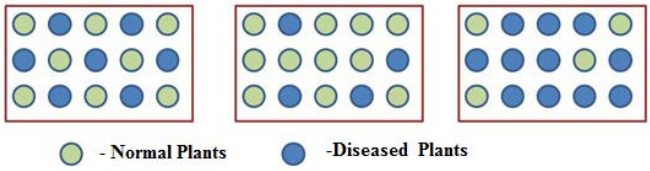
A NEW BINOMIAL MIXTURE MODEL FOR MODELING DISEASE INCIDENCE DATA

C. Manoj
Board of Study in Statistics and Computer Science

Binomial outcome data are commonly encountered in many ecological studies. Examples include “plant disease incidence data”, “plant species richness data”, “capture-recapture data “andmany others. Understanding the ecological characteristics of such data is an important focusing ecological studies. However, it is observed that the Binomial distribution often fails to model this data due to “overdispersion”.

In such situations, the class of Binomial mixture distributions can be used to well describe the binomial outcome data. Conditional on the success probability p, suppose Y follows a Binomial distribution given by $Bin(n,P)$, which is denoted by $Y|p \sim Bin(n,P)$. Unconditional probability mass function of the Y can be obtained by evaluating the well-known integral, $P_Y(y) = \int P_{Y|p}(y) f_P(p|\Theta) dp$ for $y = 0,1,\dots n$ and Θ is the parameter space of the mixing distribution$f_P(p|\Theta)$. The Beta-Binomial distribution is the classical member of this class of distributions. Even though several alternative and generalized Binomial mixtures distributions have been proposed in statistical methodological literature, the applications of these distributions are not yet discussed in a wide range including “Plant Pathology” studies. The major reason for this is that the likelihoods of the recently developed Binomial mixture distributions are complex and hence not much easier to handle compared to that of Beta-Binomial distribution. However, due to the availability of very flexible computer software packages, this is not a big problem nowadays. Here we focus on recently developed McDonald Generalized Beta-Binomial (McGBB) to model these types of data and comparisons are done with classical Beta-Binomial (BB) distribution to model the frequency distributions and to evaluate ecological characteristics. Similarities and Improvements of the two Binomial mixture distributions are discussed.



A Binomial experimental setup of plant disease incidence data.


TSWV infected tomato plants

Plant disease incidence data can be characterized by the number of diseased plants(Y) out of total number(n) of plants available in a sample unit. In order to model this type of data using the Binomial distribution, the location of a diseased plant should be independent of the location of the other diseased plant and the probability of being diseased should remain constant from plant to plant. But, by its nature, these assumptions are often violated in practice.

Here we consider a simulated realistic data. An experimental tomato field is divided into N=500 quadrats with each consisting n=5 plants. The number of tomato spotted wilt virus (TSWV) infected tomato plants out of n=5 in each of the quadrats were recorded in order to study the mean disease incidence and the degree of disease aggregation. The data in the form of the frequencies are given in Table 1.


Fruit and Leaf Symptoms of TSWV

Here, we briefly outline the two Binomial mixture distributions. Theoretical properties of the distributions are not presented here.

- **Beta-Binomial Distribution**

$$P_{BB}(y) = \binom{n}{y} \frac{B(a+y, n+b-y)}{B(a,b)} \text{for} y = 0,1, \ldots \text{n and } a, b > 0.$$

- **McDonald Generalized Beta-Binomial distribution** (Manoj, Wijekoon and Yapa.,2013)

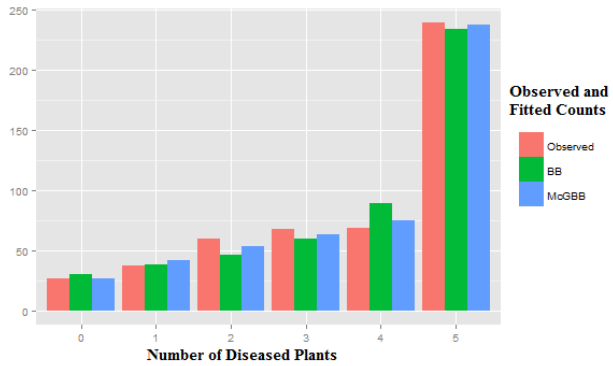
$$P_{McGBB}(y; n, \alpha, \beta, \gamma) = \binom{n}{y} \frac{1}{B(\alpha,\beta)} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma} + \alpha + \frac{j}{\gamma}, \beta\right) \text{for } y = 0,1, \ldots n$$

and $\alpha, \beta, \gamma > 0$

Maximum Likelihood Estimates of the unknown parameters of the mixture models are obtained by constructing the log-likelihood functions of the models and minimizing the negative log-likelihood function with respect to the parameter space of the particular distribution. R programs are written for estimations and comparisons.

Table 1: Modeling Results of TSWV Disease Incidence Data

Number of Diseased Plants	0	1	2	3	4	5	Total	χ^2(DF)	p-value
Observed	27	37	60	68	69	239	500	-	-
BB	30.62	38.06	46.96	60.76	89.47	234.13	500	9.7252 (3)	2.1067 (2)
McGBB	26.91	41.9	53.58	63.72	74.98	238.91	500	0.0210	0.3487



Graphical Comparison of Observed and Fitted Frequencies

Table 2: Evaluated Ecological Characteristics of TSWV Disease Incidence Data

Ecological Characteristics	BB	McGBB
Mean Disease Incidence, $\hat{\pi}_D$	0.7371	0.7339
Degree of Disease aggregation, $\hat{\rho}_D$	0.4032	0.4002

Pearson's Chi-Square goodness of fit test results indicate that McGBB (p-value=0.3487) is a better fitted model to model the plant disease distribution presented above. The BB model (p-value=0.0210) is significantly rejected to model this data based on the Chi-Square goodness of fit test. Furthermore, Analysis of Deviance (ANODEV) results indicate that the classical BB model is rejected in favour of McGBB model (p-value=0.0055). Although McGBB model provides better fit than BB model, both models result similar ecological conclusions. From the Table 2, it can be interpreted that a quite high mean diseases incidence ($\hat{\pi}_D \approx 0.73$) and a moderate degree of disease aggregation ($\hat{\rho}_D \approx 0.4$)exists in TSWV infected tomato plants in experimental fields.

A comparison study is presented for the recently developed Binomial mixture distribution, McDonald Generalized Beta-Binomial with the classical Beta-Binomial distribution for the analysis of "plant disease incidence data" which arises in many plant pathology studies. Even though Beta-Binomial model still stands as an adequate model in modeling such data, the need of an improved model in analyzing ecological data is recognized. An important point which should be noted from this study is, even though more complex models result significant improvements in modeling frequency distributions, they provide similar ecological inferences. Simulations study results do provide guidelines in identifying the parameter combinations for which a particular model performs well compared to other models. Interested readers may refer to the reference cited below for a detailed discussion on identifying a better model based on parameters combinations.

Acknowledgment

A special acknowledgement goes to the research supervisors Prof. (Mrs.) Pushpa Wijekoon and Dr. Roshan D. Yapa, Department of Statistics and Computer Science, University of Peradeniya.

References:

Manoj, C., P.Wijekoon, P. and Yapa, R.D. (2013), The McDonald Generalized Beta-Binomial Distribution: A New Binomial Mixture Distribution and Simulation Based Comparison with Its Nested Distributions in Handling Overdispersion, *International Journal of Statistics and Probability*, 2(2). DOI: 10.5539/ijsp.v2n2p24.