Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Montesinos-López et al., J Biomet Biostat 2014, 5:1
DOI: 10.4172/2155-6180.1000185

Research Article Open Access

Estimating a Proportion Based on Group Testing for Correlated Binary Response

Osval A Montesinos-López[1,2], Abelardo Montesinos-López[3], Kent Eskridge[2] and José Crossa[4]*

[1]Faculty of Telematics, University of Colima, 333 University Avenue, Col. Las Víboras, C.P. 28040 Colima, Colima, Mexico
[2]University of Nebraska, Statistics Department, Lincoln, Nebraska, USA
[3]Department of Statistics, Center for Mathematical Research (CIMAT), Guanajuato, Guanajuato, Mexico
[4]Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico, DF, Mexico

Abstract

When the sampling scheme is in clusters and when the pools (of size *k*) within a cluster are assumed not to be independent, the Dorfman model for estimating the proportion under the binomial model is incorrect. The purpose of this paper is to propose a method for analyzing correlated binary data under the group testing framework. First, assuming that the probability of an individual varies according to a beta distribution, we derived an analytic expression for the probability of a positive pool and the correlation between two pools in each cluster. Second, we derived the exact probability mass function of the number of positive pools in each cluster that should be used to obtain the maximum likelihood estimate (MLE) of the proportion of individuals with a positive outcome. However, this MLE is not efficient in terms of computational resources. For this reason, we proposed another estimator based on the beta-binomial model for obtaining the approximate MLE of the proportion of interest. Based on a simulation study, the approximate estimator produced results that are very close to the exact MLE of the proportion of interest, with the advantage that this approach is computationally more efficient.

Keywords: Pools; Correlated data; Clusters; Group testing; Beta-binomial model

Introduction

The group testing model of Dorfman [1] is effective for reducing the number of diagnostic tests because instead of performing n individual diagnostic tests, it only requires $g = \frac{n}{k}$ when retesting is not done (where k is the pool size). However, caution needs to be exercised when choosing the pool size (k), because if k is too large, the diagnostic test may be sensitive to dilution effects [2,3]. Assuming perfect testing, a pool is declared positive if at least one of the k individuals is positive, and declared free of the disease if the test is negative.

The assumption of a homogeneous distribution of transgenic maize (*Zea mays* L.) in a population, though easy to use in practice, is unrealistic [4] and therefore may affect the quality of the estimated proportion of interest. Since plant samples are taken at different locations throughout a geographical region or seed samples are taken from seed lots obtained from different regions, this means that individual plants or seed lots are inherently clustered by design and share common characteristics [5]. This clustering results in correlated samples. Therefore, it is important to develop methods for analyzing pooled data when individuals are correlated and do not require the assumption of homogeneous plant distribution, as in a binomial distribution.

When there is overdispersion (extra-binomial variation), binary data often show greater variability than predicted by the binomial model [6]. Overdispersion is said to be the norm in practice, and nominal dispersion, the exception. Hung and Swallow [7] studied the robustness of group testing in estimation problems when the underlying assumption of independent individuals is violated. They found that when defectives are clustered, as in a serial correlation model with positive serial correlation, even using a small group size offers little robustness. Group testing to estimate the proportion of defectives in a serially correlated population should be done cautiously. The recommendation is not to form groups directly from the ordered population, but to randomly assign the individuals to groups and destroy the correlation. However, group testing for classification purposes only (whether defective or non-defective) benefits from having the defectives clustered, and the clustering should be preserved and exploited [7].

Liu et al. [6] provide confidence interval procedures for estimating proportions estimated by group testing with groups of unequal size adjusted for overdispersion (extra-binomial variation). They used a quasi-likelihood approach to correct for the presence of overdispersion. However, in this case, heterogeneity in pool responses is induced by using different pool sizes (k) and may be due to the number of pools per cluster used in the group testing method. In their study, Liu et al. [6] introduced heterogeneity by assuming three clusters (m=3) and using a different pool size (k_1,k_2,k_3) in each cluster, with the following number of pools per cluster: N_1=5, N_2=10 and N_3=15. For example, when k_1=20, k_2=10 and k_3=5, they observed that if Y_1=5, Y_2=7 and Y_3=4, then $\hat{\sigma}^2 = 1.0028$, where Y_i denote the number of positive pools observed, i=1,2,3, and $\hat{\sigma}^2$ denotes the estimated dispersion parameter. However, if Y_1=1, Y_2=7 and Y_3=3, then $\hat{\sigma}^2 = 4.0733$, which indicates that the proportion of group testing varies widely for specific combinations; this also implies the presence of overdispersion. Here it is important to point out that the outcomes of the units in each cluster are assumed to be independent, identically distributed (i.i.d.) binomial distributions with N and p and that testing was conducted with no errors. However, the assumption of i.i.d. binomial distribution with N and p is not

***Corresponding author:** José Crossa, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico, DF, Mexico, Tel: 52 55 58042004 (2208); Fax: 52 55 58047559; E-mail: j.crossa@cgiar.org

Received November 05, 2013; **Accepted** January 13, 2014; **Published** January 20, 2014

Citation: Montesinos-López OA, Montesinos-López A, Eskridge K, Crossa J (2014) Estimating a Proportion Based on Group Testing for Correlated Binary Response. J Biomet Biostat 5: 185. doi:10.4172/2155-6180.1000185

Copyright: © 2014 Montesinos-López OA, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

appropriate when the sampling process is hierarchical and the plants in each cluster are correlated due to genetic factors or because the plants are spatially adjacent [6].

Regression models for pooled data have been proposed that incorporate covariates to identify which factors influence prevalence [8-10], while assuming that individual statuses (positive or negative) are independent random variables. Group testing regression models with fixed and random effects have also been developed to handle within-cluster correlation among individual latent binary responses [5], where the correlation is incorporated into the model by using the clusters as random effects; with help of covariates, it is possible to vary the prevalence between units. However, when we do not have access to covariates, it is not possible to know the unit-specific prevalences that control the correlation between units induced by this variability in the prevalence between units. Also, with these models, it is not possible to get a closed form of the likelihood function and of the correlation between pools (or individuals) induced by the random effect. For this reason, it would be useful to develop an alternative method for analyzing pooled correlated data that takes into account the correlation between individuals when estimating the proportion of interest. Such a method would provide us with an analytical expression for the likelihood function that we could use for calculating the probability of a positive pool and the correlation between two pools.

Furthermore, ignoring the correlation among individuals with cluster data under group testing produces a biased estimate of the proportion of interest; it also narrows down the confidence intervals and causes overestimated p-values for hypothesis testing. Data analysis methods are available for data with correlated responses in a non-group testing context with the correlation incorporating extra-binomial variation. One way of including extra-binomial variation is by introducing an unobserved continuous variable P_i which is independently distributed on the interval (0,1) with $E(P_i)=p_i$; $Var(P_i)=\phi[p_i(1-p_i)]$, where ϕ is the parameter of overdispersion, and by assuming that, conditional on $P_i=p_i$, R_i is binomial (m_i,p_i), $E(R_i)=m_i p_i$ and $Var(R_i)=m_i p_i(1-p_i)[1+\delta(m_i-1)]$, where δ is the intraclass correlation. However, note that when $m_i=1$, the variance does not change $Var(R_i)=p_i(1-p_i)$, but we are still introducing a correlation between the individual binary responses [11,12].

A special case of this model for extra-binomial variation, described by Williams [11], assumes that P_i has a beta distribution, which results in R_i having a beta-binomial distribution. Another distribution with the same relationship between $E(R_i)$ and $Var(R_i)$ is the correlated-binomial model [13,14], in which ϕ plays the role of a correlation between the binary components of a population.

Turechek and Madden [15] used beta-binomial distribution to estimate the proportion (p) when there is heterogeneity. The key element of their approach was to approximate the probability of a positive pool of size k in the presence of heterogeneity with the probability of a positive pool under the binomial model (assuming homogeneity, that is, assuming that p is constant across clusters) and adjust this binomial probability with the design effect ($deff=[1+\delta(k-1)]$). In this case, $deff$ was defined as the ratio of the variance of the beta-binomial model divided by the variance of the proportion under the binomial distribution. Turechek and Madden [15] then defined the effective pool size $k_{2deff}=\frac{k}{deff}$, which represents the reduction in information obtained in the pool size due to the effects of over-dispersion. Then, to correct for overdispersion when calculating the probability of a positive pool, they replaced k with k_{2deff} in the binomial model to approximate the probability of a positive pool under the beta-binomial model. However, this effective pool size $k_{2deff}=\frac{k}{deff}$ sometimes does not predict the probability of a positive pool in the presence of heterogeneity very well, so they suggested using $k_{2D}=\frac{k}{(0.98135+0.8179\theta+0.004958k+0.30387\theta k-0.3471\theta^2-0.08475\theta^2 k}$ as the effective pool size to produce better results (where $\theta=\frac{\delta}{1-\delta}$). It is important to point out that this approach works well if the correlations between pools are negligible; however, most of the time this assumption is violated in the context of plants collected from the same cluster that share genetic and environmental background.

Recent work by Lendle et al. [16] proposed group testing procedures for case identification with correlated responses for studying the efficiency of a group testing procedure when units within clusters are correlated, understanding by efficiency the expected number of diagnostic tests per unit required to classify all units as either positive or negative. In the work of Lendle et al. [16], clusters were assumed to be of equal size with the same distribution, contain exchangeable units and have a particular type of distribution. They used three models to examine how the efficiencies of group testing procedures are affected by correlated responses: a beta-binomial model where π has a beta distribution with mean p and variance $\sigma p(1-p)$; the model of Madsen [17], which is useful for modeling exchangeable binary data letting $\pi=p$ with probability $1-\sigma$, $\pi=0$ with probability $\sigma(1-p)$, and $\pi=1$ with probability σp; and the model of Morel and Neerchal [18], which is constructed by letting $\pi=p+(1-p)\sqrt{\sigma}$ with probability p and $\pi=p-p\sqrt{\sigma}$ with probability $1-p$. However, it is important to point out that the focus of the Lendle et al. [16] paper was classification, not estimation. In fact, they derived a closed-form expression for the expected number of tests per unit (i.e., efficiency) of hierarchical and matrix-based group testing procedures used for classification when units within clusters are correlated under a class of model for exchangeable binary random variables. Considering the above three models of exchangeable binary random variables in their study, they found that if units from the same cluster are tested together, the efficiency of a particular procedure can be improved, sometimes substantially, relative to random arrangements, which ignore information about cluster membership [16].

The main objective of this research is to propose a method for estimating binary responses using the Dorfman group testing model without retesting when the data were collected in clusters and the individuals within each cluster are positively and equally correlated. Negative correlations are not discussed here. To account for this correlation in the analysis, we proceed as in the standard context of the group testing binomial model, but vary the parameter p as a beta distribution, which is used to achieve a closed form of the probability mass function (pmf) of the number of positive pools in each cluster. This also allows deriving a closed form for the probability of a positive pool $(\pi_p^{(k)})$ and the correlation between two pools $(\delta_p^{(k)})$. This pmf is used to estimate the proportion of interest (π) and the correlation between two individuals (δ) [16].

It is essential to point out that with this method we get a closed expression for the probability of a positive pool and the correlation between two pools that is not available in conventional approaches for pooled correlated data. However, with the proposed model these maximum likelihood estimates (MLEs) are difficult to compute, so we

approximated them by using the beta-binomial distribution which was applied directly over the pooled correlated data to obtain estimates of $\pi_p^{(k)}$ and $\delta_p^{(k)}$ ($\hat{\pi}_p^{(k)}$and $\hat{\delta}_p^{(k)}$). Equating these two estimates with the closed-form expressions derived for $\pi_p^{(k)}$ and $\delta_p^{(k)}$, we get the approximate MLEs for π and δ while solving a system of nonlinear equations. These approximate MLEs based on the beta-binomial distribution produce results that are close to the exact MLEs derived using the proposed pmf with cheap computational resources.

Sampling Process using the Dorfman Model

Suppose that our population is composed of *I* clusters, and that *N* independent clusters are drawn from the *I* clusters in the population. Further within the *l*-th cluster, we form n_l pools of size k_l individuals, where we use the Dorfman model without retesting, with random allocation of individuals to the pools. Let Y_{ijl} denote a binary random variable that indicates whether the *i*-th individual within the pool *j* (*j*=1,2,…,n_l) in the cluster *l* (*l*=1,2,…,*N*) is diseased (Y_{ijl}=1) or not diseased (Y_{ijl}=0). Let $Z_{jl} = I_{\left(\sum_{i=1}^{k} Y_{ijl} > 0\right)}$ be the indicator variable, whether the *j*-th pool inside the cluster *l* is positive Z_{jl}=1 or negative (Z_{jl}=0).

Let us assume that all clusters are independent and that for each cluster, conditional on *p*, all individuals have a Bernoulli distribution with parameter *p*, and that *p* varies according to a beta distribution with parameters α=π/θ and β=(1−π)/θ, where π, θ>0. It is not difficult to show that for each individual, the unconditional mean and variance, respectively, are π and π(1−π), while the correlation between any two individuals within the same cluster *l*, Y_{ijl} and $Y_{i'jl}$ ($i \neq i'$), is $\delta = \frac{\theta}{\theta+1}$ (see Appendix A and Kupper and Haseman [14] for details). In this context, from Appendix A we derived that the probability that a pool of size *k* is positive, is given by

$$\pi_p^{(k)} = P\left(Z_{lj}=1\right) = 1 - \frac{B\left[\pi/\theta, (1-\pi)/\theta + k\right]}{B\left[\pi/\theta, (1-\pi)/\theta\right]} \tag{1}$$

The correlation between any two pools in the same cluster, ($\delta_p^{(k)}$), Z_{jl} and $Z_{j'l}$ ($j \neq j'$), is derived in Appendix A and given by

$$\delta_p^{(k)} = \mathrm{Cor}\left(Z_{jl}, Z_{j'l}\right) = \frac{1-\pi_p^{(2k)} - \left(1-\pi_p^{(k)}\right)^2}{\pi_p^{(k)}\left(1-\pi_p^{(k)}\right)} \tag{2}$$

where $\pi_p^{(2k)} =: 1 - \frac{B\left[\pi/\theta, (1-\pi)/\theta + 2k\right]}{B\left[\pi/\theta, (1-\pi)/\theta\right]}$ is the probability that a pool of $2k$ individuals is positive. Although we are using only pools of size *k*, $\pi_p^{(2k)}$ is a simplified notation and will be used in the proposed graphical estimator method. In this way, we can see that both the probability that a pool (of size *k*) is positive, $\pi_p^{(k)}$, and the correlation between any two pools, $\delta_p^{(k)}$, are functions of the probability that an individual is positive, π, and of the correlation between any two individuals in the same cluster, δ.

From Appendix C we have that $\pi_p^{(k)}$ increases with *k*, $\lim_{\delta\to 0}\pi_p^{(k)} = \lim_{\theta\to 0}\pi_p^{(k)} = 1-(1-\pi)^k$ and $\lim_{\delta\to 1}\pi_p^{(k)} = \lim_{\theta\to\infty}\pi_p^{(k)} = \pi$.

Deriving the Probability Mass Function of the Number of Positive Pools in a Cluster

Let $Z_l = \sum_{j=1}^{n_l} Z_{jl}$ denote the number of positive pools in cluster *l* (*l*=1,2,…,*N*) and let $f_{Z_l}(z)$ be the probability mass function (pmf) of Z_l, where

$$f_{Z_l}(z\,|\,\pi,\theta) = \frac{\binom{n_l}{z}}{B\left[\frac{\pi}{\theta}, \frac{1-\pi}{\theta}\right]} \sum_{i=0}^{z}\binom{z}{i}(-1)^i B\left[\frac{\pi}{\theta}, k_l(n_l - z + i) + \frac{1-\pi}{\theta}\right] \tag{3}$$

Details of how this pmf was derived are given in Appendix B. It is interesting to point out that $\lim_{\delta\to 0} f_{Z_l}(z|\pi,\theta) = \lim_{\theta\to 0} f_{Z_l}(z|\pi,\theta) = \binom{n_l}{z}\left[1-(1-\pi)^k\right]^z\left[(1-\pi)^k\right]^{n_l-z}$, $z \in \{0,1,\ldots,n_l\}$, that is, as $\delta \to 0$, the pmf of Z_l reduces to the binomial ($P = 1-(1-p)^k, n_l$) when there is no correlation between individuals.

Parameter Estimation

Maximum likelihood estimation

Let $z=(z_1,\ldots,z_N)$ be the vector that contains the number of positive pools of N clusters analyzed. Then, since the clusters are independent, the log-likelihood is given by

$$\ell(\pi,\theta\,|\,z) = \sum_{l=1}^{N} \log\left[f_{Z_l}(z_l\,|\,\pi,\theta)\right]$$

Thus the ML estimators $\hat{\pi}$ and $\hat{\theta}\left(\hat{\delta}\right)$ are obtained by solving the equations

$$\sum_{l=1}^{N} \frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\pi} = 0 \text{ and}$$

$$\sum_{l=1}^{N} \frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\theta} = 0$$

Where $l_l(\pi,\theta\,|\,z_l) = \log\left[f_{Z_l}(z_l\,|\,\pi,\theta)\right] = \log\left[L_l(\pi,\theta\,|\,z_l)\right]$, and $\frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\pi}$ and $\frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\theta}$ are given in Appendix D. This system of equations can be solved iteratively using the Newton-Raphson method.

Moment estimation

We first obtain moment estimates for $\pi_p^{(k)}$ and $\delta_p^{(k)}$; from this we obtain estimates for the interest parameters (π and δ) by solving a system of nonlinear equations. We define the first and second empirical moments based on the number of positive pools contained in N clusters sampled respectively by

$$m_1 = \frac{1}{N}\sum_{l=1}^{N} Z_l \text{ and}$$

$$m_2 = \frac{1}{N}\sum_{l=1}^{N} Z_l^2$$

Then, by setting these moments to their expected values and solving for $\pi_p^{(k)}$ and $\delta_p^{(k)}$, we obtain the moment estimators for the cluster *l* as

$$\tilde{\pi}_p = \frac{Nm_1}{n}$$

$$\tilde{\delta}_p = \frac{1}{n^* - n}\left[\frac{m_2 - n^*\tilde{\pi}_p^2}{\tilde{\pi}_p(1-\tilde{\pi}_p)} - n\right]$$

where $n^* = \sum_{l=1}^{N} n_l^2$ and $n = \sum_{l=1}^{n} n_l$ is the total number of pools analyzed.

Now, since $\pi_p^{(k)}$ and $\delta_p^{(k)}$ are functions of π and δ (Eq. 1 and Eq. 2), estimates of the parameters of interest can be obtained by solving the next system of nonlinear equations

$$\pi_p^{(k)} = \tilde{\pi}_p \tag{4}$$

$$\frac{1-\pi_p^{(2k)}-\left(1-\pi_p^{(k)}\right)^2}{\pi_p^{(k)}\left(1-\pi_p^{(k)}\right)} = \tilde{\delta}_p \tag{5}$$

By replacing Eq. 5 in Eq. 6, this system of equations is reduced to

$$g(\pi,\delta,k) = 1-\tilde{\pi}_p \tag{6}$$

$$g(\pi,\delta,2k) = 1-\left(1-\tilde{\pi}_p\right)\left[\tilde{\pi}_p\left(\tilde{\delta}_p-1\right)+1\right] \tag{7}$$

where $g(\pi,\delta,s) = \pi_p^{(s)} = \frac{B[\pi/\theta,(1-\pi)/\theta+s]}{B[\pi/\theta,(1-\pi)/\theta]}$.

The system of nonlinear equations given by Eq.6 and Eq.7 can be solved iteratively by the Newthon-Raphson method; alternatively, given that the right side of Equations 6 and 7 involves a quantity in the interval (0,1), and the parameters are between 0 and 1, they can be approximated by graphing the contours of g(π,δ,k) and g(π,δ,2k) at levels $1-\tilde{\pi}_p$ and $1-\left(1-\tilde{\pi}_p\right)\left[\tilde{\pi}_p\left(\tilde{\delta}_p-1\right)+1\right]$, respectively, and then observing where this intersection is located. This can be done with the R *contour* command. We will denote this solution $(\tilde{\pi},\tilde{\delta})$; it can be used as the initial value in true maximum likelihood (π and δ).

However, these MLEs are difficult to compute with the proposed model, so we approximated them using beta-binomial distribution. We applied it directly over the pooled correlated data to obtain estimates of $\pi_p^{(k)}$ and $\delta_p^{(k)}$ ($\hat{\pi}_p^{(k)}$ and $\hat{\delta}_p^{(k)}$) and, by equating these two estimates with the closed-form expressions derived for $\pi_p^{(k)}$ and $\delta_p^{(k)}$, we get approximate MLEs for π and δ by solving a system of nonlinear equations. These approximate MLEs based on beta-binomial distribution produced results that are close to exact MLEs derived using the proposed pmf with cheap computational resources.

An Alternative Approach based on Beta-binomial Distribution

Calculations of MLEs of the parameters of interest (π and δ) using the model described in the last section are difficult due to the complexity of the derived pmf. Therefore, in this section, we propose an alternative approach for estimating the parameters required with the beta-binomial model.

As shown in the previous section, the total number of positive pools in every cluster does not have a beta-binomial distribution; however, within each cluster the pool responses are binary with a probability of success $\pi_p^{(k)}$ and a positive correlation ($\delta_p^{(k)}$), which are functions of π and δ, as shown in Eq. 1 and Eq. 2. Alternative estimates of the parameters (π and δ) can be developed if we assume that the total number of positive pools in each cluster has a beta-binomial distribution with parameters n_l, $\pi_p^{(k)}$ and $\delta_p^{(k)}$, and we obtain the MLEs of $\hat{\pi}_p^{(k)}$ and $\hat{\delta}_p^{(k)}$. We can obtain the MLEs $\pi_p^{(k)}$ and $\delta_p^{(k)}$ ($\hat{\pi}_p^{(k)}$ and $\hat{\delta}_p^{(k)}$) with this approach and by solving Equations 6 and 7 for π and δ with $\tilde{\pi}_p$ and $\tilde{\delta}_p$ replaced by $\hat{\pi}_p^{(k)}$ and $\hat{\delta}_p^{(k)}$, respectively, or by directly maximizing the likelihood we estimate π and δ. We can obtain MLEs (for $\pi_p^{(k)}$ and $\delta_p^{(k)}$) using the R library VGAM and the *betabinomial* function [19]. This alternative approach based on beta-binomial distribution has computational advantages over the exact solution, since for large n_l (e.g., >20), the exact solution [Eq. 3] is unstable due to the alternative sums involved in $f_{Z_l}(z\,|\,\pi,\theta)$.

The corresponding log-likelihood using the beta-binomial model is given by

$$\ell_{BB}(\pi,\theta\,|\,\mathbf{z}) = \sum_{l=1}^{N}\log\left[f_{BB}\left(z_l\,|\,n_l,\pi_p,\theta_p\right)\right]$$

where $f_{BB}(z_l\,|\,n_l,\pi_p,\delta_p)$ is the probability function of the beta binomial with parameters n_l, π_p and θ_p evaluated at z_l; specifically,

$$f_{BB}\left(z|n_l,\pi_p,\delta_p\right) = \binom{n_l}{z}\frac{B[\pi_p/\theta_p+z,\,n_l+(1-\pi_p)/\theta_p-z_l]}{B[\pi_p/\theta_p,(1-\pi_p)/\theta_p]}$$

where π_p and θ_p (δ_p) are given by Equations 1 and 2 ignoring the superscripts.

Simulation Study

We present the results of a simulation study conducted to evaluate the performance of the approximate estimators (using the binomial or beta-binomial distribution) instead of the exact distribution (Eq. 3). The simulation study was performed using four values of π (0.025, 0.05, 0.075 and 0.1), four values of δ (0.025, 0.05, 0.075 and 0.1) and five values of N (10, 30, 50, 100, 200) with k=25 and n_l=10, l=1,…,N. For each combination of these parameters, we obtained 2000 random samples generated using the model given in Eq. 3. To estimate the relative bias (RB) and the relative mean squared error (RMSE) for each of these samples, we calculated the corresponding MLEs of the parameters using the true model, the binomial model and the beta-binomial model.

We also evaluated the results of the simulation based on the use of the beta-binomial model in order to approximate the correct distribution given in Eq. 3. This approach has an attractive computational advantage over the exact distribution. To evaluate the quality of the approximate estimators, we calculated the relative bias (RB) as

$$RB = \frac{E(\hat{\pi}_i)-\pi_0}{\pi_0}$$

and the relative mean squared error (RMSE) as:

$$RMSE = \frac{MSE(\hat{\pi}_i)}{MSE(\hat{\pi})} = \frac{E(\hat{\pi}_i-\pi_0)^2}{E(\hat{\pi}-\pi_0)^2}$$

where $\hat{\pi}$ is the MLE of π using the true model, $\hat{\pi}_i$ is the usual MLE of π using the binomial or beta-binomial model, and π_0 is the parameter for which the data were generated using the model given in Eq. 3.

Figure 1 shows the RMSE plots assuming a binomial model. All the plots show that miss-specification of the true model (Eq. 3) lowers (less than 1) the RMSE when the sample size at the cluster level is equal to 10 and larger than 1 when the sample sizes are 30, 50, 100 and 200. This means that when the number of clusters is equal to 10, the RMSE using the binomial model is smaller. However, when the sample size at the cluster level is 30, 50, 100 or 200, the RMSE with the binomial model is considerably larger and increases linearly with sample size; the performance of the binomial model is more deficient for larger values

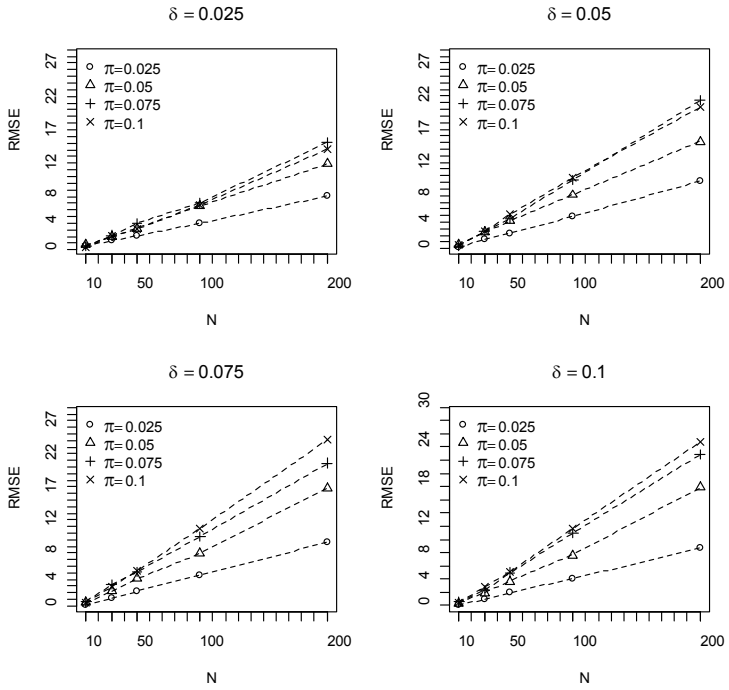

Figure 1: Relative mean squared error (RMSE) for the binomial model using various combinations of π, δ and N with k=25 and n_i=10, l=1, 2, …, N.

N		10		30		50		100		200	
δ	**π**	***RB***	***RMSE***	***RB***	***RMSE***	***RB***	***RMSE***	***RB***	***RMSE***	***RB***	***RMSE***
0.025	0.025	-0.19851	0.62720	-0.21900	1.39904	-0.21404	2.07079	-0.21983	3.99190	-0.21802	8.10129
	0.05	-0.19846	0.82584	-0.21224	2.01364	-0.21294	3.23978	-0.21866	6.61699	-0.21839	12.93352
	0.075	-0.19159	0.58552	-0.20734	2.16589	-0.21525	3.95608	-0.21557	6.99832	-0.21770	16.05542
	0.1	-0.18630	0.47723	-0.20989	2.04705	-0.21070	3.15518	-0.21594	6.79163	-0.21745	15.11770
0.05	0.025	-0.33664	0.29684	-0.34216	1.38478	-0.34687	2.22916	-0.35037	4.84176	-0.35118	10.18286
	0.05	-0.32390	0.55233	-0.34550	2.48866	-0.34682	4.21423	-0.34792	8.09808	-0.34914	16.14353
	0.075	-0.31866	0.42907	-0.33918	2.58116	-0.34041	4.53962	-0.34644	10.35087	-0.35034	22.34342
	0.1	-0.32452	0.57821	-0.33669	2.62008	-0.34382	5.20362	-0.34643	10.76827	-0.34634	21.27869
0.075	0.025	-0.41765	0.21261	-0.43070	1.14142	-0.43465	2.28080	-0.43728	4.62541	-0.43816	9.57794
	0.05	-0.41226	0.49517	-0.43016	2.22917	-0.43729	4.09072	-0.43711	7.91928	-0.44006	17.84205
	0.075	-0.41035	0.57524	-0.43316	3.24027	-0.43726	5.11599	-0.43433	10.47707	-0.43923	21.55867
	0.1	-0.40325	0.52001	-0.42756	2.83587	-0.43379	5.31205	-0.43266	11.63008	-0.43582	25.09460
0.1	0.025	-0.49251	0.16253	-0.49814	0.94624	-0.50126	1.93484	-0.50360	3.98548	-0.50508	8.79043
	0.05	-0.49530	0.39304	-0.49912	1.94190	-0.50125	3.65935	-0.50238	7.63380	-0.50425	17.88317
	0.075	-0.47931	0.54316	-0.49485	2.29268	-0.50104	4.98767	-0.50378	10.93567	-0.50427	22.90261
	0.1	-0.47330	0.44214	-0.49533	2.89089	-0.49896	5.15652	-0.50163	11.57889	-0.50280	24.79560

Table 1: Relative bias (RB) and relative mean squared error (RMSE) for the binomial model using various combinations of π, δ and N with k=25 and n_i=10, l=1,2,…,N.

of δ (Table 1). The binomial model has worse RB (Figure 2) than the beta-binomial model and the true model (Eq. 3), and underestimates the true values; this behavior is less severe as the sample size increases. Also, it is clear that increasing the correlation between individuals (δ) significantly increases the RB (Figure 2).

Figure 3 depicts the RMSE plots for the same parameters as in Figures 1 and 2. All of these plots show that, each time, the approach of the beta-binomial model in Eq. 3 performs well in RMSE. However, when δ increases, RMSE performance decreases somewhat, but is still reasonable for the larger values of δ. In addition, it is important to point out that for $N \geq 30$, performance is good and similar in all cases studied (Table 2). For the same parameters studied, Figures 3 and 4 shows that the beta-binomial model performs well in RB, except when the number of clusters is less than 30 (N<30) but comparable with the exact model; additionally, when the correlation between individuals (δ) decreases or N increases, this performance improves substantially in a similar way for both the beta-binomial approach and the exact model. Furthermore, Figure 4 shows that in all cases the approach using the beta-binomial model has a positive off-target bias, but it gradually converges to 0 as N increases, although with different patterns in each combination. The parameter that influences RB the most (in the exact and the beta-binomial approximation) is δ. For larger values of δ, RB convergence to the desired value is slower; for example, for δ=0.025, convergence is reached approximately at N>50, while for δ=0.075, it is

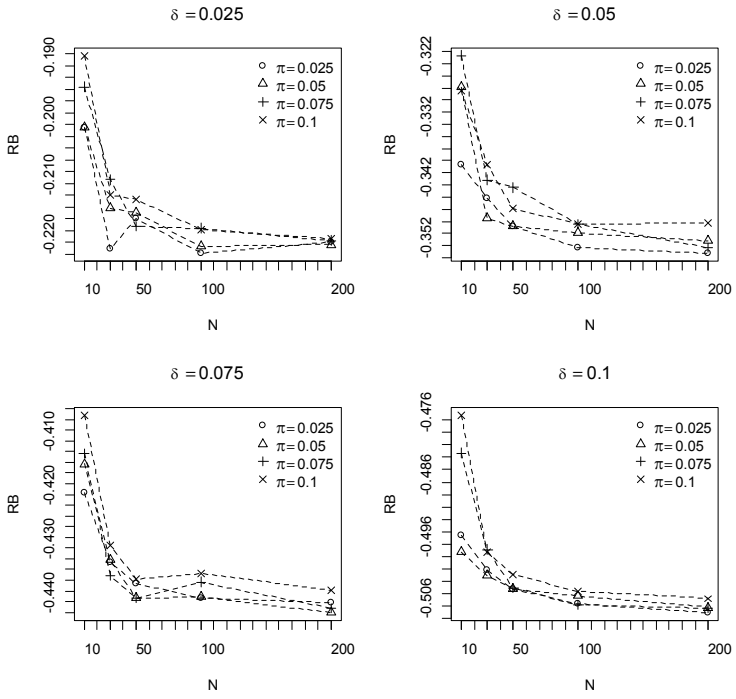

Figure 2: Relative bias (RB) for the binomial model using various combinations of π, δ and N with k=25 and n_l=10, l=1, 2, …, N.

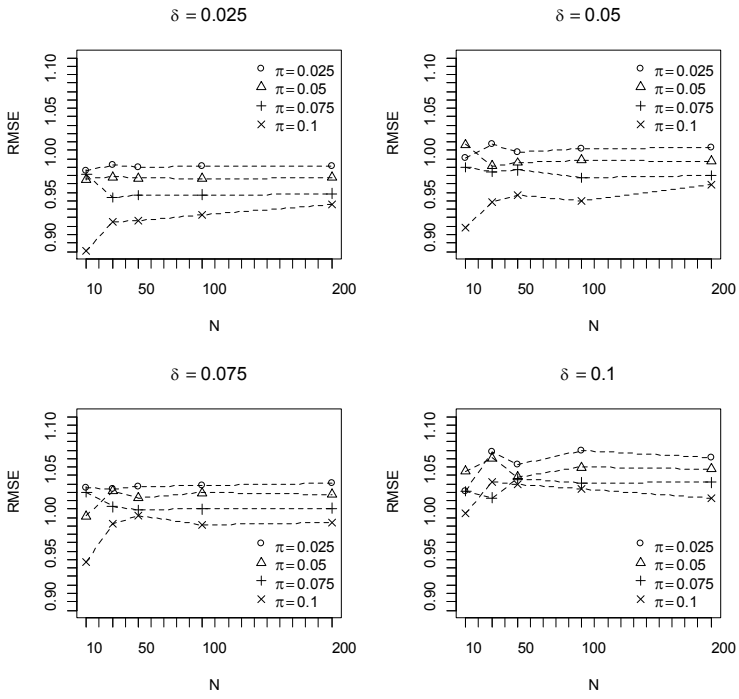

Figure 3: Relative mean squared error (RMSE) for the beta-binomial model with various combinations of π, δ and N and with k=25 and n_l=10, l=1, 2, …, N.

reached approximately at N>100. Furthermore, smaller values of π are more affected by δ because they have larger RB values, but again this is observed in both estimators of π (the beta-binomial and the exact model). Therefore, the performance in RMSE and RB of the approach based on the beta-binomial model is good and has the advantage of being more efficient than the exact distribution (Eq. 3).

Application

In this section, we give two examples to illustrate the methodology.

		δ															
		0.025				0.05				0.075				0.1			
		π				π				π				π			
N		**0.025**	**0.05**	**0.075**	**0.1**	**0.025**	**0.05**	**0.075**	**0.1**	**0.025**	**0.05**	**0.075**	**0.1**	**0.025**	**0.05**	**0.075**	**0.1**
10	*RB*	0.02947	0.02258	0.04016	0.03863	0.13102	0.07173	0.09012	0.06763	0.24288	0.12071	0.11051	0.10479	0.33570	0.18675	0.14564	0.15245
	RB_E	0.03021	0.02474	0.04221	0.04513	0.12475	0.07008	0.09192	0.07777	0.22789	0.11552	0.10892	0.11093	0.31133	0.16957	0.13704	0.15162
	RMSE	0.99578	0.98579	0.99114	0.90046	1.01075	1.02702	1.00010	0.92797	1.04519	1.01203	1.04029	0.95725	1.04125	1.06464	1.04087	1.01525
30	*RB*	0.00802	0.01036	0.01316	0.00871	0.03736	0.01877	0.01972	0.01997	0.06229	0.03655	0.02350	0.03446	0.08924	0.06558	0.06124	0.03677
	RB_E	0.00778	0.01194	0.01684	0.01474	0.03303	0.01777	0.02252	0.02682	0.05223	0.03117	0.02224	0.03797	0.07198	0.05321	0.05455	0.03376
	RMSE	1.00309	0.98850	0.96432	0.93459	1.02808	1.00170	0.99492	0.95885	1.04456	1.04185	1.02271	1.00198	1.08761	1.08077	1.03345	1.05222
50	*RB*	0.01008	0.00247	0.00309	0.00439	0.01118	0.01067	0.01058	0.00521	0.03868	0.01866	0.02416	0.01035	0.06565	0.03432	0.03297	0.03320
	RB_E	0.00997	0.00401	0.00688	0.01069	0.00796	0.01036	0.01337	0.01225	0.03002	0.01382	0.02270	0.01336	0.04995	0.02394	0.02592	0.03031
	RMSE	1.00058	0.98683	0.96670	0.93616	1.01788	1.00520	0.99666	0.96695	1.04735	1.03331	1.01907	1.01246	1.07330	1.05838	1.05647	1.04896
100	*RB*	0.00395	0.00342	-0.00092	-0.00203	0.00787	0.00838	0.00384	0.00221	0.01871	0.01858	0.01445	0.00792	0.04271	0.02876	0.01196	0.01248
	RB_E	0.00383	0.00502	0.00295	0.00423	0.00494	0.00790	0.00697	0.00957	0.01080	0.01356	0.01322	0.01130	0.02889	0.01841	0.00537	0.01005
	RMSE	1.00121	0.98672	0.96740	0.94398	1.02159	1.00838	0.98715	0.96012	1.04804	1.03945	1.01994	1.00181	1.08931	1.06980	1.05148	1.04374
200	*RB*	-0.00134	-0.00086	-0.00197	-0.00718	0.00898	-0.00057	-0.00191	-0.00822	0.01747	0.01051	0.00419	0.00133	0.02724	0.01607	0.00991	0.00703
	RB_E	-0.00140	0.00076	0.00195	-0.00099	0.00610	-0.00095	0.00124	-0.00086	0.01031	0.00597	0.00307	0.00454	0.01496	0.00642	0.00388	0.00483
	RMSE	1.00082	0.98763	0.96873	0.95629	1.02283	1.00717	0.99031	0.97972	1.05062	1.03692	1.02093	1.00423	1.08130	1.06762	1.05158	1.03280

Table 2: Relative bias (RB) and relative mean squared error (RMSE) for the beta-binomial model and Relative bias (RB_E) for the exact distribution (Eq. 3), using various combinations of π, δ and *N* with *k*=25 and n_l=10, *l*=1,2,…,*N*.

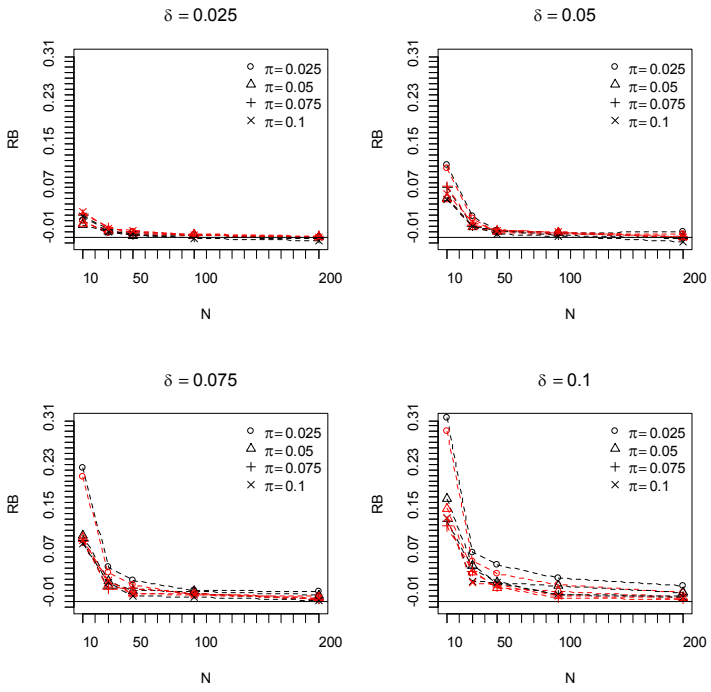

Figure 4: Relative bias (RB) for the beta-binomial model (in black) and for the true model (in red) with various combinations of π, δ and *N* and with *k*=25 and n_l=10, *l*=1, 2, …, *N*.

Example: Transgenic maize estimation

In 2009, a study was conducted to estimate the proportion of genetically modified maize plants in farmers' fields in the Sierra Juárez region of Oaxaca, Mexico (Table 3) [20]. Of an estimated total of 50 fields in the Santa María Jaltianguis locality, 30 fields were sampled; 300 leaves were collected from plants randomly chosen throughout each field. During leaf collection in each field, 4-mm leaf sections were bulked per field totaling 300 sections per bulk sampled. The remaining leaves were labelled and stored separately (a total of 9000 leaf samples were stored). The bulk samples comprising 4-mm sections of 300 leaves each were subdivided into six pools of 50 leaves each. DNA was extracted, and the presence of 35S and NOSt sequences was determined by polymerase chain reaction (PCR) (Table 3) [20].

Each 300-leaf bulk was disaggregated into 50-leaf bulks (6 per field) for DNA extraction, and PCR amplification of HSP101, 35S and NOSt sequences was performed. Data on HSP101 and NOSt amplification are not shown. Results presented in Table 3 correspond to bulks that were confirmed as positive in at least two independent PCRs [19]. Fields 6, 8, 11, 15, 25 and 27 had exactly one positive pool, field 17 had exactly 2 positive pools and field 30 had 3 positive pools.

Location:	x	0	1	2	3	4	5	6
Santa María Jaltianguis	N_x	22	6	1	1	0	0	0
	$N_{x,s}$	26	1	1	1	1	0	0

Table 3: Number of pools comprised by leaf samples from Oaxaca, Mexico (2009), with a positive 35S PCR band based on 30 fields and 300 maize leaves per field. x indicates the number of positive pools, N_x is the observed frequency of each category at this location and $N_{x,s}$ is the frequency of each category simulated assuming π=0.001324 and δ=0.045.

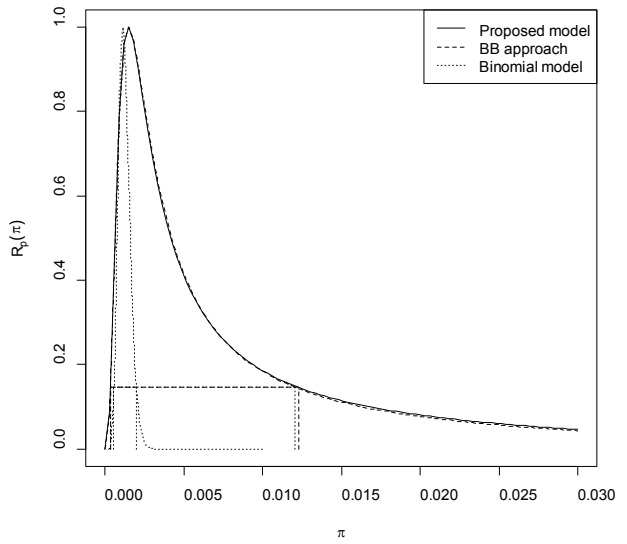

Figure 5: Relative profile likelihood function and 95% confidence interval based on the profile likelihood of π for the transgenic maize example.

The traditional binomial approach resulted in an estimated prevalence of transgenic plants of 0.001260; the exact MLEs were $\hat{\pi} = 0.001324$ and $\hat{\delta} = 0.002104$ taking into account the correlation; the beta-binomial approach gave estimates of $\hat{\pi}_{BB} = 0.001324$ and $\hat{\delta} = 0.002104$. Since the estimated correlation is low ($\hat{\delta} = 0.002104$), this data set is not appropriate for illustrating the proposed methodology. For the purpose of illustration, we assumed that 0.001324 is the true prevalence (π), and that δ=0.045 is the true correlation between individuals, and we maintained the same number of clusters and individuals per cluster (frequencies obtained now are in row $N_{x,s}$ of Table 3). Now the exact MLEs were $\hat{\pi} = 0.001486$ and $\hat{\delta} = 0.019491$, while the MLEs using the beta-binomial approach were $\hat{\pi}_{BB} = 0.001505$ and $\hat{\delta} = 0.019881$. Again, we see that the approximate MLEs based on the beta-binomial model are very close to the exact MLEs. However, assuming there is no correlation between individuals and pools, the estimated prevalence is equal to 0.001142515. The 95% Wald and profile confidence intervals for π using the exact approach were (-0.000662, 0.003635) and (0.000367, 0.012265), respectively; using the beta-binomial approach, they were (-0.000701, 0.0003711) and (0.000369, 0.012063), respectively, and using the binomial mode, they were (0.000435, 0.001850) and (0.000573, 0.002003), respectively [19]. The similarity between the results of the exact and beta-binomial models can be observed in more detail in the profile likelihood shown in Figure 5.

Example: Seed health assay

We used the data set given in Liu et al. [6] for detecting seed transmission of the cucumber green mottle mosaic virus (CGMMV). They selected seed lot (1877T-2B) of bottle gourds (*Lagenaria siceraria* L.) cv. "S-1" for testing. Test seeds of the working samples were soaked in pure water overnight; the suspensions were then used as coating antigens to initiate the indirect enzyme-linked immunosorbent assay (ELISA) for detecting the presence of CGMMV. Fifteen sub-samples were randomly taken from the seed lot. Working samples were prepared using pool sizes (k) of 1, 2, 5, 10, and 100 seeds from each sub-sample (cluster). When k=1, 2, 5, or 10, 10 replicates (n_l) of each were used in the experiment. However, if k=100 of a sample, only five replicates were used. The aim of the experiment was to estimate the proportion of infected seeds and its CI with group testing (Table 4).

The MLEs based on Eq. 3 were $\hat{\pi} = 0.020221$ and $\hat{\delta} = 0.057920$, while the approximate MLEs using the beta-binomial approach were $\hat{\pi}_{BB} = 0.020267$ and $\hat{\delta}_{BB} = 0.057821$. With the conventional binomial model, the approximate estimate was $\hat{\pi}_B = 0.006029$. The approximate MLE based on the beta-binomial approach is almost identical to the exact MLE, whereas the binomial estimate is different. Also, the estimated correlation using the beta-binomial model is very close to that given by the exact MLE. Furthermore, the 95% confidence interval based on the profile likelihood of π using the exact MLE approach is (0.006897, 0.070214) and that of the beta-binomial approach is (0.006928, 0.068484), which indicates the similarity of the results of the two approaches. Indeed, the profile likelihood of this approach overlies the profile exactly, as shown in Figure 6.

Using the traditional binomial model, (0.002724, 0.009334) and (0.003181, 0.010005) are the 95% Wald and profile confidence intervals, respectively. The similarity of these confidence intervals is due to the assumption of independence among individuals (and also among pools) in each cluster and a large sample of 135 pools. Note that these

Cluster l,	k_l	n_l	z_l	**Cluster l,**	k_l	n_l	z_l	**Cluster l,**	k_l	nl	z_l
1	1	10	0	6	2	10	0	11	10	10	4
2	1	10	1	7	5	10	0	12	10	10	0
3	1	10	0	8	5	10	1	13	100	5	0
4	2	10	1	9	5	10	1	14	100	5	0
5	2	10	2	10	10	10	2	15	100	5	0

Table 4: Data for detecting CGMMV in seed. k_l is the pool size in cluster l, n_l is the number of pools in cluster l, and z_l is the number of positive pools in cluster l.

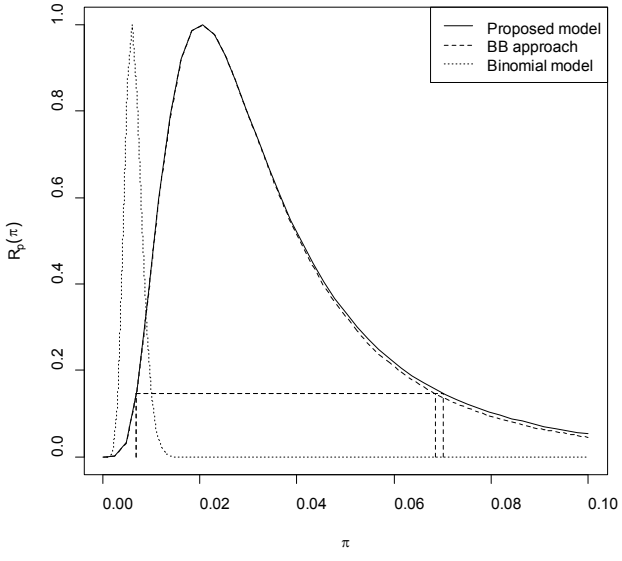

Figure 6: Relative profile likelihood function and 95% confidence interval based on the profile likelihood of π for the seed health assay data.

confidence intervals have a narrow width because they ignore extra binomial variation.

The 95% Wald confidence interval for π is (-0.002029, 0.042472) with the exact MLE approach, while with the approximate beta-binomial approach it is (-0.001769, 0.042303). As before, the exact MLE and the approximation based on the beta-binomial approach produced similar results. It is important to point out that the width of our confidence intervals is larger than the width adjusted for overdispersion that Liu et al. [6] reported. This can be explained by the fact that Liu et al. [6] used a quasi-likelihood approach to model the number of positive pools by cluster with the assumption that the individuals within each cluster are independent binary variables having the same prevalence. In contrast, our approach is based on the assumption that the responses of all individuals within each cluster are equally correlated binary variables and, as a result, we take into account the induced correlation between individuals and pools.

Conclusions

When we obtained a sample of N independent clusters from a finite population of clusters, we sampled individuals within each selected cluster and randomly allocated these individuals to n_i pools of size k_i individuals for the detection or estimation of a particular disease (positive). To produce correct estimations, in this case it is important to take into account the correlation between units and pools. For the purpose of estimation, it is important to use the probability mass function (pmf) of the number of positive pools in a cluster derived in this study to correctly estimate the proportion of interest, because it takes into account the fact that the pools formed in each cluster are correlated. Also, we showed that if we use the binomial distribution to estimate the proportion of interest, the results will present a large bias and very inflated mean square errors when $N \geq 30$. This result agrees with the paper of Hung and Swallow [7], who concluded that "for clustered and correlated individuals in each cluster even using a small pool size offers a little robustness." Since our methods (exact and approximate) induce correlations between individuals with a beta distribution, they are valid for hierarchical sampling because they take into account the correlation between individuals and pools in each cluster. This is an advantage over the approach proposed by Liu et al. [6], which is not appropriate for a hierarchical sampling process because they assumed that the individuals in each cluster are i.i.d binomial distributed and used a quasi-likelihood approach to correct for the presence of overdispersion.

For this reason, it is important to use the pmf given in Eq. 3 to obtain correct estimations of the proportion in a group testing context when the responses are correlated. However, using Eq. 3 when the sample size increases is inefficient due to the term involving the sum that it contains. For this reason, we studied an approach based on the beta-binomial model, which according to the simulation study performed, produces results that are very close to those obtained using the exact distribution [Eq. 3] with the great advantage that the approach based on the beta-binomial model is computationally more efficient, although we still need to use Eq. 1 and Eq. 2 to estimate the corresponding parameters required for the beta-binomial model. In addition, we control the induced correlation because we get a closed form of the probability of a positive pool and the correlation between any two pools.

References

1. Dorfman R (1943) The detection of defective members of large populations. The Annals of Mathematical Statistics 14: 436-440.
2. Montesinos-López OA, Montesinos-López A, Crossa J, Eskridge K, Hernández-Suárez CM (2010) Sample size for detecting and estimating the proportion of transgenic plants with narrow confidence intervals. Seed Science Research 20: 123-136.
3. Montesinos-López OA, Montesinos-López A, Crossa J, Eskridge K (2013) Sample size for detecting transgenic plants using inverse binomial group testing with dilution effect. Seed Science Research 23: 279-288.
4. Bourgeois FS, Lyman GJ (2012) Quantitative estimation of sampling uncertainties for mycotoxins in cereal shipments. Food Addit Contam Part A Chem Anal Control Expo Risk Assess 29: 1141-1156.
5. Chen P, Tebbs JM, Bilder CR (2009) Group testing regression models with fixed and random effects. Biometrics 65: 1270-1278.
6. Liu SC, Chiang KS, Lin CH, Deng TC (2011) Confidence interval procedures for proportions estimated by group testing with groups of unequal size adjusted for overdispersion. Journal of Applied Statistics 38: 1467-1482.
7. Hung M, Swallow WH (1999) Robustness of group testing in the estimation of proportions. Biometrics 55: 231-237.
8. Farrington CP (1992) Estimating prevalence by group testing using generalized linear models. Stat Med 11: 1591-1597.
9. Vansteelandt S, Goetghebeur E, Verstraeten T (2000) Regression models for disease prevalence with diagnostic tests on pools of serum samples. Biometrics 56: 1126-1133.
10. Xie M (2001) Regression analysis of group testing samples. Stat Med 20: 1957-1969.
11. Williams DA (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31: 949-952.
12. Williams DA (1982) Extra-binomial variation in logistic linear models, Journal of the Royal Statistical Society. Series C (Applied Statistics) 31: 144-148.
13. Altham PME (1978) Two generalizations of the binomial distributions. Applied Statistics 27: 162-1677.
14. Kupper LL, Haseman JK (1978) The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 34: 69-76.
15. Turechek WW, Madden LV (2003) A generalized linear modeling approach for characterizing disease incidence in a spatial hierarchy. Phytopathology 93: 458-466.
16. Lendle SD, Hudgens MG, Qaqish BF (2012) Group testing for case identification with correlated responses. Biometrics 68: 532-540.
17. Madsen T (1993) Generalized binomial distributions. Communications in Statistics-Theory and Methods 22: 3065-3086.
18. Morel JG, Neerchal NK (1997) Clustered binary logistic regression in teratology data using a finite mixture distribution. Stat Med 16: 2843-2853.
19. http://CRAN.R-project.org/package=numDeriv
20. Piñeyro-Nelson A, Van Heerwaarden J, Perales HR, Serratos-Hernández JA, Rangel A, et al. (2009) Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. Mol Ecol 18: 750-761.

Appendix A

Suppose that conditionally on p, Y has a Bernoulli distribution with parameter p, and that p has a beta distribution with parameters $\alpha = \pi/\theta$ and $\beta = (1-\pi)/\theta$. The mean and variance of Y, respectively, are:

$$E(Y) = E\left[E(Y|p)\right] = E(p) = \frac{\alpha}{\alpha+\beta} = \pi$$

and

$$Var(Y) = Var\left[E(Y|p)\right] + E\left[Var(Y \mid p)\right] = Var(p) + E\left[p(1-p)\right]]$$

$$= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha}{\alpha+\beta} - \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = \pi(1-\pi).$$

Then, if conditionally on p, Y_i and Y_j are independent Bernoulli variables with parameter p, the unconditional correlation of Y_i and Y_j is given by

$$Corr(Y_i, Y_j) = \frac{cov(Y_i, Y_j)}{\sqrt{Var(Y_i)}\sqrt{Var(Y_j)}} = \frac{1}{\pi(1-\pi)}\left[E(Y_iY_j) - E(Y_i)E(Y_j)\right]$$

$$= \frac{1}{\pi(1-\pi)}\left\{E\left[E(Y_iY_j \mid p)\right] - \pi^2\right\}$$

$$= \frac{1}{\pi(1-\pi)}\left\{E\left[E(Y_i \mid p)E(Y_j \mid p)\right] - \pi^2\right\} = \frac{1}{\pi(1-\pi)}\left[E(p^2) - \pi^2\right]$$

$$= \frac{1}{\pi(1-\pi)}\left[\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \pi^2\right] = \frac{\theta}{\theta+1} = \delta.$$

Since Z_{lj} is a binary random variable and $E(Z_{lj}|p) = 1-(1-p)^k$,

$$\pi_p^{(k)} = P(Z_{lj}=1) = E(Z_{lj}) = E\left[E(Z_{lj}|p)\right] = E\left[1-(1-p)^k\right] = 1 - E\left[(1-p)^k\right]$$

$$= 1 - \frac{B\left[\pi/\theta, (1-\pi)/\theta + k\right]}{B\left[\pi/\theta, (1-\pi)/\theta\right]}$$

This last equality is true because $E\left[(1-X)^r\right] = \frac{B(\alpha, \beta+r)}{B(\alpha,\beta)}, \forall r > -\beta$ if $X \sim Beta(\alpha, \beta)$.

Now, since all the individuals within a cluster are independent conditional on p, subsequently any two pools are as well, i.e., $Z_{lj} \perp Z_{lj'} \mid p \; \forall \; j \neq j'$

$$E(Z_{lj}Z_{lj'}) = E[E(Z_{lj}Z_{lj'} \mid p)] = E[E(Z_{lj} \mid p)E(Z_{lj'} \mid p)]$$

$$= E\left[1-(1-p)^k\right]^2 = 1 - 2E\left[(1-p)^k\right] + E\left[(1-p)^{2k}\right]$$

$$= 1 - 2\left(1-\pi_p^{(k)}\right) + 1 - \pi_p^{(2k)}$$

where $\pi_p^{(2k)} = 1 - \frac{B\left[\pi/\theta, (1-\pi)/\theta + 2k\right]}{B\left[\pi/\theta, (1-\pi)/\theta\right]}$.

Thus the correlation between two any pools in the same cluster (l) is

$$\delta_p^{(k)} = Cor\left(Z_{lj}, Z_{lj'}\right) = \frac{E\left(Z_{lj}, Z_{lj'}\right) - E\left(Z_{lj}\right)E(Z_{lj'})}{\sqrt{Var(Z_{lj})Var(Z_{lj'})}} = \frac{1-2\left(1-\pi_p^{(k)}\right)+1-\pi_p^{(2k)}-\left[\pi_p^{(k)}\right]^2}{\pi_p^{(k)}\left(1-\pi_p^{(k)}\right)}$$

$$= \frac{1-\pi_p^{(2k)}-\left(1-\pi_p^{(k)}\right)^2}{\pi_p^{(k)}\left(1-\pi_p^{(k)}\right)}$$

where $\pi_p^{(2k)} =: 1 - \frac{B\left[\pi/\theta, (1-\pi)/\theta + 2k\right]}{B\left[\pi/\theta, (1-\pi)/\theta\right]}$, which corresponds to the probability that a pool is positive as if it were made up of $2k$ individuals.

Appendix B

Note that conditionally on p, $Z_l = \sum_{j=1}^{n_l} Z_{jl}$ have binomial distribution with parameters $P = 1-(1-p)^k$ and n_l because all the individuals within a cluster are conditionally independent and $E\left(Z_{jl}|p\right) = P\left(Z_{jl}=1|\,p\right) = 1 - P(Z_{jl}=0) = 1-(1-p)^k$. Hence the marginal distribution of Z_l is given by

$$f_{Z_l}\left(z|\pi,\theta\right) = P\left(Z_l = z\right) = \int_0^1 P\left(Z_l = z|p\right) f\left(p\right) dp = \int_0^1 \binom{n_l}{z} P^z (1-P)^{n_l - z} \frac{p^{\pi/\theta-1}(1-p)^{(1-\pi)/\theta-1}}{B\left[\pi/\theta,(1-\pi)/\theta\right]} dp$$

$$= \frac{\binom{n_l}{z}}{B\left[\pi/\theta,(1-\pi)/\theta\right]} \int_0^1 \left[1-(1-p)^k\right]^z p^{\pi/\theta-1}(1-p)^{k(n_l-z)+(1-\pi)/\theta-1} dp$$

$$= \frac{\binom{n_l}{z}}{B\left[\pi/\theta,(1-\pi)/\theta\right]} \int_0^1 \sum_{i=0}^{z} \binom{z}{i} (-1)^i (1-p)^{ik} p^{\pi/\theta-1}(1-p)^{k(n_l-z)+(1-\pi)/\theta-1} dp$$

$$= \frac{\binom{n_l}{z}}{B\left[\pi/\theta,(1-\pi)/\theta\right]} \sum_{i=0}^{z} \binom{z}{i} (-1)^i \int_0^1 p^{\pi/\theta-1}(1-p)^{ik+k(n_l-z)+(1-\pi)/\theta-1} dp$$

$$= \frac{\binom{n_l}{z}}{B\left[\pi/\theta,(1-\pi)/\theta\right]} \sum_{i=0}^{z} \binom{z}{i} (-1)^i B\left[\pi/\theta, k(n_l - z + i) + (1-\pi)/\theta\right]$$

Appendix C

It is well known that $\Gamma(x+1) = x\Gamma(x)$. So by recursively applying these properties of the gamma function, we get

$$\frac{B\left[\pi/\theta, c+(1-\pi)/\theta\right]}{B\left[\pi/\theta,(1-\pi)/\theta\right]} = \frac{\Gamma\left[c+(1-\pi)/\theta\right]}{\Gamma\left[(1-\pi)/\theta\right]} \frac{\Gamma(1/\theta)}{\Gamma(c+1/\theta)}$$

$$= \prod_{j=1}^{c}\left(\frac{c-j+(1-\pi)/\theta}{c-j+1/\theta}\right) = (1-\pi)\prod_{j=1}^{c-1}\left(1-\frac{\pi}{j\theta+1}\right)$$

for $c \in \{2,3,4,\ldots\}$.

Here, it is easy to see the following:

- $\pi_p^{(k)}$ is increasing with respect to k

$$\lim_{\delta\to 0}\pi_p^{(k)} = \lim_{\theta\to 0}\pi_p^{(k)} = 1-(1-\pi)^k$$

$$\lim_{\delta\to 1}\pi_p^{(k)} = \lim_{\theta\to \infty}\pi_p^{(k)} = \pi$$

$$\lim_{\delta\to 0} f_{Z_l}(z|\pi,\theta) = \lim_{\theta\to 0} f_{Z_l}(z|\pi,\theta) = \binom{n_l}{z}\left[1-(1-\pi)^k\right]^z\left[(1-\pi)^k\right]^{n_l-z},\ \ z\in\{0,1,\ldots,n_l\}$$

Appendix D

To obtain the gradient of $l_l(\pi,\theta\,|\,z_l)$, first let c be a constant and $\psi_0(x)=\partial\Gamma(x)/\partial x$ be the first derivate of the usual gamma function. Note that

$$\frac{\partial B[\pi/\theta,(1-\pi)/\theta+c]}{\partial\pi} = \frac{\partial}{\partial\pi}\frac{\Gamma(\pi/\theta)\Gamma\left[c+(1-\pi)/\theta\right]}{\Gamma(c+1/\theta)}$$

$$= \frac{B[\pi/\theta,(1-\pi)/\theta+c]}{\theta}\{\psi_0(\pi/\theta)-\psi_0[(1-\pi)/\theta+c]\}$$

$$\frac{\partial B[\pi/\theta,(1-\pi)/\theta+c]}{\partial\theta} = \frac{\partial}{\partial\theta}\frac{\Gamma(\pi/\theta)\Gamma\left[c+(1-\pi)/\theta\right]}{\Gamma(c+1/\theta)}$$

$$= \frac{B[\pi/\theta,(1-\pi)/\theta+c]}{\theta^2}\{\psi_0(1/\theta+c)-\pi\psi_0(\pi/\theta)-(1-\pi)\psi_0[(1-\pi)/\theta+c]\}$$

Therefore

$$\frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\pi} = -\frac{\psi_0(\pi/\theta)-\psi_0[(1-\pi)/\theta]}{\theta}+$$

$$\frac{\binom{n_l}{z_l}\sum_{i=0}^{z}\binom{z_l}{i}(-1)^i B\left[\pi/\theta,c_i^l+(1-\pi)/\theta\right]\{\psi_0(\pi/\theta)-\psi_0[(1-\pi)/\theta+c_i^l]\}}{\theta B\left[\pi/\theta,(1-\pi)/\theta\right]L(z;k,n,\pi,\theta)}$$

$$= -\frac{\binom{n_l}{z_l}\sum_{i=0}^{z_l}\binom{z_l}{i}(-1)^i B\left[\pi/\theta,c_i^l+(1-\pi)/\theta\right]\Delta_0[(1-\pi)/\theta+c_i^l,(1-\pi)/\theta]}{\theta B\left[\pi/\theta,(1-\pi)/\theta\right]L_l(z_l\,|\,\pi,\theta)}$$

$$\frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\theta} = -\frac{\psi_0(1/\theta)-\pi\psi_0(\pi/\theta)-(1-\pi)\psi_0[(1-\pi)/\theta]}{\theta^2}+$$

$$\frac{\binom{n_l}{z_l}\sum_{i=0}^{z_l}\binom{z_l}{i}(-1)^i B\left[\pi/\theta,c_i^l+(1-\pi)/\theta\right]\left\{\psi_0\left(1/\theta+c_i^l\right)-\pi\psi_0(\pi/\theta)-(1-\pi)\psi_0\left[(1-\pi)/\theta+c_i^l\right]\right\}}{\theta^2 B\left[\pi/\theta,(1-\pi)/\theta\right]L_l(z_l\,|\,\pi,\theta)}$$

$$= \frac{\binom{n_l}{z_l}\sum_{i=0}^{z_l}\binom{z_l}{i}(-1)^i B\left[\pi/\theta,c_i^l+(1-\pi)/\theta\right]\{\Delta_0\left(1/\theta+c_i^l,1/\theta\right)-(1-\pi)\Delta_0[(1-\pi)/\theta+c_i^l,(1-\pi)/\theta]\}}{\theta^2 B\left[\pi/\theta,(1-\pi)/\theta\right]L_l(z_l\,|\,\pi,\theta)}$$

where $c_i^l =: k(n_l-z_l+i)$ and $\Delta_0(a,b)=\psi_0(a)-\psi_0(b)$.

Hence, using the parameterization $\theta=\delta/(1-\delta)$

$$\frac{\partial l_l(\pi,\delta\,|\,z_l)}{\partial\delta} = \frac{1}{(1-\delta)^2}\frac{\partial l_l(\pi,\theta\,|\,z_l)}{\partial\theta}\bigg|_{\theta=\delta/(1-\delta)}.$$