$\left(\begin{smallmatrix} \dot{\;}\dot{\;}\dot{\;} \\ \text{M} \\ \text{CS} \\ \dot{\;}\dot{\;} \end{smallmatrix}\right)$

An R implementation for estimating the parameters of the discrete transmuted generalized inverse Weibull distribution

Atchanut Rattanalertnusorn, Sirinapa Aryuyuen

Department of Mathematics and Computer Science
Faculty of Science and Technology
Rajamangala University of Technology Thanyaburi
Pathum Thani, Thailand

email: atchanut_r@rmutt.ac.th, sirinapa_a@rmutt.ac.th

Abstract

Recently, one discretization of a continuous distribution, the so-called discrete transmuted generalized inverse Weibull (DTGIW) distribution, has been suggested. In this article, we propose the application of the DTGIW distribution along with using the `dtgiw` package in the R language program to be analyzed the discretized data or count data. We estimate the model parameters by using the maximum likelihood (ML), Anderson-Darling (AD), and least-square (LS) methods. We compare the performances of the proposed estimators based on numerical calculations for various values of the distribution parameters, and sample sizes in terms of the root mean squared error and estimated values. Simulation results show that the ML method is most efficient for a large sample size. We analyze two applications to illustrate the applicability of the DTGIW distribution with the results showing that the DTGIW distribution is an alternative flexible in analyzing these count data.

1   Introduction

The count data sets arise in different fields such as yearly number destructive earthquakes, monthly traffic accidents, hourly bacterial growth, among others. These kind of data sets are modeled with discrete probability distributions. A Poisson distribution is the most popular distributions and is widely used to model these kind data sets.

In probability distribution theory, vast numbers of discrete distributions have specified applications. However, at times, the observable data have distinct features that these classical discrete distributions do not exhibit. So to overcome these limitations, researchers often develop new probability distributions so that these new distributions can be employed in those situations where classical distributions are not providing any good fit. In Statistics literature, different researchers have used different discretization methods to propose a discrete type of distribution analogs of continuous distribution [1].

In recent years, researchers have shown great interest in introducing new discrete distributions by discretizing a continuous failure time model. Many researchers have introduced discrete distributions by applying the discretization method to the continuous failure time models such as the discrete Lindley (DL) [2], the discrete inverse Weibull [3], etc.

One of the discretization methods has been used to find discrete analogous continuous transmuted generalized inverse Weibull distribution introduced by Rattanalertnusorn and Aryuyuen [4], the so-called the discrete transmuted generalized inverse Weibull (DTGIW) distribution. The distribution is an alternative flexible model to analyze count data. In this paper, we show the capabilities of proposing an R implementation for estimating the parameters of the DTGIW distribution along with using the `dtgiw` package in the R language program to do analysis of discretized data or count data. The researcher will use the R language program, the most popular data science language. There are many packages and libraries provided for doing different tasks. It is necessary to develop statistical tools that are convenient and quick to use and to the point. In the development process, the choice of tools and language for development must be flexible and responsive to the theory that has been developed. Three methods of parameter estimation of the DTGIW distribution are compared. Simulation and application studies are illustrated.

2　The DTGIW Distribution

Let X be a random variable distributed as the DTGIW distribution with parameter vector $\boldsymbol{\Theta}$ where $\boldsymbol{\Theta} = (\alpha, \beta, \lambda, \theta)^T$, denoted by $X \sim$ DTGIW $(\alpha, \beta, \lambda, \theta)$. Its probability density function (pdf) is

$$f(x; \boldsymbol{\Theta}) = (1 + \theta) \left\{ \exp\left[-\lambda(\beta x + \beta)^{-\alpha}\right] - \exp\left[-\lambda(\beta x)^{-\alpha}\right] \right\}$$
$$- \theta \left\{ \exp\left[-2\lambda(\beta x + \beta)^{-\alpha}\right] - \exp\left[-2\lambda(\beta x)^{-\alpha}\right] \right\}, \qquad (2.1)$$

where $x = 0, 1, 2, \ldots$, the positive parameters α, β and λ, and $-1 \leq \theta \leq 1$. Its corresponding cumulative density function (cdf) is

$$F(x; \boldsymbol{\Theta}) = \exp\left[-\lambda(\beta x + \beta)^{-\alpha}\right] \left\{ 1 + \theta - \theta \exp\left[-\lambda(\beta x + \beta)^{-\alpha}\right] \right\}. \qquad (2.2)$$

The DTGIW distribution is a flexible model that approaches different distributions when its parameters are changed [4]. It has eight sub-models as Table 1.

Table 1: Sub-models of the DTGIW distribution

Parameters	Distributions
$(\alpha, \beta, \lambda, \theta)$	DTGIW
$(\alpha, \beta, \lambda = 1, \theta)$	Discrete transmuted inverse Weibull (DTIW)
$(\alpha, \beta, \lambda = 1, \theta = 0)$	Discrete inverse Weibull (DIW) [3]
$(\alpha = 1, \beta, \lambda = 1, \theta)$	Discrete transmuted inverse exponential (DTIE)
$(\alpha = 1, \beta, \lambda = 1, \theta = 0)$	Discrete inverse exponential (DIE)
$(\alpha = 2, \beta, \lambda = 1, \theta)$	Discrete transmuted inverse Rayleigh (DTIR)
$(\alpha = 2, \beta, \lambda = 1, \theta = 0)$	Discrete inverse Rayleigh (DIR)
$(\alpha, \beta = 1, \lambda, \theta)$	Discrete transmuted Fréchet (DTFr)
$(\alpha, \beta = 1, \lambda, \theta = 0)$	Discrete Fréchet (DFr)

In terms of the distribution function F, the quantile function Q returns the value x such that $F(x; \boldsymbol{\Theta}) = p$ for $0 < p < 1$. The quantile function of the DTGIW distribution is

$$Q(p; \boldsymbol{\Theta}) = F^{-1}(p; \boldsymbol{\Theta}). \qquad (2.3)$$

From the quantile function in (2.3), a random variate generation of the DTGIW distribution is obtained as

$$X_i = Q(u_i; \boldsymbol{\Theta}); 0 \leq u_i \leq 1, i = 1, 2, 3, \ldots, n, \qquad (2.4)$$

where u_i is a value of a uniform random variable on [0,1].

3   R Package for the DTGIW Distribution

In 2022, Rattanalertnusorn and Aryuyuen [5] provided a new contributed package for R that is the `dtgiw` package. The `dtgiw` package requires R version 4.1.3 or higher. There is a function that allows users to install it directly-R program. Therefore, the step to install the `dtgiw` package on R [6] or RStudio [7] is as follows:

```
install.packages("dtgiw")
library(dtgiw)
install.packages("graphics")
library(graphics)
install.packages("stats")
library(stats)
```

After installing the `dtgiw` package, users can use "`help(dtgiw)`" command for getting an R package manual. This new package includes the eight distributions (i.e., DTIW, DIW, DTIE, DIE, DTIR, DIR, DTFr, and DFr distributions) and the corresponding calling sequences to compute the probability density function, random numbers, and distribution function for some distributions. The `dtgiw` package consists of six functions:

(1) `dDTGIW(x,alpha,beta,lambda,theta)` is the pdf of the DTGIW distribution as in (2.1),

(2) `pDTGIW(q,alpha,beta,lambda,theta)` is the cdf of the DTGIW distribution as in (2.2),

(3) `qDTGIW(p,alpha,beta,lambda,theta)` is the quantile function of the DTGIW distribution as in (2.3),

(4) `rDTGIW(n,alpha,beta,lambda,theta)` is the n random number of the DTGIW distribution as in (2.4).

(5) `loglikeDTGIW(x,alpha,beta,lambda,theta)` is the negative log-likelihood value of the DTGIW distribution.

(6) `plotDTGIW(x,alpha,beta,lambda,theta)` is the graph for the pmf of the DTGIW distribution.

In the calling sequence for using the functions `x` must be a vector of data values; `q` must be a vector of quantiles; `p` must be a vector of probabilities; `n` must be a number of observations; `alpha` must be a value of the parameter α; `beta` must be a value of the parameter β; `lambda` must be a value of the parameter λ; `theta` must be a value of the parameter θ.

4 An R Implementation for Estimating the Parameters of the DTGIW Distribution

In this section, three-parameter estimation methods, the maximum likelihood (ML), Anderson-Darling (AD), and least-square (LS) estimators, are used to estimate the parameters of the DTGIW distribution.

4.1 The ML method for the DTGIW distribution

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed (i.i.d.) sample of size n with pdf $f(x_i; \boldsymbol{\Theta})$ as (2.1), the likelihood function can be expressed as

$$L(\boldsymbol{\Theta}; x_i) = \prod_{i=1}^{n} (1 + \theta) \left\{ \exp\left[-\lambda(\beta x_i + \beta)^{-\alpha} \right] - \exp\left[-\lambda(\beta x_i)^{-\alpha} \right] \right\}$$
$$- \theta \left\{ \exp\left[-2\lambda(\beta x_i + \beta)^{-\alpha} \right] - \exp\left[-2\lambda(\beta x_i)^{-\alpha} \right] \right\},$$

where $\boldsymbol{\Theta} = (\alpha, \beta, \lambda, \theta)^T$. The components of the unit score vector are obtained by taking the partial derivatives of the $\ell(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}; x_i)$ with respect to each parameter, and the ML estimates of parameters can be obtained by setting the score functions equal to zero, i.e.,

$$\frac{\partial \ell(\boldsymbol{\Theta})}{\partial \alpha} = 0, \frac{\partial \ell(\boldsymbol{\Theta})}{\partial \beta} = 0, \frac{\partial \ell(\boldsymbol{\Theta})}{\partial \lambda} = 0, \text{ and } \frac{\partial \ell(\boldsymbol{\Theta})}{\partial \theta} = 0.$$

In this work, we employ the `optim` function in R language and the `dDTGIW` function in the `dtgiw` package [5] to obtain the ML estimates of α, β, λ and θ as the following.

```
#============ML method for the DTGIW distribution============#
logDTGIW<-function(x,par){
  alpha<-par[1]; beta<-par[2]; lambda<-par[3]; theta<-par[4];
  loglike<--sum(log(dDTGIW(x,alpha,beta,lambda,theta)))
  return(loglike)
} optim(par=t.start,fn=logDTGIW,x=x)
```

4.2 The AD method for the DTGIW distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. sample of size n, and $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is ordered random sample with cdf $F(x_i; \boldsymbol{\Theta})$ as (2.2). The estimators of each

parameter for the DTGIW distribution is obtained by minimizing the function as follows:

$$A(\Theta) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left\{\log\left[F\left(x_{(i)};\Theta\right)\right]\right.$$
$$\left. + \log\left[1 - F\left(x_{(n+1-i)};\Theta\right)\right]\right\}.$$

The AD estimators are calculated by solving the non-linear equations as follows:

$$\frac{\partial A(\Theta)}{\partial \alpha} = 0, \frac{\partial A(\Theta)}{\partial \beta} = 0, \frac{\partial A(\Theta)}{\partial \lambda} = 0, \text{ and } \frac{\partial A(\Theta)}{\partial \theta} = 0.$$

In this work, we employ the `nlm` function in the R language and the `pDTGIW` function in the `dtgiw` package [5] to obtain the AD estimates of α, β, λ and θ as the following.

```
#===========AD method for the DTGIW  distribution===========#
ADE_DTGIW<-function(x,par){
  alpha<-par[1]; beta<-par[2]; lambda<-par[3]; theta<-par[4];
  sx<-sort(x); r<-length(sx); L<-numeric()
  for(v in 1:r){
   L[v]<-((2*v)-1)*(log(pDTGIW(sx[v],alpha,beta,lambda,theta))
          +log(1-pDTGIW(sx[r+1-v],alpha,beta,lambda,theta)))
  } M<--r-(1/r)*sum(L)
  return(M)
} optim(par=t.start,fn=ADE_DTGIW,x=x)
```

4.3 The LS method for the DTGIW distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. sample of size n, and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be an ordered random sample with cdf $F(x_i;\Theta)$ as (2.2). The estimators of each parameter for the DTGIW distribution is obtained by minimizing the function as follows:
Give $R_i = F(x_{(i)}|\Theta)$ and the corresponding expected value $E(R_i) = i/(n+1)$ where $i = 1, \ldots, n$. Thus LS estimators of the parameters α, β, λ, and θ are obtained by minimizing sum of a quadratic form in the difference between the theoretical distribution function, and their expected, where

$$SSE(\Theta) = \sum_{i=1}^{n}\left[\left(F(x_{(i)}|\Theta)\right) - \frac{i}{n+1}\right]^2.$$

The LS estimators are calculated by solving the non-linear equations which are obtained by setting the partial derivatives to zero as follows:

$$\frac{\partial SSE(\boldsymbol{\Theta})}{\partial \alpha} = 0, \frac{\partial SSE(\boldsymbol{\Theta})}{\partial \beta} = 0, \frac{\partial SSE(\boldsymbol{\Theta})}{\partial \lambda} = 0, \text{ and } \frac{\partial SSE(\boldsymbol{\Theta})}{\partial \theta} = 0.$$

We employ the `optim` function in the R language and the `pDTGIW` function in the `dtgiw` package [5] to obtain the LS estimates of α, β, λ and θ as follow:

```
#============LS method for the DTGIW  distribution==========#
LS_DTGIW<-function(x,par){
  alpha<-par[1]; beta<-par[2]; lambda<-par[3]; theta<-par[4];
  sx<-sort(x); r<-length(sx); L<-numeric()
  for(v in 1:r){
  L[v]<-pDTGIW(sx[v],alpha,beta,lambda,theta)-(v/(r+1))
  } return(sum(L^2))
} optim(par=t.start,fn=LS_DTGIW,x=x)
```

5   Simulation Study for Estimating the Parameters of the DTGIW Distribution

Simulation studies are computer experiments that involve creating data by pseudo-random sampling from known probability distributions. They are invaluable for statistical research, particularly for comparing alternative methods [8]. In this section, we compare the performances of the proposed estimators (i.e., ML, AD, and LS methods) of parameters α, β, λ and θ. This comparison is carried out by taking random samples of different sizes ($n = 30$, 60, 100, and 200) with various parameters values as in Figure 1. Experiments consider each case with 1,000 replications. The estimators are compared in terms of the estimated values' root mean square errors (RMSE). The results are summarized in tables 2-3.

Based on tables 2-3, note that:
(i) The RMSE values are decreasing as the sample sizes values are increasing for all cases considered in this section.
(ii) The estimated values of the suggested estimators are close to the true parameter as the sample sizes increase for large n.
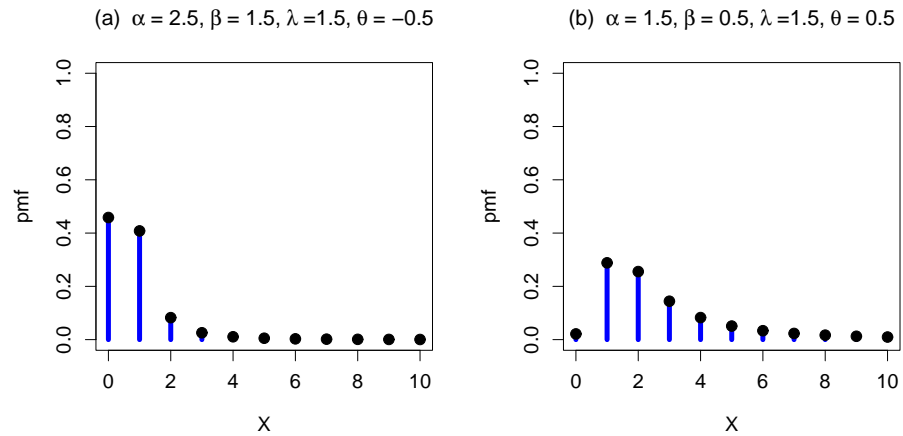(iii) Observe that the ML method gives the RMSE values less than the AD and LS methods.

Figure 1: The pmf plots for the DTGIW distribution with the specified values of α, β, λ and θ.

Table 2: Estimates (Est) and RMSE values of parameters of the DTGIW distribution for $\alpha = 2.5$, $\beta = 1.5$, $\lambda = 1.5$, and $\theta = -0.5$

n		ML		AD		LS	
		Est	RMSE	Est	RMSE	Est	RMSE
30	α	2.5815	0.5813	2.1046	0.1839	1.9752	0.3039
	β	1.4485	0.0833	1.3083	0.0477	1.3462	0.0363
	λ	1.5376	0.1964	1.9692	0.3285	1.9430	0.2283
	θ	-0.3541	0.0899	-0.3916	0.0673	-0.4586	0.0240
60	α	2.5515	0.2158	2.1073	0.1634	2.0078	0.2512
	β	1.4724	0.0173	1.2037	0.0898	1.2207	0.0804
	λ	1.5131	0.0035	1.6357	0.0189	1.6201	0.0148
	θ	-0.4641	0.0050	-0.4581	0.0088	-0.4592	0.0080
100	α	2.5595	0.1305	2.1087	0.1595	2.0178	0.2387
	β	1.4642	0.0130	1.1969	0.0933	1.2082	0.0869
	λ	1.5140	0.0023	1.6319	0.0177	1.6097	0.0122
	θ	-0.4745	0.0008	-0.4366	0.0110	-0.4638	0.0052
200	α	2.5053	0.0482	2.1011	0.1620	2.0336	0.2212
	β	1.4706	0.0069	1.1898	0.0969	1.2091	0.0854
	λ	1.5111	0.0012	1.6121	0.0127	1.6216	0.0149
	θ	-0.4912	0.0002	-0.4498	0.0098	-0.4381	0.0115

Table 3: Estimates (Est) and RMSE values of parameters of the DTGIW distribution for $\alpha = 1.5$, $\beta = 0.5$, $\lambda = 1.5$, and $\theta = 0.5$

n		ML		AD		LS	
		Est	RMSE	Est	RMSE	Est	RMSE
30	α	1.6682	0.1182	1.6088	0.0788	1.6364	0.0856
	β	0.5473	0.0110	0.4116	0.0194	0.4274	0.0146
	λ	1.5775	0.0737	1.5736	0.0507	1.5593	0.0597
	θ	0.3295	0.0811	0.5404	0.0964	0.4747	0.0928
60	α	1.5961	0.0469	1.6044	0.0644	1.6524	0.0630
	β	0.5243	0.0090	0.3990	0.0167	0.4156	0.0122
	λ	1.5594	0.0324	1.5530	0.0473	1.5381	0.0429
	θ	0.3837	0.0801	0.5340	0.0857	0.5175	0.0828
100	α	1.5307	0.0382	1.5960	0.0495	1.6407	0.0538
	β	0.5311	0.0052	0.3879	0.0136	0.4047	0.0119
	λ	1.5645	0.0276	1.5338	0.0370	1.5350	0.0362
	θ	0.4542	0.0641	0.5874	0.0694	0.5445	0.0953
200	α	1.5121	0.0288	1.5891	0.0363	1.6264	0.0449
	β	0.5249	0.0040	0.4154	0.0115	0.4027	0.0114
	λ	1.5168	0.0223	1.5630	0.0316	1.5316	0.0340
	θ	0.4962	0.0032	0.6285	0.0607	0.5173	0.0871

6 Application on Real Data

In this section, we discuss the goodness of fit of the DTGIW with two count data sets. The first data set is the number of insects from Kemp [9] as in Table 4 (see [1]). The second data set is the number of sperm in eggs in the UK coal mining industry in successive four-week periods in the years 1948-1959 from Kendall [10] as in Table 5.

We compare the proposed distribution with the DL distribution with a positive parameter γ [2], the DIW distribution [3] with positive parameters α and β, and the Poisson distribution with a positive parameter μ. The Kolmogorov-Smirnov (KS) and Cramer-von Mises (CVM) tests compare fitting distributions, where the smaller values of test statistics give the best fit for the data. From the results in Tables 4-5, the DTGIW distribution gives the lowest statistic values of KS and CVM tests. Hence, the DTGIW distribution could be a good alternative to explaining two data sets.

Table 4: Observed and expected frequencies for distribution of the number of insects (X) from Kemp [9]

	Observed	Expected frequency			
X	frequency	Poisson	DL	DIW	DTGIW
0	33	26.45	29.62	32.62	32.58
1	12	19.84	16.16	14.66	14.64
2	6	7.44	6.61	4.11	4.16
3	3	1.86	2.40	1.71	1.73
4	1	0.35	0.82	0.89	0.89
5	1	0.05	0.27	0.52	0.52
ML estimates		$\hat{\mu} = 0.7500$	$\hat{\gamma} = 1.2994$	$\hat{\alpha} = 1.6745$	$\hat{\alpha} = 1.7014$
				$\hat{\beta} = 1.4441$	$\hat{\beta} = 1.1726$
					$\hat{\lambda} = 0.5308$
					$\hat{\theta} = -0.3839$
KS test		0.1169	0.0603	0.0407	0.0397
(p-value)		(0.4282)	(0.9870)	(0.9999)	(0.9999)
CVM test		0.3357	0.0887	0.0290	0.0282
(p-value)		(0.1016)	(0.4760)	(0.8060)	(0.8124)

7   Conclusions

In this article, we proposed the R implementation for the DTGIW distribution [4] along with using the `dtgiw` package [5] in the R program to do an analysis of the count data. The `dtgiw` package includes the probability density function, distribution function, quantile function, and random generation procedure for the DTGIW distribution. We compared three methods of parameter estimation of the DTGIW distribution, including the ML, AD, and LS methods. Simulation results showed that the ML method was better than the LS and AD methods. Finally, the results of the count data analysis for two data sets showed that the DTGIW distribution was better than the DIW, DL, and Poisson distributions.

Table 5: Observed and expected frequencies for distribution of the nnumber of sperm in eggs (X) [10]

X	Observed frequency	Expected frequency			
		Poisson	DL	DIW	DTGIW
0	28	37.23	41.94	27.76	27.74
1	44	28.48	23.15	45.16	45.13
2	7	10.89	9.58	5.31	5.40
3	1	2.78	3.53	1.12	1.11
ML estimates		$\hat{\mu} = 0.7650$	$\hat{\gamma} = 1.2875$	$\hat{\alpha} = 3.5146$	$\hat{\alpha} = 3.5925$
				$\hat{\beta} = 0.9840$	$\hat{\beta} = 1.3323$
					$\hat{\lambda} = 2.2492$
					$\hat{\theta} = -0.4103$
KS test		0.1153	0.1742	0.0116	0.0109
(p-value)		(0.2377)	(0.0155)	(0.9999)	(0.9999)
CVM test		0.5663	1.1410	0.0049	0.0045
(p-value)		(0.0295)	(0.0017)	(0.9689)	(0.9720)

References

[1] B. Abebe, K. K. Shukla, A discrete Pranav distribution and its applications. Biom. Biostat. Int. J, **8**, no. 1, (2019), 33–37.

[2] B. Abebe, R. Shanker, A discrete Lindley distribution with applications in Biological Science, Biom. Biostat. Int. J., **7**, no. 1, (2018), 1–5.

[3] M. A. Jazi, C. D. Lai, M. H. Alamatsaz, A discrete inverse Weibull distribution and estimation of its parameters. Stat. Methodol., **7**, no. 2, (2010), 121–132.

[4] A. Rattanalertnusorn, S. Aryuyuen, The zero-truncated discrete transmuted generalized inverse Weibull distribution and its applications, Songklanakarin J. Sci. Tech., **43**, no. 4, (2021), 1140–1151.

[5] A. Rattanalertnusorn, S. Aryuyuen, dtgiw: Discrete transmuted generalized inverse Weibull distribution, R package version 1.0.0, https://CRAN.R-project.org/package=dtgiw, 2022.

[6] R Core Team, R: a language and environment for statistical computing.R foundation for statistical computing, Vienna, Austria, https://R-project.org/, 2022.

[7] RStudio Team, RStudio: Integrated development for R, 2022-03-10, https://rstudio.com/, 2022.

[8] T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical methods. Stat. Med., **38**, no. 11, (2019), 2074–2102.

[9] C. D. Kemp, A. W. Kemp, Some properties of the Hermite distribution, Biometrika., **52**, nos. 3-4, (1965), 381–394.

[10] M. G. Kendall, Natural law in social sciences, J. R. Stat. Soc. Ser. A Stat. Soc., **126**, (1961), 124–119,