International Conference on Distribution Theory, Order Statistics and Inference in Honor of Barry C. Arnold. June 16-18, 2004

Properties and Applications of the Family of Gaussian Discrete Distributions

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

Departamento de Estadística e I.O.,
Universidad de Jaén.

Abstract

In this work we present a study about the properties of the family of discrete distributions generated by the Gaussian hypergeometric function, named $GHD(\alpha, \beta, \gamma, \lambda)$. This family generalizes some widely known distributions, such as the Waring distribution. In this sense, results about several probabilistic aspects are included (its expression as mixture of distributions, a partition of the variance, etc.). Finally, its applicability by the modelling and the interpretation of data sets originating from fields such as Sports is shown.

Key Words: Gaussian Hypergeometric Distributions, mixture, beta function, Waring distribution, partition of variance, sport.

1 Introduction

The family of Gaussian Hypergeometric Distributions is characterized by its probability generating function, given by the Gaussian hypergeometric function (except for the constant)

$$_2F_1(\alpha, \beta, \gamma, \lambda z) = \sum_{r=0}^{\infty} \frac{(\alpha)_r (\beta)_r}{(\gamma)_r} \frac{(\lambda z)^r}{r!}. \tag{1.1}$$

The probability mass function is

$$f_r = P[X = r] = f_0 \frac{(\alpha)_r (\beta)_r}{(\gamma)_r} \frac{\lambda^r}{r!}, \qquad r = 0, 1, ..., \tag{1.2}$$

where f_0 is the constant of normalization. We denote these distributions as $GHD(\alpha, \beta, \gamma, \lambda)$.

[*] Correspondence to: José Rodríguez Avi. Departamento de Estadística e I.O. Universidad de Jaén. España

Table 1: *Classification of the GHD family (α and β are interchangeable)*

Parameters			Conditions	Range	Type
$\gamma > 0$	$0 < \lambda \leq 1$	$\alpha, \beta > 0$	$\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	I
		$\alpha, \beta \in \mathbb{C}$, $\alpha = \bar{\beta}$	$\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	II
		$\alpha, \beta < 0$, $\alpha, \beta \notin \mathbb{Z}^-$	$[\alpha] = [\beta]$	$[0, \infty)$	III
	$0 < \lambda$	$\alpha, \beta < 0$, $\alpha \in \mathbb{Z}^-$	$\lvert\beta\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	IV
	$\lambda < 0$	$\alpha \in \mathbb{Z}^-, \beta > 0$		$[0, \lvert\alpha\rvert)$	V
$\gamma < 0$	$0 < \lambda \leq 1$	$\alpha < 0, \alpha \notin \mathbb{Z}^-$, $\beta > 0$	$[\alpha] = [\gamma]$, $\gamma > \alpha + \beta$ if $\lambda = 1$	$[0, \infty)$	VI
	$0 < \lambda$	$\alpha \in \mathbb{Z}^-, \beta > 0$	$\lvert\gamma\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	VII
	$\lambda < 0$	$\alpha, \beta < 0$, $\alpha \in \mathbb{Z}^-$	$\lvert\gamma\rvert, \lvert\beta\rvert > \lvert\alpha\rvert - 1$	$[0, \lvert\alpha\rvert)$	VIII

A general summation result for computing the f_0 value is unknown. When $\lambda = 1$, the Gauss Summation Theorem establishes that

$$f_0 = \frac{\Gamma(\gamma - \alpha)\,\Gamma(\gamma - \beta)}{\Gamma(\gamma)\,\Gamma(\gamma - \alpha - \beta)}. \tag{1.3}$$

Table 1 includes the conditions of the parameters that allow the series (1.1) to converge and that guarantee the positivity of all its terms. Type II has been studied by Rodríguez-Avi et al. (2003a, 2004); likewise, the case $\lambda = 1$ has been analyzed by Rodríguez-Avi et al. (2003b), Sibuya (1979) and Sibuya and Shimizu (1981), among others.

Many well-known distributions (Poisson, Negative Binomial, Binomial, Beta-Binomial, Hypergeometric, Waring, etc.) belong to the GHD family or they are limit cases of that distributions, as it is shown in Table 2.

2 Properties

Moreover, f_r is the solution of the extended Pearson system

$$G(r)\,f_{r+1} - L(r)\,f_r = 0, \ r = 0, 1, 2, ..., \tag{2.1}$$

Table 2: Some distributions and their p.g.f. in relation with the Gauss function

Distribution	P.g.f.
Binomial	$\dfrac{{}_2F_1\left(-n,\beta;\beta;\lambda z\right)}{{}_2F_1\left(-n,\beta;\beta;\lambda\right)},\ \lambda=-\dfrac{p}{1-p}$
Poisson	$\lim_{\substack{n\to\infty\\ p\to 0}}\dfrac{{}_2F_1\left(-n,\beta;\beta;\lambda z\right)}{{}_2F_1\left(-n,\beta;\beta;\lambda\right)},\ \lambda=np$
Negative Binomial	$\dfrac{{}_2F_1\left(k,\beta;\beta;\left(1-p\right)z\right)}{{}_2F_1\left(k,\beta;\beta;1-p\right)}$
Crow-Bardwell	$\lim_{\beta\to\infty}\dfrac{{}_2F_1\left(1,b;\lambda;\dfrac{\theta z}{b}\right)}{{}_2F_1\left(1,b;\lambda;\dfrac{\theta}{b}\right)}$
Extended Crow-Bardwell	$\lim_{\beta\to\infty}\dfrac{{}_2F_1\left(\beta,b;\lambda;\dfrac{\theta z}{b}\right)}{{}_2F_1\left(\beta,b;\lambda;\dfrac{\theta}{b}\right)}$
Hypergeometric	$\dfrac{{}_2F_1\left(-n,-M;N-M-n+1;z\right)}{{}_2F_1\left(-n,-M;N-M-n+1;z\right)}$
Negative Hypergeometric	$\dfrac{{}_2F_1\left(-n,M;M-N-n+1;z\right)}{{}_2F_1\left(-n,M;M-N-n+1;1\right)}$
Waring	$\dfrac{{}_2F_1\left(1,k;k+\rho+1;z\right)}{{}_2F_1\left(1,k;k+\rho+1;1\right)}$
Generalized Waring	$\dfrac{{}_2F_1\left(a,k;a+k+\rho;z\right)}{{}_2F_1\left(a,k;a+k+\rho;1\right)}$
CBPD	$\dfrac{{}_2F_1\left(bi,-bi;\gamma;z\right)}{{}_2F_1\left(bi,-bi;\gamma;1\right)}$
CTPD	$\dfrac{{}_2F_1\left(a+bi,a-bi;\gamma;z\right)}{{}_2F_1\left(a+bi,a-bi;\gamma;1\right)}$

where functions L and G are second-degree polynomials

$$\begin{aligned} L\left(r\right) &= \left(\alpha+r\right)\left(\beta+r\right)\lambda \\ G\left(r\right) &= \left(\gamma+r\right)\left(r+1\right). \end{aligned} \tag{2.2}$$

From (2.1) Fajardo (1986) proved that non-central moments verify the recurrence equation

$$\mu'_{h+2} + (\gamma - 1)\,\mu'_{h+1} - \lambda \sum_{m=0}^{h} \binom{h}{m} \left[\mu'_{m+2} + (\alpha+\beta)\,\mu'_{m+1} + \alpha\beta\mu'_m\right] = 0, \tag{2.3}$$

for $h = 0, 1, 2, ...$ if $\lambda < 1$ or the distribution is finite, and for $h = 0, 1, 2, ..., k-2$ if $\lambda = 1$ and $\gamma > \alpha + \beta + k$ with $k \geq 2$. It is of note that this recurrence relation can not generally provide explicit expressions of moments from parameters, because n equations involve $n+1$ moments; nevertheless, if $\lambda = 1$, n equations involve n moments which may be calculated as solutions of the corresponding linear system.

On the other hand, as Johnson et al. (1992) suggest, the GHD distributions, generated by the ${}_2F_1$ function, may be seen as mixtures of distributions generated by the ${}_1F_0$ and the ${}_1F_1$ functions where the mixing distributions are generalized Beta and Gamma distributions, respectively. A clearer methodology in order to make explicit these results is given by the conditional specification of these distributions: Arnold et al. (1999) provide a result in exponential families that determines the most general marginal distributions when the conditional distributions are assumed.

As an example of this type of results, it can be proved that if $\gamma > \beta$, the $GHD\,\mathrm{I}\,(\alpha, \beta, \gamma, \lambda)$ is the mixture

$$Poisson\,(\Lambda) \underset{\Lambda}{\wedge} Gamma\left(\alpha, \frac{\lambda\,(1-P)}{1-\lambda\,(1-P)}\right) \underset{P}{\wedge} GBeta\,(\gamma-\alpha-\beta, \beta, \alpha, \lambda)\,, \tag{2.4}$$

where $GBeta\,(\gamma-\alpha-\beta, \beta, \alpha, \lambda)$ denotes a generalization of the Beta distribution whose density function is

$$f_P\,(p) = \frac{1}{{}_2F_1\,(\alpha, \beta; \gamma; \lambda)} \frac{\Gamma(\gamma)}{\Gamma(\gamma-\beta)\Gamma(\beta)} \frac{p^{\gamma-\beta-1}\,(1-p)^{\beta-1}}{(1-\lambda\,(1-p))^{\alpha}}, \quad 0 \leq p \leq 1. \tag{2.5}$$

3 Applications

This type of results about mixtures allows us to obtain interesting conclusions about the variability of data.

Thus, if $\gamma > \beta$ and $X \rightsquigarrow GHD\,\mathrm{I}\,(\alpha, \beta, \gamma, \lambda)$,

$$VarX = \alpha E_P\,[V] + \alpha E_P\left[V^2\right] + \alpha^2 Var_P\,(V)\,, \tag{3.1}$$

where $V = \lambda(1-P)/(1-\lambda(1-P))$ and $P \rightsquigarrow GBeta(\gamma-\alpha-\beta, \beta, \alpha, \lambda)$. The first of these addends is related to random factors, the second to the variability due to external factors that affect the population (liability), and the third to the differences in the internal conditions of the individuals (proneness). The result is an extension of that known in the case $\lambda = 1$, corresponding to the Waring distribution.

One of the main drawbacks of the GHD is that the parameters α and β are interchangeable. Thus, in practice, the identification of the latter two components in (3.1) is not clear without some additional information about the framework of application. Irwin (1968) suggests that "the statistician will usually know from studying the data in various ways whether the proneness or the liability component should be the greater".

Next, we carry out the fit of two data sets by the GHD I distributions and we obtain and interpret the partition of the variance given by the above result.

3.1 Number of goals scored by the footballers

Data correspond to the number of goals scored by the footballers that have played at least one match in the 2000/01 football season of the Spanish League.[1]

The method of maximum likelihood provides the fit given by the GHD $(0.5468, 8.1142, 9.0597, 0.8621)$. It should be emphasized that none fit can be found by the Waring distribution. Table 3 contains the observed and expected values, while the obtained fit may be graphically seen in Figure 1.

Regarding the decomposition of the model variability, the involved factors may be interpreted as follows:

- *Proneness*: There are footballers who are more prone to score goals than others, independently of the position they have in the ground. V.gr., there exist defenses who are more prone to go on the offensive and to score. In fact, this factor may be defined as *goal intuition.*
- *Liability*: The probability of scoring is related to the place that the

[1] Excluding the goalkeepers.

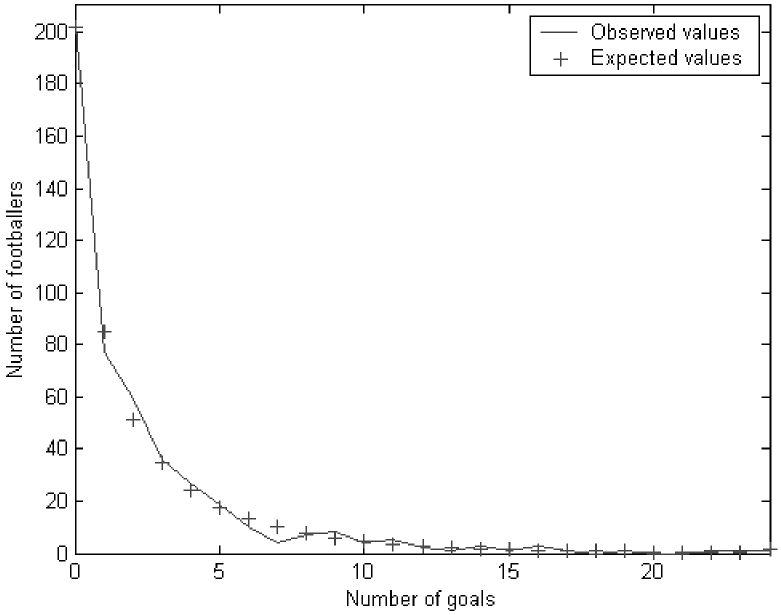

Figure 1: *Number of goals scored. Season 2000/01*

footballer has in the ground. Thus, a forward has much chance to score in a match.

- *Randomness*: There are random circumstances that make a footballer score in a match.

The quantifying of the variability corresponding to these three factors in the proposed model for data appears in Table 4.

3.2 Yellow cards

In this case the variable is *Number of yellow cards shown to footballers*[2]*. Season 2000/01*. It should be noted the steep shape the variable presents, as it can be seen in Figure 2.

[2] excluding the goalkeepers

Table 3: *Fit for goals scored. Season 2000/01*

x_i	O_i	$E_i(a)$
0	203	201.7353
1	77	85.1727
2	59	51.4513
3	36	34.4363
4	27	24.2601
5	19	17.6420
6	10	13.1148
7	4	9.9104
8	7	7.5852
9	8	5.8658
10	4	8.1690
11	5	
12	2	5.0984
13	1	
14	3	9.5587
15	1	
16	3	
17	1	
19	1	
22	1	
23	1	
24	1	
Total	474	474
χ^2		9.0427
d.f.		8
p-value		0.3387

The method of maximum likelihood provides the fit given by the GHD $(797.5941, 0.5386, 13.4859, 0.0217)$. The results are shown in Table 5 and the goodness of fit can be graphically seen in Figure 2. In this case, fits accuracy enough with any other discrete distribution with fewer parameters have not been found.

Provided that the proposed model is a GHD I distribution, the vari-

Table 4: *Partition of the variance*

Factor	**Variability**	**% of variability**
Randomness	2.24669	17.2068
Liability	10.2401	78.4957
Proneness	0.5606	4.2975
Total	13.0454	100

Table 5: *Fit for number of yellow cars by footballer. Season 2000/01*

x_i	O_i	$E_i(a)$
0	82	80.9469
1	54	55.9536
2	54	51.4949
3	47	48.8237
4	50	45.5157
5	45	41.0998
6	27	35.7440
7	32	29.8790
8	23	23.9965
9	14	18.5231
10	15	13.7532
11	10	9.8320
12	12	6.7746
13	2	11.6630
14	4	
15	1	
16	1	
17	1	
Total	474	474
χ^2		9.2736
d.f.		9
p-value		0.4124

ance can also be split into three components. At this point, an empirical interpretation of the factors may be established:

Table 6: *Partition of the variance*

Factor	**Variability**	**% of variability**
Randomness	4.15767	32.5206
Liability	8.59461	67.2256
Proneness	0.0324489	0.2538
Total	12.7847	100

- *Proneness*: There are footballers more prone to be shown yellow cards than others because they are more violent or given to protesting.
- *Liability*: There are places in the ground occupying by the footballers more prone to have a yellow card, as it happens with the centre-forwards or the centre-halves.
- *Randomness*: A yellow card can be shown to a footballer because of random circumstances with independence of his place in the ground or his character.

The components of the variance in the proposed model for data are included in Table 6.

References

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*. Springer, New York.

Fajardo, M. A. (1986). *Generalizaciones de los sistemas pearsonianos discretos*. Ph.D. thesis, Departamento de Estadística e Investigación Operativa, Universidad de Granada.

Irwin, J. O. (1968). The generalized waring distribution applied to accident theory. *Journal of the Royal Statistical Society A*, 131:205–227.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions*. John Wiley and Sons, Inc., New York.

Rodríguez-Avi, J., Conde-Sánchez, A., Olmo-Jiménez, M. J., and Sáez-Castillo, A. J. (2004). A triparametric discrete distribution with complex parameters. *Statistical Papers*, 45(1):81–98.

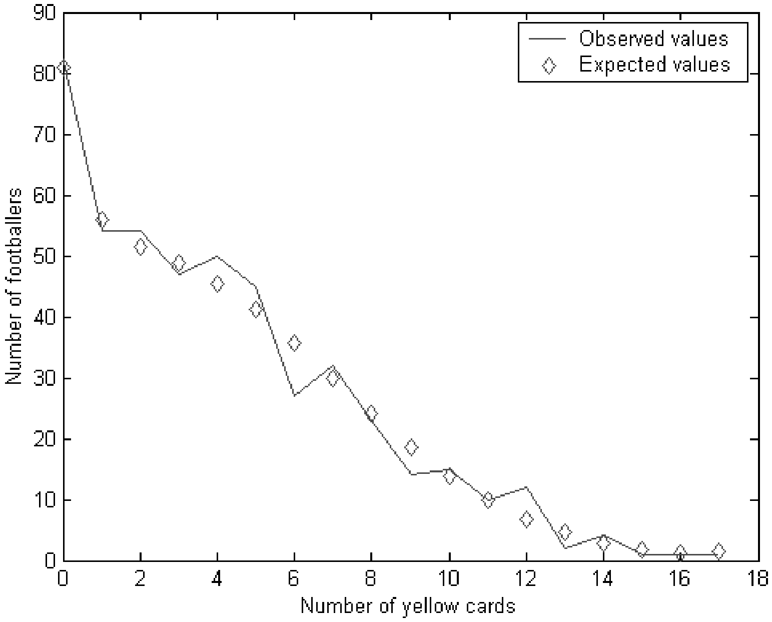

Figure 2: *Observed and expected values for Table 5*

Rodríguez-Avi, J., Conde-Sánchez, A., and Sáez-Castillo, A. J. (2003a). A new class of discrete distributions with complex parameters. *Statistical Papers*, 44(1):67–88.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo Jiménez, M. J. (2003b). Estimation of parameters in gaussian hypergeometric distributions. *Communications in Statistics: Theory and Methods*, 32:1101–1118.

Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distributions. *Annals of the Institute of Statistical Mathematics*, 31, Part A:373–390.

Sibuya, M. and Shimizu, R. (1981). Classification of the generalized hypergeometric family of distributions. *KEIO Science and Technology Reports*, 34:1–38.