A Full Likelihood Procedure of Exchangeable Negative Binomials for Modelling Correlated and Overdispersed Count Data

Running title: Exchangeable Negative Binomials

Fei Tan [a], Gibson Johnston Rayner [b], Xiaodong Wang [c], and Hanxiang Peng [d,*,1]

[a] *Institute of Public Health, College of Pharmacy and Pharmaceutical Sciences, Florida A&M University, Tallahassee, FL 32307, USA*

[b] *Department of Mathematics, University of Mississippi, University, MS 38677-1848, USA*

[c] *Sanofi-Aventis, Bridgewater, NJ 08807, USA*

[d] *Department of Mathematical Sciences, Indiana University – Purdue University At Indianapolis, Indianapolis, IN 46202-3216, USA*

SUMMARY. This paper introduces an exchangeable negative binomial distribution resulting from relaxing the independence of the Bernoulli sequence associated with a negative binomial distribution to exchangeability. It is demonstrated that the introduced distribution is a mixture of negative binomial distributions and can be characterized by infinitely many parameters that form a completely monotone sequence. The moments of the distribution are

* Corresponding author: hpeng@math.iupui.edu (Hanxiang Peng)

[1] This research is supported by the US National Science Foundation under Grant No. DMS-0940365

derived and a small simulation is conducted to illustrate the distribution. For data analytic purposes, two methods, truncation and completely-monotone links, are given for converting the saturated distribution of infinitely many parameters to parsimonious distributions of finitely many parameters. A full likelihood procedure is described which can be used to investigate correlated and overdispersed count data common in biomedical sciences and teratology. In the end, the introduced distribution is applied to analyze a real clinical data of burn wounds on patients.

KEY WORDS: Beta-binomial; complete monotonicity; exchangeability; negative binomial; overdisperson.
AMS 2000 subject classification: Primary: 62M09, 62M10; Secondary: 62G05, 62G20.

1 Modeling Correlated and Overdispersed Count Data

Poisson regression is the standard method used to model count data. However, the Poisson distribution requires the equality of its mean and variance, an assumption which is rarely met in real data. What often happens is that the variance of data is bigger than the mean, which is termed as *Poisson overdisperson* in the literature. The standard parametric model to account for Poisson overdisperson is the negative binomial distribution. Since most real count data situations appear to involve overdispersion, the negative binomial regression has now been finding increased use in many fields of science including biomedical sciences, econometrics and teratology. For a systematic discussion, see the first book on the negative binomial regression by Hilbe (2007).

A negative binomial random variable can be viewed as the count to get the desired number of successes in a series of independent and identically distributed Bernoulli trials. The

independence, nevertheless, cannot be assumed in many real situations. The perception of exchangeability, intensely studied over the past century, is meant to capture the notion of symmetry in a collection of random variables and is often used as an alternative to independence. In this article, by relaxing independence to exchangeability, we introduce the exchangeable negative binomial distribution, based upon which we propose a full likelihood procedure for investigating correlated and overdispersed binary response data.

The introduced exchangeable negative binomial distribution has many advantages over some existing models. In models like those various marginal models and generalized estimating equation (GEE) procedures, only the marginal means and variances or second order correlations are modeled, while higher order correlations are usually so intractable that they are replaced by "Gaussian working matrices" and are estimated by the method of moments. In an exchangeable model, the joint distribution is expressed in terms of the marginal probabilities. Hence, correlations of all orders are given by these probabilities, so that an exchangeable model incorporates higher order moments and makes the full use of information in them. Unlike a mixture model in which an arbitrary mixing distribution is chosen, an exchangeable model does not utilize an arbitrary mixing distribution but makes full use of the structure that characterizes the associated exchangeable binary sequence based on the elegant de Finetti theorem. It should be noted that the exchangeable negative binomial generalizes the negative binomial. Since the negative binomial model is very useful both in theory and application when the associated Bernoulli sequence is independent, so would be the exchangeable negative binomial model when it is more appropriate or safer to assume exchangeability for the associated Bernoulli sequence.

Statistical analysis of correlated data has been increasingly important because correlation is often evident in many fields including biomedical sciences and teratology. Take teratology for example, which is the study of the estimation of risk in developmental toxicity. The

endpoints are malformations, intrauterine deaths, resorptions, growth retardation, etc. The recorded responses in many teratogenic studies are binary, denoting, for example, death or survival of a fetus, or the presence or absence of some fetal abnormality due to the toxic substance. The offspring in a litter are *not independent but correlated*. Indeed, offspring from the same litter may respond more similarly (call this *similarity response*) to a stimulus than fetuses from different litters. This within-litter correlation also causes overdispersion (extra-binomial variation). That is, the variance of the responses exceeds the nominal variance when the binomial model is used.

Similarity responses also often occur in other situations. Consider burn wounds on different body locations of a patient. The responses to a treatment of burn wounds on different body locations of the same patient are certainly correlated. Further, burn wounds from the same patient would, seemingly, respond more similarly to a treatment than burn wounds from different patients. Another example in medical science is arthritis on different body locations of a patient. In modeling these similarity responses it seems more realistic to assume the less restrictive assumption of exchangeability rather than independence. In fact, similarity responses are positively correlated, which is carried on by exchangeability because the latter of course implies nonnegative correlation.

In the past several decades, many statistical procedures have developed to analyze teratogenic data and account for the litter effect. These procedures include the beta-binomial model by Williams (1975), the correlated-binomial model by Kupper and Haseman (1978), the non-parametric jackknife procedure by Gladen (1979), and the adjusted chi-square tests by Donner and Donald (1988). Using the relationship between exchangeability and complete monotonicity, George and Bowman (1995) initiated an investigation on correlated binary data. George and Kodell (1996) gave the tests of independence, treatment heterogeneity, and dose-related trend with exchangeable binary data. Kuk (2004) introduced a litter-based ap-

proach to risk assessment in developmental toxicity studies via a power family of completely monotone functions. Both George and Bowman and Kuk applied their approaches in clinical and developmental toxicity studies and compared their results with existing models. Xu and Prorok (2003) investigated modelling and analyzing exchangeable binary data with random cluster sizes. Stefanescu and Turnbull (2003) used the EM algorithm to model exchangeable binary data with varying cluster sizes. Recently, Yu and Zelterman (2007) investigated the sums of exchangeable Bernoulli random variables for family and litter frequency data. Dang, Keeton and Peng (2009) proposed a unified approach for analyzing exchangeable binary data and applied the approach to developmental toxicity studies. Their approach generalizes the methodology in generalized linear models. In particular, the generalization of *one* link function in GLM's to *a sequence* of link functions. Specifically, *one* link specifies one model in GLM's, but in their models *many* links are employed to specify one model. Resulting from completely monotonic functions, they introduced a rich family of parametric parsimonious binomial mixtures. With such a rich family, one can perform statistical inference on correlated binary data and, in particular, overdispersed data.

In this article, we apply the proposed procedure to fit a real clinical burn wounds dataset. Our results indicate that the proposed exchangeable negative binomial (ENB) fitting improves the negative binomial (NB) fitting. The ENB fitting yields the estimate 52.3% (33.7%) of the rate of one wound location (two wound locations at the same time respectively) of no need of surgical treatment after using the debriding agent. We have not yet found in the literature that the number of burn wounds on a patient as a response is fitted this way.

We organize our article as follows. In section 2, we introduce the ENB and give the moments, simulations and graphs. Completely monotone links are discussed in section 3. The truncated ENB is given in section 4. Section 5 is devoted to maximum likelihood and empirical estimation. In section 6, we apply the proposed procedure to fit a real clinical data set. Some

technical details are collected in the Appendix.

2 The Exchangeable Negative Binomial

In this section, we introduce the ENB and demonstrate that it is a "parametric distribution" with infinitely many parameters, followed by the characterization. We also discuss simulation and give graphs.

2.1 Definition and Moments. In modeling correlated binary data with the binomial distribution such as in a developmental study, independence is not an adequate assumption. Relaxing independence to exchangeability, George and Bowman (1995) initiated a full likelihood procedure. Dang, Keeton and Peng (2009) further developed the procedure and proposed a unified approach. Recall that a finite sequence of binary random variables X_1, X_2, ..., X_n are *exchangeable* if

$$\mathbb{P}(X_{\pi(1)} = x_1, ..., X_{\pi(n)} = x_n) = \mathbb{P}(X_1 = x_1, ..., X_n = x_n), \quad x_1, ..., x_n \in \{0, 1\}$$

for every permutation $\pi(1)$, $\pi(2)$, ..., $\pi(n)$ of 1, 2, ..., n. A sequence of infinitely many binary random variables are *exchangeable* if any finite subsequence is *exchangeable*.

Let R be the total number of successes in X_1, ..., X_n, so that $R = X_1 + X_2 + \cdots + X_n$. George and Bowman (1995) among others derived the distribution of R, which is given by

$$\mathbb{P}(R = r) = \binom{n}{r} \sum_{k=0}^{n-r} (-1)^k \binom{n-r}{k} \lambda_{r+k}, \quad r = 0, 1, 2, ..., n, \tag{2.1}$$

where $\lambda_0 = 1$, $\lambda_k = \mathbb{P}(X_1 = 1, X_2 = 1, ..., X_k = 1)$, $k = 1, 2, ..., n$. See also Dang, Keeton and Peng (2009) for more details about the distribution. From now on, we shall assume that X_1, X_2, ... is an infinite sequence of exchangeable Bernoulli r.v.'s. Let r be the desired number of successes. Let Y be an integer-valued random variable that counts the number of

trials until the first r successes are realized. Then it can be seen that the probability that the first r successes are realized in y trials is given by

$$\mathbb{P}(Y = y) = \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k}, \quad y = r, r+1, \ldots \tag{2.2}$$

Note that the exchangeable Bernoulli sequence $X_1, X_2, \ldots$ may be hidden and unobservable. What we can observe is the value of Y – the number of trials to get the first r successes; or equivalently, the number of failures to get the first r successes. Let S be the number of failures. Then $Y = S + r$ and the above probability (2.2) can be rewritten as

$$\mathbb{P}(S = s) = \binom{s+r-1}{r-1} \sum_{k=0}^{s} (-1)^k \binom{s}{k} \lambda_{r+k}, \quad s = 0, 1, \ldots \tag{2.3}$$

This latter form is used in our application of the distribution to the real data.

It is interesting to observe that there are infinitely many parameters, $\{\lambda_k : k = r, r+1, \ldots\}$, in this distribution. Further, these parameters are a part of a *completely monotone* sequence $\{\lambda_k : k = 0, 1, \ldots\}$ ($\lambda_0 = 1$) in the sense that

$$(-1)^l \Delta^l \lambda_k \geq 0, \quad k, l = 0, 1, 2, \ldots, \tag{2.4}$$

where Δ is the difference operator defined inductively by $\Delta a_i = a_{i+1} - a_i$, $\Delta^l = \Delta \Delta^{l-1}$, $\Delta^0 = I$, $l = 1, 2, \ldots$ for a sequence $\{a_1, a_2, \ldots\}$ with I the identity operator. For a given completely monotone sequence $\{\lambda_k : k = 0, 1, \ldots\}$, one has (2.2) and $\mathbb{P}(Y = y) = \binom{y-1}{r-1}(-1)^{y-r} \Delta^{y-r} \lambda_r \geq 0$ (for details, see Feller (1971)). Accordingly, to justify that (2.2) indeed defines a legitimate probability distribution, one only has to show

$$\sum_{y=r}^{\infty} \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+y} = 1. \tag{2.5}$$

Using the elegant de Finetti theorem, this can be verified and the details are given in the Appendix.

Clearly the above distribution generalizes the negative binomial distribution. Specifically, if the associated Bernoulli sequence is independent and identically distributed, so that $\lambda_k = {\lambda_1}^k$ for $k = 1, 2, ...$, then Y has a negative binomial with parameters λ_1 and r, i.e., $Y \sim \mathbf{NB}(\lambda_1, r)$. Further, the above probability (2.2) is a mixture of negative binomial distributions, since from the de Finetti theorem it follows

$$\mathbb{P}(Y = y) = \int_0^1 \binom{y-1}{r-1} \lambda_1^r (1-\lambda_1)^{y-r} \, dQ(\lambda_1), \quad y = r, r+1, ...$$

where Q is the probability measure on $[0, 1]$ uniquely determined by the infinite Bernoulli sequence. Thus, we have obtained an interesting fact that the mixture of negative binomial distributions is equivalently to a "parametric distribution" with infinitely many parameters.

Let us denote $Y \sim \mathbf{ENB}(\boldsymbol{\lambda}, r)$ with $\boldsymbol{\lambda} = \{\lambda_r, \lambda_{r+1}, ...\}$. The special case $r = 1$ corresponds to the *exchangeable geometric distribution* and we write $Z \sim \mathbf{EG}(\boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, ...\}$. The probability distribution of Z is

$$\mathbb{P}(Z = z) = \sum_{k=0}^{z-1} (-1)^k \binom{z-1}{k} \lambda_{1+k}, \quad z = 1, 2, ... \tag{2.6}$$

The exchangeable geometric distribution can be used to *efficiently* estimate the mixing measure Q, which is still under our investigation.

We now give the moment generating function (m.g.f.) with the proof delayed to the Appendix.

Theorem 1. *Let $Y \sim \mathbf{ENB}(\boldsymbol{\lambda}, r)$. Then the moment generating function of Y is given by*

$$M_Y(t) = e^{tr} \int_0^1 u^r \left(1 - (1-u)e^t\right)^{-r} dQ(u), \quad t \in B \tag{2.7}$$

for some neighborhood B of the origin in which the above integral converges.

We now derive the formulas for the mean and variance. To guarantee the existence of the first and second moment, we need to assume the convergence of certain improper integrals.

We formally define

$$\lambda_{-k} = \int_0^1 u^{-k}\, dQ(u), \quad k = 1,\, 2,\, ... \tag{2.8}$$

Differentiating the m.g.f. and evaluating the derivatives at zero, we obtain the following results.

Corollary 1. *Let $Y \sim \mathbf{ENB}(\boldsymbol{\lambda}, r)$. (1) If $\lambda_{-1} < \infty$, then the mean of Y exists and is given by $\mathbb{E}(Y) = M_Y'(0) = r\lambda_{-1}$. (2) If $\lambda_{-2} < \infty$, then the second moment of Y exists and is given by $\mathbb{E}(Y^2) = M_Y''(0) = r(r+1)\lambda_{-2} - r\lambda_{-1}$. Then, of course, the variance of Y is simply $\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y^2) - \Big(\mathbb{E}(Y)\Big)^2 = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2$.*

By Hausdorff's theorem (see e.g. Feller (1971)), the completely monotone sequence $\{\lambda_k\}$ can be expressed as the moments of the measure Q, namely,

$$\lambda_k = \int_0^1 u^k\, dQ(u), \quad k = 0,\, 1,\, 2,\, ... \tag{2.9}$$

Interestingly (2.8) extends the definition of λ_k in (2.9) from positive $k = 1,\, 2,\, ...$ to negative $k = -1,\, -2,\, ...$, so that we have $\{..., \lambda_{-1},\, \lambda_0,\, \lambda_1,\, ...\}$.

In terms of $\{\lambda_k, k = 1,\, 2,\, ...\}$ of positive indices, $\{\lambda_{-k},\, k = 1,\, 2,\, ...\}$ of negative indices and therefore the first and second moment can be expressed as follows, using the negative binomial expansion.

$$\mathbb{E}(Y) = r \sum_{j=0}^{\infty} \sum_{k=0}^{j} (-1)^k \binom{j}{k} \lambda_k, \tag{2.10}$$

and

$$\begin{aligned} \mathbb{V}\mathrm{ar}(Y) = {} & r(r+1) \sum_{k=0}^{\infty} k \sum_{j=0}^{k} (-1)^j \binom{k}{j} \lambda_j \\ & + r^2 \sum_{k=0}^{\infty} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \lambda_j - \left(r \sum_{k=0}^{\infty} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \lambda_j \right)^2. \end{aligned} \tag{2.11}$$

We take some special cases of the Beta-binomial distribution to illustrate the behavior of

the distribution. Suppose now that the distribution Q has a density, say $q(u)$, with respect to the point-mass measure (Case 1) and the Lebesgue measure (Cases 2-4).

Case 1. Q is a point mass concentrated on $p \in (0,1)$. Then the resulting distribution is the negative binomial distribution $\mathbf{NB}(p,r)$. Here we have

$$\lambda_k = \int_0^1 u^k \; dQ(u) = p^k, \quad k = 0, \pm 1, \pm 2, \ldots$$

Hence, we recover the familiar mean and variance of the negative binomial, which are

$$\mathbb{E}(Y) = r\lambda_{-1} = r/p, \; \mathbb{V}\mathrm{ar}(Y) = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2 = r(1-p)/p^2.$$

In this case all the moments exist.

Case 2. Suppose $q(u) = 1$. We have $\lambda_k = \int_0^1 u^k(1) \; du = 1/(k+1)$, $k = 0, 1, \ldots$ In this case $\lambda_{-k} = \int_0^1 u^{-k} \, dQ(u) = \int_0^1 u^{-k} \; du$ does not exist for all $k = 1, 2, \ldots$; therefore, none of the moments exists either.

Case 3. Now suppose $q(u) = 2u$. Then we have $\lambda_k = \int_0^1 u^k(2u) \; du = 2/(k+2)$, $k = 0, 1, 2, \ldots$ In this case, $\lambda_{-1} = \int_0^1 2u/u \; du = 2$. Consequently, the mean of Y is given by $\mathbb{E}(Y) = r\lambda_{-1} = 2r$. However, the variance and higher moments still do not exist.

Case 4. Finally, suppose $q(u) = 4u^3$. Then we have $\lambda_k = \int_0^1 u^k(4u^3) \; du = 4/(k+4)$, $k = 0, 1, 2, \ldots$ In this case, $\lambda_{-1} = \int_0^1 \frac{4u^3}{u} \; du = \frac{4}{3}, \lambda_{-2} = \int_0^1 \frac{4u^3}{u^2} \; du = 4\int_0^1 u \; du = 2$. Consequently, the mean and variance of Y are given by

$$\mathbb{E}(Y) = r\lambda_{-1} = 4r/3, \quad \mathbb{V}\mathrm{ar}(Y) = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2 = 2r^2/9 + 2r/3.$$

2.2 Simulations and Probability Curves. Here we give a method to simulate ENB random variables. Some histograms and probability curves are drawn to illustrate the introduced distribution.

First, we discuss a direct method of generating ENB random variables Y_1, Y_2, ..., Y_n. A probability measure Q concentrated on [0,1] can be given by specifying its density q with respect to the Lebesgue measure. Then, by (2.9), values of λ_1, λ_2, ..., λ_m can be calculated for some m. Next, values of $\mathbb{P}(Y = j)$ are computed with the calculated λ_j's. Let $F(j)$ be the cumulative distribution function given by

$$F(j) = \sum_{y=r}^{j} \mathbb{P}(Y = y) = \sum_{y=r}^{j} \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k}, \quad j = r,\, r+2,\, ...$$

Calculate $F(j)$ for $j = 1,\, 2,\, ...,\, m$. Generate n random variables, U_1, U_2, ..., U_n, from the uniform distribution on $(0, 1)$. Define $Y_j = k + r - 1$ if $F(k-1) \leq U_j < F(k)$ so that $\mathbb{P}(Y_j = k + r - 1) = \mathbb{P}(F(k-1) \leq U_j < F(k)) = F(k) - F(k-1) = \mathbb{P}(Y = k + r - 1)$. This is equivalent to defining $Y_j = F^{-1}(U_j)$. The resulting Y_1, Y_2, ..., Y_n are simulated random variables from Y having distribution $\mathbf{ENB}(\lambda_r,\, \lambda_{r+1},\, ...,\, r)$ with the selected sequence λ_j's. Note that for finite sample size n only finitely many parameters λ_j's are involved. We consider two cases $q(u) = 5u^4, r = 1$ and $q(u) = 7u^6, r = 2$ both with sample size $n = 200$ and $m = 30$ (large enough). For $q(u) = (a+1)u^a, u \in (0, 1)$ with $a \geq 0$, it is easy to compute

$$\mathbb{P}(Y = y) = \frac{(a+1)(r+a)(r+a-1)\cdots r}{(y+a+1)(y+a)\cdots y}, \quad y = r,\, r+1,\, ... \tag{2.12}$$

The histograms in Fig. 1 are generated by the simulated random variables Y_j's, superimposed with the corresponding ENB and NB probability curves respectively. We observe clear difference between the two curves, especially in the right graph. Consequently, if the Bernoulli sequence associated with a data set is *exchangeable* but not *independent*, and if we fit it with the *independence*-based negative binomial, then there would be such a difference between the true and fitted probabilities. Indeed, we have observed this difference in our application of the ENB to the real clinical data in the last section.

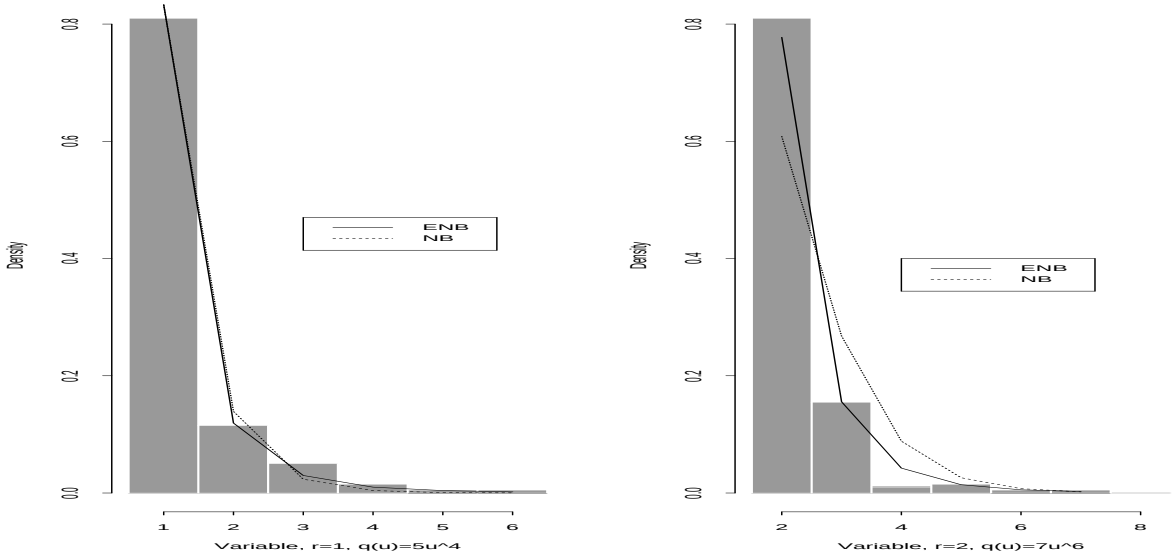

Fig. 1. The histograms superimposed with the probability curves for the ENB with $q(u) = 5u^4$, $r = 1$ and $q(u) = 7u^6$, $r = 2$ and the corresponding NB probability curves. The ENB curve fits better than the NB curve.

3 Completely Monotone Links

An ENB distribution has infinitely many parameters, which may complicate statistical inference and the interpretation of the parameters. Either theoretically or practically, we may wish to work with a distribution of finitely many parameters, because such a distribution is succinct and may allow pleasant interpretation of the parameters. There are many approaches to converting a distribution of infinitely many parameters to parsimonious distributions of finitely many parameters. Here we mention a few of them: (1) Re-parameterization. (2) Bayesian approach. (3) Truncation. (4) Completely monotone links. The last two approaches will be discussed below. The approach (4) was first appeared in George and Bowman (1995), continued in Kuk (2004), and systematically studied in Dang, Keeton and Peng (2009). We discuss (4) first and then (3) in the next section.

Note that $\mathbf{ENB}(\boldsymbol{\lambda}, r)$ has infinitely many parameters and is the saturated model with parameter space $\Lambda = \{\boldsymbol{\lambda} = (\lambda_k : k = r, r+1, ...) : \boldsymbol{\lambda} \text{ satisfies } (2.4)\}$. Following Dang, Keeton and Peng (2009), we will seek parsimonious distributions of finitely many parameters of the saturated distribution. By mapping a lower d-dimensional subset $\Theta \subset \mathbb{R}^d$ into Λ, we obtain a parsimonious model. Consider such a map from Θ into Λ defined by $\boldsymbol{\lambda} = \mathbf{h}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is the parameter space of the parsimonious model. Write $\mathbf{h} = \{h_k : k = r, r+1, ...\}$ so that $\lambda_j = h_j(\boldsymbol{\theta})$, $j = r, r+1, ...$ Substitution of these representations in (2.2) results in a parsimonious model,which can be expressed as

$$f(y; \boldsymbol{\theta}) = \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} h_{y+k}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta, \, y = r, \, r+1, \, ... \tag{3.13}$$

where $h_0(\boldsymbol{\theta}) = 1$ for every $\boldsymbol{\theta} \in \Theta$. In order to ensure that the above expression is a legitimate probability mass function, these $\{h_j(\boldsymbol{\theta})\}$ have to be *completely monotone* (2.4):

$$(-1)^k \Delta^j h_k(\boldsymbol{\theta}) \geq 0, \quad \boldsymbol{\theta} \in \Theta, \quad j = 0, \, 1, \, ..., \, k = r, \, r+1, \, ... \tag{3.14}$$

Such $\mathbf{h}$ is called a *completely monotone (inverse) link*. Dang, Keeton and Peng (2009) have given a wide variety of completely monotone link functions and demonstrated with real data how the above approach can be used to conduct statistical inference for correlated binary data. Here we quote several links from them and more can be found in their paper: the Independence-Binomial link $h_t(\theta) = \theta^t$, $\theta \in (0, 1)$; the MM-Binomial link $h_t(\theta) = \theta/(\theta + t)$, $\theta \in (0, \infty)$; the Beta-Binomial link $h_t(\boldsymbol{\theta}) = B(\theta_1 + t, \theta_2)/B(\theta_1, \theta_2)$, $\boldsymbol{\theta} \in (0, \infty)^2$; the Gamma-Binomial link $h_t(\theta) = (1 + \theta_2 t)^{-\theta_1}$, $\boldsymbol{\theta} \in (0, \infty)^2$; and the Poisson-Binomial link $h_t(\theta) = \exp(\theta(e^{-t} - 1))$, $\theta \in (0, \infty)$. We will not further discuss this approach and refer interested readers to Dang, Keeton and Peng (2009) for a full description of the approach.

4 The Truncated ENB Distribution

Now we propose a truncated ENB, which converts the distribution of infinitely many parameters into a distribution of finitely many parameters. In fact, we may view it as an approximation to the ENB because it converges to the ENB as the truncation grows to infinity. Consider a sequence of infinitely many exchangeable Bernoulli trials. Let $Y \sim \mathbf{ENB}(\boldsymbol{\lambda}, r)$ associated with the sequence, and let m be a fixed positive integer. Let $W^{(m)}$ be the number of trials necessary to get the first r successes, given that the total number of trials does not exceed m, *i.e.*, $W^{(m)} = Y|Y \leq m$. The resulting distribution is called *the truncated exchangeable negative binomial (TENB).*

We now derive the distribution of $W^{(m)}$. By definition, the distribution of $W^{(m)}$ is

$$\mathbb{P}(W^{(m)} = w) = \mathbb{P}(Y = w|Y \leq m) = \mathbb{P}(Y = w)/\mathbb{P}(Y \leq m), \quad w = r, r+1, ..., m,$$

provided $\mathbb{P}(Y \leq m) > 0$; otherwise $\mathbb{P}(W^{(m)} = w) = 0$. One issue that must be addressed is the selection of the truncation value m in real applications. We may choose m to be the largest observed value of Y. Besides, it can be seen that the limiting distribution of TENB as $m \to \infty$ is the ENB, because

$$\lim_{m\to\infty} \mathbb{P}(Y \leq m) = \mathbb{P}(Y < \infty) = 1,$$

which is demonstrated in the Appendix. Observe that Y occurring before m means that there are at least r successes in the m exchangeable Bernoulli trials. That is, the two events $\{Y \leq m\}$ and $\{R_m \geq r\}$ are equivalent, where R_m has the exchangeable binomial distribution $\mathbf{EB}(\lambda_1, ..., \lambda_m, m)$, and the formula of the probability $\mathbb{P}(R_m = r)$ is given in (2.1). Thus

$$\mathbb{P}(Y \leq m) = \mathbb{P}(r \leq R_m \leq m). \tag{4.15}$$

An algebraic proof of the above equality can be found in the Appendix. Thus we have a nice

formula:

$$\mathbb{P}(W^{(m)} = w) = \mathbb{P}(Y = w)/\mathbb{P}(r \leq R_m \leq m), \quad w = r,\, r+1,\, ...,\, m.$$

The above discussion results in a relation between the cumulative distribution of the ENB and the EB.

Remark 1. *Let $Y \sim \mathbf{ENB}(\boldsymbol{\lambda}, r)$. For a fixed positive integer m, let $R_m \sim \mathbf{EB}(\lambda_1,\, ...,\, \lambda_m,\, m)$. Then the cumulative distribution function of Y given r is equal to the probability of having at least r successes in m trials as given in (4.15).*

5 Maximum Likelihood and Empirical Estimates

This section gives an estimating procedure based on maximum likelihood and empirical estimation.

The ENB distribution has infinitely many parameters; thus it is not a parametric distribution in the strict sense. There exists well-developed theory in the literature on the estimation of parameters with infinite dimensionality, see e.g. the monograph by Bickel, Klaassen, Ritov and Wellner (1993). Here we shall directly apply maximum likelihood estimation to "estimate" the infinitely many parameters. It is interesting to observe that even though there are infinitely many parameters contained in the ENB, for a finite sample, the number of parameters contained in the (joint) likelihood is finite. To estimate all the parameters, we need to have infinitely many observations. In fact, the number of parameters in the joint likelihood for a random sample increases with the number of observations.

Note that unlike λ_i's which are continuous variables in $[0, 1]$, r is integer-valued. Both λ_i's and r are unknown in real applications and have to be estimated. We first address estimating

λ_i's. Let the observations be collected in terms of the numbers S_1, ..., S_n of failures to get the first r successes. The maximum likelihood estimators $\hat{\boldsymbol{\lambda}}_{S_n^*+r}$ are the maximizers of the average of the log-likelihood

$$\ell_n(\boldsymbol{\lambda}_{S_n^*+r}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=0}^{S_i}(-1)^k\binom{S_i}{k}\lambda_{r+k}\right) + \frac{1}{n}\sum_{i=1}^{n}\log\binom{S_i+r-1}{r-1}$$

subject to the constraints

$$\lambda_r \le 1, \quad (-1)^l\Delta^l\lambda_i \ge 0, \quad i \ge r,\, l \ge 0,\, i+l \le Y_n^*, \tag{5.16}$$

where $Y_n^* = \max(Y_1, ..., Y_n)$. A numerical solution $\hat{\boldsymbol{\lambda}}^*$ of $\hat{\boldsymbol{\lambda}}_{Y_n^*}$ can be found by the Newton-Raphson iteration scheme,

$$\hat{\boldsymbol{\lambda}}^{k+1} = \hat{\boldsymbol{\lambda}}^k - [H_n(\hat{\boldsymbol{\lambda}}^k)]^{-1}T_n(\hat{\boldsymbol{\lambda}}^k), \quad \hat{\boldsymbol{\lambda}}^k \in C_n, \tag{5.17}$$

where C_n is a feasible set, $T_n = (\partial\ell_n/\partial\lambda_i)^\top$ is the score and $H_n = (\partial^2\ell_n/\partial\lambda_i\partial\lambda_j)$ is the Hessian matrix. As for the initial value $\hat{\boldsymbol{\lambda}}^0$, we may start from the case where the hidden Bernoulli sequence is independent. That is, we may choose the initial value $\hat{\boldsymbol{\lambda}}^0 = \bar{\lambda}_r \times (1, (\bar{\lambda}_{r+1}/\bar{\lambda}_r), ..., (\bar{\lambda}_{r+1}/\bar{\lambda}_r)^{1+S_n^*})$, where $\bar{\lambda}_r$, $\bar{\lambda}_{r+1}$ are the first and second coordinate of the empirical estimate of $\boldsymbol{\lambda}_{Y_n^*}$ which is given below. The SAS subroutine *NLPNRA* performs constrained nonlinear optimization by the newton-raphson method, and we used it to carry out the numerical computation in our application.

The Estimating Procedure. When fitting a dataset with an ENB model in practice, r is usually an unknown integer-valued parameter, and has to be estimated. Here we propose a procedure to estimate both r and λ_i's, based on the maximum likelihood. To stress the dependence of $\ell_n(\boldsymbol{\lambda}_{r+S_n^*})$ on r, let us write $\ell_n(\boldsymbol{\lambda}_{r+S_n^*}) = \ell_n(r, \boldsymbol{\lambda}_{r+S_n^*})$. Note that for a finite sample, the possible values for r are $r = 1, 2, ..., r^*$, where $r^* = \max(r_i : i = 1, ..., n)$ with $r_i = Y_i - S_i$. Theoretically, possible values of r can be 1, 2, ..., Y_n^*; practically, however, we may take 1, 2, ..., r^*. Hence the MLE $(\hat{r}, \hat{\boldsymbol{\lambda}}_{r+S_n^*})$ of the parameters $(r, \boldsymbol{\lambda}_{r+S_n^*})$ are the

maximizers of the log-likelihood $\ell_n(r, \boldsymbol{\lambda}_{r+S_n^*})$ for $r = 1, ..., r_n^*$ subject to the constraints $\boldsymbol{\lambda}_{r+S_n^*} \in C_n$. Therefore, we carry out the maximization for each $r = 1, 2, ..., r_n^*$, so we obtain the maximized log-likelihood values $M_{n,r} = \ell_n(r, \boldsymbol{\lambda}_r)$. Then $(\hat{r}, \hat{\boldsymbol{\lambda}}_{r+S_n^*})$ is the argument that results in the largest value of $M_{n,1}, ..., M_{n,r_n^*}$. Our application below to the real clinical data of burn wounds uses the above procedure, where $n = 153$, $S_n^* = 8$ and $r_n^* = 9$, so that there are 9 parameters $\lambda_r, ..., \lambda_{r+S_n^*}$ to be estimated based on the 153 observations $S_1, ..., S_{153}$.

It is worth to mention that there are infinitely many parameters contained in the ENB distribution, but the number of parameters to be estimated for a random sample is finite, i.e., S_n^*+1 parameters. Besides, the number of constraints are also finite. The following result tells that we can estimate almost all the parameters asymptotically if the sample size grows to infinity. The proof can be found in the Appendix.

Proposition 1. *If* $0 < \lambda_1 < 1$, *then* $\mathbb{P}(\lim_{n\to\infty} Y_n^* = \infty) = 1$.

With the observations $S_1, ..., S_n$, we can only estimate $\lambda_r, ..., \lambda_{r+S_n^*}$. Substitution of the estimates $\hat{\boldsymbol{\lambda}}_{S_n^*} = (\hat{\lambda}_r, ..., \hat{\lambda}_{r+S_n^*})$ in (2.3) gives an estimate $\hat{\mathbb{P}}(S = s)$ of the probability $\mathbb{P}(S = s)$ as follows,

$$\hat{\mathbb{P}}(S = s) = \binom{s+r-1}{r-1} \sum_{k=0}^{s} (-1)^k \binom{s}{k} \hat{\lambda}_{r+k}. \tag{5.18}$$

This probability can only be evaluated at $s = 0, 1, ..., S_n^*$, so that any statistical analysis relating to this probability can only be carried out in the range $s \in [0, S_n^*]$.

The Empirical Estimates. There are explicit formulas for the empirical estimates, which can be used as initial values for obtaining efficient estimates. We now briefly describe the empirical estimates and the details can be found in Rayner (2005). Denote $A_y = \sum_{i=1}^n \mathbf{1}[Y_i = y]$. Then the empirical estimates of the marginal probabilities are

$$\bar{\lambda}_t = \frac{1}{n} \sum_{i=r}^{t} (-1)^{i-r} c_{ti} A_i, \quad t = r, r+1, ..., Y_n^*, \tag{5.19}$$

where $c_{ij} = \binom{i-r}{j-r}/\binom{j-1}{r-1}$. For small n, the above estimates $\bar{\lambda}_t$'s may not satisfy complete monotonicity (2.4). However, $\bar{\lambda}_t$'s will satisfy (2.4) almost surely for large n by the law of large numbers. Denote the frequency of failures by $B_s = \sum_{j=1}^{n} \mathbf{1}[S_j = s]$. Then the empirical estimates can be rewritten in terms of the failures S_1, ..., S_n as

$$\bar{\lambda}_t = \sum_{i=0}^{t-r} (-1)^i \frac{\binom{t-r}{i}}{\binom{i+r-1}{r-1}} \frac{B_i}{n}, \quad t = r,\, r+1,\, ...,\, r + S_n^*.$$

This latter form of estimates is useful when we observe the numbers S_1, ..., S_n of failures. Our application to the real data uses this latter form.

Under the exchangeable geometric distribution, the correlation of the associated binary sequence is positive and calculated by $\rho = \mathbb{C}\mathrm{ov}(X_1, X_2)/\mathbb{V}\mathrm{ar}(X_1) = (\lambda_2 - \lambda_1^2)/(\lambda_1 - \lambda_1^2)$. An estimate of this can be obtained from the plug-in estimate

$$\hat{\rho} = (\hat{\lambda}_2 - \hat{\lambda}_1^2)/(\hat{\lambda}_1 - \hat{\lambda}_1^2). \tag{5.20}$$

This quantity $\hat{\rho}$ is useful. It can be used to test the exchangeability of a sequence of Bernoulli trials, because it is well known that the correlation of a sequence of exchangeable Bernoulli trials is positive.

6 Application to a Real Clinical Dataset

This section applies the ENB to fit a real clinical dataset of burn wounds on patients.

The clinical dataset of retrospective study consists of 153 patients (age from 2 months to 82 years) with etiology of fire/flame, scald and contact burns, who were hospitalized from year 1985 to 2000. These patients were treated by an enzymatic debriding agent, in partial deep dermal or full thickness burn wounds. Each patient had 1-16 different burn wound

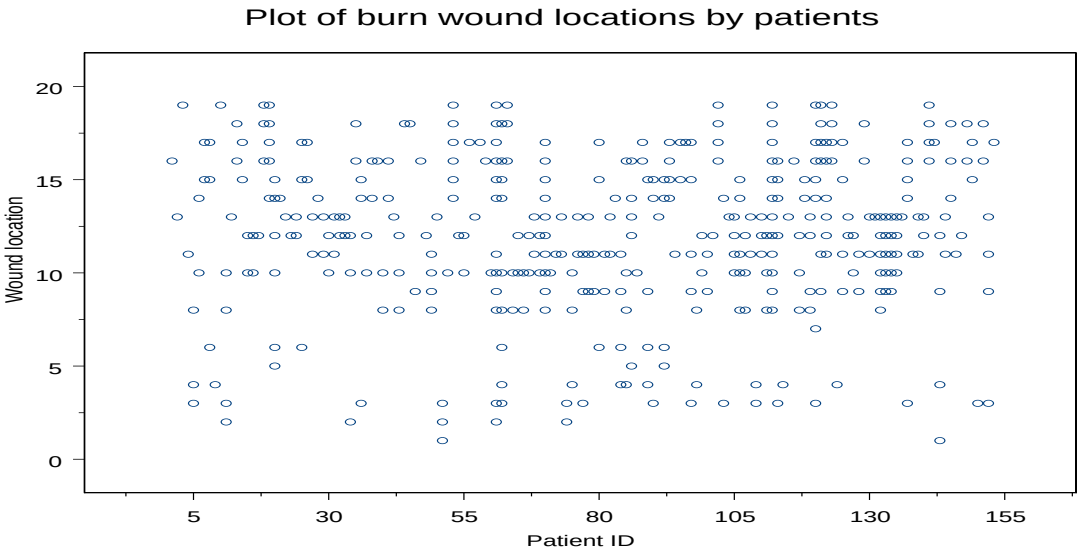

Fig. 2. The scatter plot of burn wound location against Patient ID.

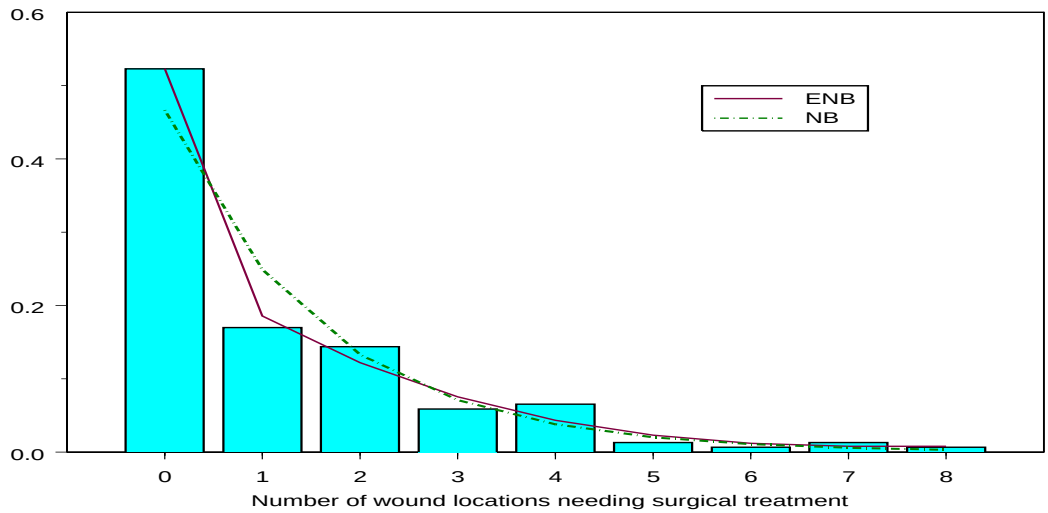

Fig. 3. The histogram of the frequencies of wound locations in need of surgical treatment superimposed with the fitted ENB and corresponding NB probability curves. The ENB fitting improves the NB fitting.

locations on his/her body such as head, neck, left/right hand, left/right leg, etc. There were 19 pre-defined burn wound locations, such as head, left/right hand, left/right leg, etc., in the study protocol and case report form. There were a total of 393 burn wound locations among the 153 patients with burn wound locations up to 12. A scatter plot of (Patient ID, Wound location) is given in Fig. 2. The mean and variance of the number of wound locations are 2.57 and 3.76 respectively. Hence a Poisson modeling is inappropriate because of the unequal mean and variance. A substitute candidate for the Poisson modeling is the negative binomial model. Here we use the proposed ENB model.

A surgical treatment such as autograft is one option for a wound area with a large full thickness defect after the enzymatic debridement. One question of interest is the rate of surgical treatment after the enzymatic debridement. Let $X_{i,j}$ be a Bernoulli random variable with $X_{i,j} = 1$ if the i^{th} wound location on j^{th} patient does not need a surgical treatment, and $X_{i,j} = 0$ otherwise for $1 \leq i \leq 19$ and $j = 1, ..., 153$. We define Y_j to be the number of wound locations for the j^{th} patient (i.e., the number of trials in the definition of the ENB) and S_j to be the number of wound locations that need a surgical treatment after the enzymatic debridement (the number of failures in the definition. Thus we have "failure=need of surgical treatment" and "success=no need of surgical treatment"), so $S_j = \sum_{1 \leq i \leq 19} \mathbf{1}[X_{i,j} = 0]$. In this setup, we may have a very nice interpretation for the parameters λ_i's. The parameter λ_1 is the probability of no need of surgical treatment and hence $1 - \lambda_1$ is the probability of surgical treatment. The parameter λ_2 is the probability of two wound locations at the same time having no need of surgical treatment.

We consider $j = 1, ..., 153$ patients to be $n = 153$ replicates from "one theoretical patient". We assume that for each of the 153 patients, the need or lack of need of surgical treatment of any of his/her burn wounds is *exchangeable*. Namely, we assume for each $j = 1, ..., 153$ the Bernoulli random variables $\{X_{i,j} : 1 \leq i \leq 19\}$ are *exchangeable*. The observations are the

numbers S_1, ..., S_{153} of failures to get the first r successes (no need of a surgical treatment). The parameters to be estimated are r and λ_r, ..., $\lambda_{r+S^*_{153}}$, where $S^*_{153} = 8$. The probability histogram of S_0, ..., S_{153} is presented in Fig. 3.

We use the estimating procedure described in Section 3 to calculate the numerical values of the MLE's of the parameters λ_r, ..., λ_{r+8} for each $r = 1, 2, ..., r^* = 9$. Using the SAS subroutine *NLPNRA* which performs nonlinear optimization by the Newton-Raphson method to optimize the log likelihood function for $r = 1, ..., 9$, we obtain the values of maximum likelihood: -223, -638, -658, -409, -407, -392, -899, -305, -393 respectively. Thus the overall maximal log likelihood value -223 is attained at $\hat{r} = 1$. The corresponding maximal likelihood estimates (MLE) and empirical estimates (EMP) of λ_1, ..., λ_9 are reported in Table 1.

From Table 1, we see $\hat{\lambda}_1 = 0.523$, indicating that the estimated rate of no need of surgical treatment after using the debriding agent is 52.3%. Hence the estimated surgical treatment rate after the enzymatic debridement is $1 - \hat{\lambda}_1 = 47.3\%$. The estimated rate of two wound locations in the same time having no need of surgical treatment after using the debriding agent is $\hat{\lambda}_2 = 33.7\%$. If *independence*, instead of *exchangeability*, of the $\{X_{i,j}\}$ is assumed, the estimated rate of no need of surgical treatment is 55.74% and the estimated rate of two wound locations in the same time having no need of surgical treatment is 30.77%. These

Table 1

MLE and Empirical Estimates (EMP) ($\hat{r} = 1$) of the Parameters for the Wound Location Data

i	1	2	3	4	5	6	7	8	9
MLE	0.523	0.337	0.274	0.257	0.254	0.254	0.254	0.254	0.254
EMP	0.523	0.353	0.327	0.386	0.540	0.837	1.392	2.340	3.824

latter two estimates are moderately different from the previous ones.

From Table 1 and by (5.20), the MLE-based and the empirical-estimate-based plug-in estimates of the correlation are $\hat{\rho}_{MLE} = 0.2557$ and $\hat{\rho}_{EMP} = 0.3188$ respectively. These two positive correlations suggest that correlation does exist but it is not very strong. The positivity of the two values also supports our assumption of *exchangeability*, because an exchangeable Bernoulli sequence admits only positive correlation. These correlations may also explain why the two groups of aforementioned estimated rates (i.e., $(52.3\%, 33.7\%)$ under *exchangeability* and $(55.74\%, 30.77\%)$ under *independence*) are different but not significantly different. Fig. 3 is the probability histogram of the wound locations superimposed with the fitted ENB and NB probability curves. We observe that the ENB model improves the NB model.

Appendix

Here we collect some of the technical details. We need the Hausdorff theorem (Feller, 1971).

Lemma 6.1. *To every infinite sequence of exchangeable binary random variables* X_1, X_2, ... *there corresponds a probability distribution* Q *concentrated on* $[0, 1]$ *such that for* $y = l + 1, l + 2, ..., l = 0, 1, 2, ...,$

$$\mathbb{P}(X_1 = 1, ..., X_l = 1, X_{l+1} = 0, ..., X_y = 0) = \int_0^1 u^l (1-u)^{y-l}\, dQ(u).$$

Proof of (2.5). We show below that $\{Y < \infty\}$ is a certainty event if $0 < \lambda_1 < 1$. It is readily verified that

$$\binom{y-1}{r-1} = O(y^{r-1}), \quad y \to \infty.$$

Then it follows from $\lim_{y\to\infty} y^{r-1}u^r(1-u)^{y-r} = 0$ for $0 \le u \le 1$ and Fatou's lemma that $\lim_{y\to\infty} \binom{y-1}{r-1} \int_0^1 u^r(1-u)^{y-r}\, dQ(u) = 0$. This and Lemma 6.1 yield $\mathbb{P}(Y = \infty) =$

$\lim_{y\to\infty} \mathbb{P}(Y = y) = 0$, which, in turn, yields that $\mathbb{P}(Y < \infty) = 1$. This shows the desired (2.5). □

Proof of Theorem 1. By definition, the moment generating function Y is

$$M_Y(t) = \mathbb{E}(e^{tY}) = \sum_{y=r}^{\infty} e^{ty}\mathbb{P}(Y = y), \quad t \in \mathbf{N}.$$

In order to derive an explicit formula, we use de Finetti theorem to obtain

$$\mathbb{P}(Y = y) = \binom{y-1}{r-1} \int_0^1 u^r (1-u)^{y-r} dQ(u), \quad y = r,\, r+1,\, ...$$

Substituting this in the formula of the m.g.f. yields an infinite series of nonnegative terms. We can swap the summation and the integration, so that

$$M_Y(t) = e^{tr} \int_0^1 u^r \left(\sum_{y=r}^{\infty} \binom{y-1}{r-1} [(1-u)e^t]^{y-r} \right) dQ(u).$$

By the Taylor expansion of the infinite negative binomial series

$$\sum_{k=0}^{\infty} \binom{r+k-1}{r-1} w^k = (1-w)^{-r}, \quad -1 < w < 1,$$

we have

$$M_Y(t) = e^{tr} \int_0^1 u^r [1-(1-u)e^t]^{-r} dQ(u).$$

This is the desired (2.7), and the proof is complete. □

Proof of (4.15). Note that

$$\mathbb{P}(Y \le m) = \sum_{y=r}^{m} \mathbb{P}(Y = y) = \sum_{y=r}^{m} \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k}.$$

Substituting $j = k + r$ and swapping the order of summation, we have

$$\mathbb{P}(Y \le m) = \sum_{y=r}^{m} \binom{y-1}{r-1} \sum_{j=r}^{y} (-1)^{j-r} \binom{y-r}{j-r} \lambda_j = \sum_{j=r}^{m} \sum_{y=j}^{m} (-1)^{j-r} \binom{y-1}{r-1} \binom{y-r}{j-r} \lambda_j$$
$$= \sum_{j=r}^{m} (-1)^{j-r} \binom{j-1}{r-1} \lambda_j \sum_{y=j}^{m} \binom{y-1}{j-1} = \sum_{j=r}^{m} (-1)^{j-r} \binom{m}{j} \binom{j-1}{r-1} \lambda_j,$$

where we used the formulas

$$\binom{y-1}{r-1}\binom{y-r}{j-r} = \binom{j-1}{y-1}\binom{y-1}{j-1} \quad \text{and} \quad \sum_{y=j}^{m} \binom{y-1}{j-1} = \binom{m}{j}.$$

Now, for $R_m \sim \mathbf{EB}(\lambda_1, ..., \lambda_m, \ m)$, by (2.1), we have

$$\mathbb{P}(r \le R_m \le m) = \sum_{i=r}^{m} \mathbb{P}(R_m = i) = \sum_{i=r}^{m} \binom{m}{i} \sum_{k=0}^{m-i} (-1)^k \binom{m-i}{k} \lambda_{i+k}.$$

Substituting $j = k + i$ and swapping the order of summation, we obtain

$$\mathbb{P}(r \le R_m \le m) = \sum_{j=r}^{m} (-1)^j \binom{m}{j} \lambda_j \sum_{i=r}^{j} (-1)^i \binom{j}{i} = \sum_{j=r}^{m} (-1)^{j-r} \binom{m}{j} \binom{j-1}{r-1} \lambda_j,$$

where we used the formulas

$$\binom{m}{i}\binom{m-i}{j-i} = \binom{m}{j}\binom{j}{i} \quad \text{and} \quad \sum_{i=r}^{j} (-1)^i \binom{j}{i} = (-1)^r \binom{j-1}{r-1}.$$

Combining the above shows the desired (4.15). □

Proof of Proposition 1. For any fixed positive integer M, we have $\mathbb{P}(Y_1 \le M) < 1$ so that $\mathbb{P}(Y_m^* \le M) = (\mathbb{P}(Y_1 \le M))^m \to 0$ as $m \to \infty$. Note for any fixed $m > 0$, we have $\mathbb{P}(\lim_{n\to\infty} Y_n^* \le M) \le \mathbb{P}(Y_m^* \le M) \to 0$, so that allowing $m \to \infty$ yields $\mathbb{P}(\lim_{n\to\infty} Y_n^* \le M) = 0$ for any fixed $M > 0$. Now letting $M \to \infty$, we have $\mathbb{P}(\lim_{n\to\infty} Y_n^* < \infty) = 0$. This is the desired result. □

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Dang, X., Keeton, S. L. and Peng, H. (2009). A unified approach for analyzing exchangeable binary data with applications to developmental toxicity studies. *Statist. Med.* **28**, 2580-2604. DOI: 10.1002/sim.3638.

Donner, A. and Donald, A. (1988). The Statistical Analysis of Repeated Binary Measurements. *J. Clinical. Epidemiology.* **41**, 899-906.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc., New York.

Gaylor, D.W. and Razzaghi, M. (1992). Process of Building Biologically-Based Dose-Response Models for Developmental Defects. *Teratology* **46**, 573-581.

Gladen, B. 1979. "The Use of the Jackknife to Estimate Proportions from Toxicological Data in the Presence of Litter Effect." *Journal of the American Statistical Association* **74**, 278-283.

George, E.O. and Bowman, D. (1995). A Full Likelihood Procedure for Analyzing Exchangeable Binary Data. *Biometrics* **51**, 512-523.

George, E.O. and Kodell, R.L. (1996). Tests of independence, treatment heterogeneity, and dose-related trend with exchangeable binary data. *J. Amer. Statist. Assoc.* **91**, 1602–1610.

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge University Press, New York.

Kuk, A.Y.C. (2004). A Litter-based Approach to Risk Assessment in Developmental Toxicity Studies Via a Power Family of Completely Monotone Functions. *Appl. Statist.* **53**, 369-386.

Kupper, L.L. and Haseman, J.K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics* **34**, 69-76.

Rayner, G.J. (2005). An Exchangeable Negative Distribution and Its Applications. *Ph.D. Dissertation*, The University of Mississippi.

Stefanescu, C. and Turnbull, B.W. (2003). Likelihood Inference for Exchangeable Binary Data with Varying Cluster Sizes. *Biometrics* **59**, 18-24.

Williams, D.A. (1975). The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenecity. *Biometrics* **31**, 949-952.

Xu, J.L. and Prorok P.C. (2003). Modelling and analyzing exchangeable binary data with random cluster sizes. *Statist. Med.* **22**, 2401-2416. DOI: 10.1002/sim.1527.

Yu, C. and Zelterman, D. (2007). Sums of exchangeable Bernoulli random variables for family and litter frequency data. *Comput. Statist. Data Analy.* **52**, 1636-1649.