Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 285-299
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Empirical Bayesian Approach to Testing Homogeneity of Several Means of Inflated Poisson Distributions (IPD)

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

[1]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Canada
[2]Knowledge Translation Department, Saudi National Institute of Health, Riyadh, Saudi Arabia
Email: *Shoukri.mohmed@gmail.com, mmshoukr@uwo.ca, maleid@snih.gov.sa

How to cite this paper: Shoukri, M.M. and Aleid, M. (2023) Empirical Bayesian Approach to Testing Homogeneity of Several Means of Inflated Poisson Distributions (IPD). *Open Journal of Statistics*, **13**, 285-299.
https://doi.org/10.4236/ojs.2023.133015

Received: May 10, 2023
Accepted: June 13, 2023
Published: June 16, 2023

Abstract

Objectives: We introduce a special form of the Generalized Poisson Distribution. The distribution has one parameter, yet it has a variance that is larger than the mean a phenomenon known as "over dispersion". We discuss potential applications of the distribution as a model of counts, and under the assumption of independence we will perform statistical inference on the ratio of two means, with generalization to testing the homogeneity of several means. **Methods:** Bayesian methods depend on the choice of the prior distributions of the population parameters. In this paper, we describe a Bayesian approach for estimation and inference on the parameters of several independent Inflated Poisson (IPD) distributions with two possible priors, the first is the reciprocal of the square root of the Poisson parameter and the other is a conjugate Gamma prior. The parameters of Gamma distribution are estimated in the empirical Bayesian framework using the maximum likelihood (ML) solution using nonlinear mixed model (NLMIXED) in SAS. With these priors we construct the highest posterior confidence intervals on the ratio of two IPD parameters and test the homogeneity of several populations. **Results:** We encountered convergence problem in estimating the hyperparameters of the posterior distribution using the NLMIXED. However, direct maximization of the predictive density produced solutions to the maximum likelihood equations. We apply the methodologies to RNA-SEQ read count data of gene expression values.

Keywords

Distributions of Over-Dispersed Counts, Lagrange Class of Distributions, Knowledge Transfer, Gamma Prior, Posterior Inference, Wilson-Hilferty

Transformation, RNA_SEQ Read Counts Data

1. Introduction

Bayesian methods are becoming a popular technique for estimating model parameters and hypothesis testing. Under certain conditions choosing the correct prior distribution is a critical step in Bayesian modeling. The traditional Bayesian approach assumes that all prior distribution parameters are known. Knowledge of these parameters may be based on observed data in similar studies with similar objectives. If such data are unavailable, a non-informative prior distribution may be used [1]. As a result, choosing an appropriate prior distribution is important. It is known that the Gamma distribution is used for the Poisson family of distributions. This distribution is considered conjugate in the sense that the posterior probability belongs to the class of Gamma distributions. Similarly, we shall employ the Gamma distribution as an alternative prior for the IPD parameter.

The method of maximum likelihood estimation of unknown prior distribution parameters can be used and in most situations, we need numerical algorithms to estimate the parameters of the prior distribution. In this paper we used NLMIXED procedure in SAS.

In Section 2 we introduce the functional form of the IPD and discuss some of its interesting properties. In Section 3 we investigate the Bayesian inference on the ratio of IPD parameters based on samples drawn from independent populations. Both non-informative priors and conjugate gamma priors will be used to achieve the main objectives. In Section 4 we extend the methodology to test the homogeneity of the several population parameters, and in Section 5 we apply the methods to published genomics data.

2. The Inflated Poisson Distribution (IPD)

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. 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 [2]. 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 [3], and in a more recent review by Hayat and Higgins [4]. 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 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 [5] reviewed the methodologies of NB regression. Joe and Zhu [6] 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 [7] and a detailed account of its properties was given by Consul [8]. 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 [9] as a queuing model [10] [11] [12] and genomics [13]. The Bayesian statistical inference on the proposed form of the GPD, which we shall call "Inflated Poisson Distribution" (IPD) is the subject of this research paper.

We introduce three random mechanisms by which the IPD is generated and discuss some of its properties. In Section 3 we consider the Bayesian inference on the parameter of the distribution and construct the exact posterior density of the ratio of two parameters for independent populations. In Section 4 we apply the Wilson-Hilferty (WH) [14] to the posterior distribution in order to test the homogeneity of parameters of several populations. We analyze published data sets related to the read counts of RNA-SEQ.

2.1. The IPD and Its Moments

A random variable x is said to have IPD if the probability function is given by:

$$P(x=x)=\frac{(1+x)^{x-1}}{x!}\lambda^{x}\bar{e}^{\lambda(1+x)} \tag{1}$$

This distribution is a special case of the Generalized Poisson Distribution (GPD). We review the literature on the derivation of the probability function given in (1).

1) Consul and Shenton [15] showed that the Lagrange expansion of implicit Probability Generating Function. If $g(t)$ and $f(t)$ are two probability generating functions, then under the transformation:

$$t=u\cdot g(t)$$

and within the circle of convergence, $f(t)$ can be expanded in powers of u by the Lagrange expansion, then the coefficient of u^{x}, produces a probability distribution given as:

$$P(x=x)=\frac{1}{x!}\frac{\partial^{x-1}}{\partial t}\left[\left(g(t)\right)^{x}\frac{\partial f(t)}{\partial t}\right]_{t=0} \tag{2}$$

making the substitutions:

$$g(t)=e^{\lambda(t-1)} \quad \text{and} \quad f(t)=e^{\lambda(t-1)}$$

in Equation (2) we get the probability distribution (1).

2) Consul and Shoukri [15] showed that in the special case of a Borel-Tanner distribution [16] conditional on a parameter η

$$P(x|\eta)=\frac{\eta}{(x-\eta)!}x^{x-\eta-1}\lambda^{x-\eta}\bar{e}^{\lambda x},\ x=\eta,\eta+1,\cdots \tag{3}$$

If η has a Poisson distribution (4)

$$P(\eta=j)=\bar{e}^{\lambda x}\lambda^{j}/j!, \tag{4}$$

then the unconditional distribution of x is

$$P(x=x)=\sum P(x=x|\eta)P(\eta=j)=\frac{(1+x)^{x-1}}{x!}\lambda^{x}\bar{e}^{\lambda(1+x)},\ x=0,1,\cdots$$

3) The IPD arises as a limiting form of the quasi-generalized negative distribution (see: Shoukri and Aleid [17]):

$$Px=P(X=x)=\frac{\beta-1}{(\beta-1)+\beta x}\cdot\frac{\Gamma(\beta+\beta x)\theta^{x}(1-\theta)^{\beta+\beta x-x-1}}{x!\Gamma(\beta+\beta x-x)} \tag{5}$$

As $\beta\to\infty,\theta\to 0$, so that $\beta\theta=\alpha$, the limiting distribution (5) becomes (1).

2.2. Moments and Variance Stabilizing Transformation

From [6], the mean and variance of the distribution are given respectively by:

$$\mu=\lambda/1-\lambda,\ 0<\lambda<1$$

$$\sigma^{2}=\lambda/(1-\lambda)^{3}$$

In terms of the mean μ, the parameter λ is given by:

$$\lambda=\frac{\mu}{1+\mu} \tag{6}$$

Therefore the variance is given in (7)

$$\sigma^{2}(\mu)=\mu(1+\mu)^{2} \tag{7}$$

Shoukri and Mian [10] established a recurrence relation among the r-th non-central moments μ_r' so that:

$$\mu_{r+1}'=\sigma^{2}(\mu)\frac{\partial\mu_r'}{\partial\mu}+\mu\mu_r' \tag{8}$$

From (8) we get:

$$\mu_0'\equiv 1,\ \mu_1'=\mu=E(y)$$

The third and fourth central moments are given respectively by

$$\mu_3=E(y-\mu)^{3}=\mu(1+\mu)^{3}(1+3\mu) \tag{9}$$

$$\mu_4=E(y-\mu)^{4}=\mu(1+\mu)^{4}(1+13\mu+15\mu^{2}) \tag{10}$$

For discrete random variables such as binomial negative binomial, and Poisson distributions a variance stabilizing transformation can be developed if the relationship between the mean and variance is known. From the definition of skewness and on using (9) and (10) we have:

skewness (*sk*) is:

$$sk = \sqrt{\frac{(1+3\mu)^2}{\mu}} = \frac{1+3\mu}{\sqrt{\mu}}$$

Or

$$sk = (1+2\lambda)\sqrt{\frac{1-\lambda}{\lambda}} \tag{11}$$

Clearly for $0<\lambda<1$ the distribution is positively skewed. From (11) clearly the degree of skewness is inversely related to the values of λ.

The distribution has an interesting property in that, within the class of discrete distributions defined on $R=(0,1,2,\cdots,\infty)$, the IPD has a single parameter and yet it possess the property of overdispersion.

Akin to the Poisson distribution we may develop a variance stabilizing transformation. We would like to find a variance stabilizing transformation $Z=\Psi(x)$ such that $\text{var}(z)=c^2$, which does not depend on the population parameter. We employ the Taylor series expansion so that:

$$\Psi(x)=\Psi(\mu)+\frac{\partial\Psi}{\partial\mu}(x-\mu)+0(x-\mu) \tag{12}$$

Solving (12) then

$$\frac{\partial\Psi}{\partial\mu} \propto c\frac{1}{(\text{var}(x))^{1/2}} \tag{13}$$

Solution of the differential Equation (13) is: $z=\tan^{-1}(\sqrt{y})$ which has variance = 1/4, similar to the variance stabilizing transformation $z_p=\sqrt{y}$ of the Poisson distribution.

This variance stabilizing transformation may be used to derive an approximate expression for the sample size needed to test the equality of two IPD means.

To test the null hypothesis $H_0:\mu_1=\mu_2$ versus $H_1:\mu_1\neq\mu_2$ at type I error rate 5%, and power 80%, the approximate sample size drawn from each of two inflated Poisson populations is:

$$N = 2.48^2/4\Delta^2 \tag{14}$$

where $\Delta=\tan^{-1}\sqrt{\mu_1}-\tan^{-1}\sqrt{\mu_2}$.

In Table 1 values of N in equation (14) for selected values of the population parameters.

3. Bayesian Analysis

Bayesian methods are becoming a popular technique for estimating model parameters and hypothesis testing. Under certain conditions choosing the correct

Table 1. Selected values of the sample sizes.

μ_1	μ_2	Δ	N
1	2	−0.17	53
1	3	−0.26	23
1	5	−0.36	12
2	7	−0.25	24
2	10	−0.31	16
2	15	−0.36	12
5	10	−0.11	118
5	15	−0.17	55
5	20	−0.20	38

prior distribution is a critical step in Bayesian modeling. The traditional Bayesian approach assumes that all prior distribution parameters are known. Knowledge of these parameters may be based on observed data in similar studies with similar objectives. If such data are unavailable, a non-informative prior distribution may be used [1]. As a result, choosing an appropriate prior distribution is important. It is known that the Gamma distribution is used for the Poisson family of distributions. This distribution is considered conjugate in the sense that the posterior probability belongs to the class of Gamma distributions. Similarly, we shall use the Gamma distribution as an alternative prior for the IPD parameter.

The method of maximum likelihood estimation of unknown prior distribution parameters can be used and in most situation we need numerical algorithms to estimate the parameters of the prior distribution. In this paper we used NLMIXED in SAS.

Let $\left(x_{i1}, x_{i2}, \cdots, x_{n_i}\right)$ denotes k random samples of size n_i from the j^{th} IPD. The likelihood is given in (15):

$$L_i = \prod_{j=1}^{n_i}\left(1+x_{ij}\right)^{x_{ij}-1}\frac{\lambda_i^{x_{ij}}}{x_{ij}!}e^{-\lambda_i\left(x_{ij}+1\right)} \tag{15}$$

Clearly

$$y_i = \sum_{j=1}^{n_i} x_{ij}$$

is the sufficient statistic for λ_i. The probability distribution of y_i is given in (16) as:

$$P\left(Y_i = y_i \mid \lambda_i\right) = \frac{n_i\lambda_i^{y_i}\left(n_i+y_i\right)^{y_i-1}}{y_i!}\exp\left[-\lambda_i\left(n_i+y_i\right)\right],\ i=1,2,\cdots,k \tag{16}$$

To achieve the main objective of this paper, we shall consider a prior distribution with a conjugacy-like distribution. Here we consider two prior specifications: the first being the vague prior and the second is the conjugate gamma prior distribution for the parameter λ_i.

Hierarchical Bayes and Empirical Bayes are related by their goals, but quite different by the methods of how these goals are achieved. The attribute hierar-

chical refers mostly to the modeling strategy, while empirical is referring to the methodology. Both methods are concerned in specifying the distribution at prior level, hierarchical via Bayes inference involving additional degrees of hierarchy (hyperpriors and hyperparameters), while empirical Bayes is using data more directly.

3.1. Vague Prior

We start by using the vague prior specification for λ_i given in (17)

$$\pi(\lambda_i) \propto 1/\sqrt{\lambda_i} \tag{17}$$

The posterior density of λ_i is therefore given by (18):

$$\pi(\lambda_i \mid y_i) = \frac{(y_i + n_i)^{y_i+1/2}}{\Gamma(y_i + 1/2)} \lambda_i^{y_i - 1/2} \exp\left[-\lambda_i (y_i + n_i)\right] \tag{18}$$

This means that the posterior density of λ_i is such that $2\lambda_i(y_i + n_i)$ has a Chi-square distribution with $v_i = 2(y_i + 1/2)$ degrees of freedom. Therefore, the posterior mean and variance of λ_i are given respectively as $(y_i + 1/2)/(y_i + n_i)$ and $(y_i + 1/2)/(y_i + n_i)^2$.

$$M(y_i) = \int_0^\infty P(y_i \mid \lambda_i)\pi(\lambda_i \mid y_i)\mathrm{d}\lambda_i \tag{19}$$

It can be easily shown that (19) has the closed form given in (20):

$$M(y_i) = \frac{n_i}{n_i + y_i} \frac{\Gamma(2y_i + 1/2)}{\Gamma(y_i + 1/2)\Gamma(y_i + 1)} \tag{20}$$

We can construct an HPD confidence interval on the ratio $R = \frac{\lambda_1}{\lambda_2}$. Since,

$$\frac{\lambda_1}{\lambda_2} = \frac{(2y_1 + 1)(y_2 + n_2)}{(2y_2 + 1)(y_1 + n_1)} \cdot Fv_1, v_2$$

Therefore, the exact posterior distribution of the ratio is that of a weighted F-distribution. We can directly construct HPD limits on the ratio using the F-distribution tables after substituting the estimated values of the hyperparameters as shown below.

An $(1-\alpha)100\%$ confidence interval on $R = \frac{\lambda_1}{\lambda_2}$ is such that:

$$\mathrm{UL} = \frac{(2y_1 + 1)(y_2 + n_2)}{(2y_2 + 1)(y_1 + n_1)} F_{1-\alpha/2}, v_1, v_2,$$

$$\mathrm{LL} = \frac{(2y_1 + 1)(y_2 + n_2)}{(2y_2 + 1)(y_1 + n_1)} F_{\alpha/2}, v_1, v_2$$

$F_{1-\alpha/2}, v_1, v_2$ $F_{\alpha/2}, v_1, v_2$ are respectively the upper and lower quantiles of the F-distribution with (v_1 and v_2) degrees of freedom.

3.2. Gamma Prior

We consider the two parameters gamma distribution as a prior for the IPD pa-

rameter. The suggested prior density is given by:

$$\pi\left(\lambda_i \mid a_i, b_i\right)=\frac{b_i^{a_i}}{\Gamma\left(a_i\right)} \lambda_i^{a_i-1} \overline{e}^{b_i \lambda_i} \tag{21}$$

where $0<\lambda_i<\infty$ and $a_i, b_i>0$. The prior mean and variance are easily obtained from (21) and are given respectively as a_i / b_i, and a_i / b_i^2. The posterior distribution of λ_i is therefore given by:

$$\pi\left(\lambda_i \mid y_i, a_i, b_i\right) \propto \lambda_i^{y_i+a_i-1} \exp\left[-\lambda_i\left(y_i+n_i+b_i\right)\right] \tag{22}$$

The exact posterior density of λ_i is thus given by (23):

$$\pi\left(\lambda_i \mid y_i, a_i, b_i\right)=\frac{\left(y_i+n_i+b_i\right)^{y_i+a_i}}{\Gamma\left(y_i+a_i\right)} \lambda_i^{y_i+a_i-1} \exp\left[-\lambda_i\left(y_i+n_i+b_i\right)\right] \tag{23}$$

This means that the posterior density of λ_i is such that, $X=2\lambda_i\left(y_i+n_i+b_i\right)$ has a Chi-square distribution with $d_i=2\left(y_i+a_i\right)$ degrees of freedom. Hence, the posterior distribution of $\lambda_i=\dfrac{X}{2\left(y_i+n_i+b_i\right)}$ is that of a weighted Chi-square variable with $2\left(y_i+a_i\right)$ degrees of freedom.

Therefore, the posterior mean and variance of λ_i are given respectively as

$\left(y_i+a_i\right) / \left(y_i+n_i+b_i\right)$ and $\left(y_i+a_i\right) / \left(y_i+n_i+b_i\right)^2$.

We conclude that the ratio $R=\dfrac{\lambda_1}{\lambda_2}$ has a weighted F distribution with v_1 and v_2 degrees of freedom, or

$$R=\frac{\lambda_1}{\lambda_2}=\frac{\left(y_1+b_1\right)\left(y_2+n_2+b_2\right)}{\left(y_2+b_2\right)\left(y_1+n_1+b_1\right)} \cdot Fd_1, d_2$$

Therefore on $(1-\alpha)100\%$ posterior confidence (24) interval on R is such that:

$$1-\alpha=P_r\left[c_1<R<c_2\right] \tag{24}$$

where

$$c_1=\frac{\left(y_1+a_1\right)\left(y_2+n_2+b_2\right)}{\left(y_2+a_2\right)\left(y_1+n_1+b_1\right)} \cdot F_{1-\alpha_{1/2}}, d_1, d_2$$

$$c_2=\frac{\left(y_1+a_1\right)\left(y_2+n_2+b_2\right)}{\left(y_2+a_2\right)\left(y_1+n_1+b_1\right)} \cdot F_{\alpha_{1/2}}, d_1, d_2$$

The confidence limits in (24) depend on $\left(a_i, b_i\right)$ and their estimates are obtained by maximizing the marginal predictive density given in (25) with respect to the target parameters. The marginal predictive density is given by:

$$\begin{aligned} M\left(y_i \mid a_i, b_i\right) &= \int_0^{\infty} P\left(y_i \mid \lambda_i\right) \pi\left(\lambda_i \mid a_i, b_i\right) \mathrm{d}\lambda_i \\ &= \frac{n_i \Gamma\left(y_i+a_i\right)}{\Gamma\left(y_i+1\right) \Gamma\left(a_i\right)}\left[\frac{b_i}{n_i+y_i+b_i}\right]^{a_i} \frac{\left(n_i+y_i\right)^{y_i-1}}{\left(n_i+y_i+b_i\right)^{y_i}} \\ &= \frac{n_i}{\left(n_i+y_i\right)} \frac{\Gamma\left(y_i+a_i\right)}{\Gamma\left(y_i+1\right) \Gamma\left(a_i\right)}\left[\frac{n_i+y_i}{n_i+y_i+b_i}\right]^{y_i}\left[\frac{b_i}{n_i+y_i+b_i}\right]^{a_i} \end{aligned} \tag{25}$$

The maximum likelihood estimates of the hyperparameters (a_i, b_i) are obtained by maximizing the log-likelihood $l(a_i, b_i)$ given by:

$$l(a_i, b_i) = \sum_{i=1}^{k} \log\left[M(y_i \mid a_i, b_i)\right] \tag{26}$$

Therefore, the maximum likelihood estimators are $(\hat{a}_i, \hat{b}_i) = \arg\max\{l(a_i, b_i)\}$.

Thus, the empirical Bayes method may be interpreted as a mixed model, and mixed model software may be used (proc NLMIXED) In the marginal likelihood function, the unknown IPD parameter is integrated out. The resulting marginal distribution is not a negative binomial (NB) distribution. The marginal predictive density is, however, a difficult vehicle for parameter estimation. As will be shown below, the convergence of the SAS optimization procedures is not always guarantied, and the solution is not necessarily a global maximum of the likelihood function even when we have large samples.

4. Application: 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 [20] [21] [22]). Before these tools can be applied to RNA-Seq data, it is worth noting that microarray data and RNA-Seq data are inherently different [20]. 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 [19]. Popular models to safeguard against over-dispersion include the negative binomial distribution, or two-stage Poisson distribution [20], as discussed below.

When over-dispersion is observed across the samples, the gene counts cannot be estimated accurately by a simple Poisson model [21]. 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 [22].

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 IPD 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.

Data Analysis Results

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:

Chrom_chr11, Chrom chr9_ra, and Chrom chrUn_ra.

For the two samples case we analyzed the read count of the Mice-Brain tissue data first using the two experimental conditions: Chrom_ chr11, Chrom chr9_ra. The data analyses were done in three steps. In the first step we apply the Chi-square test if goodness of fit where the null hypothesis is that the data is drawn from and IDP. In the second step explain the approach to estimating the problem of estimating the hyperparameters (a_i, b_i). The results based on the vague prior are not presented as they are similar to the more general gamma prior specifications.

The histogram of the read counts data show severe skewness to the right as shown in **Figure 1** & **Figure 2**.

The Chi-square goodness of fit on the hypothesis that the random mechanism generating the data is the IPD with moment estimator of $\lambda = 0.886$, has a p-value = 0.245.

The Chi-square goodness of fit on the hypothesis that the random mechanism generating the data is the IPD with moment estimator of $\lambda = 0.670$, has a p-value = 0.242.

Therefore, we conclude that the two data sets support the IPD hypothesis against a general alternative. We present the summary statistics in **Table 2**.

The problem of estimating the hyperparameters

Table 2. Summary statistics of the first two read counts experiments.

Chrom_chr11	Chrom chr9_ra
sample size = 37623	Sample size = 698
Mean = 7.84	Mean = 3.03
SD = 8.85	SD = 2.37
$a_1 = 1.00003$	$a_2 = 1.00004$
$b_1 = 1.127$	$b_2 = 1.33$

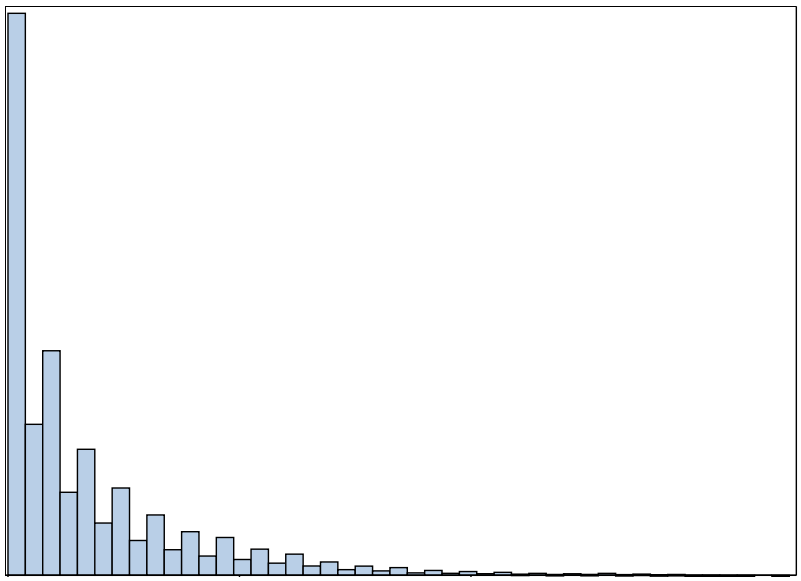

Figure 1. Histogram of the Chrom_ chr11 read data.

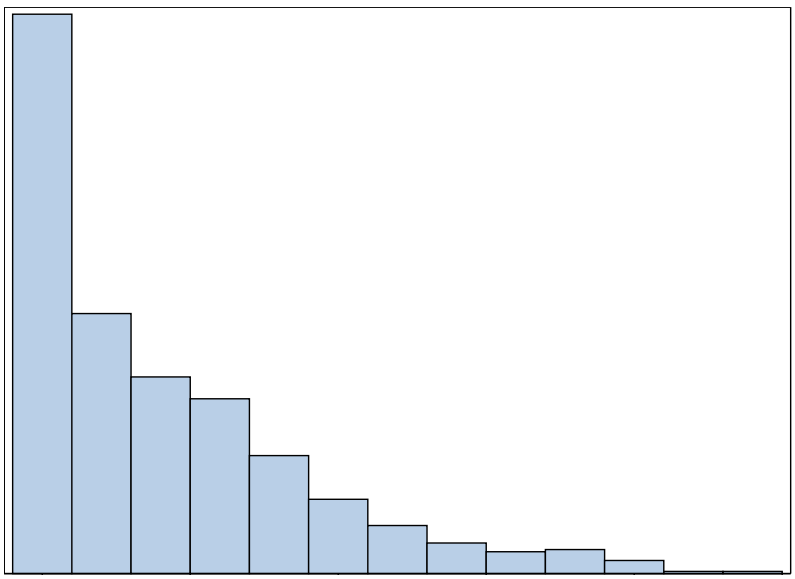

Figure 2. Histogram of the Chrom chr9_ra.

The 95% highest posterior Bayesian interval on the ratio $\frac{\lambda_1}{\lambda_2}$ is given by:

$$0.95 = \text{Probability}\left(0.391 < \frac{\lambda_1}{\lambda_2} < 0.473\right)$$

We conclude that there is no significant difference between the read counts means in the two sample.

5. Approximate Bayesian Test of Homogeneity of IPD Parameters

We shall apply the Wilson-Hilferty transformation of a Chi-square distribution with v degrees of freedom. The W-H is given by:

$$\left[\left(x^2/v\right)^{1/3} - \left(1 - \frac{2}{9v}\right)\right] \Big/ \sqrt{\frac{2}{9v}} \sim N(0,1) \tag{27}$$

Applying the transformation to the random variables:

$2\lambda_i/(y_i + n_i + b_i)$ which has a Chi-square distribution with $v_i = 2(y_i + a_i)$ degrees of freedom we have:

$$\left[\frac{2\lambda_i(y_i + n_i + b_i)}{2(y_i - a_i)}\right]^{1/3} \tag{28}$$

is approximately distributed as:

$$\sim N\left(1 - \frac{1}{9v_i}, \frac{1}{9v_i}\right)$$

Accordingly, the random variables $\xi_i = \lambda_i^{1/3}$, are approximately distribution as normal with mean m_i, and variance w_i^{-1}, where

$$m_i = \left[\frac{y_i + a_i}{y_i + n_i + b_i}\right]^{1/3}\left(1 - \frac{1}{9v_i}\right)$$

$$w_i^{-1} = \left[\frac{y_i + a_i}{y_i + n_i + b_i}\right]^{2/3}\left(\frac{1}{9v_i}\right)$$

Assuming that the data are available from independent random samples can construct a test statistic based on the Chi-square test of homogeneity of several parameters.

Thus

$$Q = \sum_{i=1}^{k} w_i\left[(\xi_i - m_i) - \sum_{i=1}^{k} w_i(\xi_i - m_i)/w\right]^2 \tag{29}$$

where

$$w = \sum_{i=1}^{k} w_i$$

Has approximately a Chi-square distribution with $k-1$ degrees of freedom.

Under the null hypothesis $H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_k$.

Q in (29) reduces to:

$$Q_0 = \sum_{i=1}^{k} w_i \left[m_i - m \right]^2 \tag{30}$$

where

$$m = \sum_{i=1}^{k} w_i m_i / w$$

Therefore, a Bayes test of equality of quasi-generalized Poisson distribution means is thus provided by treating Q_0 (30) as Chi-square with $k-1$ degrees of freedom.

6. Application

We added the data from the third experiment of 98 read counts data points from the same data source. The added read count data together with Chrom_chr11, Chrom chr9_ra will give three groups of read counts. The estimated hyperparameters are $a_3 = 1.006$, and $b_3 = 1.648$. The objective here is to test the hypothesis:

$H_0 : \lambda_1 = \lambda_2 = \lambda_3$. Direct computations give $Q_0 = 0.0009$. From the tables of Chi-square with 2 degrees of freedom we have p-value = 0.9999. We therefore conclude that there is no enough evidence in the data to reject the homogeneity hypothesis.

7. Discussion

Knowledge translation is an approach to increase the use of evidence within policy and practice decision-making contexts. When information is available from previous experiments or accumulated data, transfer of Bayesian methods research into practice is a challenge; academic articles are not always the best enabler for mathematical adoption of research. Furthermore, synthesis of knowledge from multiple research studies is needed to provide evidence-based decision-support for research [23].

The estimation of the ratio of Poisson rates is a problem of interest and arises in medical investigations. The frequentist approach does not provide an exact solution either to the problem estimation or to the hypothesis testing. The Bayesian methodology provides exact solutions to both problems. When samples are available from multiple sources, we proposed an approximate solution to the problem of testing homogeneity from multiple samples. An approximation based of the precise WH of the Chi-square solution was developed to address this problem.

It is important to note that the IPD which is the subject of the current investigation is the only discrete distribution defined on the set of non-negative integers with one parameter and exhibits the overdispersion property. Consul and Shenton [15] defined a wide class of discrete probability distributions in terms of Lagrange's expansion. It is of interest to note that this class of distributions coincides with that of the distributions of tree sizes in the Benaim-Galton Watson process. By this means the branching or cascade process with discrete time and where the probability generating function for the number of children of each in-

dividual is the same, for every individual. Since the IPD is a member of the Lagrange family of distributions it will have direct applicability in the study of branching processes as pointed out by I.J. Good [24]. Consul and Shenton ([2]: p. 239) state that the Lagrange expansion seems to be associated with queuing processes; and such an association is in fact spelled-out by Good ([24]: p. 376) with references to some 1951 literature.

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 three data sets; it was shown to provide a better fit than several other related models.

Conflicts of Interest

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

Acknowledgements

The authors acknowledge the constructive remarks made by the reviewer that have improved on the presentation of the results.

References

[1] Carlin, B.P. and Louis, T.A. (2000) Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC Press, Boca Raton.

[2] Cox, D.R. (1983) Some Remarks on Overdispersion. *Biometrika*, **70**, 269-274. https://doi.org/10.1093/biomet/70.1.269

[3] Hinde, J. and Demetrio, C.G.B. (1998) Overdispersion: Models and Estimation. *Computational Statistics and Data Analysis*, **27**, 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3

[4] Hayat, M.J. and Higgins, M. (2014) Understanding Poisson regression. *Journal of Nursing Education*, **53**, 207-215. https://doi.org/10.3928/01484834-20140325-04

[5] Hinde, J.M. (2007) Negative Binomial Regression. Cambridge University Press, Cambridge, 2011.

[6] Joe, H. and Zhu, R. (2005) Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution. *Biometrical Journal*, **47**, 219-229. https://doi.org/10.1002/bimj.200410102

[7] Consul, P.C. and Jain, G.C. (1973) A Generalization of the Poisson Distribution. *Technometrics*, **15**, 791-799. https://doi.org/10.1080/00401706.1973.10489112

[8] Consul, P.C. (1989) Generalized Poisson Distribution. Marcel Dekker Inc., New York.

[9] Janardan, K.G. and Schaeffer, D.J. (1977) Models for the Analysis of Chromosomal Aberrations in Human Leukocytes. *Biometrical Journal*, **19**, 599-612. https://doi.org/10.1002/bimj.4710190804

[10] Shoukri, M.M. and Mian, I.U.H. (1991) Some Aspects of Statistical Inference on the Lagrange (Generalized) Poisson Distribution. *Communication in Statistics: Computa-*

tions and Simulations, **20**, 1115-1137. https://doi.org/10.1080/03610919108812999

[11] Tanner, J.C. (1961) A Derivation of Borel Distribution. *Biometrika*, **48**, 222-224. https://doi.org/10.1093/biomet/48.1-2.222

[12] Consul, P.C. and Shoukri, M.M. (1988) Some Chance Mechanisms Related to a Generalized Poisson Probability Model. *American Journal of Mathematical and Management Sciences*, **8**, 181-202. https://doi.org/10.1080/01966324.1988.10737237

[13] Srivastava, S. and Chen, L. (2010) A Two-Parameter Generalized Poisson Model to Improve the Analysis of RNA-Seq Data. *Nucleic Acids Research*, **38**, e170. https://doi.org/10.1093/nar/gkq670

[14] Wilson, E.B. and Hilferty, M.M. (1931) The Distribution of Chi-Square. *Proceedings of the National Academy of Sciences of the United States of America*, **17**, 684-688. https://doi.org/10.1073/pnas.17.12.684

[15] Consul, P.C. and Shenton, L.R. (1973). Use of Lagrange Expansion for Generating Discrete Generalized Probability Distributions. *SIAM Journal of Applied Mathematics*, **23**, 239-248. https://doi.org/10.1137/0123026

[16] Haight, F.A. and Breuer, M.A. (1960) The Borel-Tanner Distribution. *Biometrika*, **47**, 143-150. https://doi.org/10.1093/biomet/47.1-2.143

[17] Shoukri, M.M. and Aleid, 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

[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/10.1165/rcmb.2017-0430TR

[19] 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

[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**, 1-26. 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] Badampudi, D. (2018) Decision-Making Support for Choosing among Different Component Origins. Blekinge Institute of Technology, Karlskrona.

[24] Good, I.J. (1975) The Lagrange Distributions and Branching Processes. *SIAM Journal on Applied Mathematics*, **28**, 270-275. https://doi.org/10.1137/0128022