Analyzing clustered count data with a cluster-specific random effect zero-inflated Conway–Maxwell–Poisson distribution

Somnath Datta*
University of Florida, Gainesville, USA – somnath.datta@ufl.edu

Hyoyoung Choo-Wosoba
National Institutes of Health, Rockville, USA – coollss07@gmail.com

Count data analysis techniques have been developed in biological and medical research areas. In particular, zero-inflated versions of parametric count distributions have been used to model excessive zeros that are often present in these assays. The most common count distributions for analyzing such data are Poisson and negative binomial. However, a Poisson distribution can only handle equi-dispersed data and a negative binomial distribution can only cope with overdispersion. However, a Conway–Maxwell–Poisson (CMP) distribution can handle a wide range of dispersion. We show, with an illustrative data set on next-generation sequencing of maize hybrids, that both underdispersion and overdispersion can be present in genomic data. Furthermore, the maize data set consists of clustered observations and, therefore, we develop inference procedures for a zero-inflated CMP regression that incorporates a cluster-specific random effect term. Unlike the Gaussian models, the underlying likelihood is computationally challenging. We use a numerical approximation via a Gaussian quadrature to circumvent this issue. A test for checking zero-inflation has also been developed in our setting. Finite sample properties of our estimators and test have been investigated by extensive simulations. Finally, the statistical methodology has been applied to analyze the maize data mentioned before.

Keywords: mixed effects model; next-generation sequencing; Poisson distribution; over-dispersions.