11-1-2011

Error Analysis on the Generalized Negative Binomial Distribution

Felix Famoye
Central Michigan University, felix.famoye@cmich.edu

Oluwakemi Aremu
University of Lagos, chemmy413@yahoo.com

Error Analysis on the Generalized Negative Binomial Distribution

Felix Famoye	Oluwakemi Aremu
Central Michigan University,	University of Lagos,
Mt. Pleasant, MI	Akoka-Yaba, Lagos, Nigeria

The generalized negative binomial distribution characterized by three parameters, has been used to fit data from various fields of study. The distribution can model data for which the variance is larger or smaller than the mean, however, it becomes truncated under certain conditions. This truncation error is investigated via a detailed error analysis that determines the parameter space when the model can be used in place of the truncated generalized negative binomial distribution. The fitting of a generalized negative binomial distribution to a data set of absenteeism among shift-workers in a steel industry is re-analyzed.

Key words: Truncation error, dispersion, maximum likelihood estimates.

Introduction

A generalized negative binomial distribution (GNBD) was defined and studied by Jain and Consul (1971). The probability mass function of the GNBD is given by

$$P_x = P(X = x)$$

$$= \begin{cases} \dfrac{m}{m+\beta x}\dbinom{m+\beta x}{x}\theta^x(1-\theta)^{m+\beta x-x}, & x = 0,1,2,\dots \\ 0, & \text{for } x > k \text{ when } \beta < 0 \text{ or } 0 < \beta < 1, \end{cases}$$

$$(1.1)$$

and zero otherwise, where $0 < \theta < 1$, $m > 0$ and $\beta = 0$ or $0 < \beta < 1/\theta$ and k is the largest positive integer for which $m + 1 + (\beta - 1)k > 0$ when $\beta < 0$ or $0 < \beta < 1$. The GNBD in (1.1) reduces to the binomial distribution when $\beta = 0$ and m is an integer, and to the negative binomial distribution when $\beta = 1$. For the non-truncated GNBD, the mean and variance are

$$\mu = m\theta / (1 - \theta\beta)$$

Felix Famoye is a Professor in the Department of Mathematics. Email him at: felix.famoye@cmich.edu. Oluwakemi Aremu is a student in the Department of Mathematics. Email him at: chemmy413@yahoo.com.

and

$$\sigma^2 = m\theta(1-\theta)/(1-\theta\beta)^3. \qquad (1.2)$$

The moments in (1.2) exist when $\theta\beta < 1$.

Famoye and Consul (1993) defined and studied the truncated GNBD. The advantage of the truncated GNBD is that the distribution is defined for all values of β. However, the truncated GNBD is more difficult to estimate than the ordinary GNBD. The major difficulty is in finding suitable initial estimates for the model parameters.

All the estimation methods suggested by Famoye and Consul (1993) involve iterative procedure like the Newton-Raphson method. Because no estimation technique can be done without iteration, it is difficult to determine an initial estimate for the iteration. One way to obtain an initial estimate is to use the moment estimate of the non-truncated GNBD as the initial estimate; however, the moment estimates of non-truncated GNBD may not provide satisfactory initial estimates.

Famoye (1997) discussed parameter estimation for the GNBD. The asymptotic relative efficiencies of the estimators were compared. The method of first two moments and proportion of zeros (MOZE) has good efficiency when compared to the maximum likelihood estimates. From the simulation results, the MOZE method performed very well when both

bias and variance of the estimators were considered.

Nelson (1975) noted that the GNBD as first defined by Jain and Consul (1971) is truncated on the right hand side when $\beta < 0$. Also, the distribution gets truncated when $0 < \beta < 1$. Nelson (1975) remarking on GNBD stated that "A rigorous error analysis has not been performed, but it appears that for $n > -3\beta$, the error resulting from having negative value of β should be tolerable for most applications" (p. 136). The parameter n was replaced with m in (1.1), and to the best of our knowledge, no such error analysis has been conducted for the GNBD. One motivation for this study is to examine the error analysis for the GNBD when $\beta < 0$ and when $0 < \beta < 1$.

Due to the truncation described above, the sum of the probabilities in (1.1) may differ from unity. The difference between 1 and the sum of the probabilities (ΣP_x) is the truncation error. The percentage truncation error is computed as $100(1 - \Sigma P_x)$. Some illustrative examples for $k \leq 3$ are presented in Table 1. For two classes only, the truncation leads to only two probabilities P_0 and P_1, and the sum of the two probabilities could be very small or very large as shown in Table 1. As the values of θ decrease, the truncation error decreases. In general, the sum of the non-negative probabilities is much closer to 1 for small values of θ. As m increases, the value of k increases and, as the value of k increases, the truncation error decreases.

Other parameter sets can be used to illustrate the same phenomena. When $\beta < 1$ many of the cases shown in Table 1 satisfy the condition $m > -3\beta$, however, these values produce the sums of probabilities that are not close to 1. The statement that the error may be tolerable when $m > -3\beta$ does not seem to hold; more conditions than this are required. This study seeks to determine these other conditions such that the error will be tolerable or negligible. For example, in row 7 for $k = 1$, the sum of the probabilities is more than 3 on the account that the $P(X = 1)$ leads to $1 - \theta$ being raised to a negative power (see Table 1).

Review of the GNBD Dispersion Property

The GNBD model in (1.1) is over-dispersed (the variance is larger than the mean) when $\theta < (2\beta - 1)/\beta^2$, under-dispersed (the variance is smaller than the mean) when $\theta > (2\beta - 1)/\beta^2$ and equi-dispersed (the variance is equal to the mean) when $\theta = (2\beta - 1)/\beta^2$. These conditions differ from those given by Jain and Consul (1971), which involve the square root of $1 - \theta$. When $\beta \geq 1$, it is known that $\theta\beta < 1$ for the existence of the moments, therefore the condition for over-dispersion is always satisfied; hence, the GNBD is over-dispersed when $\beta \geq 1$. The GNBD model is under-dispersed whenever $\beta \leq 0.5$. When $0.5 < \beta < 1$, the GNBD is over-dispersed for all values of θ satisfying $0 < \theta < (2\beta - 1)\beta^{-2}$ and under-dispersed for values of θ satisfying $(2\beta - 1)\beta^{-2} < \theta < 1$. These results for the GNBD model can be summarized as follows:

- It is over-dispersed (i) when $\beta \geq 1$ and (ii) when $0.5 < \beta < 1$ and $0 < \theta < (2\beta - 1)\beta^{-2}$.

- It is under-dispersed (i) when $\beta \leq 0.5$ and (ii) when $0.5 < \beta < 1$ and $(2\beta - 1)\beta^{-2} < \theta < 1$.

- It is equi-dispersed when $\theta = (2\beta - 1)\beta^{-2}$.

- The GNBD dispersion is independent of the parameter m.

Figure 1 shows the dispersion regions for the GNBD model: All points above the line $\theta = (2\beta - 1)/\beta^2$ represent the region where the GNBD model is over-dispersed, all points below the line represent the region where the model is under-dispersed, and all points on the line are where the GNBD model is equi-dispersed.

Table 1: Sum of Probabilities for Some GNBD Parameter Sets

k	Parameters			Probabilities				ΣP_x
	θ	β	m	P_0	P_1	P_2	P_3	
1	.95	−2	4.0	0.0000	0.1900			0.1900
	.50	−2	4.0	0.0625	1.0000			1.0625
	.05	−2	4.0	0.8145	0.1900			1.0045
	.95	−.5	1.6	0.0083	1.1265			1.1348
	.50	−.5	1.6	0.3299	0.7464			1.0763
	.05	−.5	1.6	0.9212	0.0796			1.0008
	.95	−.1	0.5	0.2236	2.8662			3.0898
	.50	−.1	0.5	0.7071	0.3789			1.0860
	.05	−.1	0.5	0.9747	0.0258			1.0005
	.95	.1	0.5	0.2236	1.5744			1.7980
	.50	.1	0.5	0.7071	0.3299			1.0370
	.05	.1	0.5	0.9747	0.0255			1.0002
2	.95	−2	7.0	0.0000	0.0000	0.3159		0.3159
	.50	−2	7.0	0.0078	0.2188	0.8750		1.1016
	.05	−2	7.0	0.6983	0.2851	0.0166		1.0000
	.95	−.5	2.6	0.0004	0.0915	2.3332		2.4251
	.50	−.5	2.6	0.1649	0.6065	0.2573		1.0287
	.05	−.5	2.6	0.8751	0.1229	0.0020		1.0000
	.95	−.1	1.5	0.0112	0.4299	1.6533		2.0944
	.50	−.1	1.5	0.3535	0.5684	0.0914		1.0133
	.05	−.1	1.5	0.9259	0.0735	0.0006		1.0000
	.95	.4	0.5	0.2236	0.6409	0.5571		1.4156
	.50	.4	0.5	0.7071	0.2679	0.0305		1.0055
	.05	.4	0.5	0.9747	0.0251	0.0002		1.0000
3	.95	−.5	3.6	0.0000	0.0063	0.4307	0.8388	1.2758
	.50	−.5	3.6	0.0825	0.4199	0.4750	0.0154	0.9928
	.05	−.5	3.6	0.8314	0.1616	0.0070	0.0000	1.0000
	.95	−.1	2.5	0.0006	0.0358	0.5970	0.9419	1.5753
	.50	−.1	2.5	0.1768	0.4737	0.3300	0.0218	1.0023
	.05	−.1	2.5	0.8796	0.1163	0.0040	0.0000	0.9999
	.95	.6	0.5	0.2236	0.3520	0.3880	0.2269	1.1905
	.50	.6	0.5	0.7071	0.2332	0.0639	0.0066	1.0008
	.05	.6	0.5	0.9747	0.0249	0.0004	0.0000	1.0000

Figure 1: Dispersion Region for the GNBD
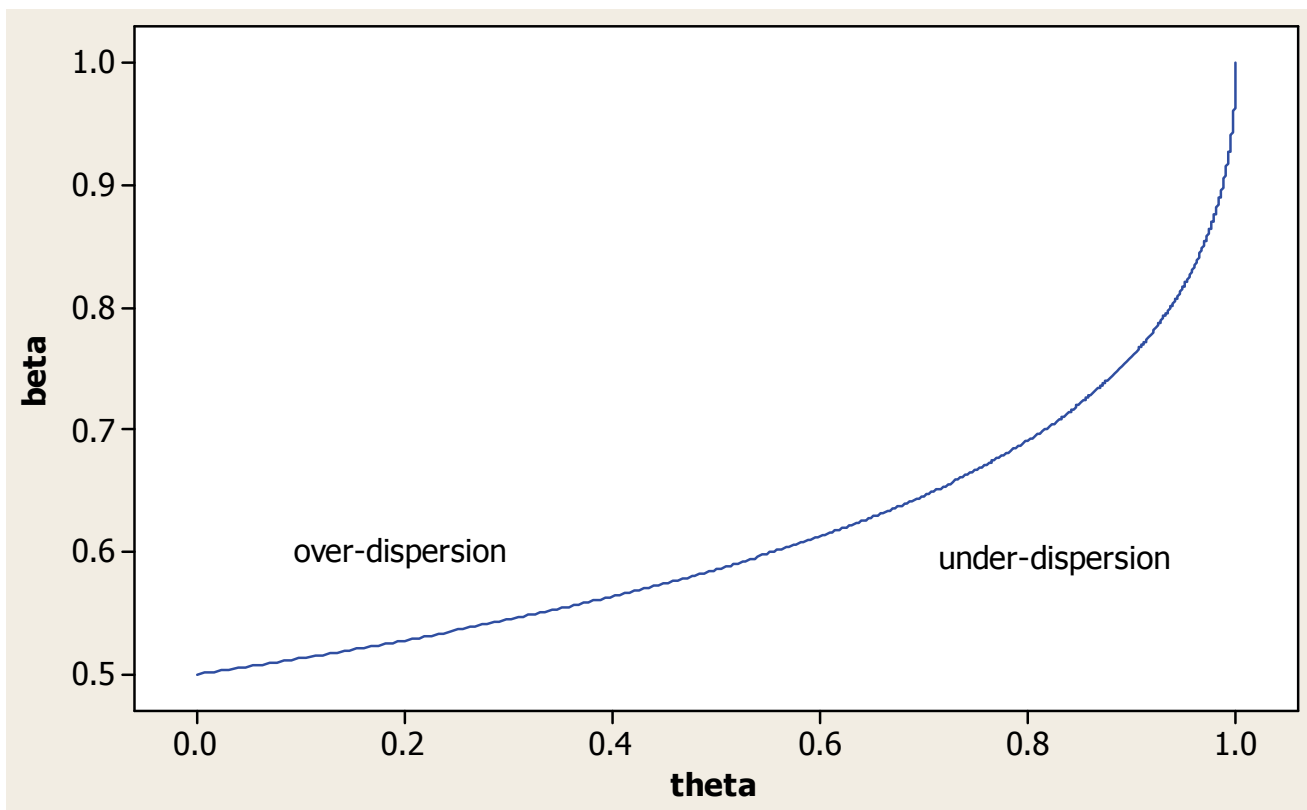
Error Analysis of the GNBD

Re-writing the GNBD in (1.1),

$$P_x = m\theta^x (1-\theta)^{m+\beta x - x} [\Pi_{i=1}^{x-1} (m + \beta x - i)] / x!.$$

When $\beta < 0$ or $0 < \beta < 1$, it is required that $m + \beta x - x + 1 \geq 0$. If this condition is not satisfied, then P_x is set to 0 as shown in (1.1). Thus, the largest x value can be obtained from $0 \leq m + 1 + (\beta - 1)x \Rightarrow (1 - \beta)x \leq m+1 \Rightarrow x \leq (m + 1)/(1 - \beta)$ because $1 - \beta > 0$. The largest x value, k, is given by the integer part of $(m + 1)/(1 - \beta)$. Through computation, a detailed error analysis can be conducted on the GNBD model when $\beta < 0$ and $0 < \beta < 1$. This analysis considers the values of m and θ in the parameter space of the model and the values of β when the truncation occurs; the values of $m > 0$, $0 < \theta < 1$, $\beta < 0$ and $0 < \beta < 1$. Observe that $\theta\beta$ is always less than 1 when truncation occurs. In the analysis, the values of $P(X = x)$ are computed for

$x = 0, 1, 2, \ldots, k$, where k is such that $k \leq (m+1)/(1-\beta)$, and where $\beta < 0$ or $0 < \beta < 1$. In addition to these probabilities, the mean and variance of the truncated model are computed using the formulas $\mu_* = \Sigma x P_x / \Sigma P_x$ and $\sigma_*^2 = \Sigma x^2 P_x / \Sigma P_x - (\mu_*)^2$. After obtaining these values, percentage truncation errors in the sum of probabilities, the means and the variances are calculated using the formulas $100(1 - \Sigma P_x)$, $100(1 - \mu_*/\mu)$, and $100(1 - \sigma_*^2/\sigma^2)$, respectively.

In fitting the GNBD to an observed data set, the three parameters θ, β, and m must be estimated. In order to have at least 1 degree of freedom for the Chi-square goodness-of-fit test, at least five non-zero probability classes are needed. Thus, it is necessary that the smallest value of x be 4; therefore, in all analyses, the smallest x value is required to be 4. The

percentage error of truncation will be said to be tolerable or negligible if it is below 0.5%; in other words, the difference between 1 and the sum of all non-negative probabilities is below 0.005. This value was used by Consul and Shoukri (1985) in their error analysis for the generalized Poisson distribution. In view of this, the error analysis for $k \geq 4$ was conducted.

The maximum truncation error for the different values of m, θ, and β are provided in Table 2. Because at least five non-zero probability classes are needed, the different errors for cases where x is at least 4 are examined. In the error analysis the values of $\theta = 0.01(0.01)0.99$, $\beta = (-2.0)(0.01)(-0.01)$ and $m = 0.1(0.1)(15.0)$ are considered.

Table 2 shows the ranges for the parameters that produce the maximum percentage error in the sum of the non-zero probabilities and specific parameter values at which the maximum truncation error occurs. The corresponding percentage errors in means and variances are also reported. For example, when $0 < \theta \leq 0.71$, $0.01 \leq \beta \leq 0.99$ and $0.1 \leq m \leq 0.5$, the maximum truncation error with at least 5

non-zero probability classes is –0.4799. When $0 < \beta < 1$, the percentage error in the means and percentage error in variances decrease as m increases. As m values increase, the range of θ values decreases in order to have a maximum truncation error of less than 0.5%. As the number of non-zero probability classes increases, the truncation error decreases.

When $0 < \beta < 1$ and $k \geq 4$, the GNBD can be used in general when $0 < \theta \leq 0.57$ for any value of $m > 0$. If $m < 1$, the range of θ values increases to $0 < \theta \leq 0.65$. When $\beta < 0$ and $k \geq 4$, the GNBD can be used in general when $0 < \theta \leq 0.36$ for $m \geq 4$. When $-1 < \beta < 0$ and $k \geq 4$, the range of θ values increases to $0 < \theta \leq 0.46$ for $4 \leq m \leq 10$.

Application to the Absenteeism Numbers among Shift-Workers

Gupta and Ong (2004) defined a new generalization of the negative binomial distribution by mixing the mean of the Poisson distribution with that of a generalized gamma distribution. The probability mass function of their generalized negative binomial distribution,

Table 2: Maximum Percentage Error and Corresponding Percentage Errors in Means and Variances ($k = 5$)

Range of Parameter Values			% Error (θ, β, m)	Percentage Errors	
θ	β	m		Means	Variances
[.01, .71]	[.01, .99]	[0.1, 0.5]	–0.4799 (0.71, 0.63, 0.5)	–3.2261	–13.8517
[.01, .65]	[.01, .99]	[0.1, 1.0]	–0.4761 (0.66, 0.53, 1.0)	–1.8264	–8.1959
[.01, .61]	[.01, .99]	[0.1, 2.0]	–0.4547 (0.61, 0.32, 2.0)	–0.9883	–4.8586
[.01, .57]	[.01, .99]	[0.1, 5.0]	–0.4536 (0.57, 0.01, 3.5)	–0.6274	–3.4805
[.01, .57]	[.01, .99]	[3.6, 5.0]	–0.4440 (0.57, 0.01, 3.6)	–0.5878	–3.1860
[.01, .57]	[.01, .99]	[5.0, 15]	–0.0947 (0.57, 0.01, 5.5)	–0.1105	–0.8318
[.01, .54]	[–.99, –.01]	[4.0, 5.0]	–0.4656 (0.54, –0.3, 5.0)	–0.4952	–3.0429
[.01, .46]	[–.99, –.01]	[5.0, 10]	0.4329 (0.46, –0.99, 7.0)	0.4981	4.1317
[.01, .39]	[–2.0, –.01]	[4.0, 10]	0.4397 (0.39, –1.66, 10)	0.4597	3.9250
[.01, .36]	[–2.0, –.01]	[10, 15]	0.4543 (0.36,–2.0,11.6)	0.4400	3.5627

characterized by four parameters, is in terms of the confluent hypergeometric function of the second kind. This new distribution is fitted to a data set on absenteeism among shift-workers in a steel industry. The data comes from Arbous and Sichel (1954). Gupta and Ong (2004) also fitted the data to the GNBD in (1.1) and obtained the following maximum likelihood estimates (MLE): $\hat{\theta}$ = 0.00010775, $\hat{\beta}$ = 5978.5288 and $\hat{m}$ = 29337.08391. They remarked that, because the parameter θ is small and both β and m are large, the fit by the GNBD corresponds to the fit by the generalized Poisson distribution. These large values of β and m and the small value of θ piqued our curiosity to re-analyze the data.

Famoye (1997) stated that the MOZE estimators are better than the moment estimators and they have good efficiency when compared to the MLE. In view of this, the moment estimates and the MOZE estimates of the GNBD in (1.1) were computed. The moment estimates of θ, β and m are respectively 0.9443, 0.9582, and 0.9058. The corresponding results for the MOZE method are $\bar{\theta}$ = 0.4590, $\bar{\beta}$ = 1.5323 and $\bar{m}$ = 5.8071.

Using the moment estimates as the initial for MLE and the Newton-Raphson method in SAS PROC NLMIXED, the ML estimates for the parameters did not reach acceptable convergence. After reaching convergence, the SAS warning that at least one of the gradients is more than 1.0e–3 (i.e. 0.001) was noted. In this analysis, two of the gradients were over 0.001 and the greater value is 0.0072. However, when the initial estimates are taken to be the MOZE estimates, there was proper convergence to the MLE (see Table 3). The maximum gradient was 1.141e–8. The MLEs in Table 3 are very far from the values given by Gupta and Ong (2004). Gupta and Ong did not report what they took as the initial estimates in finding the MLE. It appears the initial estimates might have caused their estimates to be too small or too large.

Based on the MLE result for parameter β, the negative binomial distribution (NBD) should provide an adequate fit to the data. Table 3 shows the fit by the GNBD and the NBD.

Exact MLEs reported by Gupta and Ong (2004) for the NBD were not obtained in this study, however, estimates are not far from their results.

Although Gupta and Ong (2004) found that their new GNBD provided an adequate fit to the data, the GNBD in (1.1) also provides an adequate fit. In this example, the MLEs of β ($\hat{\beta}$ = 1.0824) is in the parameter region when the sum of the probabilities is 1. This parameter estimate for β is not significantly different from β = 1.0, for which the GNBD reduces to the NBD. The log-likelihood for both the GNBD and NBD are respectively equal to –793.91 and –794.00. This also shows that the NBD provides an adequate fit to the data.

Conclusion

When β < 0 or 0 < β < 1, the truncated GNBD can be used. However, due to estimation problems with the truncated GNBD, the non-truncated GNBD should be considered if the truncation error is negligible. This study provides the region of the parameter space for which the truncation error is below 0.5%. It is important to ensure that the number of non-zero probability classes is at least five (that is, $k \geq 4$). By using the parameter region specified in Table 2, it can be determined whether the estimated parameter values are in the region where the truncation error is negligible.

Jain and Consul (1971) applied the non-truncated GNBD to four data sets. The number of non-zero frequency classes and the parameter estimates given by Jain and Consul (1971) are provided in Table 4. In all data sets, the estimated values of β are between 0 and 1. For data sets 1, 2 and 3, the number of non-zero frequency classes is over 5 and the truncation error is expected to be negligible. In data set 4, there are exactly 5 non-zero frequency classes. However, in comparing the parameter estimates with the regions in Table 2, the maximum truncation error is –0.4547. Computed truncation errors for these data sets are: 0.0351%, 0.2616%, 0.0053% and 0.0182% for data sets 1 through 4 respectively. Thus, the truncation error is negligible for all data sets considered by Jain and Consul (1971).

Table 3: Absenteeism Numbers among Shift-Workers

Count	Observed Frequency	NBD	New GNB by GO[a]	GNBD by JC[b]
0	7	11.13	9.23	10.02
1	16	15.74	16.18	15.70
2	23	17.77	19.86	18.39
3	20	18.36	21.06	19.20
4	23	18.10	20.50	18.89
5	24	17.32	18.78	17.94
6	12	16.24	16.46	16.66
7	13	15.01	14.02	15.22
8	9	13.72	11.79	13.76
9	9	12.43	9.95	12.33
10	8	11.19	8.55	10.99
11	10	10.01	7.54	9.74
12	8	8.91	6.84	8.61
13	7	7.90	6.33	7.58
14	2	6.98	5.94	6.67
15	12	6.14	5.61	5.85
16	3	5.40	5.29	5.13
17	5	4.73	4.97	4.49
18	4	4.13	4.64	3.92
19	2	3.61	4.28	3.43
20	2	3.14	3.92	2.99
21	5	2.73	3.55	2.61
22	5	2.37	3.19	2.28
23	2	2.06	2.84	1.99
24	1	1.78	2.50	1.74
25 – 48	16	11.10	14.13	11.87
Total	248	248.00		248.00
$\hat{\theta}$		0.8525 (0.0157)		0.7435 (0.3284)
$\hat{m}$		1.6792 (0.1775)		2.3580 (2.4079)
$\hat{\beta}$				1.0824 (0.3264)
[c]Chi-Square		15.97	8.27	13.27
df		17	15	16
p-value		0.5260	0.9125	0.6529

[a]Gupta and Ong (2004); [b]Jain and Consul (1971); [c]Adjacent classes for Chi-square values were combined as in Gupta and Ong (2004)

Table 4: Parameter Estimates for Data Sets Analyzed by Jain and Consul (1971)

Data Set	Number of Non-Zero Frequency Classes	Parameter Estimates		
		$\tilde{\theta}$	$\tilde{\beta}$	$\tilde{m}$
1 (in Table 1 of JC[a])	6	0.6013	0.8020	0.4006
2 (in Table 2 of JC)	8	0.7806	0.8549	0.4886
3 (in Table 3 of JC)	11	0.3531	0.0389	11.3188
4 (in Table 4 of JC)	5	0.3171	0.5496	1.5884

[a]Jain and Consul (1971)

References

Arbous, A. G., & Sichel, H. S. (1954). New techniques for the analysis of absenteeism data. *Biometrika*, *41*, 77-90.

Consul, P. C., & Shoukri, M. M. (1985). The generalized Poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics – Simulation and Computation*, *14*(*3*), 667-681.

Famoye, F. (1997). Parameter estimation for generalized negative binomial distribution. *Communications in Statistics – Simulation and Computation*, *26*(*1*), 269-279.

Famoye, F., & Consul, P. C. (1993). The truncated generalized negative binomial distribution. *Journal of Applied Statistical Science*, *1*(*2*), 141-157.

Gupta, R. C., & Ong, S. H. (2004). A new generalization of the negative binomial distribution. *Computational Statistics and Data Analysis*, *45*, 287-300.

Jain, G. C., & Consul, P. C. (1971). A generalized negative binomial distribution. *SIAM Journal of Applied Mathematics*, *21*(*4*), 501-513.

Nelson, D. L. (1975). Some remarks on generalization of the negative binomial and Poisson distributions. *Technometrics*, *17*(*1*), 135-136.