



Published in final edited form as:

J Agric Biol Environ Stat. 2011 March 1; 16(1): 70–87. doi:10.1007/s13253-010-0036-4.

Estimating Disease Prevalence Using Inverse Binomial Pooled Testing

Nicholas A. Pritchard[Assistant Professor] and

Department of Mathematics and Statistics, Coastal Carolina University, Conway, SC 29528, USA

Joshua M. Tebbs[Associate Professor]

Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

Nicholas A. Pritchard: npritcha@coastal.edu; Joshua M. Tebbs: tebbs@stat.sc.edu

Abstract

Monitoring populations of hosts as well as insect vectors is an important part of agricultural and public health risk assessment. In applications where pathogen prevalence is likely low, it is common to test pools of subjects for the presence of infection, rather than to test subjects individually. This technique is known as pooled (group) testing. In this paper, we revisit the problem of estimating the population prevalence p from pooled testing, but we consider applications where inverse binomial sampling is used. Our work is unlike previous research in pooled testing, which has largely assumed a binomial model. Inverse sampling is natural to implement when there is a need to report estimates early on in the data collection process and has been used in individual testing applications when disease incidence is low. We consider point and interval estimation procedures for p in this new pooled testing setting, and we use example data sets from the literature to describe and to illustrate our methods.

Keywords

Foot and mouth disease; Geometric distribution; Group testing; Maximum likelihood; Negative binomial distribution; Prevalence; West Nile virus

1. INTRODUCTION

Pooled testing, also known as group testing, arises naturally as a means of data collection in agricultural and other infectious disease applications; see, e.g., Swallow (1985), Block, Hill, and McGee (1999), Tebbs and Bilder (2004), Hepworth (2005), Katholi and Unnasch (2006), Peck (2006), and Biggerstaff (2008). The underlying premise is that individuals (e.g., insect vectors, ear notch samples, seed kernels, urine/fecal specimens, etc.) are first pooled together into groups (pools), and the groups are then tested for the presence of infection. When the infection rate is low, there is overwhelming evidence that pooled testing can confer substantial benefits when compared to testing subjects individually.

Statistical research in pooled testing has a rich history, dating back to Dorfman (1943). Since this seminal work, pooled testing has been applied to problems in blood bank screening, genetics, drug discovery, epidemiology, veterinary medicine, and other areas; see Hughes-Oliver (2006) for a review. In the estimation problem, the primary goal is to estimate p , the individual probability of infection. A large majority of research in the estimation problem has presumed a binomial model. However, other models have been used to achieve this goal, including the hypergeometric (Bhattacharyya, Karandinos, and DeFoliart 1979), the beta binomial (Turechek and Madden 2003), and, in the related application of limiting dilution assay, the Poisson (see, e.g., Chick 1996). Bayesian and empirical

Bayesian approaches have been proposed (see, e.g., Bilder and Tebbs 2005), and when individual covariate information is available, regression models for pooled response have been developed to produce covariate-adjusted estimates; see Vansteelandt, Goetghebeur, and Verstraeten (2000), Xie (2001), Huang and Tebbs (2009), and Chen, Tebbs, and Bilder (2009).

In this paper, we consider the problem of estimating disease prevalence from pooled samples collected using inverse binomial sampling (Haldane 1945). Our interest in this sampling model (applied to pooled testing) arises from Katholi and Unnasch (2006), who suggested that inverse sampling may be more appropriate when there is a need to report estimates early on in the data collection process. One can easily envision experimental situations where this would be the case, particularly if the pathogen under investigation is subject to constant surveillance and/or is very serious. Under the inverse sampling model, unlike the standard binomial, the number of positive pools n is fixed in advance, and testing is complete when the n th positive pool is observed. Because the binomial model presumes a fixed number of pools to be observed, inverse sampling may be preferred if the occurrence of one or a few positive pools triggers an immediate response; e.g., initiating appropriate disease intervention efforts. In addition, in emergency situations such as natural disasters, there would naturally be a need to quickly measure the amount of risk and to eliminate the spread of infection (Foppa et al. 2007). We now discuss two situations where our methods are potentially applicable.

1.1. West Nile Virus

West Nile Virus (WNV) is an arthropod-borne pathogen transmitted predominantly by mosquito vectors and is the leading cause of arboviral encephalitis in the United States. In 2002, there were about 3,500 documented cases of WNV in humans, resulting in 200 fatalities in the US (Solomon et al. 2003). WNV detection becomes critical after natural disasters, such as hurricanes and floods, due to a potential increase in the mosquito population. For this and other viral diseases vectored by mosquitos, surveillance is an important part of public health risk assessment and rapid detection is critical (Lanciotti and Kerst 2001).

1.2. Foot and Mouth Disease

Foot and mouth disease (FMD) has been described as one of the most economically important diseases (Callahan et al. 2002) because of the devastating effects an outbreak can have on the meat and milk industry. It has also been suggested that FMD could be used as a biological weapon (Zilinskas 1997; Koda 2002). FMD is extremely contagious, and although vaccinations do exist, it is common to slaughter all animals on or near infected premises if the outbreak is severe. In 2001, an outbreak occurred in Essex County, England, which led to the slaughter of six million animals, mostly cattle, sheep, and pigs (Bush et al. 2005). After this disaster, there was major incentive to develop assays to screen for FMD rapidly and reliably (Reid et al. 2002). Early detection of infection could limit the severity of an outbreak, and countermeasures could be taken to prevent further infection.

1.3. Assumptions and Related Work

We assume that individual statuses are independent and identically distributed Bernoulli random variables with mean p and that tests are performed on pools of individuals until the n th positive pool is found. Also, we assume that the testing process is perfect (i.e., pools are not misclassified). This may be reasonable for applications in arbovirus surveillance and FMD detection with assays based on nucleic acid technology (NAT). Investigators have shown empirically that NATs are extremely sensitive and specific (Lanciotti and Kerst

2001; Reid et al. 2002; Callahan et al. 2002). For imperfect testing, one could modify our methods in a manner similar to that of Tu, Litvak, and Pagano (1995) in the binomial model.

Katholi (2006) considered the negative binomial model for pooled testing estimation but examined the maximum likelihood estimator (MLE) only in the equal pool size case. In our survey of the arbovirus literature, for example, we found that pools are often of unequal size. We thus relax this assumption and treat equal pool sizes (which gives rise to the negative binomial model) as a special case. When equal sized pools are used, we also present alternative estimators that are shown to be far superior to the MLE. We examine interval estimation for p , and we use a WNV data example from the literature to illustrate the potential use of our methods.

2. POINT ESTIMATION

2.1. Maximum Likelihood

Suppose that $Y = y$ pools are tested until the first positive is observed; i.e., $n = 1$. For $j = 1, 2, \dots, y$, we let s_j denote the fixed pool size for the j th pool. The probability that pool j is positive is $1 - (1 - p)^{s_j}$ so that the log-likelihood is

$$l(p|y) = \log\{1 - (1 - p)^{s_y}\} + \sum_{j=1}^{y-1} s_j \log(1 - p).$$

A closed-form expression for the MLE exists and is given by

$$\widehat{p} = 1 - \left(\frac{s_+}{s_y + s_+} \right)^{1/s_y}, \quad (2.1)$$

where $s_+ = \sum_{j=1}^{y-1} s_j$. When the pool sizes are equal, i.e., when $s_j = s$ for all j , Y follows a geometric distribution with success probability $\theta = 1 - (1 - p)^s$, and (2.1) reduces to

$$\widehat{p} = 1 - \left(1 - \frac{1}{Y} \right)^{1/s}.$$

Now suppose that testing continues until the n th positive pool is found, where $n > 1$. In this case, we can envision the observed data as Y_1, Y_2, \dots, Y_n , where Y_i plays the role of Y above. For $i = 1, 2, \dots, n$ and for $j = 1, 2, \dots, y_i$, let s_{ij} denote the size for the j th pool in the sequence of pools tested to obtain the i th observation Y_i so that s_{iy_i} denotes the size of the i th positive pool. The log-likelihood is

$$l(p|\mathbf{y}) = \sum_{i=1}^n \left[\log\{1 - (1 - p)^{s_{iy_i}}\} + \sum_{j=1}^{y_i-1} s_{ij} \log(1 - p) \right],$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)$, and the score function, obtained through differentiation, is

$$S(p|\mathbf{y}) = \frac{\partial l(p|\mathbf{y})}{\partial p} = \sum_{i=1}^n \frac{s_{iy_i}(1-p)^{s_{iy_i}-1}}{1-(1-p)^{s_{iy_i}}} - \sum_{i=1}^n \sum_{j=1}^{y_i-1} \frac{s_{ij}}{1-p}. \quad (2.2)$$

The MLE for p , \hat{p} , is the solution to $S(p|\mathbf{y}) = 0$ and must be obtained iteratively, except when pools are of equal size, say s , in which case the MLE reduces to

$$\hat{p} = 1 - \left(1 - \frac{n}{T}\right)^{1/s}, \quad (2.3)$$

where $T = \sum_{i=1}^n Y_i$. In the equal pool size case, T follows a negative binomial distribution with waiting parameter n and success probability $\theta = 1 - (1-p)^s$; i.e., $T \sim \text{nib}(n, \theta)$. The estimator in (2.3) is the only (non-Bayesian) estimator considered by Katholi (2006). In the equal pool size case, Jensen's Inequality can be used to show that $E(\hat{p}) > p$; i.e., \hat{p} is positively biased. However, \hat{p} enjoys the usual large sample properties of MLEs, specifically, consistency and asymptotic normality.

We have performed simulation studies to assess the point estimate properties of \hat{p} using equal and unequal pool sizes for small n . Because of limitations on space, we do not present the results. However, it suffices to mention that the MLE \hat{p} is severely biased upwards, especially if n is small. For example, when $n = 5$, $p = 0.10$, and s_{ij} are sampled from a discrete uniform distribution from 5 to 15, $\mathcal{D} \mathcal{U}(5, 15)$, the positive bias in \hat{p} is approximately 0.10 and the mean squared error (MSE) is approximately 0.08! Our simulations suggest that the bias and MSE do converge to 0 as $n \rightarrow \infty$, as expected, but at a very slow rate, much too slow for the MLE to be used in application. This motivates us to find improved small sample estimators.

2.2. Alternative Estimators Using Equal Sized Pools

In the binomial model, the MLE for p is also positively biased (see, e.g., Swallow 1985). This has prompted a large literature on ways to reduce the bias; see Hepworth and Watson (2009) for an excellent review. In the negative binomial model, however, the number of pools observed is not fixed; rather, the number of observed positive pools is fixed. From a practical point of view, this complicates things significantly. For example, depending on the level of infection (and the pool size), thousands of pools may be needed to observe, say, $n = 2$ or 3 positives. In other words, in situations where inverse binomial pooled testing might be used, n may not be large enough for asymptotic considerations to apply. This makes Taylor series expansions, commonly used in the binomial problem to reduce bias (Burrows 1987), somewhat vacuous. This notwithstanding, our desire to reduce the bias in \hat{p} has led us to examine four alternative estimators when equal group sizes are used. Each estimator is motivated by the work of Colón, Patil, and Taillie (2001), who examined alternative pooled testing estimators in the binomial case.

Because $\hat{p} = 1 - \{1 - (n/T)\}^{1/s}$ is positively biased, one way to reduce the bias is to shrink the fraction of positive pools n/T towards 0. We define the "shrinkage" estimator by

$$\hat{p}_\alpha = 1 - \left\{1 - \alpha \left(\frac{n}{T}\right)\right\}^{1/s},$$

where $0 < \alpha \leq 1$. Instead of using a multiplicative constant, one can also shrink n/T by adding a positive constant to T in the denominator. We define the “shift” estimator by

$$\widehat{p}_\beta = 1 - \left(1 - \frac{n+1}{T+\beta}\right)^{1/s},$$

where $\beta \geq 1$. Using $n+1$ in the numerator of \widehat{p}_β is motivated by a Bayesian approach described in the Ph.D. dissertation of N. Pritchard. Briefly, an estimator similar to \widehat{p}_β arises from using a beta(1, ζ) prior on p and a squared-error loss function on $\theta = 1 - (1-p)^s$, the probability of observing a positive pool. By combining the attributes of the shrinkage and shift estimators, we also investigated a “combined” estimator

$$\widehat{p}_C = 1 - \left\{1 - \alpha_C \left(\frac{n+1}{T+\beta_C}\right)\right\}^{1/s},$$

where $0 < \alpha_C \leq 1$ and $\beta_C \geq 1$. Choosing α , β , α_C , and β_C has yet to be discussed. Finally, we also examined the jackknife estimator

$$\widehat{p}_J = \frac{1}{n} \sum_{i=1}^n \{n\widehat{p} - (n-1)\widehat{p}_{(i)}\},$$

where $\widehat{p}_{(i)}$ is the MLE of p when Y_i is omitted. When $n=1$, \widehat{p}_J simplifies to the MLE.

2.3. Comparisons

We compare the four alternative estimators and the MLE in the equal pool size case. Because $T \sim \text{nib}(n, \theta)$, where $\theta = 1 - (1-p)^s$, the exact bias and MSE of a generic estimator \tilde{p} can be expressed as an infinite sum. Because these sums do not reduce to anything tractable, we tried to approximate the bias and MSE using asymptotic expansions similarly to Colón, Patil, and Taillie (2001) in the binomial case. Unfortunately, this approach was not successful for the reasons mentioned in Section 2.2. We therefore approximate the bias and MSE of an estimator \tilde{p} by

$$B(\tilde{p}) \approx \sum_{t=n}^{t^*} (\tilde{p} - p) \binom{t-1}{n-1} \{1 - (1-p)^s\}^n (1-p)^{s(t-n)}$$

and

$$\text{MSE}(\tilde{p}) \approx \sum_{t=n}^{t^*} (\tilde{p} - p)^2 \binom{t-1}{n-1} \{1 - (1-p)^s\}^n (1-p)^{s(t-n)},$$

respectively, where $\text{pr}(T \leq t^*) \geq 1 - \nu$ for ν small. We take $\nu = 0.000001$ throughout, making these approximations very close to the true values of bias and MSE.

To use the alternative estimators, one must specify the additional values of α , β , α_C , and β_C . For \hat{p}_α , we define

$$\alpha^* = \arg \min_{0 < \alpha \leq 1} \text{MSE}(\hat{p}_\alpha; p),$$

the value of α that minimizes $\text{MSE}(\hat{p}_\alpha)$ for fixed n , s , and p , and we call this the “optimal” choice of α . Optimal values β^* and (α_C^*, β_C^*) are defined similarly for \hat{p}_β and \hat{p}_C , respectively. Of course, by this definition, optimal selection is not possible in application since each depends on the unknown p . We discuss this momentarily.

Figures 1 and 2 display the absolute bias and MSE of the five estimators, respectively, when $n = 5$ for pool sizes $s = 5, 10, 30$, and 50 . For \hat{p}_α , \hat{p}_β , and \hat{p}_C , optimal choices of α , β , α_C , and β_C are used. We use absolute bias in Figure 1 for cosmetic reasons only; otherwise, it is difficult to display the results. It should suffice to note that the MLE \hat{p} is the only estimator that is always positively biased, the jackknife estimator \hat{p}_J is negatively biased for small p , and \hat{p}_α , \hat{p}_β , and \hat{p}_C are always negatively biased. For the settings in Figs. 1 and 2, $\alpha_C^* \approx 1$ and $\beta_C^* \approx \beta^*$ for most values of p , making it difficult to distinguish \hat{p}_β from \hat{p}_C . Overall, the adjusted estimators \hat{p}_α , \hat{p}_β , and \hat{p}_C are clearly better choices, and the latter two are favored slightly. The jackknife estimator is not shown in Figure 2, because its absolute bias, shown in Figure 1, renders it impractical for use.

A concern arises with using the alternative estimators \hat{p}_α , \hat{p}_β , and \hat{p}_C because their optimal versions are not known. In practice, we recommend choosing the additional parameters α , β , α_C , and β_C to minimize MSE using an upper bound for p , say, $p_0 > p$. This is the same recommendation given by Swallow (1985) when choosing pool sizes in the binomial case and is not unrealistic in the applications we consider. We now illustrate with two examples.

Example 1—Data from two separate studies, Rutledge et al. (2003) and Godsey et al. (2005), suggest that an upper bound for the prevalence of WNV is around 0.01, and both studies use pool sizes rarely greater than 50 mosquitos. For illustration, we take $n = 5$, $s = 50$ and $p_0 = 0.01$, which provide $\alpha^* = 0.7050$, $\beta^* = 5.6868$, $\alpha_C^* = 1$, and $\beta_C^* = 5.6868$. Figure 3 displays the absolute bias (top left) and MSE (top right) with a vertical line at $p_0 = 0.01$ in each. Two important points are to be made. First, the adjusted estimators \hat{p}_α , \hat{p}_β , and \hat{p}_C , each based on the upper bound of $p_0 = 0.01$, outperform the MLE \hat{p} when $p < p_0$. Second, the bias and MSE of the adjusted estimators are much smaller for values of $p < p_0$ than the bias and MSE when $p > p_0$, supporting our recommendation to choose the adjusted estimators using an upper bound for p . However, the upper bound should be chosen carefully; for $p > 0.01$, the absolute biases of \hat{p}_α , \hat{p}_β , and \hat{p}_C increase until they eventually become larger than that of the MLE (not displayed in Figure 1). That is, choosing $p_0 < p$ could lead to adjusted estimators that are worse than the MLE.

Example 2—Data from the 2001 outbreak of FMD in England suggest that the individual prevalence does not exceed 0.10 (Anderson 2002). While it has been envisioned using polymerase chain reaction assays on pools of blood or serum in FMD settings (Reid et al. 2002), pooling has not been empirically evaluated. For verisimilitude, we thus take $n = 5$, $s = 30$ and $p_0 = 0.10$. These choices give $\alpha^* = 0.9603$, $\beta^* = 1.2495$, $\alpha_C^* = 1$, and $\beta_C^* = 1.2495$. Figure 3 displays the absolute bias (bottom left) and MSE (bottom right) with a vertical line at $p_0 = 0.10$ in each. The same discoveries are revealed; namely, the adjusted estimators are much better than the MLE when $p < p_0$ and choosing p_0 carefully is important to avoid estimators with high bias and MSE.

We have evaluated the performance of the MLE \hat{p} , the jackknife estimator \hat{p}_J , and the alternative estimators using many other settings for n , s , and p ; more extensive results are available from the authors. We have consistently observed that \hat{p}_β and \hat{p}_C are the best choices and have nearly the same bias and MSE, except when the pool size s or the waiting parameter n is very large. In these cases, \hat{p}_C provides the largest reduction in bias and the smallest MSE when an upper bound for p is correctly specified. As a special case, when $n = 1$, it is worth mentioning that the MLE is abhorrently bad and that the adjusted estimators \hat{p}_β and \hat{p}_C offer substantial improvements.

3. INTERVAL ESTIMATION

In this section, we derive (large sample) Wald, score, and likelihood ratio confidence intervals for p when inverse pooled testing is used. An exact interval is also presented for the equal pool size case. We focus on two-sided intervals herein; one-sided intervals are formed by making appropriate modifications.

3.1. Likelihood-Based Intervals

When n is large, the MLE \hat{p} is approximately normal with mean p and variance equal to $I(p)^{-1}$, where $I(p) = E_p \{-\partial^2 l(p|\mathbf{y})/\partial p^2\}$ is the Fisher information. When pools of equal size s are used, $l(p|\mathbf{y})$ simplifies considerably (see Section 2.1) and it can be shown that

$$I(p) = \frac{ns^2(1-p)^{s-2}}{\{1-(1-p)^s\}^2}.$$

Applying the Weak Law of Large Numbers, continuity of convergence, and Slutsky's Theorem, we have $I(\hat{p})^{-1/2}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, 1)$, as $n \rightarrow \infty$, so that

$$\hat{p} \pm z_{\gamma/2} I(\hat{p})^{-1/2}$$

is an approximate $100(1 - \gamma)$ percent Wald confidence interval for p . The symbol $z_{\gamma/2}$ denotes the upper $\gamma/2$ quantile from the $\mathcal{N}(0, 1)$ distribution. When pools are of unequal size, the distribution of $T = \sum_{i=1}^n Y_i$ is not negative binomial and $I(p)$ is not tractable. In this case, we instead calculate the observed information (Efron and Hinkley 1978) given by

$$J(p) = -\frac{\partial^2 l(p|\mathbf{y})}{\partial p^2} = \sum_{i=1}^n \frac{s_{iy_i}(1-p)^{s_{iy_i}-2} \{s_{iy_i}-1+(1-p)^{s_{iy_i}}\}}{\{1-(1-p)^{s_{iy_i}}\}^2} + \sum_{i=1}^n \sum_{j=1}^{y_i-1} \frac{s_{ij}}{(1-p)^2}, \quad (3.1)$$

and use $\hat{p} \pm z_{\gamma/2} J(\hat{p})^{-1/2}$ instead.

The score interval for p arises from the fact that the score function in (2.2), when viewed as random, is approximately normal with zero mean and variance $I(p)$ when n is large. Thus, when pools are of equal size, the solutions to

$$\frac{S(p|\mathbf{y})}{\sqrt{I(p)}} = \pm z_{\gamma/2}$$

delimit an approximate $100(1 - \gamma)$ percent confidence interval. When pools are of unequal size, we replace $I(p)$ with $J(p)$ in (3.1). The likelihood ratio (LR) interval is based on the fact that $2\{l(\hat{p}|\mathbf{y}) - l(p|\mathbf{y})\}$ is distributed approximately as a central χ_1^2 random variable when n is large. Thus, the solutions to

$$2\{l(\hat{p}|\mathbf{y}) - l(p|\mathbf{y})\} = 2 \sum_{i=1}^n \left[\log \left\{ \frac{1 - (1 - \hat{p})^{s_{i1}}}{1 - (1 - p)^{s_{i1}}} \right\} - \sum_{j=1}^{y_i-1} s_{ij} \log \left\{ \frac{1 - p}{1 - \hat{p}} \right\} \right] = \chi_{1,\gamma}^2$$

form an approximate $100(1 - \gamma)$ percent confidence interval for p , where $\chi_{1,\gamma}^2$ is the upper γ quantile of the χ_1^2 distribution.

3.2. An Exact Interval

When equal pool sizes s are used, $T \sim \text{nb}(n, \theta)$, where $\theta = 1 - (1 - p)^s$. Using the relationship between the negative binomial distribution and the incomplete beta function, Lui (1995) derived an exact interval for θ . The lower and upper confidence limits are $\theta_L =$

$B_{1-\gamma/2, n, t-n+1}$ and $\theta_U = B_{\gamma/2, n, t-n}$, respectively, where $t = \sum_{i=1}^n y_i$ and $B_{\gamma, a, b}$ denotes the upper γ quantile of the two-parameter beta(a, b) distribution. Thus, an exact $100(1 - \gamma)$ percent confidence interval for p can be obtained by suitably transforming the endpoints of the θ interval; i.e., $p_L = 1 - (1 - B_{1-\gamma/2, n, t-n+1})^{1/s}$ and $p_U = 1 - (1 - B_{\gamma/2, n, t-n})^{1/s}$. To the best of our knowledge, exact intervals for p are not available when pool sizes are unequal. This is in stark contrast to the binomial case, as demonstrated by Hepworth (1996).

3.3. Discussion

There are a myriad of other large sample approaches one could use for interval estimation. We did examine variance stabilized and suitably transformed logit-type intervals, used by Tebbs and Bilder (2004) and Hepworth (2005), respectively, in the binomial problem. Both, surprisingly, were not better than the Wald interval. When unequal pool sizes are used, Hepworth (2005) showed that the LR interval outperformed the standard score interval, while, in the equal pool size case, Tebbs and Bilder (2004) showed that the score interval was preferred in terms of coverage probability.

The lower limit of the Wald and score intervals can be negative when n is small and/or when \hat{p} is close to 0; on the other hand, LR endpoints cannot be negative. All large sample intervals have complications when $\hat{p} = 1$; i.e., every pool tests positive. In this case, $I(\hat{p})$ is undefined and the Wald interval cannot be computed; additionally, when equal sized pools are used, the score function is zero, and the upper limit for the LR interval cannot be computed. Observing $\hat{p} = 1$ would not be likely in practice, unless n is small and/or the pool size is inappropriately large. However, when it does occur, we set the Wald interval equal to the uninformative (0, 1) and make similar adjustments for the score and LR intervals.

3.4. Comparisons

We first consider the equal pool size case. For each of the confidence intervals (Wald, score, LR, and exact), the true coverage probability, for a given prevalence p , waiting parameter n , and pool size s , can be approximated in the same manner as the bias and MSE were in Section 2.3. Specifically, the exact coverage probability is approximated by

$$C(p, n, s) = \sum_{t=n}^{t^*} \mathcal{I}(\hat{p}) \binom{t-1}{n-1} \{1 - (1-p)^s\}^n \{(i-p)^s\}^{t-n},$$

where $\mathcal{I}(\hat{p}) = 1$ if the interval contains p , $\mathcal{I}(\hat{p}) = 0$ otherwise, and t^* satisfies $\text{pr}(T \leq t^*) \geq 1 - \nu$, where we take $\nu = 0.000001$. The exact mean length is approximated by using the same expression as $C(p, n, s)$, except that $D(\hat{p})$ replaces $\mathcal{I}(\hat{p})$, where $D(\hat{p})$ represents the length of the interval for a particular realization of \hat{p} .

Figure 4 displays the exact coverage probabilities for the four intervals when $\gamma = 0.05$, $s = 20$ (common pool size), $n = 1, 5$, and 100 , and values of p ranging from 0.00005 to 0.10 by increments of 0.00005 . The distinct jagged pattern in the plots is due to the discrete nature of the negative binomial distribution. For example, if $n = 1$, $t = 5$, and $s = 20$, $\mathcal{I}(\hat{p})$ for the Wald interval switches from 1 to 0 when p becomes greater than 0.0327 , which, in turn, corresponds to a drop in the coverage probability function. When $n = 1$, we find it unusual that the Wald and score intervals are somewhat centered around the nominal level, although the variability increases as the prevalence does. The LR interval is generally conservative as long as the prevalence is not small. The same general patterns hold for the $n = 5$ case, and our depiction when $n = 100$ is visual evidence that the asymptotic properties are taking hold.

To accompany Figure 4, we also calculated the exact mean length for the four intervals with equal pool sizes (results available from the authors). The exact interval, not surprisingly, possesses the greatest length, but it is generally not too much larger than the large sample intervals. The latter intervals, in turn, are comparable, with the Wald interval being slightly shorter on average. We also constructed plots of the distal and mesial non-coverage probabilities corresponding to Figure 4; these plots are available from the authors. Distal (mesial) non-coverage measures the probability content for all values of \mathbf{y} in the support producing intervals that miss p on the left (right) side. Some would argue that a desirable interval possesses symmetry in the non-coverage tails; i.e., that the distal and mesial non-coverage probabilities are approximately $\gamma/2$ (Newcombe 1998). Not surprisingly, for $n = 1$ and $n = 5$, mesial non-coverage almost exceeds distal non-coverage. In addition, for small n , it should be noted that the Wald and score lower endpoints are often negative when p is close to zero. In this situation, suitable adjustment may be warranted (e.g., use an exact interval, construct a one-sided Wald interval, etc.). When $n = 100$, the large sample intervals' non-coverage is nearly perfectly symmetrical.

For unequal pool sizes, we used simulation to compare the large sample intervals for $\gamma = 0.05$ using different n and p . The results in Table 1 are based on 10,000 Monte Carlo data sets so that the margin of error (assuming a 99 percent confidence level) is 0.006 . Pool sizes were generated from a discrete uniform distribution from 20 to 50, 10 to 30, and 5 to 15, for $p = 0.001$, $p = 0.01$, and $p = 0.1$, respectively. Perhaps most interesting is that, in terms of coverage probability, the large sample intervals are all quite good for small n , even when $n = 1$. The LR interval is notably conservative when $n = 1$, but this behavior diminishes for larger n . When $n = 20$, the estimated coverages for all procedures are all mostly within the margin of Monte Carlo error. The performance of the intervals is, of course, dependent on the pool size, and inappropriately large pool sizes can lead to very bad results. For example,

when $n = 5$ and $p = 0.1$, using $50 \leq s_{ij} \leq 100$ produced an average coverage probability of greater than 0.99 for all three methods. This is likely due to the fact that $\hat{p} = 1$ with high probability when the pool size is large (this occurs in 99.5 percent of the simulated confidence intervals for the setting just described). However, using pool sizes smaller than those seen in Table 1 does not greatly alter the procedures' coverage properties.

4. APPLICATION TO WNV PREVALENCE ESTIMATION

Rutledge et al. (2003) describe a public health study involving WNV surveillance. Immediately following a 2001 outbreak in Jefferson County, Florida, researchers set out to estimate the WNV prevalence in the mosquito population. Pools of mosquitos, formed from 11,948 individuals captured, were tested for the presence of WNV using reverse-transcription polymerase chain reaction assays, which resulted in 14 WNV positive pools. Variable pool sizes were used in the study, but nearly all of the pool sizes were less than 50. The study described by the authors was the first documented field study of WNV vectored by the North American mosquito, *Culex nigripalpus* Theobald, which is also believed to vector the St. Louis and eastern equine encephalitis viruses. WNV had been detected in Florida in early 2001; later that year, 12 human cases of WNV meningoencephalitis had been confirmed.

Inverse sampling was not used by Rutledge et al. (2003), but we use this investigation as a basis to illustrate our methodology. The authors only presented the complete results from the pools that tested positive; based on these data, the estimated prevalence was reported to be approximately 0.005 (likely an upper bound because some negative pools were ignored). We henceforth assume that 0.005 is the true value of p and simulate data sets for waiting parameters $n = 1, 5, 10$, and 20, using equal and unequal pool sizes. In the unequal pool size case, the authors reported using various sizes ranging mostly from 15 to 50 mosquitos. We thus use $15 \leq s_{ij} \leq 50$ when considering those methods based on unequal pool sizes; we use $s = 50$ in the equal pool size case.

Table 2 provides the averaged point estimates (with estimated standard errors in parentheses), based on 10,000 Monte Carlo data sets. For equal pool sizes, we present results for the MLE and the three adjusted estimators \hat{p}_α , \hat{p}_β , and \hat{p}_C using $p_0 = 0.01$ as an upper bound. The jackknife estimator is not reported because of its poor performance. The MLE from using unequal pool sizes is also reported; pool sizes s_{ij} were simulated from a $\mathcal{D} \mathcal{U}(15, 50)$ distribution, endpoints consistent with Rutledge et al. (2003). The results from Table 2 reaffirm our discovery that the adjusted estimators in the equal pool size case are preferred, especially when n is small. For $n = 1$, the MLE is exorbitantly biased in both the equal and unequal pool size cases. For larger n , all (averaged) estimates are generally on target, with standard errors generally being larger for unequal pool sizes.

Using the same simulated data, we computed 10,000 confidence intervals for p using the likelihood-based and exact methods from Section 3 and have reported in Table 3 the average confidence limits (lower and upper). The Wald and score intervals are the only ones that can return negative lower limits; when $n = 1, 77$, and 100 percent of Wald and score lower limits, respectively, were negative. The averaged upper and lower limits are not symmetrical about the average value of \hat{p} , because we assign noninformative intervals when $\hat{p} = 1$ (e.g., about 15 percent of the data sets produced $\hat{p} = 1$ when $n = 1$). When this realization is not likely, we found that all interval procedures produce roughly the same averaged endpoints. Furthermore, using the actual (binomial) data reported in the article, a 95 percent logit transformed confidence interval, suggested by Hepworth (2005), was found to be (0.003, 0.009). Most of the averaged intervals in Table 3, excluding the $n = 1$ case, are consistent with this estimate.

5. DISCUSSION

We have presented a new way of viewing the usual pooled testing prevalence estimation problem, considering situations that employ inverse binomial sampling. As we have discussed, this type of data collection procedure might be reasonable when sampling and testing occurs sequentially and when positive results instigate the need for immediate analysis. The usual binomial model requires that a fixed number of pools be used. This requirement seems to be incongruous with those applications where it is potentially critical to have timely intervention. To disseminate our work, we have written user-friendly R programs to calculate all point and interval estimates in this paper; these are located at the R-Forge website <http://r-forge.r-project.org/> under the name `invbingroup`.

It is important to note that the methods proposed herein may not be appropriate when there is a delay between the acquisition of individuals, such as mosquitos, and the actual testing of them in pools. In some research investigations, it is not uncommon for individual specimens to be stored for long periods of time before testing begins. On the other hand, an anonymous reviewer has described how mosquito testing for WNV and for other ar-boviruses occurs almost weekly in nearly every state. In addition, continuous monitoring for pathogens in livestock, such as bovine viral disease, is now a requirement to maintain healthy herds (Peck 2006). In these and related situations, our methods are applicable and may be more natural to implement because of the sequential nature of the data collection process.

Within an inverse sampling context, we have investigated point and interval estimation for p , the probability of individual infection. Future research could involve finding improved techniques in each area. For example, in the point estimation problem, Hepworth and Watson (2009) describe an iterative correction strategy to nearly remove the bias in the binomial pooled testing MLE; it may be possible to implement similar corrections in the inverse binomial problem. There may also be benefits to considering other confidence interval procedures, such as the skew-corrected score interval (Hepworth 2005) or an interval based on the mid- p correction (Hepworth 2004). The development of multiple population methods, similar to those of McCann and Tebbs (2007), Tebbs and McCann (2007), and Biggerstaff (2008) in the binomial case, would also be worthwhile extensions.

Especially when n is small, so that action is desirable early on in the data collection process, one would naturally conjecture that the development of Bayesian estimators might be a useful pursuit. Bayesian techniques could also prove to be useful in applications where there are no observed positive pools. A manuscript is currently in progress that examines these questions. We are also working on developing regression methodology, similar to that proposed by Xie (2001) in the binomial case, that can be used to incorporate covariate information on individual subjects. For example, in testing for FMD, the prevalence may be affected by animal gender, species, type of farm, and the distance from an infected premises. Early results suggest that regression estimators possess good statistical properties when the waiting parameter n is large and the prevalence is moderate.

Acknowledgments

We are grateful to the Associate Editor and the anonymous referees for their insightful comments. Implementing their suggestions has greatly strengthened this paper. We would also like to thank Drs. Christopher Bilder, Ivo Foppa, and Charles Katholi for helpful discussion. This research is supported by Grant R01 AI067373 from the National Institute of Health.

References

- Anderson, I. Foot and Mouth Disease 2001: Lessons to Be Learned Inquiry. London: The Stationery Office; 2002. available at http://archive.cabinetoffice.gov.uk/fmd/fmd_report/report/index.htm
- Bhattacharyya G, Karandinos M, DeFoliart G. Point Estimates and Confidence Intervals for Infection Rates Using Pooled Organisms in Epidemiologic Studies. *American Journal of Epidemiology*. 1979; 109:124–131. [PubMed: 34322]
- Biggerstaff B. Confidence Intervals for the Difference of Two Proportions Estimated from Pooled Samples. *Journal of Agricultural, Biological, and Environmental Statistics*. 2008; 13:478–496.
- Bilder C, Tebbs J. Empirical Bayesian Estimation of the Disease Transmission Probability in Multiple-Vector-Transfer Designs. *Biometrical Journal*. 2005; 47:502–516. [PubMed: 16161807]
- Block C, Hill J, McGee D. Relationship Between Late-Season Severity of Stewart’s Bacterial Wilt and Seed Infection in Maize. *Plant Disease*. 1999; 83:527–530.
- Burrows P. Improved Estimation of Pathogen Transmission Rates by Group Testing. *Phytopathology*. 1987; 77:363–365.
- Bush J, Phillimore P, Pless-Mulloli T, Thompson C. Carcass Disposal and Siting Controversy: Risk, Dialogue and Confrontation in the 2001 Foot-and-Mouth Outbreak. *Local Environment*. 2005; 10:649–664.
- Callahan J, Brown F, Osorio F, Sur J, Kramer E, Long G, Lubroth J, Ellis S, Shoulars K, Gaffney K, Rock D, Nelson W. Use of a Portable Real-Time Reverse Transcriptase–Polymerase Chain Reaction Assay for Rapid Detection of Foot-and-Mouth Disease Virus. *Journal of the American Veterinary Medical Association*. 2002; 220:1636–1642. [PubMed: 12051502]
- Chen P, Tebbs J, Bilder C. Group Testing Regression Models with Fixed and Random Effects. *Biometrics*. 2009; 65:1270–1278. [PubMed: 19210734]
- Chick S. Bayesian Models for Limiting Dilution Assay and Group Test Data. *Biometrics*. 1996; 52:1055–1062.
- Colón S, Patil G, Taillie C. Estimating Prevalence Using Composites. *Environmental and Ecological Statistics*. 2001; 8:213–236.
- Dorfman R. The Detection of Defective Members of Large Populations. *Annals of Mathematical Statistics*. 1943; 14:436–440.
- Efron B, Hinkley D. Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information. *Biometrika*. 1978; 65:457–482.
- Foppa I, Evans C, Wozniak A, Wills W. Mosquito Fauna and Arbovirus Surveillance in a Coastal Mississippi Community After Hurricane Katrina. *Journal of the American Mosquito Control Association*. 2007; 22:229–232. [PubMed: 17847860]
- Godsey M, Nasci R, Savage H, Aspen S, Kind R, Powers A, Burkhalter K, Colton L, Charnetzky D, Lasater S, Taylor V, Palmisano C. West Nile Virus—Infected Mosquitoes, Louisiana, 2002. *Emerging Infectious Diseases*. 2005; 9:1399–1404. [PubMed: 16229769]
- Haldane J. On a Method of Estimating Frequencies. *Biometrika*. 1945; 33:222–225. [PubMed: 21006837]
- Hepworth G. Exact Confidence Intervals for Proportions Estimated by Group Testing. *Biometrics*. 1996; 52:1134–1146.
- Hepworth G. Mid-p Confidence Intervals Based on the Likelihood Ratio for Proportions Estimated by Group Testing. *Australian and New Zealand Journal of Statistics*. 2004; 46:391–405.
- Hepworth G. Confidence Intervals for Proportions Estimated by Group Testing With Groups of Unequal Size. *Journal of Agricultural, Biological, and Environmental Statistics*. 2005; 10:478–497.
- Hepworth G, Watson R. Debiased Estimation of Proportions in Group Testing. *Applied Statistics*. 2009; 58:105–121.
- Huang X, Tebbs J. On Latent-Variable Model Misspecification in Structural Measurement Error Models for Binary Response. *Biometrics*. 2009; 65:710–718. [PubMed: 18945265]
- Hughes-Oliver, J. Pooling Experiments for Blood Screening and Drug Discovery. In: Dean, A.; Lewis, S., editors. *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*. New York: Springer; 2006.

- Katholi, C. Department of Biostatistics Technical Report. University of Alabama; Birmingham: 2006. Estimation of Prevalence by Pool Screening With Equal Sized Pools and a Negative Binomial Sampling Model. available at <http://images.main.uab.edu/isoph/BST/BST2006technicalReport.pdf>
- Katholi C, Unnasch T. Important Experimental Parameters for Determining Infection Rates in Arthropod Vectors Using Pool Screening Approaches. *American Journal of Tropical Medicine and Hygiene*. 2006; 74:779–785. [PubMed: 16687680]
- Koda E. Could Foot and Mouth Disease be a Biological Warfare Incident? *Military Medicine*. 2002; 167:91–92. [PubMed: 11873548]
- Lanciotti R, Kerst A. Nucleic Acid Sequence-Based Amplification Assays for Rapid Detection of West Nile and St. Louis Encephalitis Viruses. *Journal of Clinical Microbiology*. 2001; 39:4506–4513. [PubMed: 11724870]
- Lui K. Confidence Limits for the Population Prevalence Rate Based on the Negative Binomial Distribution. *Statistics in Medicine*. 1995; 14:1471–1477. [PubMed: 7481185]
- McCann M, Tebbs J. Pairwise Comparisons for Proportions Estimated by Pooled Testing. *Journal of Statistical Planning and Inference*. 2007; 137:1278–1290.
- Newcombe R. Two Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*. 1998; 17:857–872. [PubMed: 9595616]
- Peck C. Going after BVD. *Beef*. 2006; 42:34–44.
- Reid S, Ferris N, Hutchings G, Zhang Z, Belsham G, Alexandersen S. Detection of All Seven Serotypes of Foot-and-Mouth Disease Virus by Real-Time, Fluorogenic Reverse Transcription Polymerase Chain Reaction Assay. *Journal of Virological Methods*. 2002; 105:67–80. [PubMed: 12176143]
- Rutledge C, Day J, Lorde C, Stark L, Tabachnick W. West Nile Virus Infection Rates in *Culex nigripalpus* (Diptera: Culicidae) Do Not Reflect Transmission Rates in Florida. *Journal of Medical Entomology*. 2003; 3:253–258. [PubMed: 12943101]
- Solomon T, Ooi M, Beasley D, Mallewa M. West Nile Encephalitis. *British Medical Journal*. 2003; 326:865–869. [PubMed: 12702624]
- Swallow W. Group Testing for Estimating Infection Rates and Probabilities of Disease Transmission. *Phytopathology*. 1985; 75:882–889.
- Tebbs J, Bilder C. Confidence Interval Procedures for the Probability of Disease Transmission in Multiple-Vector Transfer Designs. *Journal of Agricultural, Biological, and Environmental Statistics*. 2004; 9:75–90.
- Tebbs J, McCann M. Large-Sample Hypothesis Tests for Stratified Group-Testing Data. *Journal of Agricultural, Biological, and Environmental Statistics*. 2007; 12:534–551.
- Tu X, Litvak E, Pagano M. On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika*. 1995; 82:289–297.
- Turechek W, Madden L. A Generalized Linear Modeling Approach for Characterizing Disease Incidence in a Spatial Hierarchy. *Phytopathology*. 2003; 93:458–466. [PubMed: 18944361]
- Vansteelandt S, Goetghebeur E, Verstraeten T. Regression Models for Disease Prevalence With Diagnostic Tests on Pools of Serum Samples. *Biometrics*. 2000; 56:1126–1133. [PubMed: 11129470]
- Xie M. Regression Analysis of Group Testing Samples. *Statistics in Medicine*. 2001; 20:1957–1969. [PubMed: 11427952]
- Zilinskas R. Iraq's Biological Weapons. *Journal of the American Medical Association*. 1997; 278:418–424. [PubMed: 9244334]

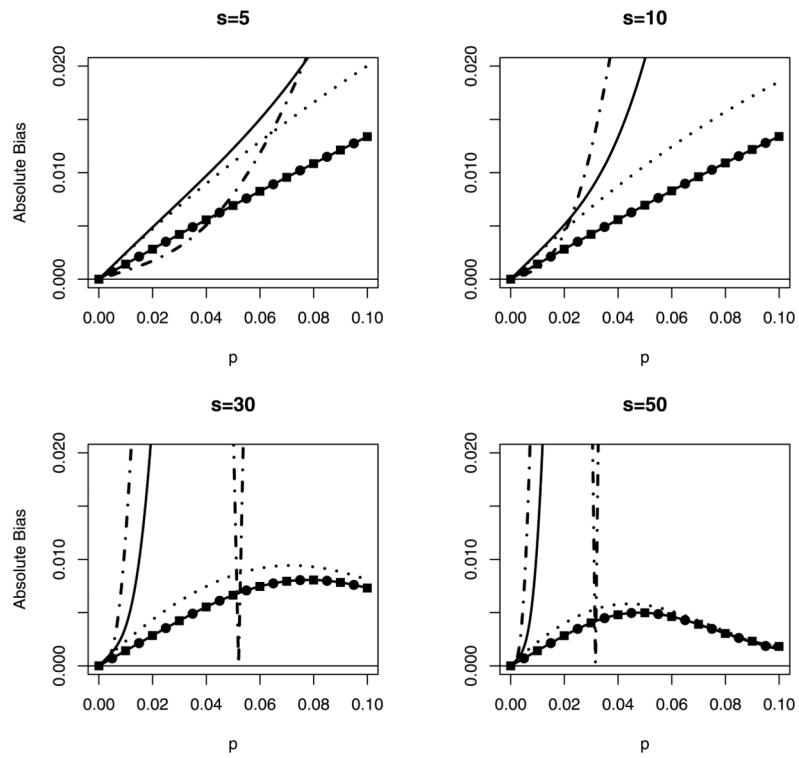


Figure 1. Absolute bias for \hat{p} (—), \hat{p}_α (⋯), \hat{p}_β (—●—), \hat{p}_C (—■—), and \hat{p}_J (---) for different pool sizes s when $n = 5$. Optimal values of α , β , α_C , and β_C are used. $B(\hat{p}_\beta)$ and $B(\hat{p}_C)$ are indistinguishable.

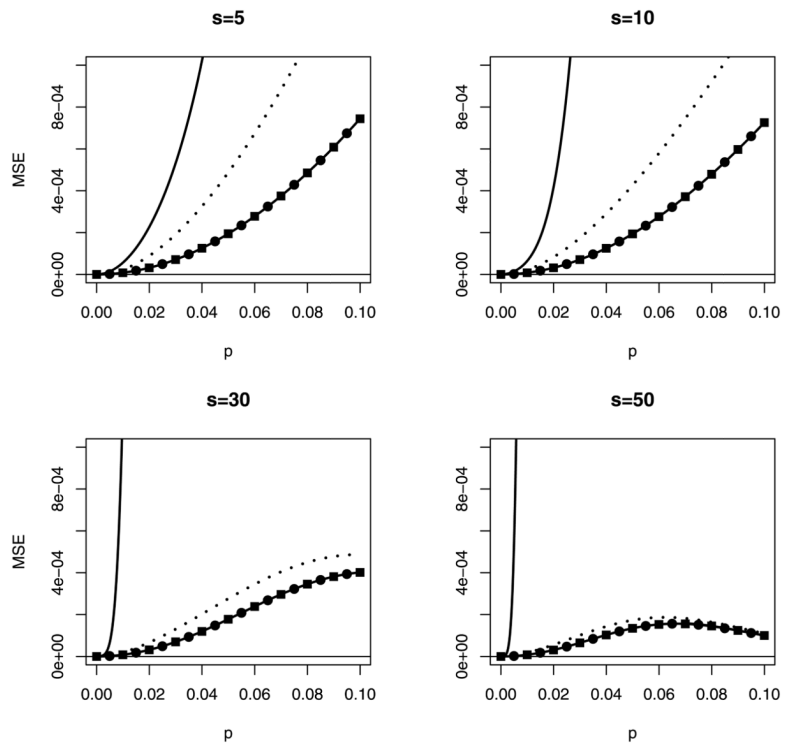


Figure 2. MSE for \hat{p} (—), \hat{p}_α (⋯), \hat{p}_β (—●—), and \hat{p}_C (—■—) for different pool sizes s when $n = 5$. Optimal values of α , β , α_C , and β_C are used. $MSE(\hat{p}_\beta)$ and $MSE(\hat{p}_C)$ are indistinguishable. The jackknife estimator is not depicted.

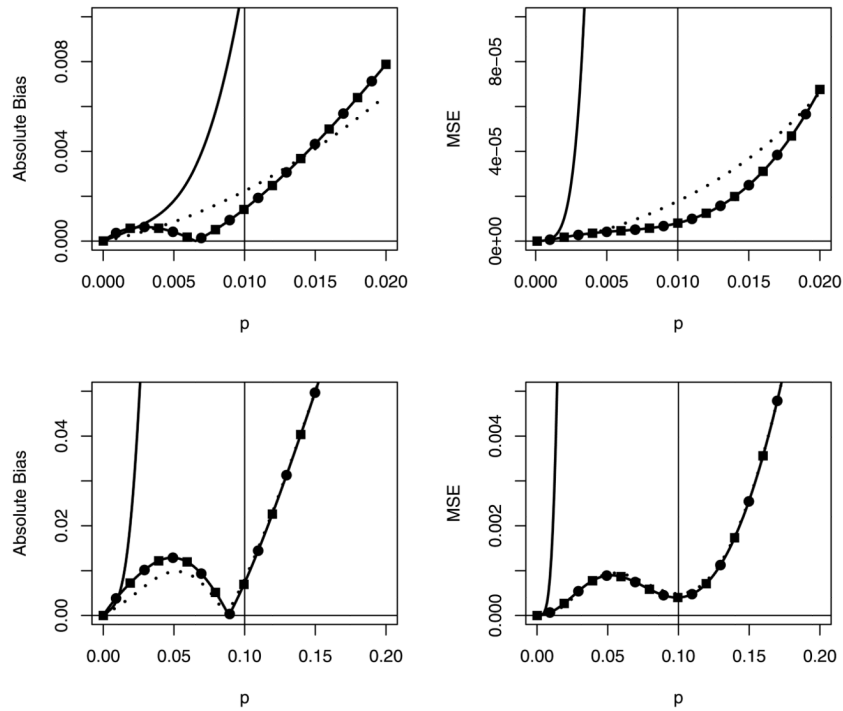


Figure 3. Absolute bias and MSE of \hat{p} (—), \hat{p}_α (⋯), \hat{p}_β (—●—), and \hat{p}_C (—■—). Top: $n = 5$, $s = 50$, and optimal values of α , β , α_C , and β_C are based on the upper bound $p_0 = 0.01$. Bottom: $n = 5$, $s = 30$, and optimal values of α , β , α_C , and β_C are based on the upper bound $p_0 = 0.10$. In both cases, \hat{p}_β and \hat{p}_C are indistinguishable.

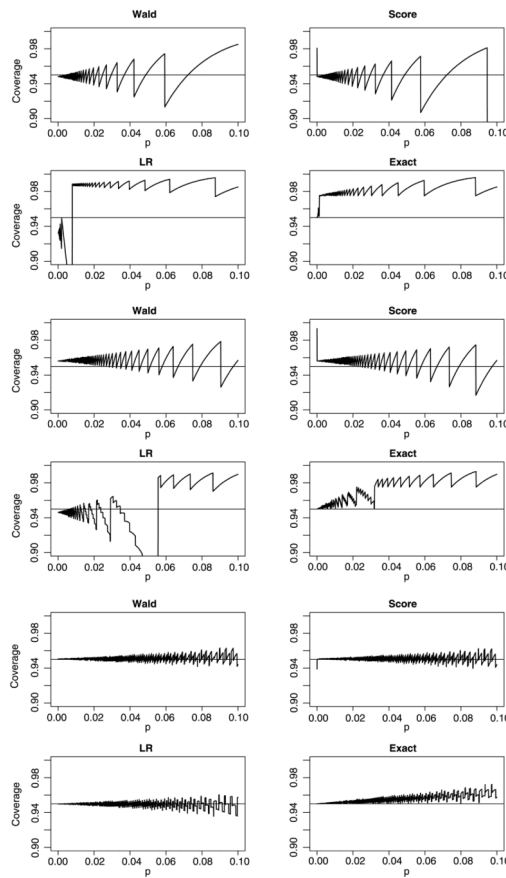


Figure 4. Exact coverage probability for 95 percent Wald, score, LR, and exact confidence intervals when $s = 20$. Top four: $n = 1$; Middle four: $n = 5$; Bottom four: $n = 100$. Within each block of four, the intervals are arranged as Wald (upper left), score (upper right), LR (lower left), and exact (lower right).

Table 1

Coverage probabilities for likelihood-based confidence intervals when pool sizes are unequal. The nominal level is $1 - \gamma = 0.95$. Coverage probabilities are approximated based on 10,000 Monte Carlo data sets. The margin of error, assuming a 99 percent confidence level, is 0.006.

# positive pools	p	Pool sizes	Wald	Score	LR
$n = 1$	0.001	$20 \leq s_{ij} \leq 50$	0.951	0.948	0.967
	0.01	$10 \leq s_{ij} \leq 30$	0.950	0.951	0.990
	0.1	$5 \leq s_{ij} \leq 15$	0.951	0.957	0.988
$n = 5$	0.001	$20 \leq s_{ij} \leq 50$	0.956	0.956	0.945
	0.01	$10 \leq s_{ij} \leq 30$	0.959	0.960	0.946
$n = 20$	0.1	$5 \leq s_{ij} \leq 15$	0.958	0.958	0.982
		$20 \leq s_{ij} \leq 50$	0.950	0.951	0.953
	0.01	$10 \leq s_{ij} \leq 30$	0.950	0.954	0.948
		$5 \leq s_{ij} \leq 15$	0.951	0.957	0.949

Table 2

Mean prevalence estimates for WNV infection using the data from Rutledge et al. (2003). Equal pool size estimates use $s = 50$. Unequal pool size estimates use $15 \leq s_{ij} \leq 50$. Estimates are based on 10,000 Monte Carlo data sets assuming $p = 0.005$. Estimated standard errors are in parentheses.

n	Equal pool sizes			Unequal pool sizes		
	\hat{p}	\hat{p}_a	\hat{p}_b	\hat{p}_c	\hat{p}	
1	0.229 (0.064)	0.005 (0.006)	0.007 (0.005)	0.007 (0.005)	0.163 (0.060)	
5	0.007 (0.015)	0.004 (0.004)	0.005 (0.006)	0.005 (0.006)	0.006 (0.010)	
10	0.006 (0.004)	0.005 (0.004)	0.005 (0.003)	0.005 (0.003)	0.006 (0.004)	
20	0.005 (0.003)	0.005 (0.003)	0.005 (0.003)	0.005 (0.003)	0.005 (0.003)	

Table 3

Mean confidence interval limits for the prevalence of WNV infection using negative binomial type sampling. Equal pool size intervals use $s = 50$. Unequal pool size intervals use $15 \leq s_{ij} \leq 50$. Lower and upper limits are averaged over 10,000 Monte Carlo confidence intervals constructed assuming $p = 0.005$.

Equal pool sizes				
n	W	S	LR	E
1	(-0.005, 0.239)	(-0.011, 0.239)	(0.001, 0.247)	(0.001, 0.248)
5	(0.001, 0.012)	(0.001, 0.012)	(0.002, 0.014)	(0.002, 0.014)
10	(0.002, 0.009)	(0.002, 0.009)	(0.003, 0.010)	(0.003, 0.010)
20	(0.003, 0.008)	(0.003, 0.008)	(0.003, 0.008)	(0.003, 0.008)
Unequal pool sizes				
n	W	S	LR	E
1	(-0.007, 0.177)	(0.000, 0.177)	(0.001, 0.188)	-
5	(0.001, 0.012)	(0.001, 0.012)	(0.002, 0.014)	-
10	(0.002, 0.009)	(0.002, 0.009)	(0.003, 0.010)	-
20	(0.003, 0.008)	(0.003, 0.008)	(0.003, 0.008)	-