Open Journal of Statistics, 2022, 12, 216-237
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Quasi-Negative Binomial: Properties, Parametric Estimation, Regression Model and Application to RNA-SEQ Data

Mohamed M. Shoukri[1]*, Maha M. Aleid[2]

[1]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Canada
[2]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
Email: *Shoukri.mohamed@gmail.com, *mmshoukr@uwo.ca, maha.aleid@yahoo.com, mahaeid@kfshec.edu.sa

How to cite this paper: Shoukri, M.M. and Aleid, M.M. (2022) Quasi-Negative Binomial: Properties, Parametric Estimation, Regression Model and Application to RNA-SEQ Data. *Open Journal of Statistics*, **12**, 216-237.
https://doi.org/10.4236/ojs.2022.122016

Received: March 16, 2022
Accepted: April 16, 2022
Published: April 19, 2022

Abstract

Background: The Poisson and the Negative Binomial distributions are commonly used to model count data. The Poisson is characterized by the equality of mean and variance whereas the Negative Binomial has a variance larger than the mean and therefore both models are appropriate to model over-dispersed count data. **Objectives:** A new two-parameter probability distribution called the Quasi-Negative Binomial Distribution (QNBD) is being studied in this paper, generalizing the well-known negative binomial distribution. This model turns out to be quite flexible for analyzing count data. Our main objectives are to estimate the parameters of the proposed distribution and to discuss its applicability to genetics data. As an application, we demonstrate that the QNBD regression representation is utilized to model genomics data sets. **Results:** The new distribution is shown to provide a good fit with respect to the "Akaike Information Criterion", AIC, considered a measure of model goodness of fit. The proposed distribution may serve as a viable alternative to other distributions available in the literature for modeling count data exhibiting overdispersion, arising in various fields of scientific investigation such as genomics and biomedicine.

Keywords

Queuing Models, Overdispersion, Moment Estimators, Delta Method, Bootstrap, Maximum Likelihood Estimation, Fisher's Information, Orthogonal Polynomials, Regression Models, RNE-Seq Data

1. Introduction

A random variable X is said to have "Quasi Negative Binomial Distribution", QNBD if the probability function is given by:

$$Px = P(X = x) = \frac{\beta - 1}{(\beta - 1) + \beta x} \cdot \frac{\Gamma(\beta + \beta x)\theta^x (1-\theta)^{\beta+\beta x - x - 1}}{x!\Gamma(\beta + \beta x - x)}$$

$$\begin{aligned} &x = 0,1,2,\cdots \\ &0 < \theta < 1 \\ &0 < \beta\theta < 1 \end{aligned} \tag{1}$$

The distribution whose probability function is given in (1) was first derived by Takács [1] as a queuing model. He assumed that we have a single server queue with independent customers arriving according to a Poisson process in batches of size $(\beta - 1)$ with traffic intensity π and exponential service time with mean $1/\alpha$. It is also assumed that the service time is independent of the interarrival time. Under these conditions, the probabilities of arrival $\theta = \pi/(\pi + \alpha)$ and departure = 1 − θ. Takács [1] and later Consul and Gupta [2] showed that the probability that a buy period has $(\beta - 1)x$ is for fixed β given by (1). The distribution is a member of the Lagrange class of distributions [3] [4] [5].

The shape of the histogram of X depends on the combination (β, θ). In **Figure 1** & **Figure 2** we can see that the distribution has a much longer tail for large values of β.

The paper is structured as follows: In Section 2 we demonstrate the connection between the QNBD and the regular exponential family of distributions [6] and derive the higher order central moments of the distribution. A limiting form of the distribution will be investigated as well. In Section 3, we derive the first order approximation of the variances and biases of the moment estimators of (β, θ). In Section 4, we derive the maximum likelihood estimators and their asymptotic variances and biases. In Section 5, we develop the regression model and establish discuss the maximum likelihood estimation of the regression parameters. In Section 6, we apply the models to real-life data arising from genomic studies. We provide general discussion in Section 7.

2. Moments of the Distribution

The simplest approach to derive the higher order central moments of the distribution is to first write (1) in the general form of the linear exponential family.

For fixed β, the QNBD belongs to the regular exponential family of discrete random variables:

$$Px = h(x)\exp\left[\eta(\theta)S(x) - \psi(\theta)\right] \tag{2}$$

with

$$h(x) \equiv \frac{\beta - 1}{(\beta - 1) + \beta x} \cdot \frac{\Gamma(\beta + \beta x)}{x!\Gamma(\beta + \beta x - x)}$$

$$S(x) \equiv x$$

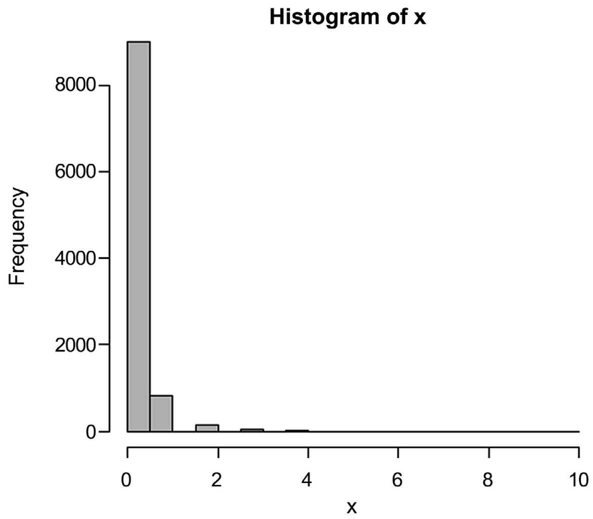

Figure 1. Histogram of QNBD for $\beta = 2$, and $\theta = 0.10$.

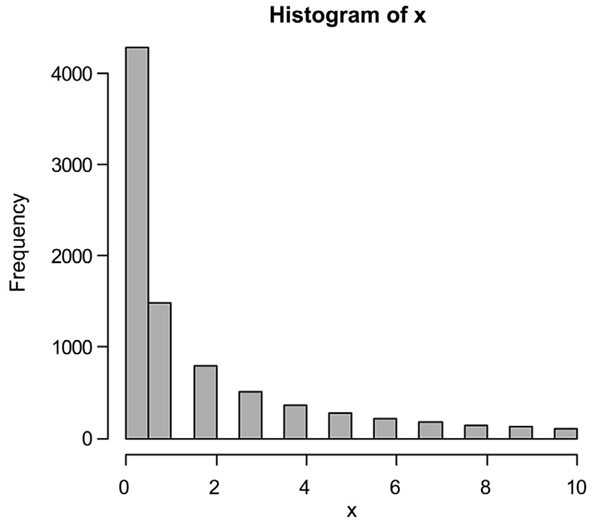

Figure 2. Histogram of X when for $\beta = 10$, and $\theta = 0.09$.

$$\eta(\theta) = \log\left[\frac{\theta}{(1-\theta)^{1-\beta}}\right]$$

$$\psi(\theta) = -\log(1-\theta)^{\beta-1}$$

The mean μ_1' and variance μ_2 of X are given respectively by:

$$\mu_1' = \frac{(\beta-1)\theta}{1-\beta\theta} \tag{3}$$

$$\mu_2 = \frac{(\beta-1)\theta(1-\theta)}{(1-\beta\theta)^3} \tag{4}$$

Writing

$g(\theta) = \exp[\eta(\theta)]$, and $f(\theta) = \exp[\psi(\theta)]$, one can establish a recurrence relationship among the central moments so that:

$$\mu_{r+1} = E\left[\left(x-\mu_1'\right)^r\right] = \frac{g(\theta)}{g'(\theta)}\left[\frac{\partial \mu_r}{\partial \theta} + \frac{\partial \mu_1'}{\partial \theta}\mu_{r-1}\right]$$

Here

$$\mu_1' = \frac{g(\theta)}{g'(\theta)} \cdot \frac{\partial \ln f}{\partial \theta} \tag{5}$$

$$r = 1, 2, \cdots$$

Therefore, the third and fourth central moments are:

$$\mu_3 = \frac{(\beta-1)\theta(1-\theta)\left(1-2\theta+2\beta\theta-\beta\theta^2\right)}{(1-\beta\theta)^5} \tag{6}$$

$$\mu_4 = 3\mu_2^2 + \frac{(\beta-1)\theta(1-\theta)}{(1-\beta\theta)^7}M \tag{7}$$

where,

$$M = 1-6\theta+6\theta^2+2\beta\theta\left(4-9\theta+4\theta^2\right)+\beta^2\theta^2\left(6-6\theta+\theta^2\right).$$

Moreover, the fifth central moment is given by:

$$\mu_5 = 10\mu_2\mu_3 + (\beta-1)\theta(1-\theta)(1-\beta\theta)^{-9} B$$

where

$$\begin{aligned} B = {} & 1-14\theta+36\theta^2+24\theta^3+2\theta\beta\left(11-42\theta+28\theta^2\right) \\ & -\theta^2\beta\left(29-96\theta+58\theta^2\right)+\theta^2\beta^2\left(58-96\theta+29\theta^2\right) \\ & -2\theta^3\beta^2\left(28-42\theta+11\theta^2\right)+2\theta^3\beta^3\left(12-9\theta+\theta^2\right) \\ & -\theta^4\beta^3\left(18-12\theta+\theta^2\right) \end{aligned}$$

3. Moment Estimators

Suppose that we have a random sample $x_1, x_2, \cdots, x_n$ with sample mean $\overline{x}$ and sample variance s^2

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Equating the sample statistics to their corresponding population parameters (3) and (4) and solving for θ and β we get

$$\hat{\theta} = 1 - \frac{\overline{x}\left(1+\overline{x}\right)^2}{s^2} \tag{8}$$

$$\hat{\beta} = 1 + \frac{\overline{x}\left(1+\overline{x}\right)}{s^2} \tag{9}$$

We use the delta method to evaluate the variances and biases of a moment estimator.

From Kendall and Ord [7] we have:

$$\operatorname{var}\left(\hat{\theta}\right)=\operatorname{var}\left(\overline{x}\right)\left(\frac{\dot{\partial}\hat{\theta}}{\partial\overline{x}}\right)^{2}+\operatorname{var}\left(s^{2}\right)\left(\frac{\dot{\partial}\hat{\theta}}{\partial s^{2}}\right)^{2}+2\operatorname{cov}\left(\overline{x},s^{2}\right)\left(\frac{\dot{\partial}\hat{\theta}}{\overline{\partial}\overline{x}}\right)\left(\frac{\dot{\partial}\hat{\theta}}{\overline{\partial}s^{2}}\right)$$

$$\operatorname{Bias}\left(\hat{\theta}\right)=\frac{1}{2!}\left[\operatorname{var}\left(\overline{x}\right)\frac{\dot{\partial}\hat{\theta}}{\partial\overline{x}^{2}}+\operatorname{var}\left(s^{2}\right)\frac{\dot{\partial}^{2}\hat{\theta}}{\partial^{2}s^{2}}+2\operatorname{cov}\left(\overline{x},s^{2}\right)\frac{\dot{\partial}^{2}\hat{\theta}}{\overline{\partial}\overline{x}\partial s^{2}}\right]$$

With similar expressions for $\operatorname{var}\left(\hat{\beta}\right)$ and $\operatorname{Bias}\left(\hat{\beta}\right)$.

One can show that:

$$V_{1}=\operatorname{var}\left(\hat{\theta}\right)=\frac{(1-\theta)}{n\theta(\beta-1)(1-\beta\theta)}[1+2\beta\theta-3\theta]^{2}$$
$$+\frac{\left(\mu_{4}-\mu_{2}^{2}\right)}{n}\cdot\frac{(1-\beta\theta)^{6}}{\theta^{2}(\beta-1)^{2}}-\frac{2\mu_{3}}{n}\frac{(1-\beta\theta)^{4}}{\theta^{2}(\beta-1)^{2}}$$

$$V_{2}=\operatorname{var}\left(\hat{\beta}\right)=\frac{(1-\theta)[1+\beta\theta-2\theta]^{2}}{n\theta(1-\theta)(\beta-1)}+\frac{\left(\mu_{4}-\mu_{2}^{2}\right)}{n}\cdot\frac{(1-\beta\theta)^{8}}{\theta^{2}(1-\theta)^{2}(\beta-1)^{2}}$$
$$-\frac{2\mu_{3}}{n}\frac{(1-\beta\theta)^{6}[1+\beta\theta-2\theta]}{\theta^{2}(1-\theta)^{2}(\beta-1)^{2}}$$

$$\operatorname{Bias}\left(\hat{\theta}\right)=-\frac{4(\beta-1)}{n(1-\beta\theta)}[2+\beta\theta-3\theta]-\frac{\mu_{4}-\mu_{2}^{2}}{n}\left[\frac{(1-\beta\theta)^{6}}{\theta^{2}(1-\theta)(\beta-1)^{2}}\right]$$
$$+\frac{\mu_{3}}{n}\left[\frac{(1-\beta\theta)^{4}(1+2\beta\theta-3\theta)}{(\beta-1)^{2}\theta^{2}(1-\theta)}\right]$$

$$\operatorname{Bias}\left(\hat{\beta}\right)=\frac{1}{n}\left[1+\left(\mu_{4}-\mu_{2}^{2}\right)\frac{(1-\beta\theta)^{7}}{\theta^{2}(1-\theta)^{2}(\beta-1)^{2}}-\mu_{3}\frac{(1-\beta\theta)^{3}}{\theta^{2}(1-\theta)^{2}}(1+\beta\theta-2\theta)\right]$$

$$\operatorname{cov}\left(\hat{\beta},\hat{\theta}\right)=\frac{\mu_{2}}{n}\left(\frac{\partial\theta}{\partial\overline{x}}\right)\left(\frac{\partial\beta}{\partial\overline{x}}\right)+\frac{\mu_{4}-\mu_{2}^{2}}{n}\left(\frac{\partial\theta}{\partial s^{2}}\right)\left(\frac{\partial\beta}{\partial s^{2}}\right)$$
$$+\frac{\mu_{3}}{n}\left[\left(\frac{\partial\theta}{\partial\overline{x}}\right)\left(\frac{\partial\beta}{\partial s^{2}}\right)+\left(\frac{\partial\theta}{\partial s^{2}}\right)\left(\frac{\partial\beta}{\partial\overline{x}}\right)\right]$$

Note that, the information matrix is the determinant of the variance covariance matrix of the moment estimators and is given by:

$$D=\operatorname{var}\left(\hat{\theta}\right)\cdot\operatorname{var}\left(\hat{\beta}\right)-\operatorname{cov}^{2}\left(\hat{\theta},\hat{\beta}\right)$$

Example: Modeling the number of brain lesions to predict Multiple Sclerosis.

The use of gadolinium (Gd) withT1 weighted imaging can identify areas of breakdown in the blood-brain barrier and increases the reliability and in detecting active Multiple Sclerosis (MS) lesions [8]. The number of new Gd enhancing lesions is a widely used end point for monitoring disease activity and for evaluating the effect of treatments in phase II clinical trials. In these studies, the results of the Magnetic Resonance Imaging (MRI) end point are in the form of counts [9]. To deal with the problem of overdispersion, the negative binomial distribution is used to model this type of data.

As application of the QNBD we simulated lesions count data like the situation described in [8] (**Table 1**).

The sample size = 116 subjects.

The histogram of the data *s* is given in Figure 3.

The *y*-axis we have the frequency of each *x*.

mean(*x*) = 3.37, and var(*x*) = 69.63.

The moment estimators are $\hat{\theta} = 0.077$ and $\hat{\beta} = 10.227$.

Bootstrapping the distribution of the moment estimators $\text{SE}\left(\hat{\theta}\right) = 0.344$, and $\text{SE}\left(\hat{\beta}\right) = 0.106$ (Figure 4).

Table 1. Distribution of the number of brain lesions.

X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	50	60
Freq	56	14	7	5	3	3	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1

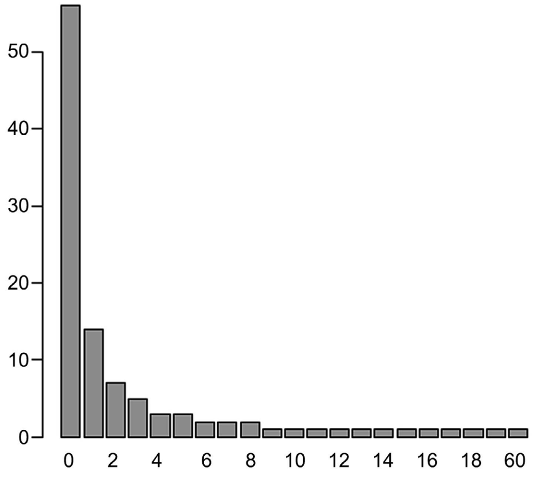

Figure 3. Histogram of the data shown in Table 1. The *x*-axis we have the *x* values.

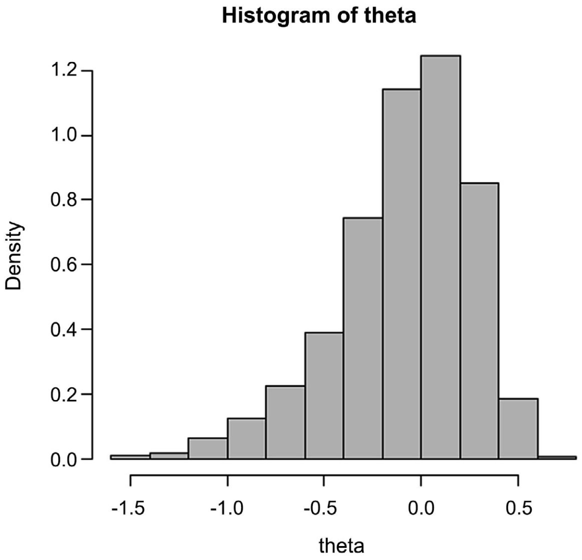

Figure 4. Histogram of distribution of $\hat{\theta}$ based on 1000 bootstrap samples.

The distribution is negatively skewed. The empirical bias in the moment estimator of $\hat{\theta}$ is $\text{bias}\left(\hat{\theta}\right) = -0.1659$.

Similarly $\text{bias}\left(\hat{\beta}\right) = -9.77$.

From Figure 5 we may infer that the distribution of $\hat{\beta}$ seems to be a mixture of two distribution or is bimodal. From these results, we may conclude that the moment estimators are not reliable unless we have extremely large sample. In the next section, we discuss the maximum likelihood estimation.

4. Maximum Likelihood Estimators (MLE)

It is well-known that the estimators obtained from application of the method of MLE possess optimal properties such asymptotic normality and efficiency. Based on a simple random sample the log-likelihood (*l*) function is given by:

$$\begin{aligned} l = {} & n\log\left(\beta-1\right) - \sum_{i=1}^{n}\log\left(\beta-1+\beta x_i\right) + \sum_{i=1}^{n}\sum_{j=1}^{x_i}\left[\beta\left(1+x_i\right)-j\right] \\ & + n\overline{x}\log\theta + n\left(\beta-1\right)\left(1+\overline{x}\right)\log\left(1-\theta\right) \end{aligned} \tag{10}$$

$$\beta\tilde{\theta} = \frac{\overline{x}}{\beta\left(1+\overline{x}\right)-1} \tag{11}$$

Similarly, setting $\frac{\partial l}{\partial \beta}$ equal to zero and solving for β we get:

$$\tilde{\beta} = \left(1+\overline{x}\right)^{-1} + \overline{x}\left(1+\overline{x}\right)^{-1}\left[1-\exp\left[\Omega_x - \left(\left(\tilde{\beta}-1\right)+\left(1-\overline{x}\right)\right)^{-1}\right]\right]^{-1} \tag{12}$$

where

$$\Omega_x = -\left(n\left(1+\overline{x}\right)\right)^{-1}\left[\sum_{i=1}^{n}\frac{\left(1+x_i\right)}{\tilde{\beta}\left(1+x_i\right)-1} - \sum_{i=1}^{n}\sum_{j=1}^{x_i}\frac{\left(1+x_i\right)}{\tilde{\beta}\left(1+x_i\right)-j}\right]$$

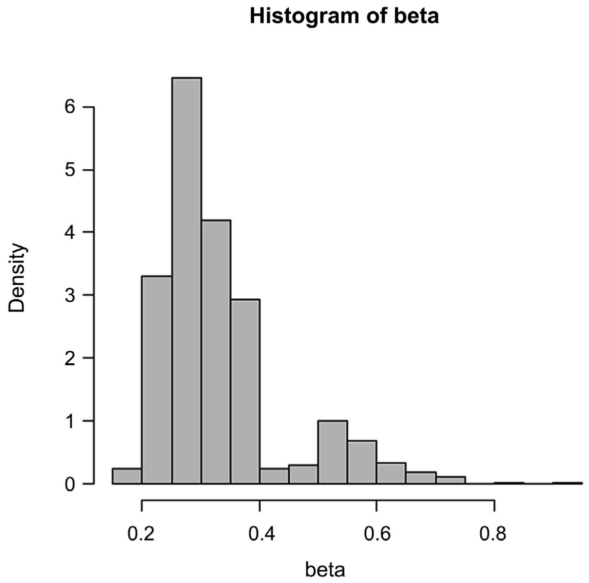

Figure 5. Histogram of distribution of $\hat{\beta}$ based on 1000 bootstrap samples.

The MLEs of θ and β are thus obtained by solving (11) and (12) iteratively, noting that (12) is in the form of $\tilde{\beta} = f\left(\tilde{\beta}\right)$ or a fixed-point equation.

Elements of the variance-covariance matrix of the $\left(\tilde{\theta},\tilde{\beta}\right)$ are obtained by inverting the Fisher's information matrix. We can show that

$$i_{\theta\theta} = -E\left[\frac{\partial^2 l}{\partial\theta^2}\right] = \frac{n(\beta-1)}{\theta(1-\theta)(1-\beta\theta)} \tag{13}$$

$$i_{\theta\beta} = -E\left[\frac{\partial^2 l}{\partial\theta\partial\beta}\right] = \frac{n}{1-\beta\theta} \tag{14}$$

$$i_{\beta\beta} = -E\left[\frac{\partial^2 l}{\partial\beta^2}\right] = \frac{n}{(\beta-1)^2} - E\left[\sum_{i=1}^{n}\frac{(1+x_i)^2}{\left[\beta(1+x_i)-1\right]^2} + \sum_{i=1}^{n}\sum_{j=1}^{x_i}\frac{(1+x_i)^2}{\left[\beta(1+x_i)-j\right]^2}\right] \tag{15}$$

$$\operatorname{var}\left(\tilde{\theta}\right) = i_{\beta\beta}/\Delta,\ \operatorname{var}\left(\tilde{\beta}\right) = i_{\theta\theta}/\Delta,$$

and

$$\operatorname{cov}\left(\tilde{\theta},\tilde{\beta}\right) = -i_{\theta\beta}/\Delta$$

where $\Delta = i_{\theta\theta}\cdot i_{\beta\beta} - i_{\theta\beta}^2$

We note that on using the digamma approximation we can write

$$i_{\beta\beta} = \frac{n}{(\beta-1)^2} - E\left[\sum_{i=1}^{n}\frac{(1+x_i)^2}{\left[\beta(1+x_i)-1\right]^2} + \sum_{i=1}^{n}\frac{x_i(1+x_i)^2}{\left[1+\beta(1+x_i)\right]\left[1+\beta(1+x_i)-x_i\right]}\right]$$

The R-Code for fitting the QNBD is given in Appendix 1.

5. Orthogonal Polynomial Approximation for $i_{\beta\beta}$

The evaluation of the asymptotic variance covariance matrix is difficult because $P_{22} = -E\left[\frac{\partial^2 \log l}{\partial\beta^2}\right]$ does not have a tractable form. To overcome this difficulty, following [5] we employ an asymptotic expansion for $\frac{\partial^2 \log P_x}{\partial\beta}$ as a linear combination of orthogonal polynomials. From Morgan *et al.* [9], if P_x is a distribution function with feint moments μ_r of all orders, then the point x_0 is a point of increase for P_x, if $P_{x0+h} > P_{x0-h}$ for every $h > 0$. If the distribution function P has atleast Y points of increase, Cramér [10] has proved that there exists a sequence of polynomials $G_0(x), G_1(x)$, uniqly determined under the following conditions:

1) $G_n(x)$ is of degree *n*, and the coefficient of x^n in $G_n(x)$ is positive
2) $G_n(x)$ satisfy the orthogonality conditions

$$\sum_{x=0}^{\infty} G_r(x)G_s(x) = E\left(G_r^2(x)\right)$$

If $r = s$

$$= 0 \quad r \neq s\left(r, s = 0,1,2,\cdots\right)$$

Szegő [11] derived the formal Fourier expansion of a continuous function $h(x)$ in terms of a set of orthogonal polynomials such that:

$$h(x)=\sum_{r=0}^{\infty}a_r G_r(x)$$

where a_r are selected so that:

$$\sum_{r=0}^{\infty}\left[\frac{\partial \log P_x}{\partial\beta}-\sum_{r=0}^{\infty}a_r G_r(x)\right]^2 P_x$$

is minimum. He showed that

$$a_0\equiv 0, a_1=\frac{\partial\mu_1'}{\partial\beta}\bigg/E\left(G_1^2(x)\right),$$

$$a_2=\left(\frac{\partial\mu_2}{\partial\beta}-\frac{\mu_3}{\mu_2}\frac{\partial\mu_1'}{\partial\beta}\right)\bigg/E\left(G_2^2(x)\right)$$

Direct calculations give:

$G_0\equiv 1, G_1(x)=x-\mu_1'$ and, $G_2(x)=(x-\mu_1')^2-\frac{\mu_3}{\mu_2}(x-\mu_1')-\mu_2$, are the orthogonal polynomials associated with the probability function P_x, where $E\left(G_1^2(x)\right)=\mu_2$. Moreover, we write

$$\Omega=E\left(G_2^2(x)\right)=E\left[(x-\mu_1')^4+\frac{\mu_3^2}{\mu_2^2}(x-\mu_1')^2+\mu_2^2-2\frac{\mu_3}{\mu_2}(x-\mu_1')^3\right.$$
$$\left.-2\mu_2(x-\mu_1')^2+2\frac{\mu_3}{\mu_2}(x-\mu_1')\mu_2\right]$$

Hence

$$\Omega=\mu_4+\frac{\mu_3^2}{\mu_2}+\mu_2^2-2\frac{\mu_3^2}{\mu_2}-2\mu_2^2=\mu_4-\frac{\mu_3^2}{\mu_2}-\mu_2^2$$

Now, since $\frac{\partial\mu_1'}{\partial\beta}=\frac{\partial}{\partial\beta}\left[\frac{(\beta-1)\theta}{1-\beta\theta}\right]=\frac{\theta(1-\theta)}{(1-\beta\theta)^2}$, and

$$\frac{\partial\mu_2}{\partial\beta}=\frac{\partial}{\partial\beta}\left[\frac{(\beta-1)\theta(1-\theta)}{(1-\beta\theta)^3}\right]=\frac{\theta(1-\theta)(1+2\beta\theta-3\theta)}{(1-\beta\theta)^4}$$

then

$$\frac{\partial\log P_x}{\partial\beta}=\frac{\frac{\partial\mu_1'}{\partial\beta}(x-\mu_1')}{E\left(G_1^2(x)\right)}+\frac{\left(\frac{\partial\mu_2}{\partial\beta}-\frac{\mu_3}{\mu_2}\cdot\frac{\partial\mu_1'}{\partial\beta}\right)}{E\left(G_2^2(x)\right)}\left[(x-\mu_1')^2-\frac{\mu_3}{\mu_2}(x-\mu_1')-\mu_2\right]$$

Since

$$nE\left[\frac{\partial\log P}{\partial\beta}\right]^2=-E\left[\frac{\partial\log l}{\partial\beta^2}\right]$$

Then

$$i_{\beta\beta} \simeq n\left[\frac{\theta(1-\theta)}{(\beta-1)(1-\beta\theta)}+\frac{\theta^4(1-\theta)^2}{(1-\beta\theta)^6\left[\mu_4-\frac{\mu_3^2}{\mu_2}-\mu_2^2\right]}\right]$$

The asymptotic relative efficiency of the moment estimators is therefore given by:

$$\text{Eff} = \frac{1}{\Delta D}$$

For the lesion data Eff = 16.6%. We interpret this number as follows: for the moment estimators to be as efficient as the maximum likelihood estimators, we need a sample size that is 16.6% larger compared to the sample size used for the maximum likelihood estimation.

3.4 Asymptotic biases of the MLE

Unlike the moment estimators, the $(\tilde{\theta},\tilde{\beta})$ do not have closed form expressions, and the applications of the delta method cannot be used to obtain their asymptotic biases. Sherton and Wallington [12] used an approach that depends on the asymptotic expansion of the log-likelihood functions. We denote the biases of $\tilde{\theta}$, and $\tilde{\beta}$ by $b_1(\tilde{\theta})$ and $b_2(\tilde{\beta})$, and these are the solutions of the system of equations

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}\begin{bmatrix} b_1(\tilde{\theta}) \\ b_2(\tilde{\beta}) \end{bmatrix} = \begin{bmatrix} -\dfrac{A_1}{2Dn} \\ -\dfrac{A_2}{2Dn} \end{bmatrix}$$

In the above system of equation we have the following notations:

$$D = P_{11}P_{22} - P_{12}^2$$

$$A_1 = P_{22} - P_{1,11} - 2P_{12}P_{1,12} + P_{11}P_{1,22}$$

$$A_2 = P_{22} - P_{2,11} - 2P_{12}P_{2,12} + P_{11}P_{2,22}$$

And

$$P_{1,11} = E\left[P^{-2}\frac{\partial P}{\partial\theta}\cdot\frac{\partial^2 P}{\partial\theta^2}\right]$$

$$P_{1,12} = E\left[P^{-2}\frac{\partial P}{\partial\theta}\cdot\frac{\partial^2 P}{\partial\theta\partial\beta}\right]$$

$$P_{1,22} = E\left[P^{-2}\frac{\partial P}{\partial\theta}\cdot\frac{\partial^2 P}{\partial\beta^2}\right]$$

$$P_{2,11} = E\left[P^{-2}\frac{\partial P}{\partial\beta}\cdot\frac{\partial^2 P}{\partial\theta^2}\right]$$

$$P_{2,12} = E\left[P^{-2}\frac{\partial P}{\partial\beta}\cdot\frac{\partial^2 P}{\partial\theta\partial\beta}\right]$$

$$P_{2,22} = E\left[P^{-2}\frac{\partial P}{\partial \beta}\cdot\frac{\partial^2 P}{\partial \beta^2}\right]$$

Since

$$\frac{\partial P}{\partial \theta} = \left(x-\mu_1'\right)\frac{\left(1-\beta\theta\right)}{\theta\left(1-\theta\right)}\cdot P$$

We can show that

$$P_{1,11} = \mu_2 w_1 + \mu_3 w_2$$

where

$$w_1 = \frac{\left(1-\beta\theta\right)\left(2\theta-\beta\theta^2-1\right)}{\theta^3\left(1-\theta\right)^3}$$

and

$$w_2 = \frac{\left(1-\beta\theta\right)^3}{\theta^3\left(1-\theta\right)^3}$$

$$P_{1,22} = \left[\theta\left(1-\theta\right)\left(1-\beta\theta\right)\right]^{-1}$$

$$P_{2,11} = \left(1-\theta\right)^{-2},\ P_{1,22} = P_{2,22} = 0$$

$$P_{2,12} = D_1 + D_2 + D_3$$

where

$$D_1 = \frac{1-2\theta+\beta\theta\left(2-0\right)}{\left(\beta-1\right)\left(1-\theta\right)^2}$$

$$D_2 = \left[\frac{\mu_2\dfrac{\partial\mu_2}{\partial\beta}-\mu_3\dfrac{\partial\mu_1'}{\partial\beta}}{\mu_2\mu_4-\mu_3^2-\mu_2^3}\right]^2\left[\mu_5+\frac{\mu_3^3}{\mu_2^2}-2\frac{\mu_3\mu_4}{\mu_2}\right]$$

$$D_3 = \frac{2\left(1-\beta\theta\right)^5}{\left(\beta-1\right)^2\theta\left(1-\theta\right)^2}\left[\mu_2-\frac{\partial\mu_2}{\partial\beta}-\mu_3\frac{\partial\mu_1'}{\partial\beta}\right]$$

where

$$\frac{\partial\mu_1'}{\partial\beta} = \frac{\theta\left(1-\theta\right)}{\left(1-\beta\theta\right)^2}\ \text{and}\ \frac{\partial\mu_2}{\partial\beta} = \frac{\theta\left(1-\theta\right)\left(1+2\beta\theta-3\theta\right)}{\left(1-\beta\theta\right)^4}$$

Finally, using the above information we can show that $P_{2,22} = 0$.

Solving the system of equations, we obtain the asymptotic biases so that

$$\text{bias}\left(\tilde{\theta}\right) = \frac{\theta^2\left(1-\theta\right)^3\left(1-\beta\theta\right)^2}{2n\left[\left(\beta-1\right)\left(1-\theta\right)P_{22}-\theta\left(1-\beta\theta\right)\right]^2} \times\left[\frac{P_{22}}{\left(1-\theta\right)^2}-\frac{2P_{1,12}}{\left(1-\theta\right)}+\frac{2P_{22}}{\theta\left(1-\theta\right)\left(1-\beta\theta\right)}-\frac{2\beta\left(\beta-1\right)P_{22}^2}{\theta\left(1-\beta\theta\right)^2}\right]$$

$$\operatorname{bias}\left(\tilde{\beta}\right)=\frac{\theta^2\left(1-\theta\right)^2\left(1-\beta\theta\right)}{2n\left[\left(\beta-1\right)\left(1-\theta\right)P_{22}-\left(1-\beta\theta\right)\right]^2}$$
$$\times\left[\frac{2\beta\left(\beta-1\right)\beta P_{22}}{\theta\left(1-\beta\theta\right)}-\frac{2}{\theta\left(1-\theta\right)}+\frac{2\left(\beta-1\right)P_{22}}{\theta}-\frac{\left(\beta-1\right)P_{22}}{1-\theta}\right]$$

For the lesion data, the biases of the maximum likelihood estimators are given by:

$$\operatorname{bias}\left(\tilde{\theta}\right)=0.003\text{, and }\operatorname{bias}\left(\tilde{\beta}\right)=0.002$$

6. Quasi Negative Binomial Regression

Our aim in this section is develop regression model based on the GNBD. The approach is facilitated by the fact that the QNBD is a member of the regular exponential family shown in [13]. We employ the transformation:

$$\tau\left(\theta_i\right)=z_i^{\mathrm{T}}\underline{\gamma},\ \ i=1,2,\cdots,k \tag{16}$$

Here we assume to $\tau\left(\theta_i\right)$ be monotone, differentiable, and positive function of θ [13]. In (16) z is a vector of $\upsilon\times1\left(\nu<n\right)$ exploratory variables and γ is a vector of regression parameters. To estimate $\gamma_1,\gamma_2,\cdots,\gamma_q$, and β, we assure that

$$x_1\sim\mathrm{QNBD}\left(\theta_i,\beta\right),i=1,2,\cdots,n$$

are independent random variables and

$$\operatorname{logit}\left[\theta_i\left(z\right)\right]=\left[z_i^{\mathrm{T}}\gamma\right] \tag{17}$$

In this section, we derive the maximum likelihood estimators of the regression parameters, the parameter β and their asymptotic properties. The log-likelihood function is given by:

$$\begin{aligned}l&=n\log\left(\beta-1\right)-\sum_{i=1}^{n}\log\left(\beta-1+\beta x_i\right)\\&\quad+\sum_{i=1}^{n}\log\Gamma\left(\beta+\beta x_i\right)-\sum_{i=1}^{n}\log\Gamma\left(\beta+\beta x_i-x_i\right)\\&\quad+\sum_{i=1}^{n}x_i\log\theta_i+\sum_{i=1}^{n}\left(\beta-1\right)\left(1-x_i\right)\log\left(1-\theta_i\right)\\&=l_1\left(\beta,x_i\right)+l_2\left(\beta,x_i,\theta_i\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial l_2}{\gamma}&=\sum_{i=1}^{n}x_i\frac{\partial}{\partial\gamma_r}\left[\log\theta_i\right]+\sum_{i=1}^{n}\left(\beta-1\right)\left(1+x_i\right)\frac{\partial}{\partial\gamma_r}\left[\log\left(1-\theta_i\right)\right]\\&=\sum_{i=1}^{n}x_iz_{ir}\left[1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right]^{-1}-\sum_{i=1}^{n}\left(\beta-1\right)\left(1+x_i\right)z_{ir}\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\left[1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right]^{-1}\end{aligned}$$

$$\sigma_{r\beta}\equiv\frac{\partial^2l_2}{\partial\beta\partial\gamma_r}=-\sum_{i=1}^{n}\left(1+x_i\right)z_{ir}\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\left[1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right]^{-1}$$

$$\begin{aligned}-I_{\beta r}&=-E\left[\frac{\partial^2l}{\partial\beta\partial\gamma_r}\right]=\sum_{i=1}^{n}z_{ir}\frac{\mathrm{e}^{z_{\mathrm{i}}^{\mathrm{T}}}}{1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}E\left[1+x_i\right]\\&=\sum_{i=1}^{n}z_{ir}\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\left[\left(1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right)\left(1-\left(\beta-1\right)\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right)\right]^{-1}\end{aligned}$$

$$\frac{\partial^2 l_2}{\partial\gamma_r\partial\gamma_s}=-\sum_{i=1}^{n}x_i z_{ir}z_{is}\frac{\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}{\left(1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right)^2}$$

$$-(\beta-1)\sum_{i=1}^{n}(1+x_i)z_{ir}z_{is}\left[\frac{\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}{1+\mathrm{e}^{z_i^{T}\gamma}}-\left[\frac{\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}{1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}\right]^2\right]$$

$$-E\left[\frac{\partial^2 l_2}{\partial\gamma_r\partial\gamma_s}\right]\equiv\sigma_{rs}$$

$$=\sum_{i=1}^{n}z_{ir}z_{is}\theta_i(1-\theta_i)\frac{(\beta-1)\theta_i}{1-\beta\theta_i}+(\beta-1)\sum_{i=1}^{n}z_{ir}z_{is}\frac{(1-\theta_i)}{1-\beta\theta_i}\left[\theta_i-\theta_i^2\right]$$

$$-E\left[\frac{\partial^2 l}{\partial\gamma_r\partial\gamma_s}\right]=(\beta-1)\sum_{i=1}^{n}z_{ir}z_{is}\frac{\theta_i(1-\theta_i)}{1-\beta\theta_i}$$

where,

$$\theta_i=\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\Big/\left(1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right)$$

$\frac{\partial l}{\partial\beta_r}$ can be approximated using the results:

$$\frac{\Gamma'(y)}{\Gamma(y)}=-\delta-\frac{1}{y}+\sum_{j=1}^{\infty}\frac{y}{j(y+j)}$$

where δ is Euler's number. Therefore

$$\frac{\partial l}{\partial\beta}\simeq\frac{n}{\beta-1}-\sum_{i=1}^{n}\frac{(1+x_i)}{\beta(1+x_i)-1}+\sum_{i=1}^{n}(1+x_i)\left[\log\left(1+\beta(1+x_i)\right)\right.$$

$$\left.-\log\left(\beta(1+x_i)+1-x_i\right)\right]+\sum_{i=1}^{n}(1+x_i)\log(1-\theta_i)$$

$$\frac{\partial l}{\partial\beta}=\frac{n}{\beta-1}-\sum_{i=1}^{n}\frac{(1+x_i)}{(\beta-1+\beta x_i)}+\sum_{i=1}^{n}\frac{\Gamma'(\beta+\beta x_i)}{\Gamma(\beta+\beta x_i)}(1+x_i)$$

$$-\sum_{i=1}^{n}\frac{\Gamma'(\beta+\beta x_i-x_i)}{\Gamma(\beta+\beta x_i-x_i)}(1+x_i)+\sum_{i=1}^{n}x_i\log\left[(1-\theta_i)\theta_i\right]+\sum_{i=1}^{n}\log(1-\theta_i)$$

$$=\frac{n}{\beta-1}+\sum_{i=1}^{n}\frac{(1+x_i)}{\beta-1+\beta x_i}+\sum_{i=1}^{n}\frac{\Gamma'(\beta+\beta x_i)}{\Gamma(\beta+\beta x_i)}(1+x_i)$$

$$-\sum_{i=1}^{n}\frac{\Gamma'(\beta+\beta x_i-x_i)}{\Gamma(\beta+\beta x_i-x_i)}(1+x_i)+\sum_{i=1}^{n}(1+x_i)\log(1-\theta_i)$$

$$\frac{\partial^2 l}{\partial\beta^2}\doteq-\frac{n}{(\beta-1)^2}-\sum_{i=1}^{n}\frac{(1+x_i)^2}{\left[\beta(1+x_i)-1\right]^2}$$

$$+\sum_{i=1}^{n}(1+x_i)\left\{\frac{(1+x_i)}{1+\beta(1+x_i)}-\frac{(1+x_i)}{1+\beta(1+x_i)-x_i}\right]$$

Simplifying we get:

$$-\frac{\partial^2 l}{\partial\beta^2}=\frac{n}{(\beta-1)^2}-\sum_{i=1}^{n}\frac{(1+x_i)^2}{\left[\beta(1+x_i)-1\right]^2}$$

$$+\sum_{i=1}^{n}\frac{x_i(1+x_i)^2}{\left[1+\beta(1+x_i)\right]^2-x_i\left[1+\beta(1+x_i)\right]}$$

$\sigma_{\beta\beta} = -E\left[\frac{\partial^2 l}{\partial \beta^2}\right]$ can be approximated by:

$$\sigma_{\beta\beta} = \frac{1}{(\beta-1)^2}\sum_{i=1}^{n}\theta_i(2-\theta_i) + (\beta-1)\sum_{i=1}^{n}\frac{\theta_i(1-\theta_i)^2}{(1-\beta\theta_i)(1-2\beta\theta_i+\beta)(1-3\beta\theta_i+\beta+\theta_i)}$$

The variance covariance matrix of the estimated parameters, and β based on the regression model is given by the inverse of Fisher's information matrix:

$$\Sigma = \begin{bmatrix} \sigma_{\gamma\gamma} & \sigma_{\gamma\beta} \\ & \sigma_{\beta\beta} \end{bmatrix}^{-1} = \begin{bmatrix} M & O \\ & C \end{bmatrix}$$

where M is and $q \times q$ symmetric matrix whose elements are m_{ij} so that $m_{ij} = \operatorname{cov}(\hat{\gamma}_i, \hat{\gamma}_j)$, and O is a $1 \times q$ matrix whose elements are $O_j = \operatorname{cov}(\hat{\gamma}_i, \hat{\beta})$ and C is a 1×1 element with $C = \operatorname{var}(\hat{\beta})$.

The simplest approach to obtain the maximum likelihood estimators of γ and β is by solving the equations;

$\frac{\partial l}{\partial \beta} = 0$ and $\frac{\partial l}{\partial \gamma_r} = 0, r = 1, 2, \cdots, r$ iteratively using a numeric technique such as Newton-Raphson. Following Cox and Hinkley [14], we have as $n \to \infty$ and under certain regularity conditions, the maximum likelihood estimators of $\hat{\phi} = (\hat{\gamma}, \hat{\beta})$ are asymptotically normal and consistent.

That is

$$V\overline{n}\left(\hat{\phi} - \phi\right) \to N_{q+1}(0, \Sigma) \text{ in law}$$

4-Limiting form of the QNBD: The Quasi-Poisson Distribution

As $\beta \to \infty, \theta \to 0$, so that $\beta\theta = \alpha$, the distribution (1) takes the following form:

$$P_x = \frac{\alpha^x}{x!}(1+x)^{x-1} \mathrm{e}^{-\alpha(1+x)}, \quad 0 < \alpha < 1$$

$$\mu = E(x) = \frac{\alpha}{1-\alpha}, \quad \operatorname{var}(x) = \frac{\alpha}{(1-\alpha)^3}$$

Therefore, $\operatorname{var}(x) = \mu(1+\mu)^2$. Expressing the distribution in terms of the mean parameter μ, the limiting distribution can be written as:

$$P_x = \frac{(1+x)^{x-1}}{x!}\left(\frac{\mu}{1+\mu}\right)^x \mathrm{e}^{-\frac{\mu}{1+\mu}(1+x)} \tag{18}$$

In a paper that follows, we shall discuss the issues of maximum likelihood estimation for the parameter μ of the probability function (18) and the regression model associated with it.

7. Data Analysis: RNA_SEQ Data: Modeling the Distribution of Read Counts

Over the past decade, various statistical analysis tools have been developed to analyze expression profiling data generated by microarrays (Reviewed in [15] [16] [17]). Before these tools can be applied to RNA-Seq data, it is worth noting that microarray data and RNA-Seq data are inherently different [16]. Microarray data is "analog" since expression levels are represented as continuous hybridization signal intensities. In contrast, RNA-Seq data is "digital", representing expression levels as discrete counts. This inherent difference leads to the difference in the parametric statistical methods that are used since they often depend on the assumptions of the random mechanism that generates the data. The Poisson, Binomial and Negative binomial distributions are more suitable for modeling discrete data in an RNA-Seq experiment. Therefore, a statistical method developed for microarray data analysis cannot be directly applied to RNA-Seq data analysis without first examining the underlying distributions. Recently several statistical methods have been developed to deal specifically with RNA-Seq count data [17]. In an RNA-Seq dataset, the expression levels of a specific gene were modeled using the Poisson distribution. This Poisson model is verified in the case where there are only technical replicates using a single source of RNA [15]. In the Poisson model, over-dispersion occurs if the sample variance is greater than the sample mean. There could be several sources that cause over-dispersion in RNA-Seq data, including the variability in biological replicates due to heterogeneity within a population of cells, possible correlation between gene expressions due to regulation, and other uncontrolled variations [18]. The existence of over-dispersion in real data was observed in several previous studies [18]. Popular models to safeguard against over-dispersion include the negative binomial distribution, or two-stage Poisson distribution [19], as discussed below.

When over-dispersion is observed across the samples, the gene counts cannot be estimated accurately by a simple Poisson model [20]. One way to handle this problem is to allow the Poisson mean to be a random variable and then model the gene counts by the marginal distribution of the mean count. Specifically, assume that the Poisson mean follows a Gamma distribution then the marginal distribution of the gene count has a Negative Binomial distribution with mean μ_i and variance = $\mu_i(1+\varepsilon\mu_i)$, where ε is the dispersion parameter [20].

Yoon and Nam [21] [22] showed that the gene dispersion value as estimated under the negative binomial modelling of read counts is the key determinant of the read count bias.

Whenever multiple samples are available and instead of modeling the raw expression, we model the gene counts as a function of the experimental sample and gene dispersion as covariates. For highly expressed genes we used the QNB regression model for published data that we downloaded from http://woldlab.caltech.edu/rnaseq/.

The published data were downloaded from http://www.ncbi.nlm.nih.gov/sra/

as the fastq files: SRA010153 for the MAQC data, SRP000727 for the human data (the two low-coverage MAQC samples were excluded), SRX000559-SRX000564 for the yeast data.

We analyzed the read count of the Mice-Brain tissue data under four experimental conditions:

Z_1 = Chrom_ chr11, Z_2 = Chrom chr9_ra, Z_3 = Chrom chrUn_ra, and Z_4 = Chrom chr13_ra, and d = the gene dispersion levels. Z_j are modeled as categorical variables with categorical with Z_4 being the reference category, and d is measured on the continuous scale. Figure 6 shows the histogram of the read counts for the 4 groups (Tables 2(a)-(d)).

We now analyze the data using three count regression models; the Poisson, the Negative binomial, and the QNB (Tables 3-5).

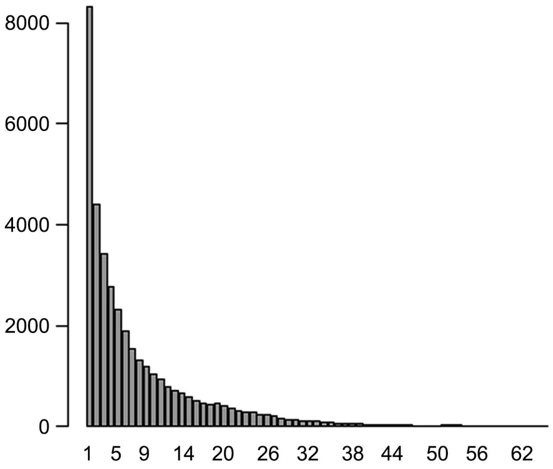

Figure 6. Histogram of the read counts data.

Table 2. (a) Summary statistics of the read count data for Chrom-Chr1 sample; (b) Summary statistics of the read count data for Chrom-Chr13 sample; (c) Summary statistics of the read count data for Chrom-Chr9_ran sample; (d) Summary statistics of the read count data for Chrom-ChrUn_ran sample.

(a) Chrom-chr1

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
d	36823	6.668	7.997	1.0	75.0
count	36823	7.99	8.905	1.0	68.0

(b) Chrom-chr13_ra

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
d	13	21.307	8.586	2.0	25.0
count	13	1.077	0.277	1.0	2.0

(c) Chrom-chr9_ran

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
d	698	10.126	9.293	1.0	50.0
count	698	3.030	2.369	1.0	13.0

(d) Chrom-chrUn_ran

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
d	89	22.843	6.626	1.0	25.0
count	89	1.157	0.541	1.0	4.0

Table 3. Fitting the data to Poisson regression model.

covariate	*Estimate*	*SE*	*P-Value*
(*Intercept*)	1.936	0.267	0.0000
Z_1	0.671	0.267	0.012*
Z_2	−0.021	0.268	0.939
Z_3	0.427	0.285	0.134
d	−0.127	0.0006	<0.000001

AIC: 314241

Table 4. Results of fitting data to the Negative Binomial regression model.

covariate	*Estimate*	*SE*	*P-Value*
(*Intercept*)	2.36746	0.33	0.06e-13***
Z_1	0.16405	0.33	0.619
Z_2	−0.39876	0.33	0.229
Z_3	0.21046	0.35	0.551
d	−0.10833	0.001	<0.00001

(Dispersion parameter for Negative Binomial (2.0488, with SE = 0.0185). **AIC: 214866**

Table 5. Results of fitting data to the QNBD.

	Estimate	*SE*	*P-Value*
(*Intercept*)	−5.1351	0.2826	0.0000
Z_1	0.1839	0.2696	0.2475
Z_2	0.1029	0.2704	0.3519
Z_3	0.1255	0.2872	0.3310
d	−0.0224	0.0005	0.0000
β *estimate*	134.8252	54.8250	0.0070

AIC = 213104

1) Modeling read count as a Poisson regression model glm(formula = $y \sim Z_1 + Z_2 + Z_3 + d$, family = poisson, data = ratdata);

2) Modeling read count using Negative Binomial to account for overdispersion;

3) Quasi negative binomial regression model.

8. Comments on the Data Fitting

We used three count regression models to fit the RNA-SEQ data. All models were fitted using the R package [23]. The first is a Poisson regression model, the second is the well-known negative binomial, and the third is the proposed QNB regression model. The Poisson model is fitted in R by applying the "GLM" while the negative binomial is fitted by using the "MASS" package in R. We provided the R-code for fitting the QNB in Appendix 2 in Appendix 2. We based the comparisons among these models on the AIC values (the smaller the better). Clearly, the Poisson model with the largest AIC = 314241, is the worst as it fails to properly account for the overdispersion in the data. Remarkable improvement is attained when the negative binomial regression model is used as its AIC = 214866. Although the QNB regression model has the smallest AIC = 213104, the improvement over the negative binomial is not tangible. We still believe that our proposed model should be a close competitor to the negative binomial model.

9. Discussion

There has been a growing interest among bioinformaticians and statisticians in constructing flexible distributions for counts that exhibit overdispersion to improve the modeling of count data. As a result, significant progress has been made towards generalizing some well-known discrete models, which have been successfully applied to problems arising in several areas of research. The proposed distribution was utilized to model two data sets; it was shown to provide a better fit than several other related models, including some with the same number of parameters. In the future paper, we shall demonstrate the applicability of the limiting form of our proposed distribution to genomics data together with inference procedures using multiple samples. Finally, we believe that the inferential results developed in this article should find numerous applications in bioinformatics, genomics, medicine, data engineering, and other areas of physical sciences.

Acknowledgements

The authors acknowledge the positive comments made by anonymous reviewers of this work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Takács, L. (1962) A Generalization of the Ballot Problem and Its Application in the Theory of Queues. *Journal of the American Statistical Association*, **57**, 327-337. https://doi.org/10.1080/01621459.1962.10480662

[2] Consul, P.C. and Gupta, H.C. (1980) The Generalized Negative Binomial Distribution and Its Characterization by Zero Regression. *SIAM Journal of Applied Mathematics*, **39**, 231-237. https://doi.org/10.1137/0139020

[3] Consul, P.C. and Shenton, L.R. (1972) Use of Lagrange Expansion for Generating Generalized Probability Distributions. *SIAM Journal of Applied Mathematics*, **23**,

239-248. https://doi.org/10.1137/0123026

[4] Consul, P.C. and Famoye, F. (2006) Lagrangian Probability Distributions. Birkhäuser, Boston.

[5] Shoukri, M.M. (1980) Estimation of Generalized Discrete Distributions. Unpublished PhD Thesis, The University of Calgary, Calgary.

[6] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*, *Series A*, **135**, 370-384. https://doi.org/10.2307/2344614

[7] Kendall, M. and Ord, K. (2009) The Advanced Theory of Statistics. Vol. 1, 6th Edition, Griffin, London.

[8] Rudick, R., Antel, J., Confavreux, C., Confavreux, C., Cutter, G., Ellison, G., *et al.* (1996) Clinical Outcomes Assessment in Multiple Sclerosis. *Annals of Neurology*, **40**, 469-479. https://doi.org/10.1002/ana.410400321

[9] Morgan, C.J., Aban, I.B., Katholi, C.R. and Cutter, G.R. (2010) Modeling Lesion Counts in Multiple Sclerosis When Patients Have Been Selected for Baseline Activity. *Multiple Sclerosis*, **16**, 926-934. https://doi.org/10.1177/1352458510373110

[10] Cramér, H. (1946) Mathematical Methods of Statistics. Princeton University Press, Princeton.

[11] Szegő, G. (1939) Orthogonal Polynomials. Vol. 23, Colloquium Publications, American Mathematical Society, New York.

[12] Shenton, L.R. and Wallington, P.A. (1962) The Bias of the Moment Estimators with an Application to the Negative Binomial Distribution. *Biometrika*, **49**, 193-204. https://doi.org/10.1093/biomet/49.1-2.193

[13] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. Chapman Hall, London.

[14] Cox, D.R. and Hinkley, D. (1974) Theoretical Statistics. Chapman and Hall, London.

[15] McCarthy, D.J., Chen, Y. and Smyth, G.K. (2021) Differential Expression Analysis of RNA-Seq Experiments with Respect to Biological Variation. *Nucleic Acids Research*, **40**, 4288-4297. https://doi.org/10.1093/nar/gks042

[16] Pan, W. (2002) A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics*, **18**, 546-554. https://doi.org/10.1093/bioinformatics/18.4.546

[17] Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays. *Genome Research*, **18**, 15-1517. https://doi.org/10.1101/gr.079558.108

[18] Koch, C.M., Chiu, S.F., Akbarpour, M., Bahart, A., Ridge, K.M., Bartom, E.T. and Winter, D.R. (2018) A Beginner's Guide to Analysis of RNA Sequencing Data. *American Journal of Respiratory Cell and Molecular Biology*, **59**, 145-157. https://doi.org/10.1101/gr.079558.108

[19] Yoon, S., Kim, S.Y. and Nam, D. (2016) Improving Gene-Set Enrichment Analysis of RNA-Seq Data with Small Replicates. *PLoS ONE*, **11**, e0165919. https://doi.org/10.1371/journal.pone.0165919

[20] Auer, P.L. and Doerge, R.W. (2011) A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*, **10**, 26. https://doi.org/10.2202/1544-6115.1627

[21] Yoon, S. and Nam, D. (2017) Gene Dispersion Is the Key Determinant of the Read

Count Bias in Differential Expression Analysis of RNA-Seq Data. *BMC Genomics*, **18**, Article No. 408. https://doi.org/10.1186/s12864-017-3809-0

[22] Robinson, M.D. and Smyth, G.K. (2008) Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data. *Biostatistics*, **9**, 321-332. https://doi.org/10.1093/biostatistics/kxm030

[23] https://cran.r-project.org/bin/windows/base/

Appendices

Appendix 1: R-CODE for Fitting the Univariate Version of the QNBD Using the Maximum Likelihood Method Applied to the "Brain Lesion" Data

```
QNBD<- function(x,theta,beta,log = FALSE){
loglik <- log((((beta-1)/(beta-1+beta*x))*(factorial(beta-1+beta*x))/
(factorial(x)*factorial(beta-1+beta*x-x))*
((theta^x)*(1-theta)^(beta*x+beta-1-x))))
if(log = = FALSE)
density <- exp(loglik)
else density<-loglik
return(density)
}
parameter <- maxlogL(x = x,dist = "QNBD",start = c(.01,2),optimizer = 'optim')
summary(parameter)
```

The fitting results by the method maximum likelihood are:

$\text{AIC} = 426.76,\ \tilde{\theta} = 0.298 \pm 0.015,\ \tilde{\beta} = 2.81 \pm 0.057$

Appendix 2: R-CODE: QNB Regression Fitting by the Method of Maximum Likelihood Applied to the RNA_SEQ Read Count Data

```
llik=function(y,par){
b0=par [1]
b1=par [2]
b2=par [3]
b3=par [4]
b4=par [5]
beta=par [6]
n=length(y)
eta=b0+b1*x1+b2*x2+b3*x3+b4*x4
mu=exp(eta)/(1+exp(eta))
ll=sum(log(beta-1)-log(beta-1+beta*y)
   +lgamma(beta+beta*y)-lgamma(1+y)-lgamma(beta+beta*y-y)
   +y*log(mu)+(beta+beta*y-1-y)*log(1-mu))
return(-ll)
}
res=optim(par=c(2,.6,-.02,.42,-.12,2.1),llik,y=y,method="BFGS",hessian=T)
theta=res$par
theta
#CALCULATING THE STANDARD ERRORS OF MLE
out3=nlm(llik,theta,y=y,hessian=TRUE)
fish=out3$hessian
solve(fish)
element=diag((solve(fish)))
```

```
se=sqrt(element)
qqnorm(y,resid(out3))
z=theta/se
p_value=1-pnorm(abs(z))
result.GNBD=data.frame(theta,se,z,p_value)
result.GNBD=round(result.GNBD,4)
```