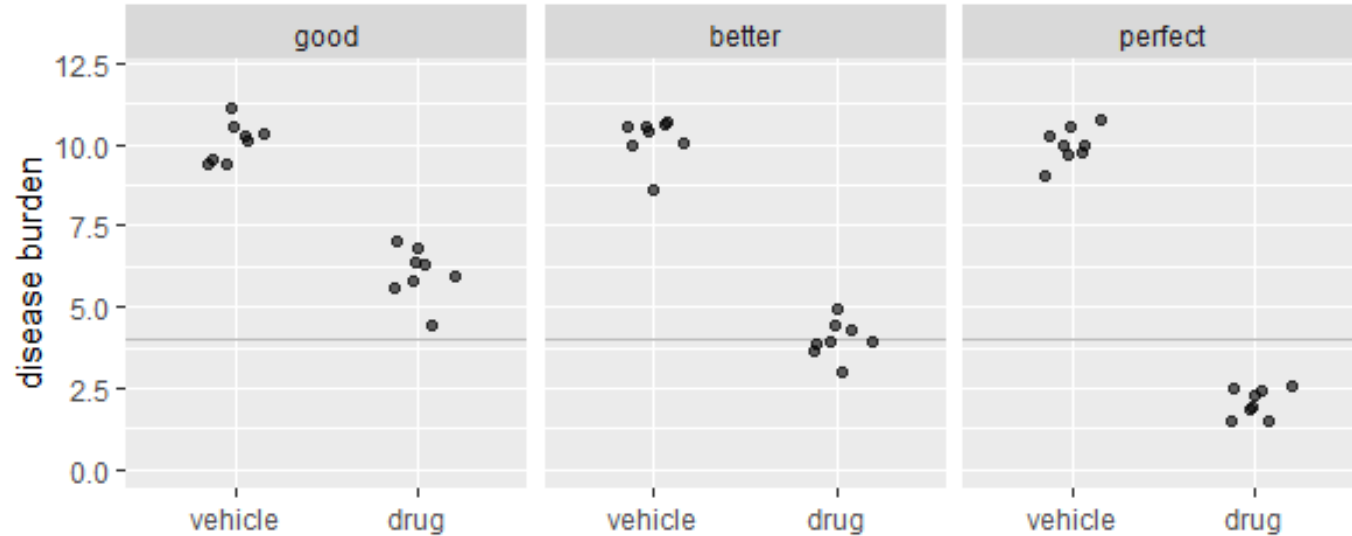


Analysing data where groups are completely censored

David R Willé and Joe Watson, GSK R&D, Stevenage

complete censorship

our motivating question

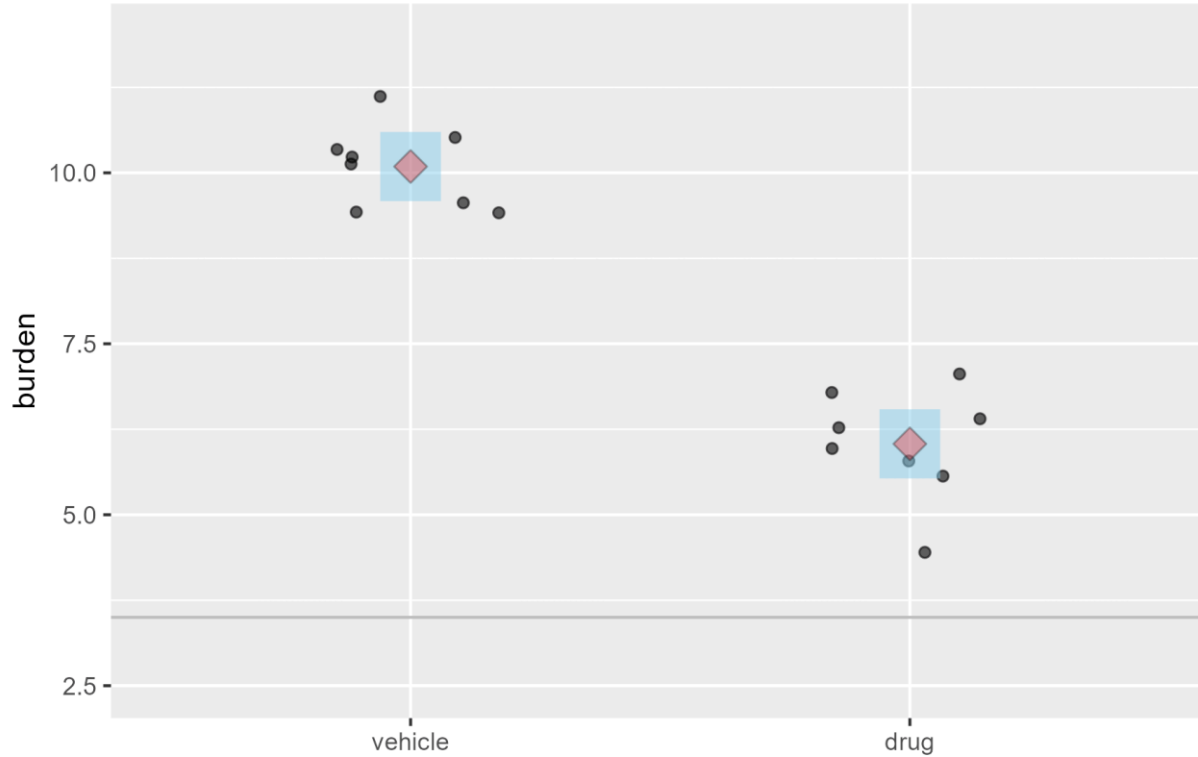


- we have often discussed what happens when data falls below* a limit of detection or quantification
- but what happens when *all the data* in in a group of interest is censored?
- what happens if our drug really really works?

* presentation focuses on left censorship, but the methods and approach are fully generalisable.

the nature of the problem ...

the power of tobit* methods



- standard tobit-type censorship methods work “surprising” well when applied to censored data
- but what happens when things go wrong?
- is this really correct – what happened in the completely censored case?

The red diamond indicates the estimated means, the blue boxes, 95% confidence intervals and the rising line, the lower censorship level. The data remains the same despite the jitter. The stability of the estimates raises questions about distributional assumptions. Estimates computed with via `mgcv::gam()` with and maximum likelihood. Underlying error modelled as normal.

* one of the most common frequentist approach. See next slide for details.

the tobit trick ...

replacing pdf's by cdf's below the censorship limit

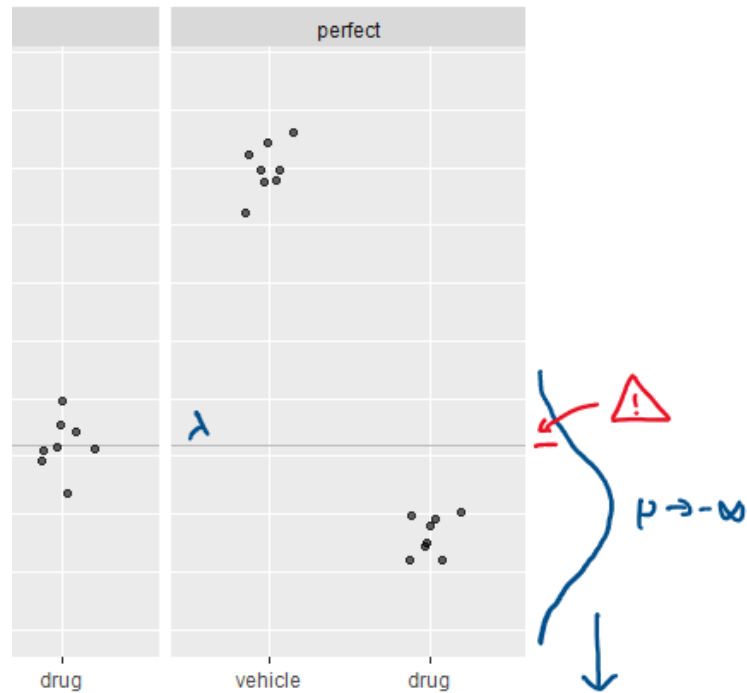
... for any θ and lower censorship limit λ , the likelihood contribution of a single point response x_i is

$$L(x_i|\theta) = \begin{cases} \text{pdf}(x_i|\theta) & : x_i > \lambda \\ \text{cdf}(\lambda|\theta) & : x_i \leq \lambda \end{cases}$$

censored data points are simply assumed to be anywhere, following the postulated distribution below the censorship limit. Similar constructs can be used for above and other censored information.

... with completely censored data

likelihood intuition



- maximum likelihood seeks to place its distribution to maximise the amount of explained variation and minimise its probability where the data is absent.
- for any postulated mean μ it can improve this by moving downwards, shifting the increasing quartile of the cdf further to the right of μ as it does so.
- continuing this in the limit* forces $\mu \rightarrow -\infty$
- all the weight of the distribution is now below the censorship limit λ giving a contributed likelihood of, up to a factor, one
- the spread of distribution is estimated solely from the completely observed data – we will return to this later

* a mean of $-\infty$ may make no sense biologically (unless the data was logged) and can arise only as an artifact of distribution used: a normal with infinitely extending tails.

more formally

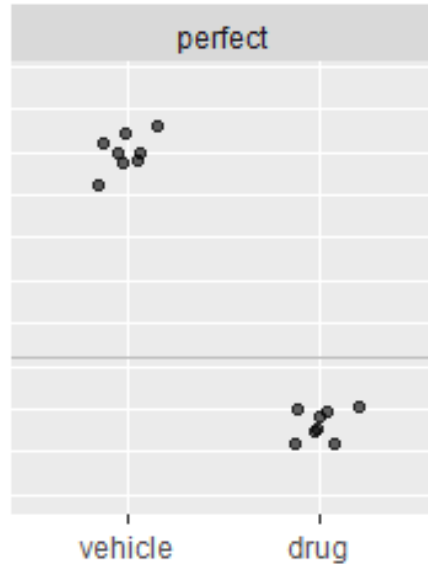
for the censored points, as the mean moves to $-\infty$

$$\lim_{\mu \rightarrow -\infty} \int_{-\infty}^{\lambda - \mu} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{z - \mu}{\sigma} \right)^2} dz = \lim_{\mu \rightarrow -\infty} \int_{-\infty}^{+\infty} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{z - \mu}{\sigma} \right)^2} dz = 1$$

the RHS of the cdf interval relative to μ becomes more extreme, and the distribution shifts to the left. In the limit $\mu \rightarrow -\infty$ this is infinitely so, $\lambda - \mu \rightarrow +\infty$, giving the expected integral probability of one.

Question:

comparing two groups?



- if one of the groups is completely censored, we can no-longer estimate the difference (only one finite mean) but is any parametric inference still possible?
- a p-value?
- a LR-test

this will be the focus
of this talk ...

the LR test

very informally

$$\lambda_{\text{LR}} = -2 \ln \frac{\sup_{\mu_1, \mu_2} L(D_1) L(D_2)}{\sup_{\mu_{12}} L(D_{12})} \rightarrow -2 \ln \frac{\sup_{\mu_1} L(D_1)}{\sup_{\mu_{12}} L(D_{12})}$$

- we compare the likelihood* contributions of the first groups $L(D_1)$ and $L(D_2)$ with separate mean parameters μ_1 and μ_2 to the contribution of their combination $L(D_{12})$ with a single shared mean estimate μ_{12}
- but note that as our second mean $\mu_2 \rightarrow -\infty$, the top second data's term $L(D_2)$ goes to one giving the simplified equation on the right
- this is the basis of a practical algorithm. The censored term simply vanishes

implementation

outline

- a simple package in R
- to compare two groups in any linear model where one is completely censored
 - run the model twice
 - first with the **censored data group completely removed**
 - then with its data included but **collapsed into the second group**
 - take the ML estimates from each and report a **LR-test**



practical issues: R does not accept factor variables with single levels. If these occur in reduced formulae, models have to be automatically re-written to remove them from all terms and interactions

implementation

details

```

censor=function(y,low=-Inf,high=+Inf){

  A=data.frame(y=y,low=low,high=high) %>%
    rowwise() %>%
    mutate(

      y1 = case_when(
        y <= low ~ low,
        y >= high ~ high,
        TRUE ~ y),

      y2 = case_when(
        y <= low ~ -Inf,
        y >= high ~ +Inf,
        TRUE ~ y)) %>%

    select(y1,y2) %>%
    as.matrix()

}

```

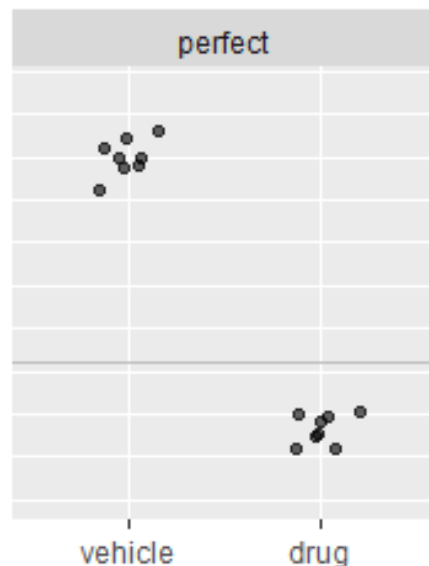
```
library(mgcv)
```

```
gam(censor(burden,lower=cutoff)~treat,
    data = data,
    family = cnorm,
    method = "ML")
```

for simplicity we currently use *mgcv*'s *gam* method for internally forward compatibility, but any ML approach or optimisation can be use.

Censorship is handled by the option (a function) family=cnorm.

applied to our starting dataset, our prototype code identified and applied an LR test*



```
dm = censoredData$new(data=M2, lower=4)
cm = censoredLRModel$new(value~treat,
                          data=dm)
cm$contrast("treat", "vehicle", "drug")
```

\$censorship_type

[1] "LR"

\$report

```
[1] ">>> comparing 'vehicle' vs 'drug' for response 'treat'"
[2] "censorship type: LR"
[3] "fitting a LR tobit based model"
[4] "fully censored groups: vehicle: FALSE, drug: TRUE"
[5] "non-informative groups: none"
[6] "no non-informative groups dropped"
[7] "using data from groups vehicle, drug"
[8] "top ML: -11.35150826578"
[9] "bottom ML: -17.4533339528423"
[10] "deviance change: 12.2036513741246"
```

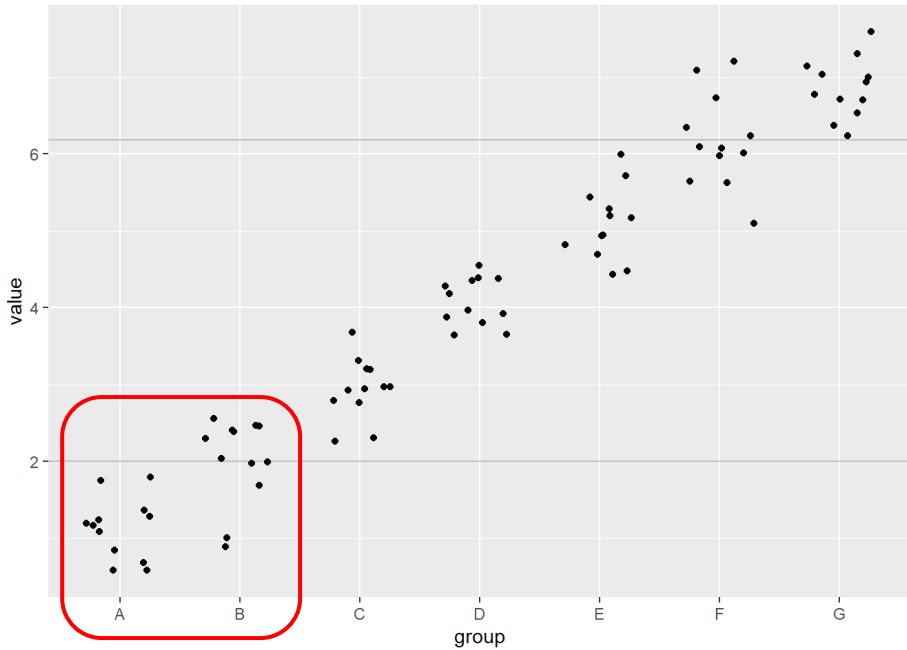
\$change.in.deviance

[1] 12.20365

\$p.value

[1] 0.0004769607

** the LR tobit model to a standard reduced LR model with censored information outside the completely censored group*



```
$censorship_type
```

```
[1] "LR-tobit"
```

```
$report
```

```
[1] ">>> comparing 'A' vs 'B' for response 'group'"
```

```
[2] "censorship type: LR-tobit"
```

```
[3] "fitting a LR tobit based model"
```

```
[4] "fully censored groups: A: TRUE, B: FALSE, C: FALSE, D: FALSE, E: FALSE, F: FALSE, G: TRUE"
```

```
[5] "non-informative groups: G"
```

```
[6] "levels 'G' of 'group' dropped as they contribute no information"
```

```
[7] "using data from groups A, B, C, D, E, F"
```

```
[8] "top ML: -78.9732372486709"
```

```
[9] "bottom ML: -83.8751359527323"
```

```
[10] "deviance change: 9.80379740812279"
```

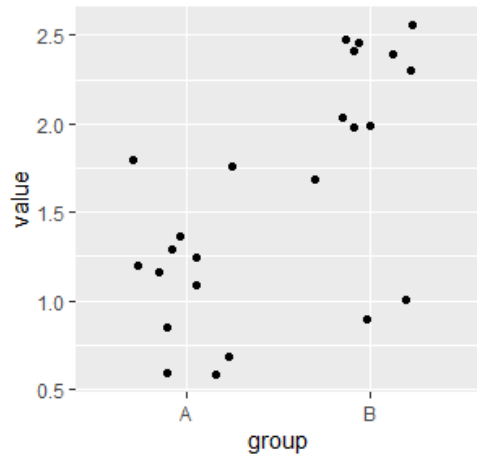
```
$change.in.deviance
```

```
[1] 9.803797
```

```
$p.value
```

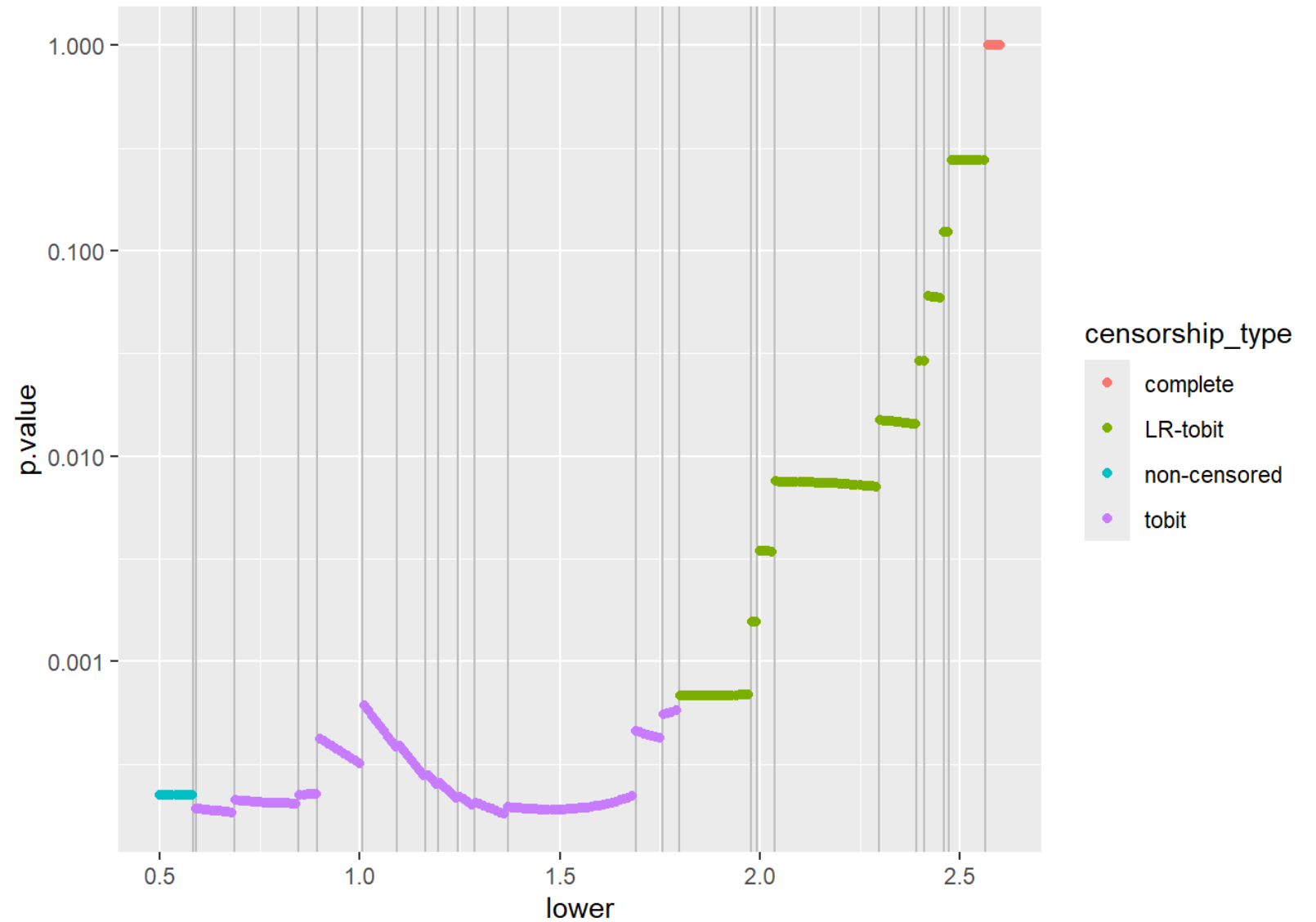
```
[1] 0.001741519
```

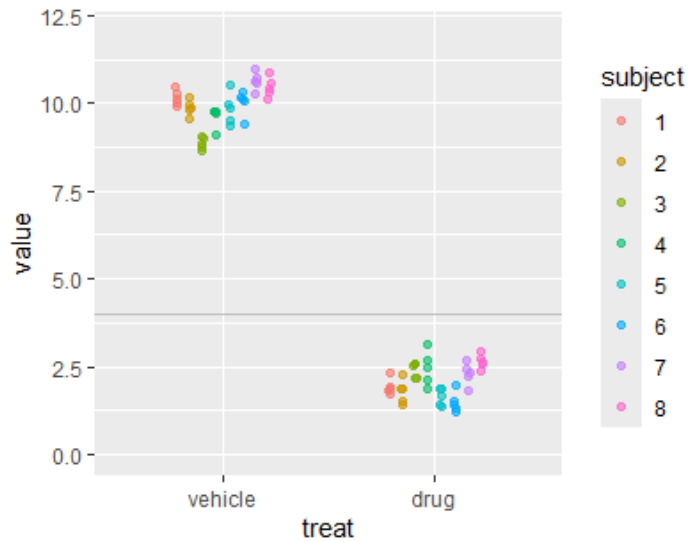
given the multiple combinations of completely-, partial- and non- lower or upper censored data, many different combinations are possible; note as in ANOVA other groups contribute to a common shared variance estimate. As indicated above for G, completely censored groups contribute *no* variance information.



investigations suggest that p-values from tobit models broadly agree with adjacent uncensored (left) and LR (right) methods although the latter worsen as less information is available

vertical lines correspond to the levels of the data response, and associated switches from pdf to cdf likelihood contributions.





extension to mixed models
follows directly through mgcv
support for random* effects

** need to use ML rather than REML for an LR-test*

```
dm = censoredData$new(data=M3, lower=4)
cm = censoredLRModel$new(value~treat + s(subject, bs="re"),
                           data=dm)
cm$contrast("treat", "vehicle", "drug")
```

\$censorship_type

[1] "LR"

\$report

```
[1] ">>> comparing 'vehicle' vs 'drug' for response 'treat'"
[2] "censorship type: LR"
[3] "fitting a LR tobit based model"
[4] "fully censored groups: vehicle: FALSE, drug: TRUE"
[5] "non-informative groups: none"
[6] "no non-informative groups dropped"
[7] "using data from groups vehicle, drug"
[8] "top ML: -53.0236997726918"
[9] "bottom ML: -87.2487597746797"
[10] "deviance change: 68.4501200039758"
```

\$change.in.deviance

[1] 68.45012

\$p.value

[1] 1.110223e-16

count models

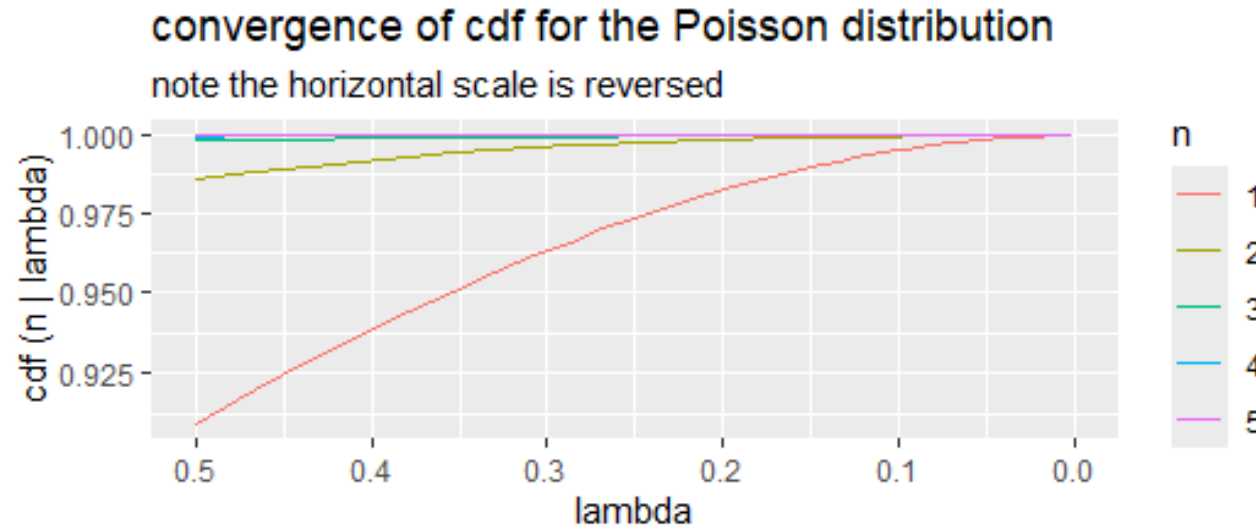
here Poisson*: applications for cases zero counts and completely censored

$$Pr(N = n) = \frac{\lambda^n e^{-\lambda}}{n!} \quad \text{for } \lambda > 0$$

- for **zeros**
 - important is only the total number of counts within a group, n
 - when $n = 0$ both λ^n and $n!$ equal one, so the probability is simply $e^{-\lambda}$ and is maximised as $\lambda \rightarrow 0$ for which the probability also tends to one, as before
- for **censorship**
 - censorship possible as before: for mgcv::gam use *family=cpois*.
 - for complete left censorship, the LR-test takes the same form as before: the probability goes to one as $\lambda \rightarrow 0$ (next slide)

** application to other models with dispersion still open but expected to be similar*

more formally

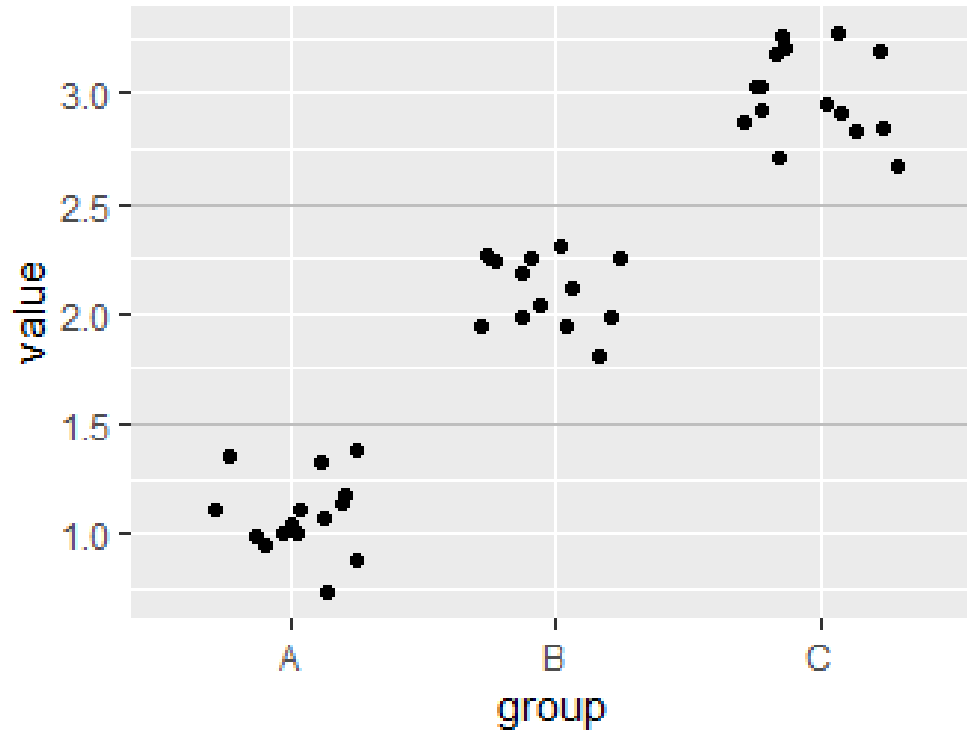


As seen by a repeated application of L'Hôpital's rule (shown **in the limit**, bottom left), the cdf of the Poisson distribution tends to one for any n as $\lambda \rightarrow 0$.

Censorship below only relevant for cutoffs $k > 0$: cdf here is strict in the sense $\Pr(N < n)$.

$$\text{cdf}(n|\lambda) = \frac{\sum_{j=1}^n \frac{\lambda^j}{j!}}{e^\lambda} = \frac{\frac{d^n}{d\lambda^n} \sum_{j=1}^n \frac{\lambda^j}{j!}}{\frac{d^n}{d\lambda^n} e^\lambda} = e^{-\lambda} \rightarrow 1$$

open questions and discussion



- unlike for normal parametric censorship models, the original variance of a completely censored group has no effect on our final inference. There is a notion of scale, but how valid are our assumptions?
- is A different to C, and what if any additional information does B add? *Answers on a post card only including your favourite method.*
- initial simulations suggest the approach may be conservative for linear and mixed models but how does the mathematics generalise to and behave for other count models and the wider exponential family?
- is any of this actually useful?



GSK