A new score test for distinguishing between the zero-inflated Poisson and the two-component Poisson mixture distribution

Anabel Blasco-Moreno[1,2], *Pere Puig*[2]

[1]anabel.blasco@uab.cat, Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona
[2]ppuig@mat.uab.cat, Department of Mathematics, Universitat Autònoma de Barcelona

Many disciplines, including medical, biology, genetics, economics, and social sciences, require an understanding of the underlying nature of the data. Data such as the number of days spent in the hospital, the number of chromosomal aberrations, or the number of standard alcoholic drinks are non-negative and often right-skewed, heavy-tailed, and multi-modal, with a point mass at zero. The presence of these features can make it difficult to choose the best distribution for fitting the data.

In biological dosimetry, for example, exposure to ionising radiation (IR) causes a variety of damage in peripheral blood limphocytes. The microscopic counting of dicentric chromosomal aberrations is used to estimate the individual absorbed radiation dose. We look at two types of IR that are important for medical diagnosis and treatment: whole body irradiation (WBI) and partial body irradiation (PBI). Furthermore, radiation exposure might be homogenous or heterogeneous. The number of dicentric chromosomes per cell in a homogeneous PBI scenario can be explained by a Zero-Inflated Poisson distribution (ZIP). In WBI, heterogeneous irradiation scenarios can be modeled by a finite mixture of Poisson distributions (MP). Knowing whether the dicentric distribution corresponds to ZIP or MP allows us to determine whether the exposure was homogeneous PBI or heterogeneous WBI and act accordingly.

In this research, we propose an exact test to contrast the null hypothesis H_0 : Data are ZIP distributed, against the alternative hypothesis H_1 : Data follow a two-component Poisson mixture distribution. Our score test was developed as a solution to a problem involving occupancy distributions. We work with the conditional probability of the data given sufficient statistics, i.e., statistics that contain all of the information about the ZIP distribution's parameters, that is, the whole sum of the observed values and the number of zero-counts. Given a sample of n independent observations of counts $\mathbf{X} = (X_1, X_2, ...X_n)$ following a ZIP distribution, where $\sum_{i=1}^{n} X_i$ and N_0 are sufficient statistics, the probability function of $\mathbf{X}$ conditioned to these sufficient statistics is given by the following expression:

$$Pr\left(\mathbf{X} \mid N_0 = n_0, \sum_{i=1}^{n} X_i = t\right) = \frac{n_0! t!}{n! S(t, n - n_0) \prod_{i=1}^{n} X_i!}\,. \tag{1}$$

This result is related to occupancy problems, and $S(t, n - n_0)$ denotes the second order Stirling number. With the values of the sufficient statistics, expression (1) allows for the generation of ZIP data. The null hypothesis can then be tested by determining the empirical distribution of the score test statistic, which is independent of the parameters. Because the obtained distribution of the test statistic is not asymptotic, it is an exact test. Finally, we applied the score test to several application examples based on in vitro data of heterogeneous WBI and homogeneous PBI.

Keywords: zero-inflated Poisson, mixture-Poisson, score test.