Notes on the modeling of cell count in Fields of Views using over-dispersed count models

Pierre Bost, Ruben Casanova, Uria Mor, Bernd Bodenmiller

February 16, 2023

1 Introduction to the Negative and Beta Binomial distributions

1.1 Introduction to the NB distribution

The negative binomial (NB) distribution is a discrete probability distribution often used to model the count of events of a certain nature. In what follows, we will discuss the properties of the NB distribution and its relations to other discrete probability distributions.

1.1.1 Basic properties of the NB distribution

Let X be a random variable which models the number of failures in a sequence of identically and independently distributed (i.i.d) random Bernoulli trials, before a number $r \in \mathbb{N}$ of successful trials are made, where the chance of success in each individual trial is $p \in [0, 1]$. Then X is said to follow NB distribution with parameters r and p.

$$X \sim \text{NB}(r, p) \ . \tag{1}$$

The probability mass function (PMF) associated with $X \sim \text{NB}(r, p)$ is defined as:

$$P(X = k) = \binom{k+r-1}{k} p^r (1-p)^k \tag{2}$$

The expectation of X is given by

$$\mathbb{E}(X) = \frac{rp}{1-p} \tag{3}$$

and the variance is

$$\text{Var}(X) = \frac{rp}{(1-p)^2} \tag{4}$$

While the traditional definition of the NB distribution considers an integer parameter r, it can be extended to any positive real r via the following

$$P(X = k) = \frac{\Gamma(k+r)}{k!\Gamma(r)} p^r (1-p)^k \tag{5}$$

1.1.2 Links with the Poisson distribution

Let $X \sim \text{NB}(r, p)$, define the size parameter $\theta := r$ and mean parameter $\mu := \theta p/(1-p)$[1], note that

$$\mu = \mathbb{E}(X) \tag{6}$$

Additionally, we have the variance of random variable X that follows NB distribution with mean μ and size θ is given by:

$$\text{Var}(X) = \mu + \frac{\mu^2}{\theta} \tag{7}$$

Note that according eq. (7) above, we have that the variance of an NB random variable X with size parameter θ tends to its expected value when $\theta \to \infty$. Formally stated, consider the sequence of NB random variables $\{X_i \sim \text{NB}(\mu, \theta_i)\}_{i=1}^{\infty}$ where the size parameters θ_i are such that for any positive M, there exists an index i_M for which $\theta_{i'} > M$ for all $i' > i_M$, then

$$\lim_{i \to \infty} X_i = Y$$

where the limit random variable Y follows a Poisson distribution centered at μ; $Y \sim \text{Pois}(\mu)$.

Another way of demonstrating the relationship between the NB and Poisson distributions is via the usual parameterization of the Negative Binomial.

First, let $X \sim \text{NB}(r, p)$ and set

$$\lambda = \frac{r(1-p)}{p} \tag{8}$$

Note note that we can write $p = r/(r + \lambda)$,thus

$$\begin{align} P(X = k) &= \binom{k+r-1}{k} p^r (1-p)^k \tag{9} \\ &= \frac{(k+r-1)!}{(r-1)!k!} p^r (1-p)^k \tag{10} \\ &= \frac{\Gamma(k+r)}{k!\Gamma(r)} p^r \left(\frac{\lambda}{r+\lambda}\right)^k \tag{11} \\ &= \frac{\lambda^k}{k!} \left(\frac{1}{1+\frac{\lambda}{r}}\right)^r \frac{\Gamma(k+r)}{\Gamma(r)(r+\lambda)^k} \tag{12} \end{align}$$

[1] Note that given μ and θ, it is possible to extract p via construction: $p = \mu/(\theta + \mu)$

Now, we have that

$$\lim_{r\to\infty}\frac{\Gamma(k+r)}{\Gamma(r)(r+\lambda)^k} \sim \lim_{r\to\infty}\frac{\Gamma(r)r^k}{\Gamma(r)(r+\lambda)^k} \tag{13}$$

$$= \lim_{r\to\infty}\frac{r^k}{(r+\lambda)^k} \tag{14}$$

$$= 1 \tag{15}$$

And since $\lim_{r\to\infty}\left(\frac{1}{1+\frac{\lambda}{r}}\right)^r = e^{-\lambda}$ we get that $\lim_{r\to\infty} P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$

1.1.3 Sum of i.i.d NB variables

Let $X_i \sim \mathrm{NB}(r_i, p)$ for $i = 1, 2, \ldots, N$ be independent NB random variables and let $Z := \sum_{i=1}^N X_i$ then $Z \sim \mathrm{NB}(\sum_{i=1}^N r_i, p)$.

1.2 Introduction to the Beta-Binomial distribution

The Beta-Binomial (BB) probability distribution models the number of successes in a sequence of n i.i.d Bernoulli trials, where the probability of success in each trial is a random variable that follows a Beta distribution.

1.2.1 Basic properties of the BB distribution

The BB distribution is usually parameterized using three parameters: $n \in \mathbb{N}$ corresponding to the number of trials, and $\alpha, \beta \in \mathbb{R}_+$ that are the parameters for the Beta distribution underlying the success rate of each trial p. Let $X \sim \mathrm{BB}(n, \alpha, \beta)$, then it has the following PMF :

$$P(X=k) = \binom{n}{k}\frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha,\beta)} \tag{16}$$

where

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt \tag{17}$$

is the beta function. Additionally, we have

$$\mathbb{E}(X) = \frac{n\alpha}{\alpha+\beta} \tag{18}$$

$$\mathrm{Var}(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta+1)(\alpha+\beta)^2} \tag{19}$$

An alternative parametrization of the Beta probability distribution, using a mean parameter $p = \alpha/(\alpha+\beta) \in (0,1)$ and sample size parameter $\phi = \alpha+\beta > 0$ enables another representation of a BB distributed random variable $X \sim \mathrm{BB}(n, p, \phi)$, which simplifies the expressions for mean and variance:

$$\mathbb{E}(X) = np \tag{20}$$

$$\mathrm{Var}(X) = n\frac{(\phi+n)}{(\phi+1)}p(1-p) \tag{21}$$

2 Formal Model Construction for Multiplexed Imaging Count Data

The scope of our sample is an 'image' of a tissue, that is indexed by spatial coordinate grid $(i, j) \in \mathbb{N}^{I \times J}$. Within this context, we are interested in much smaller entities that are the tissue's cells and their spatial interaction, i.e., the relationships between cell of different types and how these relationships vary depending on the location of the cells within the tissue.

To describe spatial relationships in a quantitative manner, we first have to establish a model for the counts of a given cell. Global models for cell count of a given type - that describe the same generative process for the entire tissue - are inadequate for the study of space-dependent interactions. Instead, we wish to be able to have models that accurately describe cell counts in a small spatial subset of the tissue. We term such small, local subsets *Regions of Interest* (ROI), and dedicate the following sections to describe and explain our model for *quadrat count* - cell count of a specific cell-type in a given ROI. The first step in construction of our models, is stating some key statistical traits of the data:

Systematic over-dispersion Quadrat counts of aggregated patterns exhibit large cell count variation between nearby location, hence, they display a strong over-dispersion, that is to say a variance bigger than the mean, a property that is compatible with NB distribution.

Finite total amount Since the number of cells in a tissue is bounded, so is the quadrat count. To recapitulate this property of the data, we utilize a count model based in the BB distribution, which is finitely supported. It is worth noting that the BB distribution can also account for over-dispersion.

2.1 The NB Model

Given a squared field of view (FoV) with sides of size w, we let X_w denote the number of cells of certain phenotype. Consider the following model

$$X_w \sim \mathrm{NB}(\mu(w), \theta(w)) \tag{22}$$

with mean and size parameters that are in linear relationship with the area and side size of the FoV (respectively). Thus, we re-write eq. (22) as

$$X_w \sim \mathrm{NB}(\lambda w^2, \theta_0 + \gamma w) \tag{23}$$

For parameters $\lambda > 0$ and θ_0, γ for which $\theta(w) > 0$, that we estimate based on our sample data.

Estimation of λ: Let $\{X_w^{(i)}\}_{i=1}^n$ be a set of n i.i.d random variables, each of which is distributed according to $X_w^{(i)} \sim \mathrm{NB}(\lambda w^2, \theta_0 + \gamma w)$. Consider the estimator for λ defined as

$$\bar{\lambda}_n = \frac{1}{nw^2} \sum_{i=1}^{n} X_w^{(i)} \tag{24}$$

Then

$$\begin{aligned} \mathbb{E}(\bar{\lambda}_n) &= \frac{1}{nw^2} \sum_{i=1}^{n} \mathbb{E}\left(X_w^{(i)}\right) && (25) \\ &= \frac{1}{nw^2} \sum_{i=1}^{n} \lambda w^2 && (26) \end{aligned}$$

where the last transition is due to eq. (6), and we have that

$$\mathbb{E}(\bar{\lambda}_n) = \lambda \tag{27}$$

thus $\bar{\lambda}_n$ is an unbiased estimator of λ. As for $\bar{\lambda}_n$'s variance, we have

$$\begin{aligned} \mathrm{Var}(\bar{\lambda}_n) &= \mathrm{Var}\left(\frac{1}{nw^2} \sum_{i=1}^{n} X_w^{(i)}\right) && (28) \\ &= \frac{1}{n^2 w^4} \sum_{i=1}^{n} \mathrm{Var}\left(X_w^{(i)}\right) && (29) \end{aligned}$$

where the last transition follows the fact that $X_w^{(i)}$ are i.i.d. Using eq. (7), we obtain

$$\begin{aligned} \mathrm{Var}(\bar{\lambda}_n) &= \frac{1}{n^2 w^4} n \left(\lambda w^2 + \frac{\lambda^2 w^4}{\theta_0 + \gamma w}\right) && (30) \\ &= \frac{1}{n}\left(\frac{\lambda}{w^2} + \frac{\lambda^2}{\theta_0 + \gamma w}\right) && (31) \end{aligned}$$

making it clear that the estimator's variance vanishes as n grows. Moreover, according to CLT (note that $X_w^{(i)}$ are i.i.d of finite second moment), we have that the limit $\bar{\lambda} - \lambda = \lim_{n\to\infty} \sqrt{n}(\bar{\lambda}_n - \lambda)$ exists and $\bar{\lambda} - \lambda \sim \mathrm{Norm}\left(0, \sqrt{\frac{\lambda}{w^2} + \frac{\lambda^2}{\theta_0 + \gamma w}}\right)$

3 Description of the BB model for MI count data

3.1 Impact of the Field of View (FoV) size on BB parameters

Let X_w be a discrete random variable corresponding to the number of cells of a given phenotype found in a single square Field of View (FoV) of width $\mathbf{w}$ containing in total n cells. We consider that X_w follows a BB distribution such that :

$$X_w \sim \mathcal{BB}(n, p, \phi_w) \tag{32}$$

Indeed, as here the μ parameter is already 'normalized' by the total number of cells within the FoV, it is independent from w. In addition the parameter n is usually known Using our empirical observations, we assume that :

$$\phi_w = \phi_0 + \delta w \tag{33}$$

Therefore :

$$X_w \sim \mathcal{BB}(n, p, \phi_0 + \delta w) \tag{34}$$

3.2 Statistical estimation of cell proportion (p parameter)

Let X_w^1, X_w^2,..., X_w^i,.., X_w^m be m i.i.d random variables following such that $\forall i \in [[1, n]]$:

$$X_w^i \sim \mathcal{BB}(n, p, \phi_0 + \delta w) \tag{35}$$

We consider $\bar{p}$, an estimator of p defined by :

$$\bar{p} = \frac{1}{m} \sum_{i=1}^{m} \frac{X_w^i}{n} \tag{36}$$

Similarly to $\bar{\lambda}$, $\bar{p}$ is unbiased :

$$\mathbb{E}(\bar{p}) = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathbb{E}(X_w^i)}{n} = \frac{1}{m} \sum_{i=1}^{m} p = p \tag{37}$$

We can also compute its variance :

$$\text{Var}(\bar{p}) = \frac{1}{m^2 n^2} \sum_{i=1}^{m} \text{Var}(X_w^i) = \frac{1}{mn} (\frac{\phi + n}{\phi + 1}) p(1-p) \tag{38}$$

4 Differential abundance testing using NB and BB models

A common goal of the analysis of MI data is the comparison of cell abundance between two samples or two group of samples. This is usually done by using non parametric rank-based methods, such as the Wilcoxon or Kruskall-Wallis test. This approach is relatively robust but lacks the sensitivity and flexibility of parametric methods. Here we will introduce statistical tests that are based on the NB and BB models.

4.1 NB based model

A fully detailed description of the NB based model used here is provided in the excellent paper [1] which we strongly encourage motivated readers to look at.

Let X^1, X^2,..., X^i,.., X^n be n i.i.d random variables such that $\forall i \in [[1, n]]$:

$$X^i \sim \mathcal{NB}(\mu, \theta) \tag{39}$$

Similarly, let Y^1, Y^2,..., Y^i,.., Y^m be m i.i.d random variables such that $\forall i \in [[1, m]]$:

$$Y^i \sim \mathcal{NB}(\gamma\mu, \theta) \tag{40}$$

We wish to know whether γ is significantly different from 1, namely the two group of samples have different means. To do so, a generalized linear model is used with a logarithmic link function to estimate the mean parameter of each sample. It is worth noting that we assume the θ parameter to be constant across samples, an important limitation of the NB model. More formally let X^1, X^2,..., X^i,.., X^n be NB distributed variables with variable mean parameters μ_i parameters and associated with an explanatory binary variable z_i describing to which experimental group the sample belongs. Then $\forall i \in [[1, n]]$:

$$\log(\mu_i) = \mu_0 + \gamma z_i \tag{41}$$

The μ_0 and γ parameters are then estimated through a classical Maximum Likelihood Estimation (MLE) iterative process. In our paper, this model is implemented using the glm.nb() function from the **MASS** R package.

4.2 BB based model

Let X^1, X^2,..., X^i,.., X^n be n i.i.d random variables such that $\forall i \in [[1, n]]$:

$$X^i \sim \mathcal{BB}(n_i, p_i, \phi_i) \tag{42}$$

An explanatory binary variable z_i is associated to each X^i and corresponds to the experimantal group to which the sample belongs to. We used a logistic link function to predict the individual p_i :

$$p_i = \frac{exp(p_0 + z_i\gamma)}{1 + exp(p_0 + z_i\gamma)} \tag{43}$$

where γ corresponds to the effect of the explanatory variable on the p parameter. Unlike the NB model, the BB regression model allows a variable over-dispersion parameter across samples. Here the ϕ_i are modeled as a function of z_i variables :

$$\phi_i = \phi_0 + \beta z_i \tag{44}$$

A more extensive description of the properties of BB regressions is available in [2]. In our paper the BB based model was implemented using the betabin() function from the **aod** (analysis of overdispersed data) R package.

4.3 Likelihood-based statistical tests

Two statistical test can be used study the significance of the γ and β parameters once the NB or BB models have been fitted :

1. Likelihood Ratio Test

2. Wald's Test

While a third test also exist (score or Rao's test), we did not used it due it lower statistical power for NB regression [1].

4.3.1 Likelihood Ratio Test (LRT)

Let Ψ be the set of possible parameter values, and Ψ_0 the restricted set of parameter values where $\gamma = 0$ and ψ a vector of parameters. In addition, $L(x, z, \psi)$ corresponds to the likelihood function of the regression model with input data vector x, explanatory data vector z and Ψ parameter vector . We define R the likelihood ratio :

$$R = \frac{\sup\{L(x, z, \psi), \psi \in \Psi_0\}}{\sup\{L(x, z, \psi), \psi \in \Psi\}} \tag{45}$$

Where the sup corresponding to maximum likelihood estimation. One can show that χ_L defined as :

$$\chi_L = -2 \log(R) \tag{46}$$

follows a χ^2 distribution with a degree of freedom set to 1 (Wilk's theorem).

4.3.2 Wald's test

Wald's test is another likelihood-based test commonly used for hypothesis testing. If one is interested to test if a parameter δ is significantly different from a hypothesized value δ_0 after estimating it values $\bar{\delta}$, he can compute the following statistic :

$$W = \frac{(\bar{\delta} - \delta_0)^2}{\text{Var}(\bar{\delta})} \tag{47}$$

Under the null hypothesis $(\bar{\delta} = \delta_0))$, W asymptotically follows a χ^2 distribution with one degree of freedom.

References

[1] Inmaculada B. Aban, Gary R. Cutter, and Nsoki Mavinga. Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data. *Computational statistics & data analysis*, 53(3):820–833, January 2008.

[2] R. L. Prentice. Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate Measurement Errors. *Journal of the American Statistical Association*, 81(394):321–327, 1986. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].