Published in final edited form as:

Biom J. 2011 February ; 53(1): 40–56. doi:10.1002/bimj.201000148.

Bayesian inference for disease prevalence using negative binomial group testing

Nicholas A. Pritchard¹ and Joshua M. Tebbs^{2,*}

¹ Department of Mathematics and Statistics, Coastal Carolina University, Conway, SC 29528, USA

² Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

Abstract

Group testing, also known as pooled testing, and inverse sampling are both widely used methods of data collection when the goal is to estimate a small proportion. Taking a Bayesian approach, we consider the new problem of estimating disease prevalence from group testing when inverse (negative binomial) sampling is used. Using different distributions to incorporate prior knowledge of disease incidence and different loss functions, we derive closed form expressions for posterior distributions and resulting point and credible interval estimators. We then evaluate our new estimators, on Bayesian and classical grounds, and apply our methods to a West Nile Virus data set.

Keywords

Bayesian estimation; Inverse sampling; Maximum likelihood; Pooled testing; West Nile Virus

1 Introduction

Accurate estimation of pathogen prevalence in insect vectors is critical when deciding proper allocation of costly and environmentally sensitive interventions. If the prevalence is small, it is far more cost effective to test insects in pools than to test them individually. This type of data collection is known as group (pooled) testing, since binary measurements (i.e. infected/not infected) are made on pools of composited individuals (e.g. insects, etc.). The intriguing notion of drawing inference about individual infection prevalence levels, while never necessarily observing the individual statuses themselves, has prompted a substantive body of statistical research in this area, and testing individuals in pools through group testing has been applied to problems in multiple areas, including genetics, drug discovery, agriculture, and infectious disease screening in humans (see, e.g. Gastwirth, 2000; Remlinger et al., 2006; Hughes-Oliver, 2006; Katholi and Unnasch, 2006; Peck, 2006; Biggerstaff, 2008; Bilder et al., 2010). This paper is motivated in part by our recent exposure to research in pooled mosquito testing for West Nile Virus (WNV).

During the active mosquito season, many municipalities and counties have mosquito control and surveillance districts that perform continuous mosquito collection and testing. For these

districts, pathogen detection usually serves as a trigger for disease intervention efforts (e.g. pesticide applications, public health outreach programs, etc.). Though the presence of a single positive WNV pool can be enough to trigger a response, it is also important to estimate pathogen prevalence since a higher prevalence may precede a significant human disease outbreak. For example, in 2008 alone, there were 1356 documented cases of WNV in multiple outbreaks across the United States, resulting in 44 fatalities (Centers for Disease Control, 2010). An accurate assessment of the infection level can help to limit the severity of an outbreak, as countermeasures can be taken to prevent further infection.

In the group testing literature, prevalence estimation and individual identification (classification) have been separate goals. We are interested in the former. Previous work has centered around point and interval estimation (Swallow, 1985; Gastwirth and Johnson, 1994; Tebbs and Bilder, 2004; Hepworth, 1996, 2004, 2005; Hepworth and Watson, 2009), group size selection (Hughes-Oliver and Swallow, 1994; McCann and Tebbs, 2007), and the use of Bayesian methods (Chick, 1996; Bilder and Tebbs, 2005; Hanson et al., 2006). What unifies this large body of work is a reliance on the binomial model, that is, individual (latent) statuses are assumed to be independent and identically distributed Bernoulli random variables with mean p , and a fixed number of pools is observed.

Pritchard and Tebbs (2010) have suggested that inverse (negative) binomial pooled sampling may be preferred when the prevalence p is known to be small, when sampling and testing occurs sequentially, and when positive pool results instigate the need for immediate analysis, such as in WNV screening. Under this model, unlike the binomial, the number of positive pools to be observed is fixed a priori, and testing stops when this number is reached. Pritchard and Tebbs (2010) use maximum likelihood as a basis to develop classical point and interval estimators for p under inverse pooled sampling. In this paper, we have the same goals, but we instead take a Bayesian approach. This is sensible, and possibly preferred, as in the context of WNV screening; for example, known information about p can be incorporated seamlessly into a prior model. In addition, the classical methods presented in Pritchard and Tebbs (2010) have serious shortcomings in small samples when p is small. In particular, the maximum likelihood estimator (MLE) of p can be severely biased, and interval estimates can extend outside the parameter space. These are unfortunate characteristics, especially because group testing and inverse sampling are most beneficial when p is small. As we demonstrate, taking a Bayesian approach can provide estimators with better properties in small samples, leading to more reliable inference in disease surveillance settings.

In Section 2, we state assumptions and briefly review the classical methodology in Pritchard and Tebbs (2010). In Section 3, we present Bayesian and empirical Bayesian estimators obtained from a variety of prior models. Credible set estimation is also discussed. In Section 4, we compare our Bayesian estimators on both frequentist and Bayesian grounds. In Section 5, we apply our methods to a real data set in WNV estimation. In Section 6, we conclude with a brief summary discussion.

2 Preliminaries

As in Pritchard and Tebbs (2010), we assume that individual (e.g. mosquito) 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. By “positive pool,” we mean that the pool contains at least one infected individual. We additionally assume that the probability of testing error is zero, that is, pools are not misclassified. For further discussion on these assumptions in the context of arbovirus surveillance, see Katholi

and Unnasch (2006). Extending our results to handle misclassified pools is discussed briefly in Section 6.

Suppose that $Y_i = y_i$ pools are tested to find the i th positive pool, $i = 1, 2, \dots, n$, where n is fixed. For $j = 1, 2, \dots, y_i$, let s_{ij} denote the number of individual insects in the j th pool in the sequence of pools tested to obtain the i th observation Y_i . The log-likelihood function 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 $\mathbf{y} = (y_1, y_2, \dots, y_n)$. The MLE for p , denoted by $\hat{p}$, is found by solving

$$S(p|\mathbf{y}) = \frac{\partial l(p|\mathbf{y})}{\partial p} = \sum_{i=1}^n \frac{s_{iy_i}(1-p)^{s_{iy_i}-1}}{1-(1-p)^{s_{iy_i}}} - \sum_{i=1}^n \sum_{j=1}^{y_i-1} \frac{s_{ij}}{1-p} = 0$$

for p . When pools are of equal size, that is, when $s_{ij} = s$, for all i and j , a closed form expression for the MLE exists and is given by

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

where $T = \sum_{i=1}^n Y_i$. As shown in Pritchard and Tebbs (2010), the MLE $\hat{p}$ can have extremely high (positive) bias when n is small.

3 Bayesian estimation

Unless otherwise stated, we henceforth assume that a common pool size $s > 1$ is used for all pools. In the group testing literature, using common pool sizes has been recommended for disease transmission studies involving insects (Swallow, 1985), likely for reasons related to the ease of implementation and testing. We make this assumption for no reason other than the desire to present closed-form expressions for posterior distributions and for all point and interval estimators. Using unequal pool sizes is not prohibitive, but requires computational methods to simulate the necessary posterior distributions; see Section 6.

3.1 Classical Bayes

When pool sizes are equal, the sufficient statistic $T = \sum_{i=1}^n Y_i$ follows a negative binomial distribution with waiting parameter n and success probability $\theta = 1 - (1-p)^s$, where s is the common pool size, so that the probability mass function of T , conditional on $P = p$, is

$$f_{T|P}(t|p) = \binom{t-1}{n-1} \{1 - (1-p)^s\}^n (1-p)^{s(t-n)}.$$

For all distributions involving T in this paper, it is understood that $t \in \mathcal{T} \equiv \{t: t=n, n+1, \dots\}$. Similarly, $0 < p < 1$ in all distributions.

The use of Bayesian negative binomial group testing has been envisioned previously. In a technical report, Katholi (2006) focuses mostly on classical inference but suggests a Bayesian approach using a Jeffreys prior, which we derive to be

$$f_p(p) \propto \frac{s(1-p)^{s/2-1}}{1-(1-p)^s}.$$

This choice of prior may be appropriate when there is a genuine lack of knowledge about the prevalence p . Even though $f_p(p)$ is improper, we show in Appendix A.1 that the posterior distribution is proper and is given by

$$f_{p|T}(p|t) = \frac{s\Gamma(t+1/2)}{\Gamma(n)\Gamma(t-n+1/2)} \{1-(1-p)^s\}^{n-1} (1-p)^{s(t-n+1/2)-1}, \quad (2)$$

where $\Gamma(\cdot)$ is the usual gamma function. With $f_{p|T}(p|t)$ and a given loss function, say $L(p, a)$, the Bayes estimate of p is the value of a that minimizes

$$E\{L(P, a) | T=t\} = \int_0^1 L(p, a) f_{p|T}(p|t) dp.$$

Unless otherwise stated, we assume that $L(p, a) = (p-a)^2$, so that the Bayes estimate of p is $E(P|T=t)$, the posterior mean. In Sections 3.2 and 3.3, we consider other loss functions; see also Section 6. For the Jeffreys prior, we show in Appendix A.1 that

$$\hat{p}_J \equiv E(P|T=t) = \int_0^1 p f_{p|T}(p|t) dp = 1 - \frac{\Gamma(t+1/2)\Gamma(t-n+1/2)}{\Gamma(t-n+1/2)\Gamma(t+1/2)}. \quad (3)$$

Unfortunately, $\hat{p}_J$ can perform poorly when p is small (see Section 4). Therefore, it is of interest to consider more informative priors.

When there is a priori knowledge that the pathogen prevalence p is small, a more sensible class of prior distributions is the beta(1, β) class; i.e.

$$f_p(p|\beta) = \beta(1-p)^{\beta-1}.$$

Note that taking $\beta = 1$ produces a $\mathcal{U}(0, 1)$ distribution, which is also uninformative. However, when $\beta > 1$, $f_p(p|\beta)$ is concentrated towards $p = 0$ and is highly right skewed. For this class, because $E_P(P) = 1/(1+\beta)$, taking β to be large is consistent with knowledge that p is small. Using a beta(1, β) prior, the posterior distribution is

$$f_{p|T}(p|t) = \frac{s\Gamma(t+\beta/s+1)}{\Gamma(n+1)\Gamma(t-n+\beta/s)} \{1-(1-p)^s\}^n (1-p)^{s(t-n)+\beta-1} \quad (4)$$

and the Bayes estimate is

$$\hat{p}_B \equiv E(P|T=t) = 1 - \frac{\Gamma(t+\beta/s+1)\Gamma(t-n+\beta/s+1/s)}{\Gamma(t-n+\beta/s)\Gamma(t+\beta/s+1/s+1)}. \quad (5)$$

Complete details are given in Appendix A.2.

While the $\text{beta}(1, \beta)$ class is flexible enough to encapsulate a large number of realistic situations involving low prevalence pathogens, an even more flexible class of priors is the two-parameter $\text{beta}(\alpha, \beta)$ class. For this choice, the two-parameter posterior distribution is

$$f_{p|T}(p|t) = c_{\text{BB}}^{-1} \{1 - (1-p)^s\}^n p^{\alpha-1} (1-p)^{s(t-n)+\beta-1}, \quad (6)$$

where the inverse normalizing constant

$$c_{\text{BB}} = \Gamma(\alpha) \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st - sn + sk + \beta)}{\Gamma(st - sn + sk + \alpha + \beta)}.$$

The Bayes posterior mean estimate also exists in closed form and is given by

$$\hat{p}_{\text{BB}} \equiv E(P|T=t) = c_{\text{BB}}^{-1} \Gamma(\alpha+1) \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st - sn + sk + \beta)}{\Gamma(st - sn + sk + \alpha + \beta + 1)}. \quad (7)$$

When $\alpha = 1$, (6) reduces to (4) and (7) reduces to (5), as expected. Complete details are given in Appendix A.2. However, it is worth noting that all calculations involving the two-parameter beta posterior are highly numerically unstable (due to the alternating sums), except when n is very small (e.g. $n = 1, 2$). We therefore avoid considering it further.

3.2 Indirect Bayes

Up until now, we have considered only the use of $L(p, a) = (p-a)^2$ when calculating Bayes estimators. A potential drawback with this loss function is that overestimation and underestimation of p are penalized equally. Clearly, this may not be desirable when p is close to zero, the situation in which group testing and inverse sampling are most useful. As an alternative, we consider a squared error loss function on $\theta = 1-(1-p)^s$; i.e. $L(\theta, a) = (\theta-a)^2$, while still choosing a prior for p . For this loss function, the problem of penalizing close to the zero boundary is generally avoided.

We now derive an estimator from using $L(\theta, a) = (\theta-a)^2$ and a $\text{beta}(1, \beta)$ prior for p . Starting with (4), note that $\theta = 1-(1-p)^s$ is a continuous, strictly increasing function of p over $(0, 1)$. Performing a univariate transformation allows us to write the conditional density of Θ , given $T = t$, as

$$f_{\Theta|T}(\theta|t, \beta) = \frac{\Gamma(t+\beta/s+1)}{\Gamma(n+1)\Gamma(t-n+\beta/s)} \theta^n (1-\theta)^{t-n+\beta/s-1},$$

for $0 < \theta < 1$. Because $f_{\Theta|T}(\theta|t, \beta)$ is the $\text{beta}(n+1, t-n+\beta/s)$ probability density function, it follows that

$$E(\Theta|T=t) = \frac{n+1}{t+\beta/s+1}.$$

Using the inverse transformation $p = 1 - (1 - \theta)^{1/s}$, we are able to create an “indirect estimate,” given by

$$\widehat{p}_I = 1 - \left(1 - \frac{n+1}{t+\beta/s+1}\right)^{1/s}. \quad (8)$$

We use the term “indirect” because $\widehat{p}_I$ is technically not a Bayes estimate. However, it is interesting that $\widehat{p}_I$ is similar to the classical “shift estimate” in Pritchard and Tebbs (2010); see Section 4.

3.3 Empirical Bayes

Using the classical or indirect Bayes estimates, $\widehat{p}_B$ or $\widehat{p}_I$, respectively, requires the investigator to elicit a value of the hyperparameter β . Consequently, a poor choice could shift the calculated posterior distribution away from the truth, thereby degrading inference, particularly when n is not large enough to dominate the prior. We, therefore, also investigate empirical Bayesian analogues of $\widehat{p}_B$ or $\widehat{p}_I$, using the observed data T to estimate β .

When using a beta(1, β) prior, we show in Appendix A.3 that the marginal distribution of T is given by

$$f_T(t|\beta) = \int_0^1 f_{T|P}(t|p) f_P(p|\beta) dp = \frac{n\beta \Gamma(t) \Gamma(t-n+\beta/s)}{s \Gamma(t-n+1) \Gamma(t+\beta/s+1)}. \quad (9)$$

Therefore, in the complete absence of a priori knowledge, there is nothing to prevent one from maximizing $f_T(t|\beta)$ to produce an estimate of β . This can be done by solving

$$\frac{\partial \log f_T(t|\beta)}{\partial \beta} = \beta^{-1} + s^{-1} \{\psi(t-n+\beta/s) - \psi(t+\beta/s+1)\} = 0,$$

for β , where $\Psi(\cdot)$ represents the digamma function. Let $\widehat{\beta}$ denote the value of β that maximizes $f_T(t|\beta)$; i.e. the MLE of β . Replacing β with $\widehat{\beta}$ in (5) gives the empirical Bayes estimate

$$\widehat{p}_B^{(E)} = 1 - \frac{\Gamma(t+\widehat{\beta}/s+1) \Gamma(t-n+\widehat{\beta}/s+1/s)}{\Gamma(t-n+\widehat{\beta}/s) \Gamma(t+\widehat{\beta}/s+1/s+1)}.$$

Replacing β with $\widehat{\beta}$ in (8) gives the empirical indirect Bayes estimate

$$\widehat{p}_I^{(E)} = 1 - \left(1 - \frac{n+1}{t+\widehat{\beta}/s+1}\right)^{1/s}.$$

In examining the properties of $\widehat{p}_B^{(E)}$ and $\widehat{p}_I^{(E)}$, we have found that the MLE of β is unique for all $t \in \mathcal{T}$ except when $t = n$, that is, when every pool tests positive. In this case, $f_T(t|\beta)$ is strictly decreasing, and an estimate of β cannot be found using the method described above. Of course, observing $t = n$ would be extremely unlikely if the pool size s is chosen sensibly, but when it does occur, we take $\widehat{\beta} = 1$; i.e. we use a $\mathcal{U}(0, 1)$ prior for P . This is not unreasonable, especially because our empirical evidence suggests that $\widehat{\beta}$ approaches unity as t approaches n from above.

3.4 Credible intervals

To estimate p from inverse (negative) binomial group testing, we have proposed five Bayesian point estimators: $\hat{p}_J$, $\hat{p}_B$, $\hat{p}_I$, $\hat{p}_B^{(E)}$, and $\hat{p}_I^{(E)}$. For a succinct summary of the estimators, and their origin of calculation, see Table 1. In Section 4, we compare these estimators among themselves and to the best classical estimator from Pritchard and Tebbs (2010). In this subsection, we present closed form formulas for posterior credible intervals. Complete details are found in Appendix A.4.

A $100(1-\alpha^*)$ percent credible interval for p is found by determining the values of p_L and p_U such that $\text{pr}(p_L < P < p_U | T = t) = 1 - \alpha^*$. For ease of exposition, we restrict attention to equal tail credible intervals so that p_L and p_U solve

$$\alpha^*/2 = \int_0^{p_L} f_{PT}(p|t) dp \quad \text{and} \quad \alpha^*/2 = \int_{p_U}^1 f_{PT}(p|t) dp.$$

When using a Jeffreys prior, closed form expressions are $p_L = 1 - (b_{\alpha^*/2, t-n+1/2, n})^{1/s}$ and $p_U = 1 - (b_{1-\alpha^*/2, t-n+1/2, n})^{1/s}$, where $b_{\gamma, a, b}$ represents the upper γ quantile of the beta(a, b) distribution. When using a beta(1, β) prior distribution, closed form expressions are $p_L = 1 - (b_{\alpha^*/2, t-n+\beta/s, n+1})^{1/s}$ and $p_U = 1 - (b_{1-\alpha^*/2, t-n+\beta/s, n+1})^{1/s}$. An empirical Bayes version of the latter credible interval results from replacing β with the marginal MLE $\hat{\beta}$.

In some cases, it may be preferred to construct a one-sided credible interval of the form $(0, p_U)$, particularly if the goal is to report an upper limit on the pathogen prevalence. This can be done by making the obvious modification above.

4 Comparisons

We now compare the proposed point estimators on both Bayesian and classical grounds, similarly to Bilder and Tebbs (2005) who compared Bayesian and non-Bayesian group testing estimators under the binomial model. We use Bayes risk to compare the Bayesian estimators. A classical comparison is made in terms of absolute bias and mean-squared error (MSE).

4.1 Bayesian comparison

When the loss function $L(p, a) = (p-a)^2$ is used, the Bayes risk for an estimator generically denoted by $\tilde{p}$ is $r(\tilde{p}) = E_P\{E_{T|p}(\tilde{p} - p)^2\}$. Note that $r(\hat{p}_J)$ cannot be computed because the Jeffreys prior is improper. For the beta(1, β) prior distribution, we show in Appendix A.5 that the Bayes risk of $\tilde{p}$ is

$$r(\tilde{p}) = \beta s^{-1} \sum_{t=n}^{\infty} \binom{t-1}{n-1} \left\{ (\tilde{p}-1)^2 B(t-n+\beta/s, n+1) + 2(\tilde{p}-1) B(t-n+\beta/s+1/s, n+1) + B(t-n+\beta/s+2/s, n+1) \right\},$$

where $B(\cdot, \cdot)$ is the beta function. The empirical Bayes risk for $\hat{p}_B^{(E)}$ and $\hat{p}_I^{(E)}$ is computed by replacing β with the marginal MLE $\hat{\beta}$ in $r(\tilde{p})$, with the same modifications as described in Section 3.3 if all pools test positive. For practical purposes, when computing $r(\tilde{p})$ for any estimator, we replace the infinite upper summand index with the finite value t^* , where t^* satisfies $\text{pr}(T \leq t^*) \geq 1 - v$ for v small. We take $v = 0.000001$ throughout, making our approximation extremely close to the true value of $r(\tilde{p})$. For purposes of comparison, we define the relative Bayes risk (RBR) of the estimator $\tilde{p}$ by

$$\text{RBR}(\tilde{p}) = \frac{r(\tilde{p})}{r(\tilde{p}_U)},$$

where $\hat{p}_U$ is the Bayes estimator that results when a $\mathcal{U}(0, 1)$ prior distribution is used, that is, the noninformative estimator $\hat{p}_U$ serves as a baseline measure for all comparisons. By this definition, note that $\tilde{p}$ has a smaller Bayes risk than $\hat{p}_U$ when $\text{RBR}(\tilde{p}) < 1$.

Using a beta(1, 100) prior, Fig. 1 displays the RBR for the estimators $\hat{p}_B$, $\hat{p}_I$, $\hat{p}_B^{(E)}$, and $\hat{p}_I^{(E)}$ for waiting parameters $n = 1, 5, 10, 20$ and pool sizes $s = 1, 2, \dots, 50$. We used other sensible prior models for $\hat{p}_B$ and $\hat{p}_I$, and the same general patterns among the estimators were realized; additional results are available from the authors. The classical Bayesian estimator $\hat{p}_B$ should serve as the benchmark in this comparison; i.e. it is the estimator that minimizes Bayes risk. However, it is interesting to note that the indirect estimator $\hat{p}_I$ always competes strongly with $\hat{p}_B$. In fact, it is hard to discern from Fig. 1 that $\text{RBR}(\hat{p}_B) < \text{RBR}(\hat{p}_I)$, but our calculations reveal this to be true when $s > 1$. Perhaps surprisingly, the empirical Bayes estimator $\hat{p}_B^{(E)}$ performs similarly to the non-informative estimator $\hat{p}_U$, and the empirical Bayes version of the indirect estimator, $\hat{p}_I^{(E)}$, performs terribly. A closer look reveals that the Bayes risk of $\hat{p}_I^{(E)}$ is dominated by the first term in the sum, that is, the term when $t = n$ (all pools test positive).

4.2 Classical comparison

We now assess the performance of the Bayesian estimators $\hat{p}_I$, $\hat{p}_B$, $\hat{p}_I^{(E)}$, and $\hat{p}_B^{(E)}$, on classical (frequentist) grounds. In doing so, we use the loss function only to determine the form of the estimators; we ignore the loss function otherwise. Specifically, for an estimator generically denoted as $\tilde{p}$, we compute the absolute relative bias and the relative efficiency,

$$\text{RB}(\tilde{p}) = \frac{|\text{Bias}(\tilde{p})|}{|\text{Bias}(\hat{p})|} \quad \text{and} \quad \text{RE}(\tilde{p}) = \frac{\text{MSE}(\hat{p})}{\text{MSE}(\tilde{p})},$$

for fixed n, p , and s , where $\hat{p}$ is the MLE in (1). Therefore, values of $\text{RB}(\tilde{p})$ [$\text{RE}(\tilde{p})$] less than unity indicate that the estimator $\tilde{p}$ has smaller absolute bias (MSE) than the MLE. We also include in the comparison the best classical estimate from Pritchard and Tebbs (2010), the so-called “shift estimate,” given by

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

where $\delta \geq 1$. The constant δ is chosen to minimize the MSE of $\hat{p}_\delta$ based on p_0 , a known upper bound of p , which is specified by the investigator. In this paper, unless otherwise stated, we compute $\hat{p}_\delta$ using an upper bound of $p_0 = 0.01$, that is,

$$\delta = \arg \min_{\delta \geq 1} \text{MSE}(\hat{p}_\delta; p=0.01),$$

See Pritchard and Tebbs (2010, Section 2.2) for complete details.

Figures 2 and 3 display the absolute relative bias and the relative efficiency, respectively, when $n = 1, 5, 10, 20$ and $s = 50$, for p ranging from 0.0001 to 0.02 by increments of 0.0001. Other combinations of s and p yield similar findings; details are available from the authors. Note that when $n = 1$; i.e. when testing stops after the first positive pool is found, all of the estimators confer substantial benefits when compared to the MLE, especially $\hat{p}_B$, $\hat{p}_I$, and $\hat{p}_\delta$.

Close inspection reveals that the empirical Bayes estimators $\hat{p}_B^{(E)}$ and $\hat{p}_I^{(E)}$ perform almost identically in terms of bias and MSE and that these estimators are much better than the MLE when n is small. In terms of bias, the shift estimate $\hat{p}_\delta$ performs best for values of p less than the upper bound $p_0 = 0.01$, a finding that is well documented in Pritchard and Tebbs (2010). For values of $p > 0.01$, the Bayesian estimates $\hat{p}_B$ and $\hat{p}_I$ are preferred in terms of reducing bias. As expected, the performance of the Jeffreys estimator $\hat{p}_J$ deteriorates as n increases.

On classical grounds, we have consistently observed that $\hat{p}_B$, $\hat{p}_I$, and $\hat{p}_\delta$ are the best estimators. The main drawback with using the shift estimate $\hat{p}_\delta$ from Pritchard and Tebbs (2010) is that the investigator must elicit an upper bound for the true p . When this upper bound is correctly specified, the shift estimate is usually preferred. However, the upper bound should be chosen carefully; for $p > 0.01$, the absolute bias of $\hat{p}_\delta$ in Fig. 2 increases and it eventually can become larger than that of the MLE. Therefore, if one is not willing to specify a correct upper bound needed to use $\hat{p}_\delta$, the Bayesian estimators $\hat{p}_B$ and $\hat{p}_I$ are preferred, as long as the $\text{beta}(1, \beta)$ prior model is not grossly misspecified. In fact, we have found that the penalty for misspecifying the $\text{beta}(1, \beta)$ prior with $\hat{p}_B$ and $\hat{p}_I$ is generally far less than the penalty for using $\hat{p}_\delta$ with an incorrect upper bound p_0 .

To illustrate the last point, we refer the reader to Fig. 4, where we consider the two cases:

- i. Case 1 (Top): $n = 5$, $s = 20$, $p_0 = 0.01$, and $p = 0.05$
- ii. Case 2 (Bottom): $n = 5$, $s = 50$, $p_0 = 0.01$, and $p = 0.01$.

For example, Case 1 states that the true prevalence is $p = 0.05$, but one uses $\hat{p}_\delta$ with an (incorrect) upper bound of $p_0 = 0.01$. Case 2 is interpreted similarly. For Case 1, we see in Fig. 4 that the absolute bias (left) and MSE (right) of $\hat{p}_B$ and $\hat{p}_I$ are smaller than that of $\hat{p}_\delta$ for most values of β , except when β is inappropriately large or small. Note that when $p = 0.05$ (Case 1), the “correct” prior model is the $\text{beta}(1, 19)$ distribution, yet, the Bayesian estimators $\hat{p}_B$ and $\hat{p}_I$ are preferred even when this distribution is misspecified badly. In fact, we have determined that $|\text{Bias}(\hat{p}_B)| < |\text{Bias}(\hat{p}_\delta)|$ when $5 \leq \beta \leq 258$ and that $|\text{Bias}(\hat{p}_I)| < |\text{Bias}(\hat{p}_\delta)|$ when $1 \leq \beta \leq 248$. When $p = 0.01$ (Case 2), the correct prior model is the $\text{beta}(1, 99)$ distribution, yet the Bayesian estimators $\hat{p}_B$ and $\hat{p}_I$ have a smaller absolute bias for almost all $\beta \leq 500$ and a smaller MSE when $41 \leq \beta \leq 500$. Note that we did not display values of $\beta > 500$ in the figures; for Case 2, the values of β for which $\hat{p}_B$ and $\hat{p}_I$ are preferred extend well beyond 500.

5 WNV example

Rutledge et al. (2003) summarize the results of a field trial to study the transmission of WNV following a 2001 outbreak in northern Florida. About 12 000 individual mosquitoes were captured and tested in pools using reverse-transcription polymerase chain reaction assays. Most of the pools used by Rutledge et al. (2003) contained $s = 50$ mosquitoes, so we use this in our investigation. The authors presented the complete results only from the mosquito collections that provided 14 positive pools; the authors ignored those collections that yielded all negative pools. The MLE of p under the binomial model, based on the available data, is $\hat{p} = 0.00536$. This is most likely an upper bound since negative pool information was not presented. Inverse sampling was not used by Rutledge et al. (2003), but we use their data as a basis to illustrate the performance of our Bayesian methods.

Using the logit transformed procedure in Hepworth (2005), a 95% confidence interval for p , based on the binomial-type data presented in Rutledge et al. (2003), is (0.003, 0.009). In this investigation, we use this interval to harness our true knowledge about p . In particular, we generate 10 000 values of p from a $\mathcal{U}(0.003, 0.009)$ distribution; with these values, we then simulate 10 000 data sets when $n = 1, 5, 10, 20$. These simulated data sets represent realizations of data that could have been observed by Rutledge et al. (2003) had negative binomial pooled sampling been used. For each data set, we compute all the estimates presented in this paper. The classical Bayes estimators use $\text{beta}(1, 100)$ and $\text{beta}(1, 200)$ priors.

In Table 2, we report the values of the point estimates when $s = 50$, averaged over the 10 000 simulations. We also include the classical shift estimator $\hat{p}_\delta$, using an upper bound of $p_0 = 0.01$, for comparison. Even when $n = 1$, the classical estimators $\hat{p}_B$ and $\hat{p}_I$ are generally on target, whereas the other Bayesian estimators overestimate the true (prior) mean value of $p = 0.006$, the midpoint of 0.003 and 0.009. For larger values of n , all the Bayesian estimates do approach this target value from above; when $n = 20$, all averaged Bayesian estimates are practically identical. The classical estimator $\hat{p}_\delta$ performs well also, but tends to slightly underestimate the true p on average for larger n . Table 3 presents the averaged lower and upper 95% credible interval endpoints using the same simulated data. When $n = 1$, credible intervals arising from the classical $\text{beta}(1, \beta)$ prior perform best. As n increases, all averaged credible interval endpoints converge to the observed (0.003, 0.009) interval based on the actual data.

6 Discussion

We have combined the methods of group testing and inverse sampling to estimate disease prevalence from a Bayesian perspective. Our focus has been on WNV applications, but the methods herein could be used for any arbovirus that is subject to continuous surveillance, especially those pathogens where positive pools trigger the need for immediate analysis. Our Bayesian approach may be preferred because of the sequential nature of the data collection process in these settings, and it allows one to incorporate prior information. We have written user-friendly R programs to calculate all point and interval estimates presented in this paper and to reproduce the misspecification analysis in Section 4.2. We also have written a WinBugs program to calculate the analogous estimates when unequal pool sizes are used. A small investigation on our part reveals that unequal pool size posterior mean estimates are very similar to those in this paper when the pool sizes are not markedly different.

We close with three brief remarks. First, in this paper, we have assumed that the probability of pool misclassification is zero. This may be a satisfactory assumption in some cases, but an unreasonable one in others. When the testing process is not perfect, one could modify our estimation methods similarly to that of Hanson et al. (2006) in the binomial setting by treating the assay sensitivity and specificity as parameters, assigning prior distributions to them, and then integrating to obtain the posterior distribution of p . Closed form expressions for posteriors and estimators would be lost, but this may not create large problems. Second, the techniques in this paper would not be directly applicable if there was a significant spatial correlation among the individual insects collected. This might occur when the pathogen under investigation is more prevalent in clusters and insects are sampled from diverse geographical areas. Our methods could be extended to handle this situation without too much additional difficulty, as long as the investigator is willing to elicit appropriate prior models for the clusters. Finally, we have generally found that the shape of the posterior distribution, $f_{P|T}(p|t)$, computed from any prior distribution, is generally unimodal and often skewed right. In the light of this, one might prefer the posterior mean $E(P|T = t)$, which we have used, to, say, the median or mode of $f_{P|T}(p|t)$ if underestimation is more severe, as it

often is in arbovirus surveillance. Alternatively, one could penalize underestimation using a linear loss function of the form $L(p, a) = \varepsilon_0(p - a)I(p \geq a) + \varepsilon_1(a - p)I(p < a)$, for suitably chosen values of $\varepsilon_0 > \varepsilon_1$. The resulting Bayes estimate the $\varepsilon_0/(\varepsilon_0 + \varepsilon_1)$ quantile of $f_{p|T}(p|t)$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and two anonymous referees for their comments on an earlier version of this article. The authors 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 Institutes of Health.

Appendix A

A.1 Jeffreys prior details

It is straightforward to show that the Jeffreys prior

$$f_p(p) \propto \frac{s(1-p)^{s/2-1}}{1-(1-p)^s},$$

is improper; i.e. $\int_0^1 f_p(p) dp = +\infty$. The posterior distribution, which is proper, is given by

$$f_{p|T}(p|t) \propto f_{T|p}(t|p) f_p(p) = c_J \{1 - (1-p)^s\}^{n-1} (1-p)^{s(t-n+1/2)-1},$$

where the constant c_J solves $\int_0^1 f_{p|T}(p|t) dp = 1$. Using the change of variable $u = (1-p)^s$, so that $p = 1 - u^{1/s}$ and $dp = -s^{-1}(1-p)^{1-s} du$, it follows that

$$c_J^{-1} = s^{-1} \int_0^1 u^{t-n-1/2} (1-u)^{n-1} du = \frac{\Gamma(t-n+1/2) \Gamma(n)}{s \Gamma(t+1/2)},$$

which establishes (2). The posterior mean estimate $\hat{p}_J$ is given by

$$\begin{aligned} \hat{p}_J &\equiv E(P|T=t) = \int_0^1 p f_{p|T}(p|t) dp \\ &= \int_0^1 \frac{s \Gamma(t+1/2)}{\Gamma(n) \Gamma(t-n+1/2)} p \{1 - (1-p)^s\}^{n-1} (1-p)^{s(t-n+1/2)-1} dp. \end{aligned}$$

Using the change of variable $u = (1-p)^s$ again, we have

$$\begin{aligned} \hat{p}_J &= \frac{\Gamma(t+1/2)}{\Gamma(n) \Gamma(t-n+1/2)} \int_0^1 (1-u^{1/s}) (1-u)^{n-1} u^{t-n-1/2} du \\ &= 1 - \frac{\Gamma(t+1/2)}{\Gamma(n) \Gamma(t-n+1/2)} \int_0^1 (1-u)^{n-1} u^{t-n+1/s-1/2} du \\ &= 1 - \frac{\Gamma(t+1/2) \Gamma(t-n+1/s+1/2)}{\Gamma(t-n+1/2) \Gamma(t+1/s+1/2)}, \end{aligned}$$

which is the Bayes estimate in (3).

A.2 Beta(α , β) prior details

We derive general results for the beta(α , β) prior distribution. Results for the beta(1, β) prior follow from taking $\alpha = 1$. The probability density function for $p \sim \text{beta}(\alpha, \beta)$ is

$$f_p(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1},$$

for $\alpha > 0$ and $\beta > 0$, so that the joint distribution of T and p is

$$f_{T,p}(t, p) = f_{T|p}(t|p) f_p(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{t-1}{n-1} \{1 - (1-p)^s\}^n p^{\alpha-1} (1-p)^{s(t-n)+\beta-1}.$$

Integrating $f_{T,p}(t, p)$ over p yields the marginal distribution of T , given by

$$f_T(t|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{t-1}{n-1} \int_0^1 \{1 - (1-p)^s\}^n p^{\alpha-1} (1-p)^{s(t-n)+\beta-1} dp.$$

Writing $\{1 - (1-p)^s\}^n$ in its binomial expansion, that is,

$$\{1 - (1-p)^s\}^n = \sum_{k=0}^n \binom{n}{k} (-1)^k (1-p)^{sk},$$

allows us to rewrite $f_T(t|\alpha, \beta)$ as

$$\begin{aligned} f_T(t|\alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{t-1}{n-1} \sum_{k=0}^n \binom{n}{k} (-1)^k \int_0^1 p^{\alpha-1} (1-p)^{s(t-n+k)+\beta-1} dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \binom{t-1}{n-1} \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st-sn+sk+\beta)}{\Gamma(st-sn+sk+\alpha+\beta)}. \end{aligned}$$

Therefore, the posterior distribution is

$$f_{p|T}(p|t, \alpha, \beta) = \frac{f_{T,p}(t, p|\alpha, \beta)}{f_T(t|\alpha, \beta)} = c_{\text{BB}}^{-1} \{1 - (1-p)^s\}^n p^{\alpha-1} (1-p)^{s(t-n)+\beta-1},$$

where

$$c_{\text{BB}} = \Gamma(\alpha) \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st-sn+sk+\beta)}{\Gamma(st-sn+sk+\alpha+\beta)}.$$

This establishes (6). The posterior mean estimate, after applying the same binomial expansion, is

$$\begin{aligned} \hat{p}_{\text{BB}} &= c_{\text{BB}}^{-1} \sum_{k=0}^n \binom{n}{k} (-1)^k \int_0^1 p^{\alpha} (1-p)^{s(t-n+k)+\beta-1} dp \\ &= c_{\text{BB}}^{-1} \Gamma(\alpha+1) \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st-sn+sk+\beta)}{\Gamma(st-sn+sk+\alpha+\beta+1)}. \end{aligned}$$

This establishes (7). Now, take $\alpha = 1$ so that

$$c_{\text{BB}} = \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st - sn + sk + \beta)}{\Gamma(st - sn + sk + \beta + 1)} = \frac{1}{s} \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{1}{t - n + \beta/s + k}.$$

Using Identity (26.5.6) in Abramowitz and Stegun (1965), it is known that the beta function

$$B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)} = \sum_{k=0}^{a-1} \binom{a-1}{k} (-1)^k \frac{1}{b+k},$$

for integer values of $a \geq 1$. Applying this identity, we can write

$$c_{\text{BB}} = \frac{\Gamma(n+1) \Gamma(t - n + \beta/s)}{s \Gamma(t + \beta/s + 1)}$$

so that the (one parameter) posterior distribution becomes

$$f_{\pi_T}(p|t) = \frac{s \Gamma(t + \beta/s + 1)}{\Gamma(n+1) \Gamma(t - n + \beta/s)} \{1 - (1-p)^s\}^n (1-p)^{s(t-n)+\beta-1}.$$

This establishes (4). When $\alpha = 1$, $\hat{p}_{\text{BB}}$ reduces to

$$\begin{aligned} \hat{p}_{\text{BB}} &= \frac{s \Gamma(t + \beta/s + 1) \Gamma(2)}{\Gamma(n+1) \Gamma(t - n + \beta/s)} \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{\Gamma(st - sn + sk + \beta)}{\Gamma(st - sn + sk + \beta + 2)} \\ &= \frac{\Gamma(t + \beta/s + 1)}{\Gamma(n+1) \Gamma(t - n + \beta/s)} \sum_{k=0}^n \binom{n}{k} (-1)^k \left\{ \frac{1}{t - n + \beta/s + k} - \frac{1}{t - n + \beta/s + 1 + k} \right\} \\ &= \frac{\Gamma(t + \beta/s + 1)}{\Gamma(n+1) \Gamma(t - n + \beta/s)} \left\{ \frac{\Gamma(n+1) \Gamma(t - n + \beta/s)}{\Gamma(t + \beta/s + 1)} - \frac{\Gamma(n+1) \Gamma(t - n + \beta/s + 1/s)}{\Gamma(t + \beta/s + 1 + 1/s)} \right\} \\ &= 1 - \frac{\Gamma(t + \beta/s + 1) \Gamma(t - n + \beta/s + 1/s)}{\Gamma(t - n + \beta/s) \Gamma(t + \beta/s + 1 + 1/s)} = \hat{p}_{\text{B}}, \end{aligned}$$

where the penultimate step follows from using the same identity in Abramowitz and Stegun (1965) applied twice. This establishes (5).

A.3 Derivation of $f_T(t|\beta)$

Recall the two-parameter marginal distribution $f_T(t|\alpha, \beta)$ in Appendix A.2. When $\alpha = 1$, $f_T(t|\alpha, \beta)$ reduces to

$$\begin{aligned} f_T(t|\beta) &= \beta \binom{t-1}{n-1} \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{1}{st - sn + sk + \beta} \\ &= \frac{\beta \Gamma(t)}{s \Gamma(n) \Gamma(t - n + 1)} \sum_{k=0}^n \binom{n}{k} (-1)^k \frac{1}{t - n + \beta/s + k} \\ &= \frac{n \beta \Gamma(t) \Gamma(t - n + \beta/s)}{s \Gamma(t - n + 1) \Gamma(t + \beta/s + 1)}, \end{aligned}$$

where, again, the last step follows from using Identity (26.5.6) in Abramowitz and Stegun (1965). This establishes (9).

A.4 Credible interval details

For the Jeffreys prior, the upper bound p_U solves

$$\alpha^*/2 = \int_{p_U}^1 \frac{s\Gamma(t+1/2)}{\Gamma(n)\Gamma(t-n+1/2)} \{1 - (1-p)^s\}^{n-1} (1-p)^{s(t-n+1/2)-1} dp.$$

Using a change of variable $u = (1-p)^s$, it is straightforward to show that p_U equivalently solves

$$\alpha^*/2 = \int_0^{(1-p_U)^s} \frac{\Gamma(t+1/2)}{\Gamma(n)\Gamma(t-n+1/2)} u^{t-n+1/2-1} (1-u)^{n-1} du,$$

so that $(1-p_U)^s = b_{1-\alpha^*/2, t-n+1/2, n}$. Solving for p_U gives $p_U = 1 - (b_{\alpha^*/2, t-n+1/2, n})^{1/s}$. For the beta(1, β) prior, the approach is identical.

A.5 Derivation of Bayes risk

For a beta(1, β) prior and $L(p, a) = (p-a)^2$, the Bayes risk $r(\tilde{p}) = E_P\{E_{T|P}(\tilde{p} - p)^2\}$ is

$$\begin{aligned} r(\tilde{p}) &= \int_0^1 \sum_{t=n}^{\infty} (\tilde{p} - p)^2 \binom{t-1}{n-1} \{1 - (1-p)^s\}^n (1-p)^{s(t-n)} \beta(1-p)^{\beta-1} dp \\ &= \beta \sum_{t=n}^{\infty} \binom{t-1}{n-1} \int_0^1 (\tilde{p} - p)^2 1 - (1-p)^s \}^n (1-p)^{s(t-n)+\beta-1} dp. \end{aligned}$$

Using the change of variable $u = (1-p)^s$, we have

$$\begin{aligned} r(\tilde{p}) &= \beta s^{-1} \sum_{t=n}^{\infty} \binom{t-1}{n-1} \int_0^1 \left\{ \tilde{p} - (1-u^{1/s}) \right\}^2 u^{t-n+\beta/s-1} (1-u)^n du \\ &= \beta s^{-1} \sum_{t=n}^{\infty} \binom{t-1}{n-1} B(t-n+\beta/s, n+1) E \left[\left\{ (\tilde{p}-1) + u^{1/s} \right\}^2 \right], \end{aligned} \quad (\text{A1})$$

where $B(\cdot, \cdot)$ is the beta function, and the expectation is taken with respect to a beta($t-n+\beta/s$, $n+1$) distribution, i.e.

$$E \left[\left\{ (\tilde{p}-1) + u^{1/s} \right\}^2 \right] = (\tilde{p}-1)^2 + 2(\tilde{p}-1) \frac{\Gamma(t+\beta/s+1)\Gamma(t-n+\beta/s+1/s)}{\Gamma(t-n+\beta/s)\Gamma(t+\beta/s+1/s+1)} + \frac{\Gamma(t+\beta/s+1)\Gamma(t-n+\beta/s+2/s)}{\Gamma(t-n+\beta/s)\Gamma(t+\beta/s+2/s+1)}.$$

Substituting this expression for $E[\{(\tilde{p}-1)+u^{1/s}\}^2]$ into (A1), and simplifying, we get

$$r(\tilde{p}) = \beta s^{-1} \sum_{t=n}^{\infty} \binom{t-1}{n-1} \left\{ (\tilde{p}-1)^2 B(t-n+\beta/s, n+1) + 2(\tilde{p}-1) B(t-n+\beta/s+1/s, n+1) + B(t-n+\beta/s+2/s, n+1) \right\}.$$

References

- Abramowitz, M.; Stegun, I. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Bureau of National Standards; Washington DC: 1965.
- 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]
- Bilder C, Tebbs J, Chen P. Informative retesting. *Journal of the American Statistical Association*. 2010; 105:942–955. [PubMed: 21113353]
- Centers for Disease Control. Final 2008 West Nile Virus activity in the United States. 2010. <http://www.cdc.gov/ncidod/dvbid/westnile/>
- Chick S. Bayesian models for limiting dilution assay and group test data. *Biometrics*. 1996; 52:1055–1062.
- Gastwirth J. The efficiency of pooling in the detection of rare mutations. *American Journal of Human Genetics*. 2000; 67:1036–1039. [PubMed: 10986050]
- Gastwirth J, Johnson W. Screening with cost effective quality control: potential applications to HIV and drug testing. *Journal of the American Statistical Association*. 1994; 89:972–981.
- Hanson T, Johnson W, Gastwirth J. Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening. *Biostatistics*. 2006; 7:41–57. [PubMed: 15947009]
- 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. Debaised estimation of proportions in group testing. *Applied Statistics*. 2009; 58:105–121.
- 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*. Springer; New York: 2006.
- Hughes-Oliver J, Swallow W. A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*. 1994; 89:982–993.
- Katholi, C. Estimation of prevalence by pool screening with equal sized pools and a negative binomial sampling model.. University of Alabama at Birmingham Biostatistics Technical Report. 2006. <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]
- McCann M, Tebbs J. Pairwise comparisons for proportions estimated by pooled testing. *Journal of Statistical Planning and Inference*. 2007; 137:1278–1290.
- Peck C. Going after BVD. *Beef*. 2006; 42:34–44.
- Pritchard N, Tebbs J. Estimating disease prevalence using inverse binomial pooled testing. *Journal of Agricultural, Biological, and Environmental Statistics*. 2010 in press. DOI:10.1007/s13253-010-0036-4.
- Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*. 2006; 48:133–143.
- Rutledge C, Day J, Lorde C, Stark L, Tabachnick W. West Nile Virus infection rates in *Culex nigripalpus* (Dipteria: Culicidae) do not reflect transmission rates in Florida. *Journal of Medical Entomology*. 2003; 3:253–258. [PubMed: 12943101]
- 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.

Figure 1. RBR for $n = 1, 5, 10, 20$ and pool sizes $s = 1, 2, \dots, 50$. The estimators depicted are $\hat{p}_B$ (—●—), $\hat{p}_I$ (—■—), $\hat{p}_B^{(E)}$ (---), and $\hat{p}_I^{(E)}$ (····). A beta(1, 100) prior is used for $\hat{p}_B$ and $\hat{p}_I$.

Figure 2.

Relative absolute bias for $n = 1, 5, 10, 20$ and $s = 50$. The estimators depicted are

$\hat{p}_J$ (—), $\hat{p}_B$ (—●—), $\hat{p}_I$ (—■—), $\hat{p}_R^{(E)}$ (---), and $\hat{p}_I^{(E)}$ (····), and the classical estimator $\hat{p}_\delta$ (—▲—) from Pritchard and Tebbs (2010). A beta(1, 100) prior is used for $\hat{p}_B$ and $\hat{p}_I$.

Figure 3.

Relative efficiency for $n = 1, 5, 10, 20$ and $s = 50$. The estimators depicted are

$\hat{p}_I$ (—), $\hat{p}_B$ (—●—), $\hat{p}_B^{(E)}$ (---), and $\hat{p}_I^{(E)}$ (···), and the classical estimator $\hat{p}_\delta$ (—■—) from Pritchard and Tebbs (2010). A beta(1, 100) prior is used for $\hat{p}_B$ and $\hat{p}_I$.

Figure 4. Absolute bias and MSE of $\hat{p}_B(\text{---})$, $\hat{p}_T(\text{---} \bullet \text{---})$, and $\hat{p}_\delta(\cdots)$ under misspecification. Top: $n = 5$, $s = 20$, $p_0 = 0.01$, and $p = 0.05$. Bottom: $n = 5$, $s = 50$, $p_0 = 0.001$, and $p = 0.01$.

Table 1

Summary of the Bayesian point estimates. The two-parameter Bayes estimate $\hat{p}_{BB}$ is not shown.

Prior	Loss function	Estimate formula
Jeffreys	$L(p-a) = (p-a)^2$	$\hat{p}_J = 1 - \frac{\Gamma(t+1/2)\Gamma(t-n+1/s+1/2)}{\Gamma(t-n+1/2)\Gamma(t+1/s+1/2)}$
Beta(1, β)	$L(p-a) = (p-a)^2$	$\hat{p}_B = 1 - \frac{\Gamma(t+\beta/s+1)\Gamma(t-n+\beta/s+1/s)}{\Gamma(t-n+\beta/s)\Gamma(t+\beta/s+1/s+1)}$
Beta(1, β)	$L(\theta-a) = (\theta-a)^2$	$\hat{p}_I = 1 - \left(1 - \frac{n+1}{t+\beta/s+1}\right)^{1/s}$
Beta(1, β)	$L(p-a) = (p-a)^2$	$\hat{p}_B^{(E)} = 1 - \frac{\Gamma(t+\hat{\beta}/s+1)\Gamma(t-n+\hat{\beta}/s+1/s)}{\Gamma(t-n+\hat{\beta}/s)\Gamma(t+\hat{\beta}/s+1/s+1)}$
Beta(1, β)	$L(\theta-a) = (\theta-a)^2$	$\hat{p}_I^{(E)} = 1 - \left(1 - \frac{n+1}{t+\hat{\beta}/s+1}\right)^{1/s}$

Table 2

Rutledge (2003) WNV data.

Estimator	Beta(1, 100) prior				Beta(1, 200) prior			
	n = 1	n = 5	n = 10	n = 20	n = 1	n = 5	n = 10	n = 20
$\hat{p}_J$	0.015	0.008	0.007	0.006	—	—	—	—
$\hat{p}_B$	0.010	0.008	0.007	0.006	0.006	0.007	0.006	0.006
$\hat{p}_I$	0.009	0.007	0.007	0.006	0.006	0.006	0.006	0.006
$\hat{p}_B^{(E)}$	0.131	0.009	0.007	0.006	—	—	—	—
$\hat{p}_I^{(E)}$	0.026	0.007	0.007	0.006	—	—	—	—
$\hat{p}_U$	0.136	0.010	0.007	0.007	—	—	—	—
$\hat{p}_\delta$	0.007	0.004	0.005	0.005	—	—	—	—

Mean values for $\hat{p}_J$, $\hat{p}_B$, $\hat{p}_I$, $\hat{p}_U$, and $\hat{p}_\delta$, based on 10000 simulated negative binomial data sets with pool size $s = 50$. Beta(1, 100) and beta(1, 200) prior models were used for $\hat{p}_B$ and $\hat{p}_I$; the other estimators do not use these priors. The classical “shift” estimate $\hat{p}_\delta$ from Pritchard and Tebbs (2010), using an upper bound $p_0 = 0.01$, is also included.

Table 3

Rutledge (2003) WNV data.

Prior	n = 1		n = 5		n = 10		n = 20	
	p_L	p_U	p_L	p_U	p_L	p_U	p_L	p_U
Jeffreys	0.001	0.052	0.002	0.016	0.003	0.012	0.004	0.009
Beta(1, 100)	0.001	0.027	0.003	0.015	0.003	0.011	0.004	0.009
Beta(1, 200)	0.001	0.017	0.002	0.013	0.003	0.011	0.004	0.009
Beta(1, $\hat{p}$)	0.011	0.255	0.003	0.017	0.003	0.011	0.004	0.009
U(0, 1)	0.012	0.270	0.003	0.020	0.004	0.012	0.004	0.010

Averaged lower and upper 95 percent credible interval endpoints, based on 10 000 simulated negative binomial data sets with $s = 50$.