Inference Procedures on the Ratio of Modified Generalized Poisson Distribution Means: Applications to RNA_SEQ Data

M.M. Shoukri[1,*] and Maha Al-Eid[2]

[1]Department of Epidemiology and Biostatistics, Schulich of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada

[2]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

Abstract: The Poisson and the Negative Binomial distributions are commonly used as analytic tools 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 is appropriate to model over-dispersed count data. The Generalized Poisson Distribution is becoming a popular alternative to the Negative Binomial. We have considered inference procedures on a modified form of this distribution when two samples are available from two independent populations and the target effect size of interest is the ratio of the two population means. The statistical objective is to construct confidence limits on the ratio. We first test the presence of over dispersion and derive several estimators in the single sample situation. When two samples are available, our interest is focused on the estimation of an effect size measured by the ratio of the respective population means. We have compared two methods; namely the Fieller's and the delta methods in terms of coverage probabilities. We have illustrated the methodologies on published genomic datasets.

Keywords: Overdispersion, Parameter orthogonality, Fieller's theorem, Mixed estimator, Delta method, Coverage probabilities.

1. INTRODUCTION

The Poisson distribution is commonly used to model count data. However, a restriction of this distribution is that the response variable must have a mean equal to the variance. This restriction does not often hold for many biological and epidemiological data. In reality, the variance can be much larger than the mean, a phenomenon known as "overdispersion". This overdispersion may occur due to population heterogeneity, or the presence of outliers in the data [1]. An analysis of data with overly dispersed counts can lead to the underestimation of parameter standard error if overdispersion is ignored. A review of the issue of overdispersion in both binary and count data was reviewed by Hinde and Demetrio [2], and in a more recent review by Hayat and Higgins [3]. Diagnosing and accounting for overdispersion is not a simple issue and should be appropriately dealt with to avoid bias in interpreting the results.

When overdispersion is suspected, the Negative-Binomial (NB) distribution has been adopted as a common alternative to the Poisson distribution. The NB has two parameters and a variance that is a quadratic function of the mean and has therefore has been the model of choice to model count data that exhibit overdispersion. Since accounting for measured covariates is one of the methods used to address the issue of over dispersion by including them in a regression model, Hinde [4] reviewed the methodologies of NB regression. Joe and Zhu [5] drew a comparison between the NB and a mixture-based generalization of the Poisson distribution.

In this paper, we discuss several inferential statistical issues related to a modified form of the Generalized Poisson Distribution (GPD). The GPD distribution was introduced to the statistical literature by Consul and Jain [6] and a detailed account of its properties was given by Consul [7]. The distribution has two parameters and has variance larger than the mean. This makes the GPD an attractive competitor of the Negative Binomial Distribution (NBD). The distribution has been used to analyze data in the fields of genetics [8] as a queuing model [9,10,11] and genomics [12]. The modified form of the GPD, which we shall call "Modified Poisson Distribution" (MGPD)was first discussed in [9]. The modification was the result of a double parametric transformation on the original parameters of the GPD. The main purpose of the transformation is to achieve parameters orthogonality [13], which will improve the statistical properties of the maximum likelihood estimators, and make the MGPD a member of the "Generalized Linear Models" [14].

There are situations when the researchers have the opportunity to study count data under two experimental conditions. One of the questions of interest is to conduct statistical inference on the ratio of the mean counts. To the best of our knowledge, the issue of constructing a confidence interval on the ratio of means of two MGPD's has not been discussed before.

The paper is divided into 6 sections. In Section 1, we discuss the issue of parameters estimation in single

*Address correspondence to this author at the Department of Epidemiology and Biostatistics, Schulich of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada; Tel: +12269778651;
E-mail: mmshoukr@uwo.ca, Shoukri.mohamed@gmail.com

E-ISSN: 1929-6029/20

samples. We study several estimators and evaluate their asymptotic efficiencies relative to the method of maximum likelihood. In Section 2 we introduce the score test for overdispersion. In Section 3 we consider the problem of testing for over dispersion, and in section 4 we deal with the problem of constructing confidence intervals on the ratio of means, where we compare two approaches; the Fieller's interval and the delta method. In section 5 illustrate the methodology on published data arising from reading counts of and RNA sequencing of gene expressions data. The conventional abbreviation is RNA_SEQ. General discussion will be given in section 6.

2. MODIFIED GENERALIZED POISSON DISTRIBUTION

The GPD was introduced by Consul and Jain [6]

$$P(X=x)=\frac{\lambda_1(\lambda_1+\lambda_2 x)^{x-1}}{x!}\exp[-\lambda_1-\lambda_2 x] \quad (1)$$

$$\lambda_1>0$$

$$0 \leqslant \lambda_2<1$$

The GPD whose probability function is given in (1) reduces to the well-known Poisson distribution when $\lambda_2=0$. Therefore the parameter λ_2 with the above restriction on its range, is considered the dispersion parameter. Shoukri and Mian [9] employed the parametric transformations:

$$\lambda_1=\mu/(1+\epsilon\mu)$$

$$\lambda_2=\epsilon\lambda_1 \quad (2)$$

Therefore, equation (1) reduces to:

$$P(X=x)=\frac{(1+\epsilon x)^{x-1}}{x!}\ g^x(\mu,\epsilon)\exp\left[-\frac{\mu}{1+\epsilon\mu}\right] \quad (3)$$

where $g(\mu,\epsilon)=\frac{\mu}{1+\epsilon\mu}\exp\left[\frac{-\epsilon\mu}{1+\epsilon\mu}\right]$

For fixed ϵ, the function g(.) is the natural parameter the transformation which renders the GPD a member of the linear family of exponential class (see; McCaullagh and Nelder [14]):

$$f(x)=h(x)\exp[\phi T(x)-A(\phi)]$$

We call the transformed GPD, the "Modified Generalized Poisson Distribution" or MGPD

Shoukri and Mian [9] showed that a recurrence relation among the *r*th non-central moments ϑ_r' is such that:

$$\vartheta_{r+1}'=\sigma^2(\mu)\frac{\partial\vartheta_r'}{\partial\mu}+\mu\vartheta_r' \quad (4)$$

Here, $\vartheta_0'\equiv 1, \vartheta_1'\equiv\mu=E(Y)$.

Moreover, $\sigma^2(\mu)=\mu(1+\epsilon\mu)^2\equiv \text{var}(Y)$ (5)

Equation (5) shows that variance is a cubic function of the population mean. Of interest to us is the situation when $\epsilon>0$.

Using the recurrence relation (4) one can show that the higher central moments are given by:

$$\vartheta_3=E(Y-\mu)^3=\mu(1+\epsilon\mu)^3(1+3\epsilon\mu) \quad (6)$$

$$\vartheta_4=E(Y-\mu)^4=\mu(1+\epsilon\mu)^4(1+3\mu+10\epsilon\mu+15\epsilon^2\mu^2) \quad (7)$$

2.2. Estimation of the Model Parameters

2.2.1. Maximum Likelihood Estimators

Let $y_1, y_2, \ldots . y_n$ denote a random sample from the MGPD.

The likelihood function is given by:

$$L=\prod_{i=1}^{n}\left[\frac{(1+\epsilon y_i)^{y_i-1}}{y_i!}g^{y_i}(\epsilon,\mu)\right]\exp\left[\frac{-n\mu}{1+\epsilon\mu}\right]$$

where $g(\epsilon,\mu)=\frac{\mu}{1+\epsilon\mu}\exp\left[\frac{-\epsilon\mu}{1+\epsilon\mu}\right]$

The log-likelihood function:

$$l=\sum_{i=1}^{n}(y_i-1)\ln(1+\epsilon y_i)+n\bar{y}\left[\frac{-\epsilon\mu}{1+\epsilon\mu}+\log\frac{\mu}{1+\epsilon\mu}\right]-\frac{n\mu}{1+\epsilon\mu}$$

The first partial derivatives of the log-likelihood function with respect to the model parameters are:

$$\frac{\partial l}{\partial\mu}=+n\bar{y}\left(\frac{1}{\mu(1+\epsilon\mu)^2}\right)-n\frac{1}{(1+\epsilon\mu)^2} \quad (7)$$

Equating $\frac{\partial l}{\partial\mu}$ to zero we get, explicit unique solution for μ as $\hat{\mu}=\bar{y}$. On the other hand;

$$\frac{\partial l}{\partial\epsilon}=\sum_{i=1}^{n}\frac{y_i(y_i-1)}{1+\epsilon y_i}-\frac{\bar{y}}{1+\epsilon\bar{y}}=0 \quad (8)$$

Consul and Shoukri [15] showed that equation (8) has a unique root in $\hat{\epsilon}$, in (0,1) if and only if $s^2>\bar{y}$.

We can also show that:

$$-E\left(\frac{\partial^2 l}{\partial\mu^2}\right)=\frac{n}{\mu(1+\epsilon\mu)^2} \quad (9)$$

$$-E\left(\frac{\partial^2 l}{\partial\mu\partial\varepsilon}\right)=0 \quad (10)$$

Equation (10) indicates that the model parameters are orthogonal. Moreover;

$$-E\left(\frac{\partial^2 l}{\partial\epsilon^2}\right)=\frac{2n\mu^2}{(1+2\epsilon+4\epsilon^2\mu+\epsilon\mu+\epsilon^2\mu^2+2\mu^2\epsilon^3)} \quad (11)$$

Hence, the variance of the maximum likelihood estimators are:

$var(\hat{\mu})$ = 1/ $-E\left(\frac{\partial^2 l}{\partial \mu^2}\right)$, $var(\hat{\epsilon})$ = 1/ $-E\left(\frac{\partial^2 l}{\partial \epsilon^2}\right)$, and the two estimators ($\hat{\mu}$, $\hat{\epsilon}$) are stochastically independent because they are orthogonal to each other as shown in equation (10).

2.2.2. Moment Estimators

Equating the first two sample moments to their corresponding population moments

$$\bar{y} = \mu$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mu(1+\epsilon\mu)^2$$

and solving for the parameters we get:

$$\tilde{\mu} = \bar{y}$$

$$\tilde{\epsilon} = (s^2)^{1/2}(\bar{y})^{-3/2} - (\bar{y})^{-1}$$

The variance of the ml estimators are (due to the parameters orthogonality)

$$\text{var}(\hat{\mu}) = \mu\,(1+\epsilon\mu)^2/n \qquad (12)$$

$$\text{var}(\hat{\epsilon}) = \frac{1}{2n\mu^2}[1 + 2\epsilon + 4\epsilon^2\mu + \epsilon\mu + \epsilon^2\mu^2 + 2\mu^2\epsilon^3] \quad (13)$$

And for the moment estimator

$$\text{var}(\tilde{\epsilon}) = \frac{(1+\epsilon\mu)^2}{2n\mu^2}[1 + 2\epsilon + 3\epsilon^2\mu] \qquad (14)$$

The relative efficiency of the moment estimator of the dispersion parameter is measured by the ratio of the variance of the maximum likelihood estimator to the variance of the corresponding estimator.

Table 1: Efficiency of the Moment Estimator for the Dispersion Parameter

e	m	eff_moment
0.00	1	1.00
0.01	2	0.98
0.01	3	0.97
0.01	4	0.96
0.01	5	0.95
0.20	10	0.45
0.20	20	0.32
0.20	50	0.17
0.20	100	0.10
0.50	10	0.19
0.50	20	0.11
0.50	50	0.04
0.50	100	0.02

RELEFF = $\text{var}(\hat{\epsilon})/\text{var}(\tilde{\epsilon})$. Calculations are given in Table **1** for a few values of the model parameters. The relative efficiency (eff_moment) of the moment estimator is quite high for small values of the mean and the dispersion parameter but declines rapidly as both parameters increase.

Another type of estimator for the dispersion parameter (**ε**) which has not been discussed before is the so-called mixed estimator. We consider this estimator in the next sub-section.

2.2.3. Mixed Estimators

Here we use two-sample statistics to estimate the model parameters. Let $y_1, y_2, \ldots y_n$ be the outcomes of a simple random sample let n_0 denote the count of zeros. Clearly $(n_0/n\,, \bar{y})$ are sufficient statistics for (P_0, μ), where

$$P_0 = P_r[y_i = 0] = \exp\left[\frac{-\mu}{1+\epsilon\mu}\right]$$

Solving the equations:

$\hat{\mu} = \bar{y}$, and $\hat{P}_0 = \frac{n_0}{n} = \exp\left[\frac{-\hat{\mu}}{1+\hat{\epsilon}\hat{\mu}}\right]$, for $\hat{\epsilon}$ we get:

$$\hat{\epsilon}_0 = -1/\log\hat{P}_0 - 1/\bar{y}$$

To find the variance of $\hat{\epsilon}_0$, we use the delta method so that to the first order of approximation we have

$$\text{var}(\hat{\epsilon}_0) = \text{var}\left(\log\hat{P}_0\right)\left(\frac{\partial\hat{\epsilon}_0}{\partial\log\hat{P}_0}\right)^2 +$$

$$\text{var}(\bar{y})\left(\frac{\partial\hat{\epsilon}_0}{\partial\bar{y}}\right)^2 + 2\text{cov}\left(\log\hat{P}_0\,, \bar{y}\right)\left(\frac{\partial\hat{\epsilon}_0}{\partial\log P_0}\right)\left(\frac{\partial\epsilon_0}{\partial\bar{y}}\right)$$

It is known that $\text{var}\left(\log\hat{P}_0\right) = \frac{1-P_0}{nP_0}$, and $\text{var}(\hat{\mu}) = \frac{\mu(1+\epsilon\mu)^2}{n}$, however, the derivation of

$cov\left(\log\hat{P}_0\,, \bar{y}\right)$ is not straight forward. To derive the covariance between the sample mean and the fraction of zeros in the sample we proceed as follows:

Since $n_0 = n\hat{P}_0$ has binomial distribution $n_0 \sim \text{bin}(n, P_0), E(n_0) = nP_0$ and y the sample total has expected value $n\mu$, one can show that the joint MGF of (y, n_0);

$$M(t_1, t_2) = E\left[\begin{matrix} t_1 y \\ e \end{matrix} \quad + \quad t_2 n_0\right]$$

$$= [M(t_1) + P_0(e^{t_2} - 1)]^n$$

The function $M(t_1)$ is the mgf of the MGPD, which does not have an explicit expression.

Differentiating $M(t_1, t_2)$ with respect to t_1 and t_2, setting $t_1 = t_2 = 0$, we get

$$E\left(\bar{y}\hat{P}_0\right) = \frac{n+1}{n}P_0\mu$$

Therefore

$$\text{cov}\left(\bar{y}, \hat{P}_0\right) = \frac{P_0\mu}{n}$$

Using the delta method, we can show that

$$\text{cov}(\bar{y}, \log \hat{P}_0) = E\left[(\bar{y}-\mu)(\hat{P}_0 - P_0)\cdot \frac{\partial \log \hat{P}_0}{\partial P_0}\right]$$

$$= -\frac{\mu}{n}$$

Direct substitution gives:

$$var(\hat{\epsilon}_0) = \frac{(1+\epsilon\mu)^4}{n\mu^4}\left[\exp\left(\frac{\mu}{1+\epsilon\mu}\right) - 1\right] - \frac{(1+\epsilon\mu)^2}{n\mu^3} \quad (15)$$

Similar to the calculation of the efficiency of the moment estimator we measure the efficiency of the mixed estimator by the ratio:

$$\text{var}(\hat{\epsilon})/var(\hat{\epsilon}_0)$$

Table **2** shows the relative efficiency of the mixed estimator (eff_mixed). The behavior of the relative efficiency of the mixed estimator is similar to that of the moment estimator in terms of variations in the parameter values. However, the relative efficiency of the moment estimator of the dispersion parameter is lower than that of the mixed estimator, for large values of the population mean and the dispersion parameter.

Table 2: Efficiency of the Mixed Estimator for the Dispersion Parameter

eps	mu	eff_mixed
0.01	2	0.45
0.01	3	0.28
0.01	4	0.17
0.01	5	0.10
0.20	10	0.25
0.20	20	0.18
0.20	50	0.15
0.20	100	0.13
0.50	10	0.64
0.50	20	0.63
0.50	50	0.63
0.50	100	0.63

3. TESTING FOR OVERDISPERSION: SAMPLE SIZE REQUIREMENTS TO DETECT OVER-DISPERSION USING THE SCORE TEST

As mentioned, the MGPD reduces to the Poisson distribution when the dispersion parameter ϵ is set equal to zero. Therefore, to construct a goodness of fit test where the null hypothesis is that the available data is drawn from a Poisson distribution against an alternative in the direction of the MGPD, our best approach is to use the score-testing. The score function is obtained by differentiating the log-likelihood function with respect to the dispersion parameter, and setting the value of the dispersion parameter equal to zero. The advantage of the score test is that the test statistic is evaluated only under the null hypothesis [16]. We proceed as follows:

Based on a $srs, (y_1, y_2, \ldots y_n)$ the score test on the null hypothesis $H_0: \epsilon = 0$ against one-sided alternative $H_a: \epsilon > 0$ is given by:

$$T = \frac{\partial l}{\partial \epsilon}\Big|_{\epsilon=0} = n[s^2 - \bar{y}] \quad (16)$$

In equation (16) s^2 and $\bar{y}$ are respectively the sample variance and the sample mean.

The mean and variance of T are given as:

$$E(T) = n[\mu(1+\epsilon\mu)^2 - \mu] \equiv nM(\epsilon)$$

$$\text{var}(T) = n\mu(1+\epsilon\mu)^2[2\mu + 4\epsilon\mu + 4\epsilon\mu^2 + 30\epsilon^2\mu^2$$

$$+2\epsilon^2\mu^3 + 40\epsilon^3\mu^3 + 15\epsilon^4\mu^4]$$

$$\equiv nV(\epsilon)$$

For Type I error rate α, and power $1-\beta$ the approximate sample size *n* to detect the departure from the Poisson (i.e. $\epsilon = 0$) in the direction of MGPD is:

$$n \simeq \left\{\frac{z_\alpha \mu\sqrt{2} + z_\beta\sqrt{V(z_\alpha\mu\sqrt{2})}}{M(\epsilon)}\right\}^2 \quad (17)$$

For example, when $\alpha = 0.05, \beta = 0.20$

$$\mu = 1, \epsilon = .01, \text{then } n = 13314$$

$$\mu = 2, \epsilon = .01, \text{then } n = 823$$

$$\mu = 1, \epsilon = .05, \text{then } n = 512$$

$$\mu = 2, \epsilon = 0.05, \text{then } n = 30$$

In Table **3** we show the empirical power for small, moderate, and large samples at a 5% level of significance. The power increases the farther away ϵ from its null value, when the sample size is large and when the population means is large as well.

In the previous sections we dealt with the problem of statistical estimation in a single sample. This was inevitable to properly deal with the two samples situation.

3.1. Estimation of the Ratio of Two Means

Let Y1, Y2 be random variables with expected values E(Y1) and E(Y2). Of interest is the ratio of the two means; R= E (Y1)/E(Y2) = μ_1/μ_2. We know that the unbiased maximum likelihood estimators of the population mean, are the sample means. We assume that we have two independent sample $y_{11}, y_{12}, \ldots\ldots y_{1n1}$ from a MGPD with parameters (μ_1, ϵ_1), and another independent sample $y_{21}, y_{22}, \ldots\ldots y_{2n2}$ from independent MGPD with parameters (μ_2, ϵ_2). We shall discuss two methods for constructing confidence levels on the ratio of mean R.

Table 3: Table of Empirical Power of the Overdispersion Test

μ	$n = 20$				$n = 500$				$n = 100$			
	ϵ				ϵ							
	.001	.01	.02	.05	.001	.01	.02	.05	.001	.01	.02	.05
1	.052	.063	.077	.13	.052	.068	.088	.168	.052	.073	.101	.22
2	.053	.075	.110	.23	.053	.086	.133	.325	.054	.099	.169	.452
3	.054	.089	.14	.33	.055	.110	.188	.492	.056	.130	.253	.672
4	.055	.103	.18	.44	.056	.130	.250	.639	.058	.166	.349	.824
5	.056	.120	.22	.53	.058	.160	.318	.751	.060	.210	.450	.912

4. FIELLER'S METHOD

The approach to construct confidence limits on the ratio of means is the Fieller's method [16,17], applied to independent samples with unequal variances as was shown for the normal distribution [18].

We denote the maximum likelihood estimator of R by $\delta_0 = \bar{Y}_1/\bar{Y}_2$. Furthermore, we denote

$$V_i = var(\bar{Y}_i) = \mu_i(1 + \varepsilon_i\mu_i)^2/n_i \ , \ i = 1,2$$

Let $U = \bar{Y}_1 - \delta_0\bar{Y}_2$

$$\text{Var}(U) = V_1 + \delta_0^2 V_2$$

For a Type I error rate α, we have:

$$\alpha = \Pr\left[\frac{(\bar{Y}_1 - \delta_0\bar{Y}_2)^2}{V(u)} \geqslant Z_{\alpha/2}^2\right] \quad (18)$$

The inequality in the square bracket in equation (18) may be written as:

$$\bar{Y}_1^{\ 2} + \delta_0^2\overline{Y_2^2} - 2\delta_0\bar{Y}_1\bar{Y}_2 > Z_{\alpha/2}^2[V_1 + \delta_0^2 V_2]$$

$$or\ \bar{Y}_1^{\ 2} + \delta_0^2\overline{Y_2^2} - 2\delta_0\,\bar{Y}_1\bar{Y}_2 - V_1 Z_{\alpha/2}^2 - \delta_0^2 V_2 Z_{\alpha/2}^2 \geqslant 0$$

$$\delta_0^2\left[\overline{Y_2^2} - V_2 Z_{\alpha/2}^2\right] - 2\delta_0\bar{Y}_1\bar{Y}_2 + \bar{Y}_1^{\ 2} - V_1 Z_{\alpha/2}^2 \geqslant 0$$

Solving the quadratic for δ_0 we get:

$$\delta_0 = \frac{2[\bar{Y}_1\bar{Y}_2] \pm \sqrt{4(\bar{Y}_1\bar{Y}_2)^2 - 4\left(\overline{Y_2^2} - V_2 Z_{\alpha/2}^2\right)\left(\bar{Y}_1^{\ 2} - V_1 Z_{\alpha/2}^2\right)}}{2\left(\overline{Y_2^2} - V_2 Z_{\alpha/2}^2\right)}$$

Simplifying we get:

$$\delta_0 = \frac{\bar{Y}_1\bar{Y}_2 \ \pm\left[\left(V_1\overline{Y_2^2} + V_2\bar{Y}_1^{\ 2}\right)Z_{\alpha/2}^2 \ - (V_1)(V_2)Z_{\alpha/2}^4\right]^{1/2}}{\left(Y_2^2 - V_2 Z_{\alpha/2}^2\right)} \quad (19)$$

$$= A \pm B$$

where

$$A = \frac{\bar{Y}_1\bar{Y}_2}{Y_2^2 - V_2 Z_{\alpha/2}^2}$$

$$B = \frac{Z_{\alpha/2}\left[\left(V_1\overline{Y_2^2} + V_2\bar{Y}_1^{\ 2}\right) - (V_1)(V_2)Z_{\alpha/2}^2\right]^{1/2}}{Y_2^2 - V_2 Z_{\alpha/2}^2}$$

When Fieller's confidence set for the ratio is finite, then it is given by $I = (A - B, A + B)$.

We can establish bioequivalence using the Fieller's theorem. The confidence set is an interval if $Y_2^2 - V_2 Z_{\alpha/2}^2 > 0$ and I is contained in $(\delta_0, \delta_0^{-1})$, that is if Fieller's confidence interval is included in the equivalence range. Usually $\delta_0 = 0.80 \Rightarrow \delta_0^{-1} = 1.25$. Hence as shown in [19] and [20] the equivalence range is $(0.8, 1.25)$.

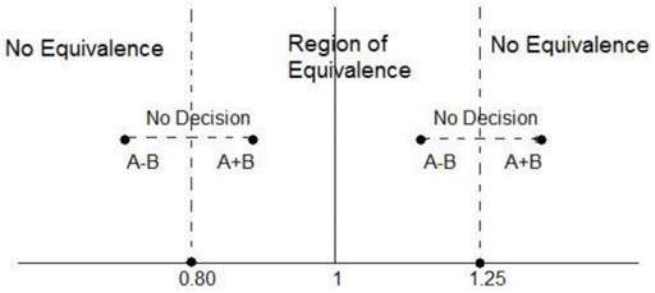


Figure 1: Region of Bioequivalence of the Fieller's interval.

4.1. Delta Method

From [21], the variance of the ratio of two random variables is, to the first order of approximation given by:

$$\text{Var}(\rho) = \left(\frac{\partial\rho}{\partial\bar{y}_1}\right)^2 \text{Var}(\bar{y}_1) + \left(\frac{\partial\rho}{\partial\bar{y}_2}\right)^2 \text{Var}(\bar{y}_2)$$

$$Var\ (\hat{\rho}) = \frac{\rho(1+\epsilon_1\,\mu_1)^2}{n_1\mu_2} + \frac{\rho 2(1+\epsilon_2\,\mu_2)^2}{n_2\mu_2} \quad (20)$$

Therefore, an approximate $(1 - \alpha)100\%$ confidencein the ratio of means is given by:

$$\rho \pm Z_{\alpha/2}\ \sqrt{Var\ (\hat{\rho})}$$

In general, the estimated ratio of two means is biased. The magnitude of bias can be obtained again by using the Delta method and is given to the first order of approximation as:

Table 4: Coverage Percentage of the Delta and Fieller's Methods. Nominal Coverage is 95%

n	ρ	μ_2	ϵ	Delta	Fieller
10	.1	1	.01	16	21
10	.1	2	.01	11	13
10	.1	3	.01	9	10
10	.1	5	.01	7	8
20	2	1	.05	20	13
20	2	2	.05	13	10
20	2	3	.05	10	8
20	2	5	.05	8	7
100	1	5	.00	6	6
100	1	5	.05	7	7
100	1	5	.06	8	8
100	1	5	.10	9	9
100	1	5	.90	33	30

$$\text{Bias}(\hat{\rho}) = \frac{1}{2!}\left[\text{Var}(\bar{y}_1)\frac{\dot{\partial}^2\hat{\rho}}{\partial\bar{y}_1{}^2} + \text{Var}(\bar{y}_2\ )\frac{\dot{\partial}^2\hat{\rho}}{\partial\bar{y}_2{}^2}\right]$$

We simplify the above expression to get:

$$\text{Bias}(\hat{\rho}) = \frac{\rho(1+\epsilon_2\,\mu_2)^2}{n_2\mu_2}$$

In Table **4**, we compare the empirical coverage probabilities of Feiller's to those of the delta method, for selected values of the population parameters, and nominal level of significance 5%. To simplify the table, we assume the homogeneity of the dispersion parameters of the two populations. As can be seen, for small values of the ratio and small values of the common dispersion parameter, the Fieller's theorem gives better coverage. However, for larger values of the dispersion parameter, both methods seem to have similar coverage probabilities.

5. APPLICATIONS (RNA-SEQ DATA)

Gene expression is the process by which information from a gene is used in the synthesis of a the functional gene product, which may be proteins. A gene is declared differentially expressed if an observed difference or change in reading counts or expression levels between two experimental conditions is statistically significant. To identify differentially expressed genes between two conditions, it is important to find the statistical distributional property of the data to approximate the nature of differential genes. As we have already indicated, the Poisson distribution is ubiquitous in the analysis of count data. It is usually assumed that the position-level read count follows a Poisson distribution with a rate λ. But evidence from a large body of data do not support the Poisson assumption of the equality of mean and variance [22,23]. Robinson and Smythe [24] used the negative binomial distribution to analyze tag abundance to account for overdispersion in this type of data.

In the present study, the focus is mainly is on investigating the differential gene expression analysis for sequence data based on MGPD. This approach was applied in RNA-seq read count data [12] where the authors used the original form the model given in equation (1). Thus, fitting of appropriate distribution to gene expression data provides statistically sound cutoff values for identifying differentially expressed genes. One of the basic questions while analyzing genomic data is related to the identification of the appropriate distribution of the position level read counts. This distribution if proven appropriate allows:

i. Better estimation of gene expressions

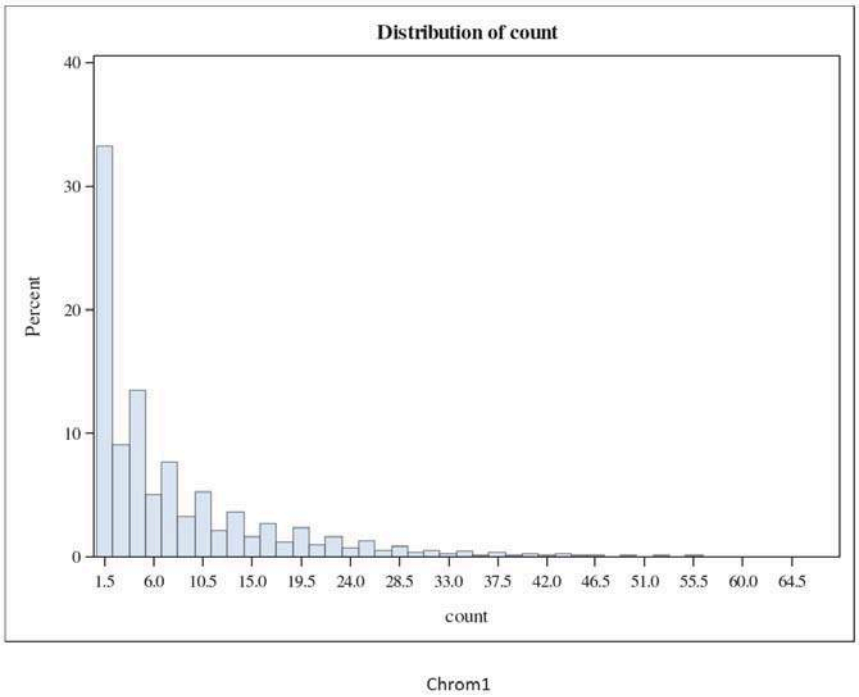
ii. Improving the identification of differentially expressed genes

The proposed MGPD will be used to re-analyze the data (Sudeep & Chen [12]) for some highly expressed genes. 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 two experimental conditions named (Chrom1, and Chrom9) using the MGPD.

Figures **2** and **3** show the histograms of the read counts for the Chrom1 and Chrom9 respectively.

As can be seen from Figures **1** and **2**, the data are heavily skewed due to the presence of outliers. This may be one of the reasons for overdispersion in the data. In Tables **5** and **6** we present the summary statistics under the two experimental conditions.

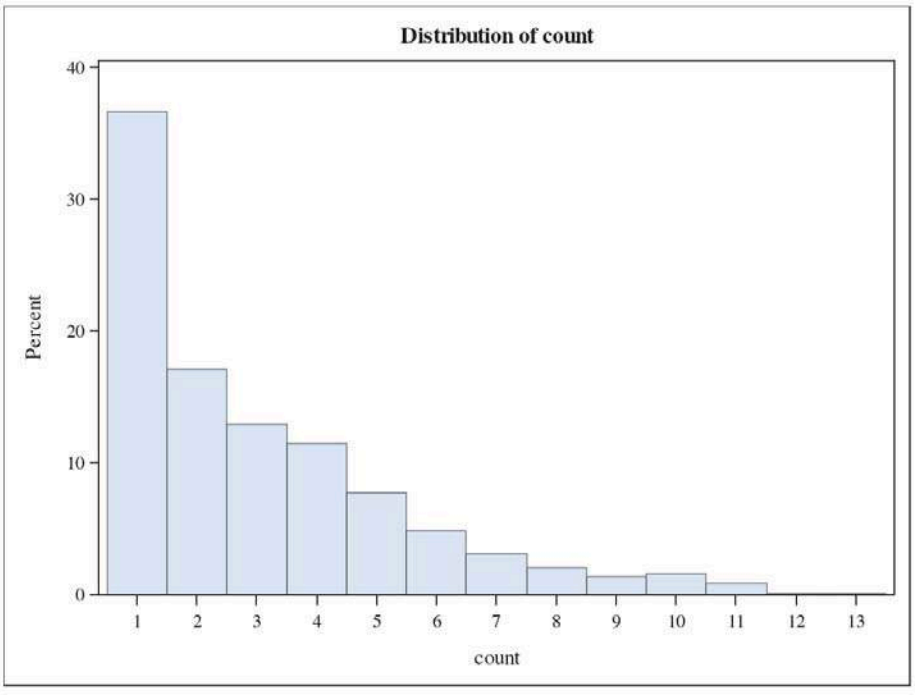


Chrom1

Figure 2: Histogram of the read count for the first sample (Chrom 1) from Mice-Brain tissues.

Table 5: Summary Statistics for the Chrom1 Sample

Moments			
N	36823	**Sum Weights**	36823
Mean	7.94	**Sum Observations**	292689
Std Deviation	8.90	**Variance**	79.31
Skewness	2.10	**Kurtosis**	5.41
Coeff Variation	112.04	**Std Error Mean**	0.046



Chrom9

Figure 3: Histogram of the read counts for the second sample from Mice-brain tissues.

Table 6: Summary Statistics for the Chrom9 Samples

Moments			
N	698	**Sum Weights**	698
Mean	3.03	**Sum Observations**	2115
Std Deviation	2.36	**Variance**	5.61
Skewness	1.38	**Kurtosis**	1.64
Coeff Variation	78.17	**Std Error Mean**	0.090

Table 7: Results of the Data Analyses using the Maximum Likelihood for Point Estimation

Data	Sample size	Sample mean	SE	ε
Chrom1	36823	7.94	0.046	0.519±0.0024
Chrom9	698	3.03	0.090	0.270±0.0013
		$\hat{\rho}$= 2.62		

The likelihood estimators of the dispersion parameters and their standard errors are given in Table **7**.

The construction of the 95% confidence intervals on the ratio of means based on the delta method and Fieller's theorems show that:

Delta method: $\{2.411 < \rho < 2.829\}$ & Fieller's method $\{2.473 < \rho < 2.785\}$

Because the sample sizes are large the two methods give almost the upper and lower limits for the same 95 % confidence level. Moreover, the Fieller's limits show the non-equivalence of the two population means as indicated in Figure **1**.

6. DISCUSSION

In this paper, we demonstrated the applicability of the modified form of the generalized Poisson distribution. The modification is, in fact, a double transformation on the original model parameters. We used the score testing to assess the departure of the model from the Poisson distribution, and provided sample size justifications, and evaluated the power of this test. The inference procedure on the ratio of two means was evaluated by estimating the coverage probabilities using simulations.

There are situations however when data may be available from multiple samples. The two questions of interest are:

i. how to test the homogeneity of the dispersion parameters in two or more MGPD models, and

ii. how to test the homogeneity of several MGPD means in the presence of covariates. This is equivalent to the ANCOVA model

Both questions are under investigation by the authors of this paper.

CONFLICT OF INTEREST

None declared by both authors.

ACKNOWLEDGEMENT

The authors acknowledge the constructive comments made by an anonymous reviewer and the Editorial Manager.

REFERENCES

[1] Cox DR. Some remarks on overdispersion. Biometrika 1983; 70: 269-274. https://doi.org/10.1093/biomet/70.1.269

[2] Hinde J, Demetrio CGB. Overdispersion: Models and estimation. Computational statistics and Data Analysis 1998; 27: 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3

[3] Hayat MJ, Higgins M. Understanding Poisson regression. Journal of Nursing Education 2014; 53: 207-215. https://doi.org/10.3928/01484834-20140325-04

[4] Hinde JM. Negative binomial regression. Cambridge University Press 2007; P2011.

[5] Joe H, Zhu R. Generalized Poisson distribution: The property of mixture of Poisson and comparison with the negative binomial distribution. Biometrical Journal 2005; 47: 219-229. https://doi.org/10.1002/bimj.200410102

[6] Consul PC, Jain GC. Ona generalization of Poisson distribution. ABSTRACT, Annals of Mathematical Statistics 1970; 41: 1387.

[7] Consul PC. Generalized Poisson Distribution. Marcel Dekker Inc., New York 1989.

[8] Janardan KG, Schaeffer DJ. Models for the analysis of chromosomal aberrations in human leukocytes. Biometrical J 1977; 19: 599-612. https://doi.org/10.1002/bimj.4710190804

[9] Shoukri MM, Mian IUH. Some aspects of statistical inference on the Lagrange (generalized) Poisson distribution. Communication in Statistics: Computations and Simulations 1991; 20(4): 1115-1137. https://doi.org/10.1080/03610919108812999

[10] Tanner JC. A derivation of Borel distribution. Biometrika 1961; 40: 222-224. https://doi.org/10.1093/biomet/48.1-2.222

[11] Consul PC, Shoukri MM. Some Chance Mechanisms Related to a Generalized Poisson Probability Model. American Journal of Mathematical and Management Sciences 1988; 8. https://doi.org/10.1080/01966324.1988.10737237

[12] Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Research 2010; 38(17): e170. https://doi.org/10.1093/nar/gkq670

[13] Cox DR, Reid N. Parameter Orthogonality and Approximate Conditional Inference Journal of the Royal Statistical Society. Series B (Methodological) 1987; 49(1): 1-39. https://doi.org/10.1111/j.2517-6161.1987.tb01422.x

[14] McCullagh P, Nelder JA. Generalized Linear Models. Chapman and Hall. London 1989. https://doi.org/10.1007/978-1-4899-3242-6

[15] Consul PC, Shoukri MM. Maximum likelihood estimation of the generalized Poisson distribution. Communications in Statistics, Theory and Methods 1984; 13(2): 1533-1547. https://doi.org/10.1080/03610928408828776

[16] Cox DR, Hinkley D. Theoretical Statistics. Chapman and Hall, London, UK 1974. https://doi.org/10.1007/978-1-4899-2887-0

[17] Fieller EC. A fundamental formula in the statistics of biological assays and some applications. Quarterly Journal of Pharmacy and Pharmacology 1944; 17: 117-123.

[18] Fieller EC. Some problems in interval estimation. Journal of the Royal Statistical Society (B) 1954; 16(2): 175-185. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x

[19] Wu J, Jiang G. Small sample likelihood inference for the ratio of means. Computational Statistics & Data Analysis 2001; 38: 181-190. https://doi.org/10.1016/S0167-9473(01)00025-1

[20] Sheiner LB. Bioequivalence revisited. Statistics in Medicine 1992; 11: 1777-1788. https://doi.org/10.1002/sim.4780111311

[21] Wu J, Wong ACM, Ng KW. Likelihood-based confidence interval for the ratio of scale parameters of two independent Weibull distributions. Journal of Statistical Planning and Inference 2005; 135: 487-497. https://doi.org/10.1016/j.jspi.2004.05.012

[22] Kendall M, Ord K. The Advanced Theory of Statistics, Sixth edition, 2009; Vol. 1.

[23] Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 2009; 25: 1026-1032. https://doi.org/10.1093/bioinformatics/btp113

[24] Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010; 11: 94. https://doi.org/10.1186/1471-2105-11-94

[25] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007; 23: 2881-2887. https://doi.org/10.1093/bioinformatics/btm453

Received on 23-04-2020 Accepted on 16-05-2020 Published on 04-06-2020

https://doi.org/10.6000/1929-6029.2020.09.05