Are my counts Poisson?

Jacobo de Uña-Álvarez[1], *María Dolores Jiménez-Gamero*[2]

[1]jacobo@uvigo.gal, CINBIO, Universidade de Vigo

[2]dolores@us.es, Department of Statistics and Operations Research, Universidad de Sevilla

The Poisson model is often used to analyze count data. For instance, sequencing experiments in genetics report a large number of read counts along a DNA or RNA region for a (typically) small number of individuals. Several authors have defended the general validity of the Poisson model for the read counts in the referred sequencing setups. Accordingly, multiple testing and other inferences are often performed from Poisson P-values, these are, tail probabilities based on a Poisson model. Still, it is a matter of fact that counts deviate from Poisson in particular applications. In such a case, Poisson tail probabilities are inaccurate and nominal significance levels may be violated.

In this work a test for the null hypothesis that a large number k of (possibly small) samples follow Poisson distributions with arbitrary rates is introduced. The test is based on a differential equation that involves the probability generating function, and that characterizes the Poisson model. The individual test statistics pertaining to the many samples are aggregated into a single measure for which a null Gaussian distribution as $k \to \infty$ is derived. The test may detect any deviation from Poisson when k is large enough (that is, it is omnibus), even for small sample sizes and a vanishing proportion of non-Poisson populations. This is proved both theoretically and through simulations. The method allows for dependences among the samples, which may happen in genome-wide studies. Illustrative applications to real sequencing experiments are provided.

Work supported by the grant PID2020-118101GB-I00, Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033).

Keywords: Goodness-of-fit, High-dimensional data, Multiple testing.