Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6138 – 6150
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A new generalization of the Waring distribution

J. Rodríguez-Avi*, A. Conde-Sánchez, A.J. Sáez-Castillo, M.J. Olmo-Jiménez

Department of Statistics and Operations Research, University of Jaén, Despacho B3-058, Campus Universitario de Jaén, 23071 Jaén, Spain

Received 2 March 2005; received in revised form 15 December 2006; accepted 16 December 2006
Available online 22 December 2006

Abstract

A tetraparametric univariate distribution generated by the Gaussian hypergeometric function that includes the Waring and the generalized Waring distributions as particular cases is presented. This distribution is expressed as a generalized beta type I mixture of a negative binomial distribution, in such a way that the variance of the tetraparametric model can be split into three components: *randomness*, *proneness* and *liability*. These results are extensions of known analogous properties of the generalized Waring distribution. Two applications in the fields of sport and economy are included in order to illustrate the utility of the new distribution compared with the generalized Waring distribution.

Keywords: Accident theory; Gaussian hypergeometric function; Mixture of distributions; Waring distribution

1. Introduction

In accident theory, the generalized Waring distribution (Irwin, 1968; Xekalaki, 1983a) provides a model for the number of accidents in which the variance can be split into three components: the first of these is the random component (randomness) and the other two will often be identifiable as the separate variances due to internal differences (proneness) and differences in exposure to external risk of accident (liability).

However, the utility of the model may be extended to any field where the occurrence of an event depends not only on chance, but also on the particular characteristics and the exposure to risk of the observed individual. In fact, applications of the univariate generalized Waring distribution (UGWD) can be found in a range of fields such as lexicology (Tesitelova, 1967), the number of authors of scientific articles (Ajiferuke, 1991), the evolution of the number of links in the World Wide Web (Levene et al., 2002) and the study of sexual behaviours (Handcock and Jones, 2003; Handcock et al., 2003). Here two new applications in the area of sport and economy are presented. In both examples the discussion about the origin of variability of the data may be interesting to understand their nature.

Firstly, the number of yellow cards that were shown to footballers (excluding goalkeepers) who have played at least one football match, is a variable that is subject to various and diffuse sources of variation, such as:

- the number of matches played;

* Corresponding author. Tel. :+34 953 212207; fax: +34 953 212034.
E-mail address: jravi@ujaen.es (J. Rodríguez-Avi).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.12.029

- the position of the player in the field: a defender usually runs more risk of being penalized than a midfielder, who in turn has a greater risk than a forward;
- the personality of the footballer, which becomes evident in his behaviour during the match.

The first two sources of variation can be associated with the differences in exposure to risk of receiving a yellow card, whereas the third is a factor inherent to the footballer. In this paper we focus on the study of two consecutive seasons, 2001/2002 and 2002/2003, of the Spanish football league (Table 2). It could be interesting to apply a model like the UGWD to quantify the percentage of the total variability that is due to internal conditions (proneness), to the different exposure to risk (liability) and to unexplained factors.

Secondly, the number of hotels by municipality in the region of Andalusia, Spain, between 1998 and 2003 has been considered. The data for 2003 are presented in Table 6. These consist of counts for a variable with great overdispersion that depends on several types of factors:

- External factors (liability), relative to demographic and geographical aspects (number of inhabitants per municipality, extension, location, etc.), which affect equally all municipalities with similar characteristics.
- Internal factors (proneness), relative to intrinsic, socioeconomic aspects of the municipality (business initiative, ability to exploit the added value of the milieu, etc.).

Again, a fit using a distribution with the properties of the UGWD would allow us to detect and quantify the extent of influence of each of these sources of variation.

In the first example none of the methods of estimation implemented in the UGWD provides an appropriate fit, whereas, in the second example, the fit of this distribution is acceptable but does not provide information about the origin of the variability of the data. There are other distributions, such as the negative binomial (NB) distribution (Neyman, 1939), that are commonly used in these type of applications. Nevertheless, the NB distribution or any generalization based on simple Poisson mixtures (see, for instance, Gupta and Ong, 2004) only distinguish between the variability due to random and non-random factors.

In this respect, the aim of this work is to propose a new generalization of the Waring distribution, which is analogous to the UGWD, and that contains both of them as particular cases. Moreover, this new generalization extends all the properties of the UGWD, provides an excellent fit for the presented data and describes the variability of every situation according to the aforementioned factors, related to randomness, liability and proneness.

2. The generalized Waring distribution

From a historical point of view, the origin of the UGWD is the classical Waring distribution that arises from the expansion of the Waring series

$$\frac{1}{x-a}=\sum_{r=0}^{\infty}\frac{(a)_r}{(x)_{r+1}}, \tag{1}$$

where $(\alpha)_k=\alpha(\alpha+1)\dots(\alpha+k-1)$; if $\alpha>0$ then $(\alpha)_k=\Gamma(\alpha+k)/\Gamma(\alpha)$. Putting $\rho=x-a$, the probability mass function (p.m.f.) of the distribution that is associated with series (1) is given by

$$f_r=\rho\frac{(a)_r}{(a+\rho)_{r+1}},\quad r=0,1,\dots,$$

where $a,\rho>0$. Series (1) can be generalized in the form

$$\frac{1}{(x-a)_k}=\sum_{r=0}^{\infty}\frac{(a)_r(k)_r}{(x)_{r+k}}\frac{1}{r!}, \tag{2}$$

with $k>0$ (Irwin, 1963). Considering again $\rho=x-a$, the p.m.f. associated with series (2) is expressed as

$$f_r=\frac{(\rho)_k}{(a+\rho)_k}\frac{(a)_r(k)_r}{(a+k+\rho)_r}\frac{1}{r!},\quad r=0,1,\dots, \tag{3}$$