Leibniz Universität Hannover

Approximate Simultaneous Confidence Intervals for Overdispersed Count Data

Daniel Gerhard, Frank Schaarschmidt and Ludwig A. Hothorn

Institute of Biostatistics, Leibniz University of Hannover, Germany

1. Rationale

Count data occur in various research applications, e.g.

– in clinical and non-clinical trials, e.g. tumor counts, . . .
– in ecological field trials, e.g. insect abundance, . . .

A widespread distributional assumption for a sample of counts is the Poisson distribution, where the variance entirely depends on the sample mean. Frequently, overdispersion occurs for counts. Due to the consideration of these deviations from the Poisson assumption, the negative binomial distribution offers a more robust alternative to model count data in general. As in most cases multiple treatment groups are investigated, our objective is to calculate simultaneous confidence intervals for any linear combination of means with parameter estimation in a generalized linear model framework.

2. Negative binomial model

Parameters are estimated by a log-linear model [3], defined as

$$\log \mu_i = \eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i. \tag{1}$$

The estimation is performed by maximizing the negative binomial likelihood on the logarithmic scale, allowing the parameter space to contain only positive values. For the variance dependence on the mean, we assume

$$\operatorname{var}(Y) = \mu + \frac{1}{\phi}\mu^2. \tag{2}$$

The constant ϕ accounts for overdispersion occurring in every sample, which is estimated from the data in our case.

3. Approximate simultaneous confidence intervals

A set of linear combinations of parameter estimates from the generalized linear model is represented by a contrast matrix $\mathbf{C}$. An exemplary contrast matrix for the comparison of three groups against a control is written as

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

where each row should add up to zero. For linear combinations of the mean vector on the log scale $\widehat{\boldsymbol{\eta}}$ with variance covariance matrix $\widehat{\boldsymbol{\Sigma}}$, approximate two-sided Wald intervals are calculated by

$$\boldsymbol{C\eta} \in \left[\boldsymbol{C}\widehat{\boldsymbol{\eta}} \pm z^{\text{2-sided}}_{k,1-\alpha,\widehat{\boldsymbol{R}}} \sqrt{\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}'} \right]. \tag{3}$$

Here, z is the two-sided quantile of the k-variate normal distribution with corresponding correlation matrix $\widehat{\boldsymbol{R}}$ based on the GLM variance covariance estimates $\widehat{\boldsymbol{\Sigma}}$ and the given contrast matrix $\boldsymbol{C}$ [2]. As the parameters are estimated in a GLM framework, the inclusion of additional covariates is possible.

4. Transformation by the inverse link function

The confidence intervals are computed on the log scale; therefore one might want to take the exponent for the transformation back to the original scale. In Equation (4) we show this transformation for one row vector of the contrast matrix $\boldsymbol{C}$. This single contrast is divided into the contrasts c_a and c_b, which cover either the absolute positive or negative contrast coefficients, with $a = 1, \ldots, m$ and $b = 1, \ldots, n$.

$$\exp\left(\sum_{a=1}^{m} c_a \eta_a - \sum_{b=1}^{n} c_b \eta_b\right) = \exp\left(\sum_{a=1}^{m} c_a \log(\mu_a) - \sum_{b=1}^{n} c_b \log(\mu_b)\right) = \frac{\prod_{a=1}^{m} \mu_a^{c_a}}{\prod_{b=1}^{n} \mu_b^{c_b}}, \tag{4}$$

where $\sum_{a=1}^{m} c_a = 1$ and $\sum_{b=1}^{n} c_b = 1$. The result is a ratio of geometric means, which reduces to a simple ratio if only two particular means are compared.

5. Example

In a mutagenicity experiment the number of micronuclei per 2000 cells were counted at four doses (30, 50, 75, 100) of hydroquinone, a vehicle control and a positive control of 25 mg/kg cyclophosphamide [1].

Treatment group [mg/kg]	Number of micronuclei						
Vehicle control C^-	1	2	2	2	3	3	5
30	2	4	4	4	5		
50	4	6	6	7	8		
75	9	12	13	18	18		
100	13	20	22	22	23		
Positive control C^+	15	20	32	33			

Calculation in R

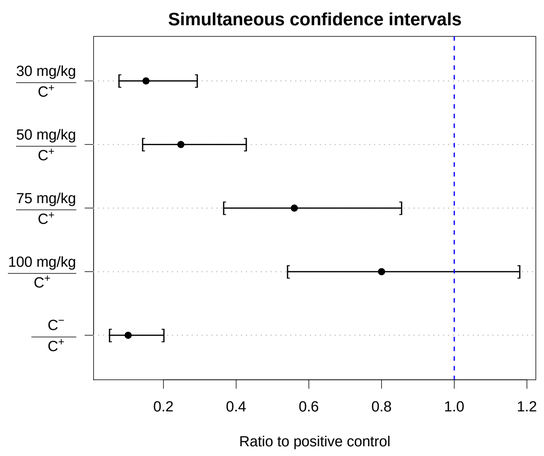
There is a straightforward way for calculating simultaneous confidence intervals in R [5] with the add-on package *multcomp* [4]. First, the package *gamlss* [6] is used to fit the generalized linear model.

```
> library(gamlss)
> glmfit <- gamlss(response ~ trt, data=example, family=NBI)
```

We want to compare each treatment and the vehicle control to the positive control.

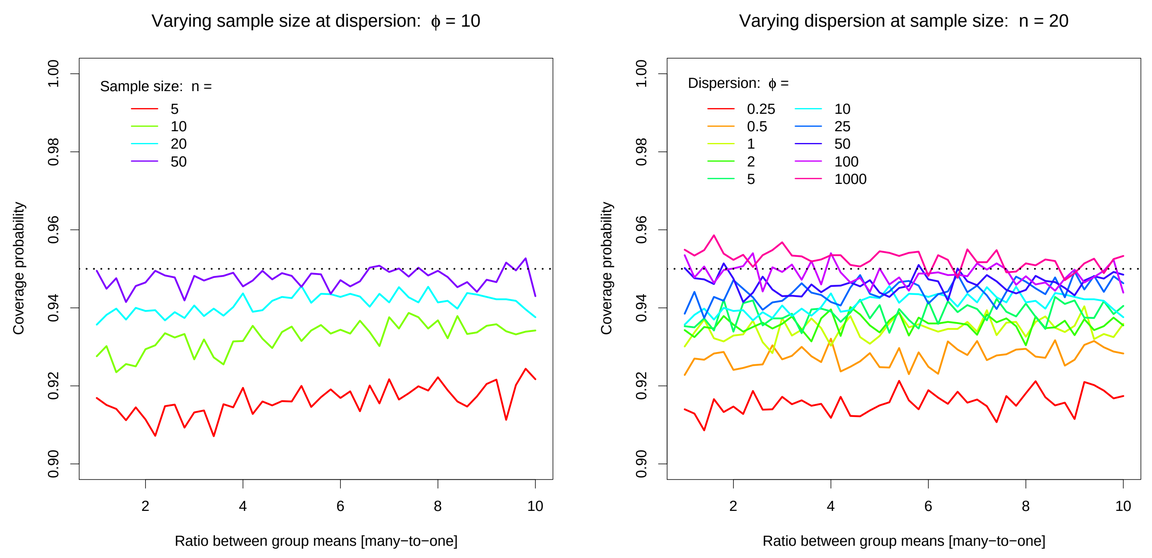
```
> library(multcomp)
> Cmat <- rbind(c(0,1,0,0,0,-1),
+               c(0,0,1,0,0,-1),
+               c(0,0,0,1,0,-1),
+               c(0,0,0,0,1,-1),
+               c(1,0,0,0,0,-1))
> MCP <- glht(glmfit,
+     linfct = mcp(trt = Cmat))
> exp(confint(MCP)$confint)

 Estimate   lwr     upr
1 0.152    0.079   0.292
2 0.248    0.144   0.427
3 0.560    0.367   0.854
4 0.800    0.543   1.179
5 0.103    0.053   0.200
```



6. Simulation study

We observed coverage probabilities of simultaneous 95% confidence intervals in a many-to-one setting of four samples. Random data were generated from a negative binomial distribution with varying sample size and dispersion parameters. The smallest sample mean was fixed at $\mu_i = 5$, to avoid problems with large amounts of zeros. For every simulation step 10000 repeats have been conducted.



7. Discussion

In cases, where we can assume count data to be negative binomial distributed, approximate simultaneous confidence intervals for linear combinations of parameters from a GLM provide a straightforward way of statistical inference for a wide area of research problems. For sufficiently large sample sizes ($n \geq 20$) and slight overdispersion ($\phi \geq 10$) a satisfactory coverage probability (≥ 0.94) can be achieved. The method shown here is already implemented in the statistical software package R [5], especially with the *multcomp* [4] extension.

References

[1] Adler, I.D. and Kliesch, U. (1990): Comparison of single and multiple treatment regiments in the mouse bone marrow micronucleus assay for hydroquinone and cyclophosphamide. *Mutation Research* **234**: 115-123.

[2] Bretz, F., Genz, A., Hothorn, L.A. (2001): On the numerical availability of multiple comparison procedures. *Biometrical Journal* **43**: 645-656.

[3] McCulloch, C.E. and Searle, S.R. (2001): *Generalized, linear and mixed models*. John Wiley & Sons, Inc.

[4] Hothorn, T., Bretz, F., Westfall, P., Heiberger, R.M.(2007). *multcomp*: Simultaneous inference for general linear hypotheses. R package version 0.991-9.

[5] R Development Core Team (2007). *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[6] Rigby, R.A. and Stasinopoulos, D.M. (2004): Generalized additive models for location, scale and shape. *Applied Statistics* **54**: 1-38.