STATISTICA, anno LXXVIII, n. 1, 2018

A COUNTING PROCESS WITH GENERALIZED EXPONENTIAL INTER-ARRIVAL TIMES

Sahana Bhattacharjee [1]
Department of Statistics, Gauhati University, Guwahati, Assam, India

1. INTRODUCTION

The Poisson process is one of the most popular counting processes, which is based on the postulates viz. stationary and independent increments and the number of events in any interval of length 't' is Poisson distributed with mean 'λt', where λ is the mean number of occurrences per unit time (Ross, 1995). Another important characteristic of the Poisson process is that the inter-arrival times are exponentially distributed. It has found extensive applications in queuing theory (Kingman, 1963), modeling scoring in a hockey game (DeJardine, 2013), modeling vehicles involved in road toll (Vong, 2013) etc.

However, the Poisson count model suffers from a major drawback of being valid only when the underlying data is equi-dispersed, i.e. when the mean and variance of the count data are the same (McShane *et al.*, 2008). This limitation has been addressed by many statisticians and over the years, count models have been developed which allow modeling of over-dispersed data (variance greater than the mean) and under-dispersed data (mean greater than the variance). A heterogeneous gamma Poisson or the negative binomial count model is the oldest model which addressed the issue of over-dispersion. Out of the several models developed for handling under-dispersed data, the one proposed by Winkelmann (1995) is based on gamma distributed inter-arrival times, which beautifully explores the Poisson-exponential connection. McShane *et al.* (2008) proposed a count model assuming Weibull distributed inter-arrival times, which handles both over-dispersed and under-dispersed data. Also, this model nests the Poisson model and the negative binomial count model as special cases.

The inter-arrival time of the Poisson count model is exponentially distributed, and so, has a constant hazard function. But in real-life, the hazard may not remain constant over time. In such a case, the Poisson model will not be appropriate. The hazard function which expresses the instantaneous exit probability, conditional on survival,

[1] Corresponding Author. E-mail: sahana.bhattacharjee@hotmail.com

captures the underlying time dependence of the count model (Jose and Abraham, 2013). Winkelmann (1995) carried out an extensive analysis of the timing model hazard function and the dispersion in the equivalent count model and found that a decreasing hazard function (corresponding to negative duration dependence of the inter-arrival time distribution) leads to an over-dispersed count model, an increasing hazard function (corresponding to positive duration dependence) leads to an under-dispersed count model and a constant hazard function (corresponding to no duration dependence) leads to an equi-dispersed count model (Jose and Abraham, 2011). Thus, a new count model can be derived by assuming some other inter-arrival time distribution, which possesses a non-constant hazard function.

In this paper, a new generalized count model is developed assuming the generalized exponential (GE) distribution as the inter-arrival time distribution, of which the exponential distribution is a special case. A corresponding count model is formulated which nests the traditional Poisson count model. The advantage of using the GE distribution over the exponential distribution is that the hazard function is non-constant and hence, the distribution is duration dependent. This allows the corresponding count model to account for over-dispersed and under-dispersed data. Some properties of the new proposed model are explored and simulation from the model is carried out. Finally, the application of the new model is illustrated with the help of two real-life data sets.

2. GENERALIZED EXPONENTIAL DISTRIBUTION

The three-parameter generalized exponential (GE) distribution (the parameters being location, scale and shape parameter) was proposed by Gupta and Kundu (1999), as an alternative to the gamma and Weibull distribution. The gamma distribution has the limitation of not having a closed form of the cumulative distribution function (and hence, the survival functions and the hazard function). The Weibull distribution, although has a easily computable form of the cumulative distribution, survival and hazard function, does not possess the likelihood ratio ordering property. In addition, there does not exist a UMP test for testing the one-sided hypothesis on the shape parameter when the other two parameters are known. The GE distribution takes care of these limitations and have properties similar to those of gamma and Weibull distribution. It has found use in analyzing lifetime data and in the field of medical science, among numerous other applications.

The random variable X has a GE distribution with parameters α, λ and μ if it has the distribution function

$$F_{GE}(x;\alpha,\lambda,\mu)=\left[1-\exp\left\{-\left(\frac{x-\mu}{\lambda}\right)\right\}\right]^{\alpha} \qquad x>\mu,\alpha,\lambda>0. \tag{1}$$

The density function corresponding to the distribution function (1) is

$$f_{GE}(x;\alpha,\lambda,\mu)=\frac{\alpha}{\lambda}\left[1-\exp\left\{-\left(\frac{x-\mu}{\lambda}\right)\right\}\right]^{\alpha-1}\exp\left\{-\left(\frac{x-\mu}{\lambda}\right)\right\}\ (x>\mu;\alpha,\lambda>0). \tag{2}$$

Here, α, λ, μ is the shape parameter, scale parameter and location parameter respectively and it is denoted by GE(α, λ, μ).

Putting $\alpha = 1$ and $\mu = 0$ in (2) yields the p.d.f of the exponential distribution with scale parameter λ, which is

$$f(x;\lambda) = \frac{1}{\lambda}\exp\left(\frac{-x}{\lambda}\right). \tag{3}$$

As discussed in Gupta and Kundu (1999), the hazard function of the GE($\alpha, \lambda, 0$) distribution is given by

$$h(x;\alpha,\lambda,0) = \frac{\alpha}{\lambda}\frac{\left(1-e^{-\frac{x}{\lambda}}\right)^{\alpha-1}e^{-\frac{x}{\lambda}}}{1-\left(1-e^{-\frac{x}{\lambda}}\right)^{\alpha}}.$$

$h(x;\alpha,\lambda,0)$ is an increasing function if $\alpha < 1$, a decreasing function for $\alpha > 1$ and constant for $\alpha = 1$. The same authors have displayed the use of the GE distribution in modeling the endurance of deep groove ball bearings.

3. GENERALIZED EXPONENTIAL COUNT MODEL

Let Z_n denotes the interval between the $(n-1)^{th}$ and n^{th} occurrence of a process $\{N(t), t \geq 0\}$ and let the sequence $Z_1, Z_2, \ldots, Z_n$ be independently and identically distributed random variables having the GE(α,1,0) distribution. Then, the sum $W_n = Z_1 + Z_2 + \ldots + Z_n$ represents the waiting time up to the n^{th} occurrence or the time from the origin of the process to the n^{th} subsequent occurrence.

If $Z_i's$ are independently and identically distributed such that $Z_i \sim GE(\alpha, 1, 0)$, then it can be seen that $Z = \sum_{i=1}^{n} Z_i$ has the density function given by (Gupta and Kundu, 1999)

$$\begin{aligned} g_Z(z) &= \sum_{j=0}^{\infty} C_j (n\alpha + j)\exp(-z)\{1-\exp(-z)\}^{n\alpha+j-1} \\ &= \sum_{j=0}^{\infty} C_j f_{GE}(z; n\alpha + j, 1, 0) \end{aligned} \tag{4}$$

where the constants C_j are defined as

$C_0 = \frac{[\Gamma(\alpha+1)]^n}{\Gamma(1+n\alpha)}, C_j = \frac{C_0 n\alpha}{(n\alpha+j)} C_j^{(n)}; j = 1,2,\ldots, C_j^{(2)} = \frac{\left[(\alpha)_j\right]^2}{j!(2\alpha)_j},$

$C_j^{(k)} = \frac{(\{k-1\}\alpha)_j}{(k\alpha)_j} \sum_{i=0}^{j} \frac{(\alpha)_i}{i!} C_{j-i}^{(k-1)}; k = 3, \ldots n.$

Here, $(\alpha)_j = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}$.

Thus, W_n, the waiting time up to the n^{th} occurrence of the process $\{N(t), t \geq 0\}$ has the density function given in (4). The distribution function of the waiting time W_n is given by

$$\begin{aligned} F_{w_n}(t) &= P\{W_n \leq t\} \\ &= 1 - P(W_n(t)\} \\ &= 1 - P\{N(t) < n\} \\ &= 1 - P\{N(t) \leq (n-1)\} \\ &= 1 - F_{N(t)}(n-1). \end{aligned}$$

Therefore

$$\begin{aligned} F_{N(t)}(n-1) &= 1 - F_{w_n}(t) \\ &= 1 - \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{n\alpha+j}. \end{aligned} \tag{5}$$

Finally, the probability law of $N(t)$ is

$$\begin{aligned} p_n(t) &= P\{N(t) = n\} \\ &= F_{N(t)}(n) - F_{N(t)}(n-1) \\ &= 1 - \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{(n+1)\alpha+j} - 1 + \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{n\alpha+j} \\ &= \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{n\alpha+j} [1 - \{1-\exp(-t)\}^{\alpha}]. \end{aligned}$$

THEOREM 1. *If the inter-arrival times are independently and identically distributed as generalized exponential distribution with parameters α, $\lambda = 1$ and $\mu = 0$, then the count model probabilities are given by*

$$p_n(t) = P\{N(t) = n\} = \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{n\alpha+j} [1 - \{1-\exp(-t)\}^{\alpha}], \tag{6}$$

where $C_0 = \frac{[\Gamma(\alpha+1)]^n}{\Gamma(1+n\alpha)}$, $C_j = \frac{C_0 n\alpha}{(n\alpha+j)} C_j^{(n)}; j = 1, 2, \ldots, C_j^{(2)} = \frac{[(\alpha)_j]^2}{j!(2\alpha)_j}$,
$C_j^{(k)} = \frac{(\{k-1\}\alpha)_j}{(k\alpha)_j} \sum_{i=0}^{j} \frac{(\alpha)_i}{i!} C_{j-i}^{(k-1)}; k = 3, \ldots n.$

In particular, when $\alpha = 1$, the count model probabilities in (6) reduce to

$$p_n(t) = P\{N(t) = n\} = \sum_{j=0}^{\infty} C_j \{1-\exp(-t)\}^{n+j} \exp(-t),$$

where

$$C_0 = \frac{1}{\Gamma(1+n)}, \quad C_j = \frac{C_0 n}{(n+j)} C_j^{(n)}; \qquad j = 1,2,\ldots,$$

$$C_j^{(2)} = \frac{[\Gamma(1+j)]^2}{j!(2)_j}, \quad C_j^{(k)} = \frac{(k-1)_j}{(k)_j} \sum_{i=0}^{j} \frac{\Gamma(1+i)}{i!} C_{j-i}^{(k-1)}, \qquad k = 3,\ldots n,$$

which is the probability function of Poisson distribution with unit rate parameter.

3.1. Characteristics of the generalized exponential count model

1. The model handles both over-dispersed and under-dispersed data
The hazard function of the GE distribution is a decreasing function of time when $\alpha < 1$ and so, the distribution displays negative duration dependence. This, in turn, causes over-dispersion in the generalized exponential count model. For $\alpha > 1$, the hazard function is an increasing function of time, so that the distribution displays positive duration dependence and causes under-dispersion in the generalized exponential count model. There is no duration dependence when $\alpha = 1$, which gives the Poisson count model having equal mean and variance. Figure 1 shows the hazard rate plot of GE(α, λ,0) for different values of α and $\lambda = 1$, which supports this result. As a validation of these findings, Figures 2 and 3 display the probability functions for the generalized exponential and Poisson count models for different parameter values. In both the cases, the generalized exponential and Poisson models are chosen so as to have identical means.

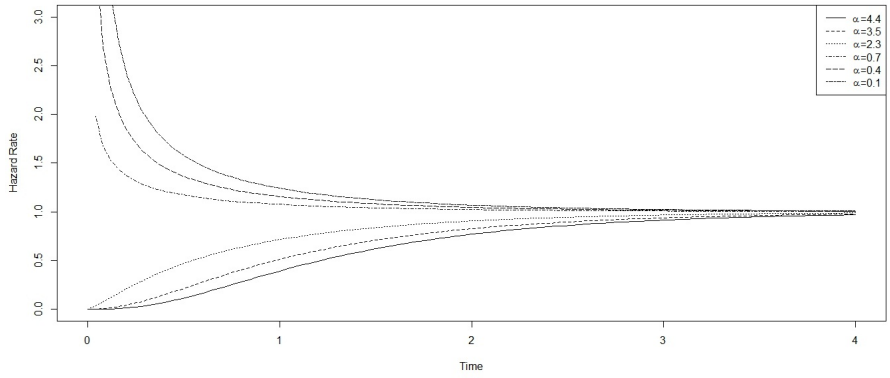

Figure 1 – Hazard rate plot of GE(α, λ,0) for different values of α and $\lambda = 1$.

In Figure 2, the probability function for an under-dispersed generalized exponential model with $\alpha = 1.5$ and Poisson model with mean 0.566 is displayed, and the variance

of the model in this case is smaller than the mean. Figure 3 displays the over-dispersed generalized exponential model with $\alpha = 0.5$ and Poisson model with mean 0.655 and the variance of the model in this case is greater than the mean, as expected.

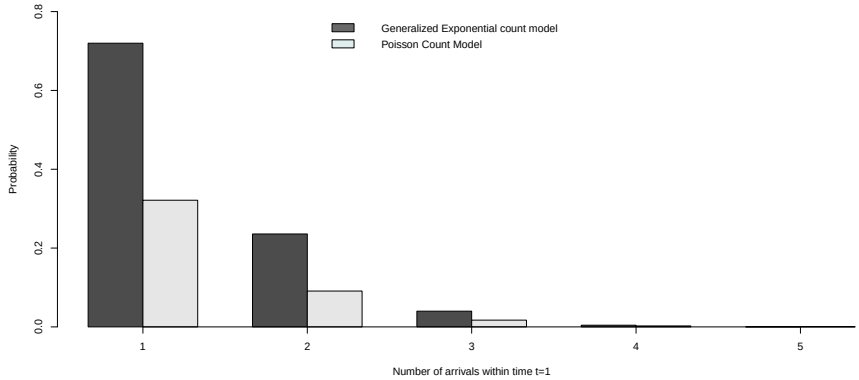

Figure 2 – Probability function for the generalized exponential model ($\alpha = 1.5$) and Poisson count model (mean=0.566), displaying under dispersion.

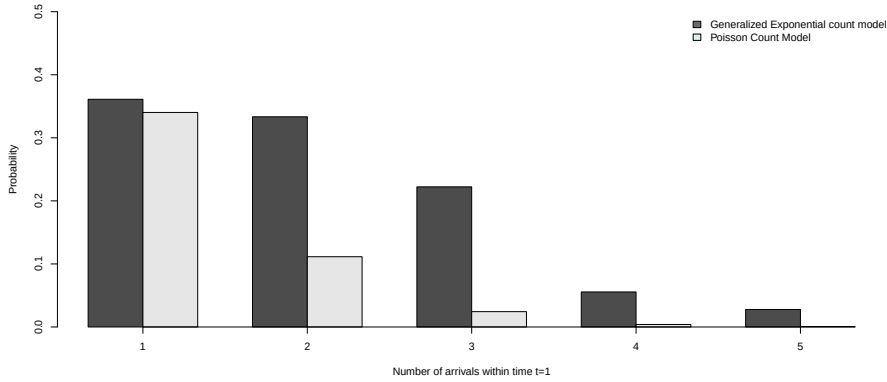

Figure 3 – Probability function for the generalized exponential model ($\alpha = 0.5$) and Poisson count model (mean=0.655) displaying over dispersion.

2. The model is computationally feasible to work with and the probabilities and moments can be estimated without taking resort to time-consuming simulation based methods
The summations which appear in the expressions for the count model probabilities, its mean and variance, are not quite sensitive to the number of terms which are used in summations to approximate the values. The required number of terms is easily identifiable through empirical testing. Therefore, the count model probabilities and mean, variance of the count model can be conveniently estimated.

3. Researchers who deal with data having GE inter arrival times now have a corresponding count model to use
Equation (6) shows the count model probabilities when the inter arrival time is $GE(\alpha, 1, 0, t)$, and thus, the link between the timing model and its count model is maintained. Using this link between the timing model and the count model, one can also predict the next inter arrival time, when only the count data is available.

4. The mean and variance of the generalized exponential count model exist
The mean and variance of the model exist and are given by

$$\begin{aligned} \text{Mean} &= E[N(t)] \\ &= \sum_{n=1}^{\infty}\sum_{j=0}^{\infty} nC_j \{1-\exp(-t)\}^{n\alpha+j}[1-\{1-\exp(-t)\}^{\alpha}] \end{aligned}$$

$$\begin{aligned} \text{Variance} &= \text{Var}[N(t)] \\ &= \sum_{n=1}^{\infty}\sum_{j=0}^{\infty} n^2 C_j \{1-\exp(-t)\}^{n\alpha+j}[1-\{1-\exp(-t)\}^{\alpha}] - \\ &\quad \left(\sum_{n=1}^{\infty}\sum_{j=0}^{\infty} nC_j \{1-\exp(-t)\}^{n\alpha+j}[1-\{1-\exp(-t)\}^{\alpha}]\right)^2. \end{aligned}$$

Table 1 gives the generalized exponential count model probabilities for different values of α at $t = 1, 2$. The values are approximated by retaining 10 terms in the summation of the expression for count model probabilities in (6).

Through simulation of the generalized exponential count model, it has been verified that for $\alpha < 1$, the variance of the model exceeds the mean, thus representing over dispersion. For $\alpha > 1$, the mean exceeds the variance, which represents under dispersion whereas for $\alpha = 1$, the mean equals the variance, which corresponds to equidispersion. Table 2, which displays the mean and variance of the generalized exponential count model, provides evidence of this intuitive fact.

TABLE 1
Values of the generalized exponential count model probabilities for different values of α at $t=1, 2$.

	t=1			t=2		
α	$P_1(t)$	$P_2(t)$	$P_3(t)$	$P_1(t)$	$P_2(t)$	$P_3(t)$
0.3	0.1380	0.0928	0.0717	0.0569	0.0385	0.0336
0.6	0.2543	0.1172	0.0615	0.1284	0.0653	0.0444
0.8	0.3138	0.1126	0.0440	0.1786	0.0764	0.0423
1	0.3382	0.1070	0.0358	0.2287	0.0827	0.0366
1.3	0.4033	0.0768	0.0133	0.3016	0.0849	0.0259
1.7	0.4271	0.0485	0.0042	0.3914	0.0784	0.0139
2	0.4259	0.0326	0.0016	0.4519	0.0696	0.0080

TABLE 2
Values of mean and variance of the generalized exponential count model for different values of α at $t=1, 2$.

	t=1		t=2	
α	Mean	Variance	Mean	Variance
0.3	0.5390	0.8648	0.2348	0.4584
0.6	0.6737	0.8239	0.3923	0.6355
0.8	0.6713	0.7103	0.4587	0.6556
1	0.6599	0.6593	0.5042	0.5042
1.3	0.5971	0.4745	0.5796	0.5136
1.7	0.5369	0.3710	0.5900	0.4822
2	0.4962	0.3252	0.6152	0.4241

4. Application to real data sets

In this section, the application of the generalized exponential count model to two real data sets is shown. The simulation of the generalized exponential count model probabilities and calculation of its mean and variance are carried out using the **R** software, version 3.4.0, through the user-contributed packages viz. *reliaR* (Kumar and Ligges, 2015) with the help of self-programmed codes. The *maxLik* package (Toomet and Henningsen, 2015) is used to obtain the maximum likelihood estimates of the parameters of the inter-arrival time distribution.

4.1. Data set-I: Arrival of patients at a clinic

Data set-I is comprised of inter arrival times of patients arriving at a clinic situated at Adabari Tiniali, Guwahati, Assam, India on a given day. The time period of the collection of data is from 07:00 PM to 09:00 PM, during which the concerned doctor attends to the patients. The inter-arrival times are positively skewed, having a long tail towards the right side of the peak and they are expressed in minutes. The number of patients arriving in the clinic on the randomly selected day is 32. It is further found that the there is usually very little gap between the arrival of consecutive customers, but in a very few instances, there is a considerable gap between the successive arrivals. In the data set considered, inter-arrival time exceeding 10 minutes or more is highly improbable.

The mean of the data set is found to be 0.53845 whereas the variance is found to be 0.57048. Given these information extracted from the data set, the goodness of fit test of the $GE(\alpha,1,0)$ distribution to the given data set is carried out. Assuming that the data set is from $GE(\alpha,1,0)$, the m.l.e of α is found to be $\hat{\alpha} = 0.50987$. Now, to test the hypothesis $H_{01} : GE(\alpha,1,0)$ with $\hat{\alpha} = 0.50987$ is a good fit to the given data, the Kolmogorov-Smirnov one sample test is used. The p-value of the test is found to be $0.7075 > 0.05$. Hence, H_{01} is accepted at 5% level of significance and it is concluded that the assumption of $GE(\alpha,1,0)$ distributed inter arrival times with $\hat{\alpha} = 0.50987$ is valid. Therefore, the number of patients arriving in an interval can be estimated using the over-dispersed generalized exponential count model.

Figure 4 shows the probability of observed counts and predicted generalized exponential model counts of patients arriving in the clinic in a time interval of length 1 minute.

4.2. Data set-II: Arrival of customers in a departmental store

Data set-II is comprised of inter arrival times of customers arriving at a departmental store, situated at Adabari Tiniali, Guwahati, Assam, India on a given day. The time period of the collection of data is from 10:00 AM to 09:00 PM, during which the store remains operative. The inter-arrival times are positively skewed, having a long tail towards the right of the peak and they are expressed in minutes. The number of customers arriving in the store on the randomly selected day is 235. Further, it is seen that there

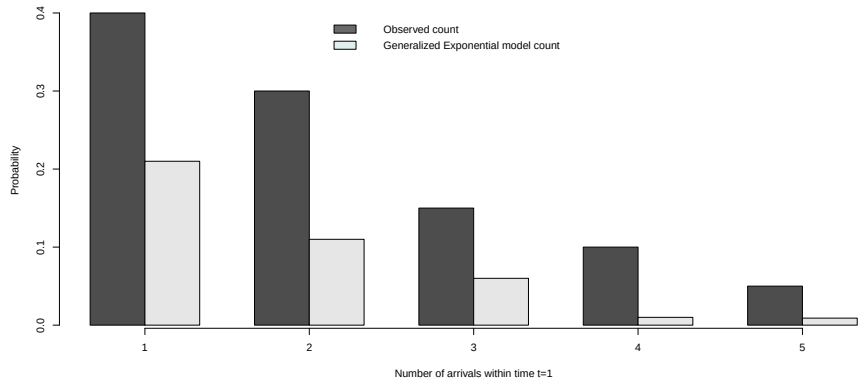

Figure 4 – Probability of observed counts and predicted generalized exponential model counts of patients arriving in the clinic.

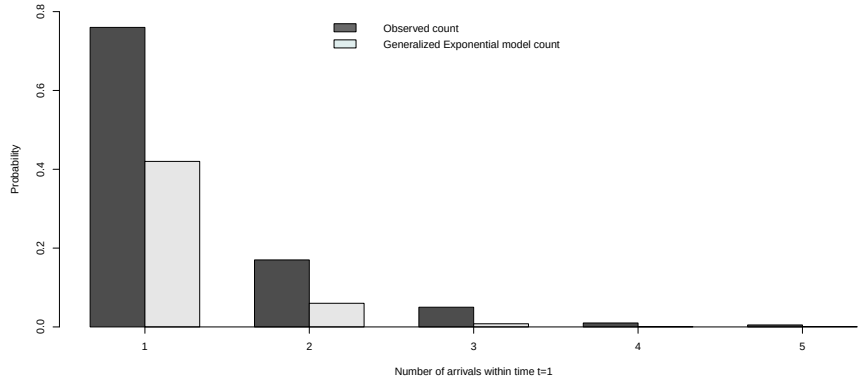

Figure 5 – Probability of observed counts and predicted generalized exponential model counts of customers arriving in the departmental store.

is not much gap between the arrival of the consecutive customers and only in a few instances, there is a fairly good gap between the successive arrivals. In the data set, inter-arrival time exceeding 8 minutes of more is highly unlikely.

The mean of the data set is 1.29377 whereas its variance is 1.16099. Given these information obtained from the data set, the goodness of fit test of the $GE(\alpha, 1, 0)$ distribution to the given data set is performed. Assuming that the data set is from $GE(\alpha, 1, 0)$, the m.l.e of α is found to be $\hat{\alpha} = 1.52752$. Now, to test the hypothesis $H_{02} : GE(\alpha, 1, 0)$ with

$\hat{\alpha} = 1.52752$ is a good fit to the given data, the Kolmogorov-Smirnov one sample test is used. The p-value of the test is found to be $0.3536 > 0.05$. Hence, H_{02} is accepted at 5% level of significance and it is concluded that the assumption of $GE(\alpha, 1, 0)$ distributed inter arrival times with $\hat{\alpha} = 1.52752$ is valid. Therefore, the number of customers arriving in an interval can be estimated using the under-dispersed generalized exponential count model.

Figure 5 shows the probability of observed counts and predicted generalized exponential model counts of customers arriving in the departmental store in a time interval of length 1 minute.

5. Conclusion

In this article, a new count model based on generalized exponential (GE) inter arrival time process is introduced. This model is based on generalized exponentially distributed inter arrival times and is a generalization of the traditional Poisson count model. Another advantage of this new model lies in its ability to model under-dispersed, equi-dispersed as well as over-dispersed count data, owing to the non-constant hazard function of the corresponding GE inter arrival time distribution. The simulation of count model probabilities and calculation of the mean and variance of the model can be carried out using the **R** software. Finally, the proposed model is applied to two real life data sets, where the inter arrival times are generalized exponentially distributed. It is seen that the generalized exponential count model is able to model both over dispersed and under dispersed count data, in addition to the equi-dispersed count data.

6. Acknowledgements

I would like to thank the reviewers for their valuable comments which helped in significantly improving the contents of the paper.

References

Z. V. C. DeJardine (2013). *Poisson Processes and Applications in Hockey*. Lakehead University, Thunder Bay, Ontario, Canada. URL `https://www.lakeheadu.ca/sites/default/files/uploads/77/docs/DejardineFinal.pdf`.

R. D. Gupta, D. Kundu (1999). *Generalized exponential distributions*. Australian & New Zealand Journal of Statistics, 41, no. 2, pp. 173–188.

K. K. Jose, B. Abraham (2011). *A count model based on Mittag-Leffler interarrival times*. Statistica, 71, no. 4, pp. 501–514.

K. K. Jose, B. Abraham (2013). *A counting process with Gumbel inter-arrival times for modeling climate data*. Journal of Environmental Statistics, 4, no. 5, pp. 1–13.

J. F. C. KINGMAN (1963). *Poisson counts for random sequences of events*. The Annals of Mathematical Statistics, 34, no. 4, pp. 1217–1232.

V. KUMAR, U. LIGGES (2015). *Package for some probability distributions*. R package version 0.01, URL: https://cran.r-project.org/web/packages/reliaR/reliaR.pdf.

B. MCSHANE, M. ADRIAN, E. T. BRADLOW, P. S. FADER (2008). *Count models based on Weibull interarrival times*. Journal of Business and Economic Statistics, 26, no. 3, pp. 369–378.

S. M. ROSS (1995). *Stochastic Processes*. John Wiley & Sons, New York.

O. TOOMET, A. HENNINGSEN (2015). *Maximum Likelihood Estimation and Related Tools*. R package version 1.3-4, URL: https://cran.r-project.org/web/packages/maxLik/maxLik.pdf.

I. K. VONG (2013). *Theory of Poisson Point Process and its Application to Traffic Modelling*. Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia. URL `https://pdfs.semanticscholar.org/0044/7f6f9f3c2809934604c9f99c583313e3f270.pdf`.

R. WINKELMANN (1995). *Duration dependence and dispersion in count data models*. Journal of Business and Economic Statistics, 13, pp. 467–474.

SUMMARY

This paper introduces a new counting process which is based on generalized exponentially distributed inter-arrival times. The advantage of this new count model over the existing Poisson count model is that the hazard function of the inter arrival time distribution is non-constant, so that the distribution is duration dependent and hence, is able to model both under dispersed and over dispersed count data, as opposed to the exponentially distributed inter arrival time of the Poisson count model, which is not duration dependent and the corresponding count model is able to model only equi-dispersed data. Further, some properties of this model are explored. Simulation from this new model is performed to study the behavior of count probabilities, mean and variance of the model for different values of the parameter. Use of the proposed model is illustrated with the help of real life data sets on arrival times of patients at a clinic and on arrival times of customers at a departmental store.

Keywords: Inter-arrival times; generalized exponential distribution; Counting process; Hazard function; Dispersion.