Distributions Families in Counting Bacteria for Compound Sampling

Miguel Felgueiras, João Paulo Martins, and Rui Santos

School of Technology and Management, Polytechnic Institute of Leiria and Center of Statistics and Applications of University of Lisbon, Leiria, Portugal
{mfelg,jpmartins,rui.santos}@ipleiria.pt

Abstract. The sensitivity and the specificity of a compound test depend on the distribution underlying the phenomenon. In this paper we consider count distributions unified under Panjer recursive formula and belonging to the Morris family, which verifies useful properties. The influence of the tail weight of the count distributions (that varies among infected and uninfected elements), evaluated in terms of the dispersion index, is investigated in the sensitivity and the specificity of the compound test.

Keywords: Compound tests, Panjer system, NEF–QVF.

1 Introduction

Dorfman seminal work [2] in the identification of the American soldiers infected with syphilis during World War II was the first work published in compound analysis. Dorfman pioneer idea was minimizing the costs of performing a large number of blood tests by mixing the blood of n individuals. A negative result of a compound test implies that no one is infected and therefore $n-1$ tests are spared. By the other hand, a positive result means that at least one of the members is infected and therefore individual tests should be performed, increasing in one the total number of performed tests for the group ($n+1$ tests are performed). The optimal batch size minimizes the total number of tests, and therefore the total cost since the mixing cost is usually irrelevant. However, Dorfman methodology does not take into account classification errors (measured by the test sensitivity and specificity) and it is restricted to a binary classification of the individuals (presence or absence of some substance in the analysed fluid). To overcome the above limitations, different methodologies were proposed by several authors. Boswell *et al* [1] presents an extensive work on this subject, including different applications of these methodologies. Later, Kim *et al* [6] also tackles the group testing application with a more modern approach. Finally, Hwang [5] and Santos *et al* [13] deal with the dilution problem. The dilution problem is quite relevant since, when measuring a quantitative variable, the concentration of some type of bacteria in blood (for instance) will be diminished by the mixing process (cf. [15] for the HIV problem).

Accordingly, and after a brief revision of some families of distributions (section 2), the connection to the dispersion index concept is explained (section 3).

B. Murgante et al. (Eds.): ICCSA 2014, Part III, LNCS 8581, pp. 539–551, 2014.

Afterwards, the appropriate models for counting bacteria are described (section 4), and the application of tests for over and under dispersion data is discussed in section 5. Section 6 provides a revision of the most applied measures of misclassification in the compound analysis context. The performance of the proposed score test is analysed for different count distributions considering healthy and non healthy populations (section 7). Finally, the main conclusions are displayed in section 8.

2 Panjer and Natural Exponential with Quadratic Variance Functions Families of Distributions

In this section we present an overview of some families of distribution which unify standard counting distributions.

2.1 Panjer Family

After some preliminary work on the aggregate claims distribution, the Panjer recursive formula was simultaneously presented in Panjer [11] and in Sundt and Jewell [14], again in the aggregate claims context and when the recursion formula holds for $n \geq 0$. Later, Willmot *et al* [16] extended the Panjer family for distributions where the formula can only be applied when $n \geq 1$, and finally Hess *et al* [4] identify all distributions from the Panjer class of order k with arbitrary k, $n \geq k$. The Panjer family is therefore characterized by the Panjer recursion formula,

$$\mathrm{P}(N = n+1) = \left(a + \frac{b}{n+1}\right)\mathrm{P}(N = n),\ n = k,\ k+1, \dots,$$

where N is a counting variable, $k \in \mathbb{N}_0$ and $a+b > 0$. When $k = 0$, we obtain the binomial (B), the negative binomial (NB) and the Poisson distributions (Po). For these distributions, the mean value is

$$\begin{aligned}\mathrm{E}(N) &= \sum_{n=0}^{\infty} n\mathrm{P}(N = n) = \sum_{n=1}^{\infty} n\left(a + \frac{b}{n}\right)\mathrm{P}(N = n-1) = \\ &= a\sum_{n=1}^{\infty}\mathrm{P}(N = n-1)(n-1+1) + b = a\mathrm{E}(N) + a + b.\end{aligned}$$

Note that $a = 1$ leads to $b = -1$, which isn't an interesting combination of parameters (see Table 1). In order to simplify we will consider from now on $a \neq 1$. Therefore,

$$\mathrm{E}(N) = \frac{a+b}{1-a}\,.$$

A similar approach can be applied in order to obtain the variance, leading to

$$\mathrm{V}(N) = \frac{a+b}{(1-a)^2}.$$