Statistical analysis for a distribution of a random walk on the plane

Shin-Zhu Sim
Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia–
shinzhusim@gmail.com

Seng-Huat Ong*
Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia–
ongsh@um.edu.my

Abstract

A distribution that arises as a first-passage time distribution of a modified random walk on the half plane is considered. It is also a member of Kemp's family of convolutions of binomial and pseudo-binomial variables. Thus it has a simple probabilistic structure and computer generation of random samples from this distribution is straightforward. Some probabilistic properties, like log-concavity, unimodality, and reliability properties have been derived. The distribution can cater for under, equi and over dispersion in count data without requiring restriction in the support and computation of normalizing constant. It has a simple three-term recurrence formula for computing the probabilities which facilitates its applications. For equi-dispersion the distribution is non-Poisson. This provides an alternative to the popular Poisson model for empirical modelling of data exhibiting equi-dispersion. Test of hypothesis for equi-dispersion by the likelihood ratio test and simulation study of the power has been conducted. Parameter estimation by maximum likelihood, minimum Hellinger distance and a method based on the probability generating function has been considered. As an illustration of its application, a fit to a real data set is given.

Keywords: Under-, equi- and over-dispersion; Log-concavity, Unimodality and reliability properties; Parameter estimation; Goodnesss-of-fit.

1. Introduction

Recently Aoyama et al. (2008) derived the generalized inverse trinomial (GIT) distribution by considering a first-passage time distribution of a modified random walk on the half plane with five transition probabilities $p_{1,} p_2, p_3, p_4, p_5$ ($p_i \geq 0$ for $i = 1, 2, \ldots, 5$; $\sum_{i=1}^{5} p_i = 1$) with barrier at $y = n$ (positive integer). Let X be a random variable which represents the number of steps xi the random walk. The GIT distribution arises from the movement of a particle from the origin with steps according to the transition probabilities until it first reaches the barrier n at the x-th step. The pmf of the GIT distribution is provided by

$$f_n(x) = \sum_{k=0}^{x} \sum_{t=0}^{\infty} \sum_{i=0}^{m} \frac{n}{n+x-i+k+2l} \binom{n+x-i+k+2l}{n-i+k+l, i, x-i-k, k, l} p_1^{n-i+k+l} p_2^i p_3^{x-i-k} p_4^k p_5^l$$

where $m = \min(n+k+l, x-k)$ and $\binom{x}{x_1, x_2, x_3, x_4, x_5} = x!/(x_1!x_2!x_3!x_4!x_5)$ is the multinomial coefficient . The GIT distribution generalized the inverse trinomial distribution considered by Shimizu and Yanagimoto (1991)

2. Properties of GIT distribution

The GIT family contains twenty-two possible distributions. The objective of this paper is to consider a particular class designated as $\text{GIT}_{3,1}$ for statistical analysis of count data. Since the $\text{GIT}_{3,1}$ distribution can cater for under, equi and over dispersion, it is compared with two popular distributions with this capability: generalized Poisson (GPD) (Consul, 1989) and COM-Poisson (Conway and Maxwell, 1962) distributions.The $\mathrm{GIT}_{3,1}$ pmf is given by

$$f_n(x)=\sum_{i=0}^{\min(n,x)}\frac{n}{n+x-i}\binom{n+x-i}{n-i,i,x-i}p_1^{n-i}p_2^{i}p_3^{x-i} \tag{1}$$

for $x=0,1,2,\dots,n$ is positive integer and $p_i\geq 0$ for $i=1,2,3$; $\sum_{i=1}^{3}p_i=1$.
Equation (1) may also be expressed as

$$f_n(x)=\binom{n+x-1}{x}p_1^n p_3^x\,{}_2F_1\left(-n,-x;-n-x+1;-\frac{p_2}{p_1p_3}\right) \tag{2}$$

in terms of the Gauss hypergeometric function ${}_2F_1$. The pmf in (2) is a negative binomial pmf weighted by ${}_2F_1$. Another alternative expression shows it as a binomial pmf weighted by ${}_2F_1$:

$$f_n(x)=\binom{n}{x}p_1^{n-x}p_2^{x}\,{}_2F_1\left(n,-x;n-x+1;-\frac{p_1p_3}{p_2}\right)\quad\text{if } n\geq x$$

$$=\binom{x-1}{n-1}p_2^{n}p_3^{x-n}\,{}_2F_1\left(x,-n;x-n+1;-\frac{p_1p_3}{p_2}\right)\quad\text{if } x>n.$$

The probability generating function (pgf) is

$$G_n(t)=\left(\frac{p_1+p_2t}{1-p_3t}\right)^n=\left(\frac{p_1+p_2t}{p_1+p_2}\right)^n\left(\frac{1-p_3}{1-p_3t}\right)^n,\quad 1-p_3=p_1+p_2 \tag{3}$$

Thus it is a member of Kemp's (1989) family of convolutions of binomial and pseudo-binomial variables. It reduces to a binomial pgf if $p_3=0$, negative binomial pgf if $p_2=0$ and shifted negative binomial (shifted *n* steps to the right) pgf if $p_1=0$.

The pmf $f_n(x)$ satisfies the following recurrence relation in *x*

$$f_n(x)=\left(a+\frac{b}{x}\right)f_n(x-1)+c\left(1-\frac{2}{x}\right)f_n(x-2),\quad x\geq 2 \tag{4}$$

with $f_n(0)=p_1^n$, $f_n(1)=np_1^{n-1}(p_1p_3+p_2)$, where

$a=(p_1p_3-p_2)/p_1$, $b=\{n(p_1p_3+p_2)-(p_1p_3-p_2)\}/p_1$, and $c=p_2p_3/p_1$, $p_1>0$. Recurrence relation (4) may be used to facilitate computation of the probabilities. For a discussion of computation by recurrence formulae see Ong (1995). The *r*th descending factorial moment of *X* is

$$\mu'_{[r]}=E(X(X-1)\dots(X-r+1)$$

$$=\frac{n}{(p_1+p_2)^r}\sum_{i=0}^{r}\binom{r}{i}\frac{(n+r-i-1)!}{(n-i)!}p_2^{i}p_3^{r-i}$$

for $r\geq 1$ and it satisfies the recursion formula

$$(p_1+p_2)^2\mu'_{[r+1]}+\{r(p_1+p_2)(p_2-p_3)-n(p_1+p_2)(p_2+p_3)\}\mu'_{[r]}-r(r-1)p_2p_3\mu'_{[r-1]}=0$$

with initial conditions $\mu'_{[0]}=1$, $\mu'_{[1]}=n\frac{p_2+p_3}{p_1+p_2}$. The mean and variance are found to be

$$E(X)=n\frac{p_2+p_3}{p_1+p_2},\;V(X)=n\frac{p_1p_2+p_3}{(p_1+p_2)^2}.$$

These lead to the Index of dispersion (*ID*),

$$ID = \frac{V(X)}{E(X)} = \frac{p_1 p_2 + p_3}{(p_1 + p_2)(p_2 + p_3)} \begin{cases} >1, & p_3 > p_2 \\ <1, & p_3 < p_2. \end{cases}$$

ID >1, $=1$ and <1 corresponds to over , equi- and under dispersion.

Note that if $p_2 = p_3 = p$, then *ID*=1 but $\mathrm{GIT}_{3,1}(n;1-2p,p,p)$ with $0<p<1/2$ is not a Poisson distribution. The $\mathrm{GIT}_{3,1}(n;1-2p,p,p)$ has a limiting Poisson distribution if $p \to 0, n \to \infty,\ np = \mathit{fixed}$.

In the limit the $\mathrm{GIT}_{3,1}$ distribution goes to a Poisson distribution with parameter $\lambda_2 + \lambda_3$ as *n* tends to infinity, provided $np_2 = \lambda_2$ and $np_3 = \lambda_3$.

<u>Log-concavity, Unimodality and Reliability Properties</u>

A distribution is said to be log-concave if its pmf $\{f_k\}$, $f_k > 0,\ \forall k$ satisfies $f_k^2 \ge f_{k+1} f_{k-1}\ \ \forall k$. The failure rate is defined by $r(k) = f_k / \sum_{i \ge k} f_i$.

From the results of Keilson and Geber (1971, page 388), the $\mathrm{GIT}_{3,1}$ distribution is log-concave. A number of reliability properties follow from log-concavity. The $\mathrm{GIT}_{3,1}$ distribution has an increasing failure rate (*IFR*) (Gupta et al. 2008). Furthermore the ensuing implications hold

$$IFR \Rightarrow IFRA \Rightarrow NBU \Rightarrow NBUE \Rightarrow HNBUE$$

where *IFRA* (increasing failure rate average), *NBU* (new better than used), *NBUE* (new better than used in expectation) and *HNBUE* (harmonic new better than used in expectation). Hence the $\mathrm{GIT}_{3,1}$ distribution is *IFR*, *IFRA*, *NBU*, *NBUE* and *HNBUE*.

From Theorem 3 of Keilson and Geber (1971, page 386), which states that a necessary and sufficient condition for pmf $\{f_k\}$ be strongly unimodal is that f_k be log-concave for all *k*, it follows that the $\mathrm{GIT}_{3,1}$ distribution is strongly unimodal.

3. Test for equi-dispersion

The index of dispersion shows that when $p_2 = p_3$, the $\mathrm{GIT}_{3,1}$ distribution is equi-dispersed. Thus, to test for equi-dispersion we consider the following set of hypotheses:

$$H_0 : p_2 = p_3$$

$$H_1 : p_2 \ne p_3$$

The test of hypothesis may be based upon Rao's score test or the likelihood ratio (LR) test where the test statistics have asymptotically a χ^2 distribution. A comparative study of the power of these two tests for the $\mathrm{GIT}_{3,1}$ distribution through a Monte Carlo simulation has been conducted. Some results of the simulation study are given in Table 1. It is observed that the power of the score test is marginally higher than that of the LR test for any tested sample sizes.

Table 1 Simulated power of Score and LR tests for $\mathbf{GIT_{3,1}}$

p_1	0.7	0.6	0.5	0.4	0.3	0.2	0.1
p_2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
p_3	0.2	0.3	0.4	0.5	0.6	0.7	0.8

		ID	1.125	1.286	1.500	1.800	2.250	3.000	4.500
N	α	Method				Power			
100	0.05	score	0.146	0.454	0.826	0.983	0.999	1.000	1.000
		LR	0.118	0.402	0.794	0.977	0.999	1.000	1.000
	0.1	score	0.220	0.563	0.882	0.992	1.000	1.000	1.000
		LR	0.196	0.521	0.862	0.989	1.000	1.000	1.000
500	0.05	score	0.475	0.975	1.000	1.000	1.000	1.000	1.000
		LR	0.445	0.971	1.000	1.000	1.000	1.000	1.000
	0.1	score	0.584	0.987	1.000	1.000	1.000	1.000	1.000
		LR	0.564	0.986	1.000	1.000	1.000	1.000	1.000
1000	0.05	score	0.744	1.000	1.000	1.000	1.000	1.000	1.000
		LR	0.727	1.000	1.000	1.000	1.000	1.000	1.000
	0.1	score	0.826	1.000	1.000	1.000	1.000	1.000	1.000
		LR	0.817	1.000	1.000	1.000	1.000	1.000	1.000

4. An Application

We illustrate an application of the distribution to the frequency distribution of 102 spiders under 240 boards (Consul, 1989, page 122). For parameter estimation, maximum likelihood, minimum Hellinger distance and pgf- based methods have been considered. The latter estimation method is considered because the $\text{GIT}_{3,1}$ distribution has a simple form for its pgf. The pgf-based estimators (Sim and Ong, 2010) used here are

$$T_1 = \int_0^1 \left(\sqrt{G_N(t)} - \sqrt{G(t)}\right)^2 dt \text{ and } T_2 = \int_0^1 \left(G_N(t) - G(t)\right)^2 dt \,,$$

where $G_N(t) = \frac{1}{N}\sum_{i=1}^{N} t^{x_i}$ and $G(t) = E\left[t^X\right]$ are respectively the empirical and theoretical pgf. The pgf-based estimators are consistent (Ng et al, 2013).

The fit of the distribution is compared with the GPD (Consul, 1989) and COM-Poisson (Conway and Maxwell, 1962) distributions. The GPD has pmf

$$f(X = x) = \begin{cases} \dfrac{\theta(\theta + x\lambda)^{x-1} e^{-\theta - x\lambda}}{x!} & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{for } x > m, \text{ when } \lambda < 0 \end{cases}$$

where $\theta > 0$, max $(-1, -\theta/m) < \lambda \le 1$ and m (≥ 4) is the largest positive integer satisfying $\theta + m\lambda > 0$ when λ is negative. Note that for λ < 0 the probabilities do not sum to 1. The pmf of COM-Poisson distribution is given by

$$f(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}$$

where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ for $\lambda > 0$ and $\nu \ge 0$. The COM-Poisson distribution is under dispersed or over dispersed when $\nu > 1$ or $\nu < 1$ respectively. The case $\nu = 1$ gives the Poisson distribution.

Table 2 Frequency distribution of 102 spiders under 240 boards (Consul, 1989).

		Expected frequency			
Number of spiders	Observed frequency	GPD	COM-Poisson	$\text{GIT}_{3,1}$	$\text{GIT}_{3,1}$ pgf-based

		MLE	MLE	MLE	MHD	T_1	T_2
0	159	159.05	159.15	159.00	158.42	158.98	158.98
1	64	65.23	63.11	64.32	63.07	64.08	64.10
2	13	13.58	14.80	13.24	14.31	13.40	13.39
3	4	2.14	2.95	3.43	4.20	3.54	3.54
χ^2		1.67	0.60	0.10	0.15	0.07	0.07
p-value		0.19	0.44	0.75	0.70	0.79	0.79
				$\hat{p}_2$	$\hat{p}_2$	$\hat{p}_2$	$\hat{p}_2$
Parameter	$n=1$	$\hat{\theta}$=0.411	$\hat{v}$=0.758	=0.132	=0.113	=0.1286	=0.1287
				$\hat{p}_3$	$\hat{p}_3$	$\hat{p}_3$	$\hat{p}_3$
estimates		$\hat{\lambda}$=0.032	$\hat{\lambda}$=0.396	=0.206	=0.227	=0.2090	=0.2089
ID				1.07			
LR test				Do not reject H_0 with T= 0.81			

With an *ID* slightly larger than 1, the LR test does not to reject the null hypothesis of equi-dispersion. The p-values showed that the $\text{GIT}_{3,1}$ distribution fits significantly better than GPD and COM-Poisson distribution.

5. Conclusions

The $\text{GIT}_{3,1}$ distribution has a stochastic origin as a first-passage time distribution of a modified random walk on the half plane. The $\text{GIT}_{3,1}$ distribution is able to cater for under-, equi- and over-dispersion. Some important probabilistic properties have been derived. Although the $\text{GIT}_{3,1}$ distribution has a complicated pmf in terms of the Gauss hypergeometric function, it has a simple three-term recurrence formula to facilitate computation without requiring the computation of normalizing constant. Furthermore it also has a simple pgf which allows parameter estimation by a pgf-based minimum Hellinger-type distance estimation which is simple to implement. The good fit to a real data when compared with the well-known GPD and COM-Poisson distribution justifies its inclusion as a viable and flexible model for count data analysis.

References

Aoyama K., Shimizu K., Ong S.H. (2008). A first–passage time random walk distribution with five transition probabilities: a generalization of the shifted inverse trinomial. *Annals Inst. Stat. Math.*, 60, 1-20

Consul, P.C. (1989). Generalized Poisson Distributions: Properties and Applications. Marcel Dekker Inc., New York/Basel

Conway, R. W. and Maxwell, W. L. (1962). A queueing model with state dependent service rates. *J. Indust. Eng.*, 12, 132–136.

Gupta, P.L., Gupta, R.C., Ong, S.H., Srivastava, H.M. (2008). A class of Hurwitz-Lerch-Zeta distribution and their applications in reliability. *Applied Maths. and Computation*, 196, 521-531.

Keilson, J., Geber H. (1971). Some Results for Discrete Unimodality. *Journal of the American Statistical Association*, 66, 386-389.

Kemp, A.W. (1979). Convolutions involving binomial pseudo-variables. *Sankya*, 41, 232-243

Ng, C.M., Ong, S.H., Srivastava, H.M. (2013). Parameter Estimation by Hellinger Type Distance for Multivariate Distributions based on Probability Generating Functions. *Applied Mathematical Modelling*, 37, 7374–7385.

Ong, S.H. (1995). Computation of probabilities of a generalized log-series and related distributions. *Communications in Statistics-Theory and Methods*, 24, 253-271.

Shimizu, K., Yanagimoto, T. (1991). The inverse trinomial distribution. *Jap. J. Appl. Stat.*, 20, 89-96, In Japanese

Sim, S. Z., Ong, S. H. (2010). Parameter Estimation for Discrete Distributions by Generalized Hellinger-Type Divergence Based on Probability Generating Function. *Communications in Statistics - Simulation and Computation*, 39, 305- 314.