Published in final edited form as:

Commun Stat Theory Methods. 2013 ; 42(19): . doi:10.1080/03610926.2011.633732.

Evaluation of a Frequentist Hierarchical Model to Estimate Prevalence when sampling from a large geographic area using Pool Screening

Thomas Birkner, PhD^a, Inmaculada B. Aban, PhD^a, and Charles R. Katholi, PhD^a

^aDepartment of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama

Abstract

We present a frequentist Bernoulli-Beta hierarchical model to relax the constant prevalence assumption underlying the traditional prevalence estimation approach based on pooled data. This assumption is called into question when sampling from a large geographic area. Pool screening is a method that combines individual items into pools. Each pool will either test positive (at least one of the items is positive) or negative (all items are negative). Pool screening is commonly applied to the study of tropical diseases where pools consist of vectors (e.g. black flies) that can transmit the disease. The goal is to estimate the proportion of infected vectors.

Intermediate estimators (model parameters) and estimators of ultimate interest (pertaining to prevalence) are evaluated by standard measures of merit, such as bias, variance and mean squared error making extensive use of expansions. Using the hierarchical model an investigator can determine the probability of the prevalence being below a prespecified threshold value, a value at which no reemergence of the disease is expected. An investigation into the least biased choice of the α parameter in the Beta (α, β) prevalence distribution leads to the choice of $\alpha = 1$.

1. INTRODUCTION

In this paper we propose a frequentist hierarchical model, which accounts for different prevalence values within a broader geographic area, for the estimation of infection prevalence. Interval estimates derived under this formulation will be more realistic (wider) compared to the traditional approach which assumes constant prevalence throughout the sampling region. This model will also allow the calculation of the probability that the prevalence is below a postulated or established threshold value for disease recrudescence. In this respect our model offers advantages over the classical frequentist point estimate and confidence interval approach.

Pool screening, also known as group testing or composite sampling, is a method that combines individual items into several groups, or pools of a certain size. The pools are then tested in place of the individual items. The test is based on chemical, biological or other properties of the sample. The result is binary, e.g. positive versus negative, infected versus non-infected, polluted versus not polluted.

Fundamental contributions to the statistical analysis of pool screening results and the experimental design itself were made by Dorfman (1943), Chiang and Reeves (1962), and Thompson (1962). Methodological developments have been made for the unequal pool size case by Hepworth (1996, 2005) and Hepworth and Watson (2009) with respect to exact and

asymptotic interval estimation and debiasing the Maximum Likelihood prevalence estimate. Tebbs and Bilder (2004) compared seven interval estimators in the equal pool size case in terms of coverage probability and mean length. The most recent development is the incorporation of covariates through regression models (Vansteelandt, Goetghebeur and Verstraeten, 2000, Xie, 2001 and Chen, Tebbs and Bilder, 2009). The method is currently applied in various fields including environmental studies (e.g. Garner *et al.*, 1989), safety of blood products (e.g. Gastwirth and Johnson, 1994), disease screening in humans (e.g. Tu, Litvak and Pagano, 1995), animals (e.g. Messam *et al.*, 2008) or plants (e.g. Swallow, 1985 and Rodoni *et al.*, 1994).

The purpose of group testing is either to detect each positive/defective member in the sample or to estimate the prevalence of positive/defective members in the population. The rationale behind the use of pool screening versus the testing of each individual item is efficiency and cost reduction. This is especially true when the defective rate or disease prevalence is low and the cost of conducting each test is substantial.

Our motivation is estimating infection prevalence in arthropod vector populations such as mosquitoes or black flies, which transmit viral and parasitic diseases (e.g. West Nile virus, St. Louis encephalitis, Onchocerciasis, Malaria, Filariasis). As a concrete example of programs that rely on accurate prevalence estimates we refer to the “African Programme for Onchocerciasis Control” (APOC) and the “Onchocerciasis Elimination Program of the Americas” (OEPA). Onchocerciasis, also known as River Blindness, was a major cause of blindness and skin disease predominantly in African countries. The disease is caused by a parasitic worm and is spread by the bite of an infected black fly. The number of infected people in Africa alone is estimated to be 37 million (Amazigo *et al.*, 2006). The aforementioned programs distribute a medication (Ivermectin) which kills the microfilariae (infant stage of the worm) within the infected human. This eliminates most of the disease symptoms and minimizes further transmission of the disease.

In order to assess the progress of control and elimination programs, insects are collected at different locations within a broader area. Since the advent of Polymerase Chain Reaction (PCR) methodology in the 1990s (for PCR method in black flies see: Katholi *et al.*, 1995) pools of homogenized flies are evaluated for the presence of the parasite’s DNA. A positive pool implies that it contains at least one infected insect. This method cannot determine the number of infected insects in a pool.

The traditional frequentist prevalence estimation approach assumes constant infection prevalence across all sampling sites. But when sampling from a large geographic area such as a state or country, it appears plausible that the disease prevalence varies from sampling site to sampling site above and beyond random variation introduced by the sampling process. The following arguments lend some support to this claim:

1. Black flies rely on fast flowing streams and rivers for breeding. Their numbers will be lower in less favorable habitats. This leads to fewer host-vector contacts thereby lowering the potential of disease transmission from host to vector and vector to next host.
2. Control and treatment programs might not have been implemented equally with respect to coverage, frequency, and duration across a large geographic area which could also lead to locally different values of the prevalence.
3. Transmission efficiencies vary between different vector species.
4. Different exposure to human and vector migration can impact the chances of reintroduction of the disease after successful elimination.

Hence, instead of treating the prevalence as a constant it appears to be more realistic to view it as a random variable having a distribution. This distribution has to be estimated given pool size, number of pools and number of positive pools. Bayesian group testing procedures model the prevalence as a random variable (Chaubey and Li, 1995; Chick 1996; Tebbs *et al.*, 2003) and present one option which incorporates possible variations in prevalence. An alternative and novel approach is the frequentist hierarchical model.

By using a hierarchical model, the non-constant prevalence assumption can be incorporated within the frequentist framework. Practitioners in the field appear to still prefer the frequentist method and are reluctant to adopt the Bayesian paradigm. The hierarchical model uses the data at hand and maximum likelihood estimates (MLE) of intermediate parameters to estimate the prevalence distribution. All statistics obtained under this approach are frequentist statistics and are to be evaluated by frequentist measures of merit, such as variance, bias and mean squared error. A point estimate for the prevalence as derived from the MLE of an intermediate variable is also a MLE by the invariance property of the MLE. As new data becomes available the intermediate parameters and the prevalence distribution have to be estimated based on the new data only. However, as longitudinal data is collected, prevalence distributions for several years of the same geographic area can be compared (see Table 1 for a summary of differences and similarities of different prevalence estimation approaches). The advantage of estimating a distribution for the prevalence over a mere point estimate and/or confidence interval lies in the ability to calculate the probability that the prevalence is below a pre-specified threshold value. This value could be established as the threshold below which further disease transmission ceases.

The outline of the paper is as follows: In section 2 we describe our proposed model, in section 3 we evaluate the properties of the estimator(s) of this model, and in section 4 we explore the choice of parameters for the prevalence distribution.

2. HIERARCHICAL MODEL

We let M be the number of pools of equal size n . We assume that the screening assay/test has perfect sensitivity and specificity, a realistic assumption when testing for the parasite causing river blindness (see Katholi *et al.*, 1995). In the *traditional frequentist model* (for development see Chiang and Reeves, 1962 or Thompson, 1962) the probability of a pool being negative ($X = 0$) or positive ($X = 1$) is distributed as Bernoulli $(1 - (1 - p)^n)$, where p is the probability of an insect being infected (carrier of virus or parasite).

$$f_{x_i}(X|n, p) = \begin{cases} (1 - p)^n & \text{when } x=0; (1 - p)^n \text{ is probability of pool being negative} \\ 1 - (1 - p)^n & \text{when } x=1; 1 - (1 - p)^n \text{ is probability of pool being positive} \end{cases} \quad (1)$$

In the proposed *hierarchical model*, p is no longer a constant, but a random variable itself

$$X|P \sim \text{Bernoulli}[1 - (1 - p)^n] = \begin{cases} (1 - p)^n & \text{when } x=0 \\ 1 - (1 - p)^n & \text{when } x=1 \end{cases} \quad P \sim \text{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}. \quad (2)$$

The $\text{Beta}(\alpha, \beta)$ pdf is defined as $f(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$, $0 < p < 1$, $\alpha > 0$, $\beta > 0$,
and $B(\alpha, \beta)$ denotes the beta function, $B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp$.

We are assuming p follows a $Beta(\alpha, \beta)$ density for the following reasons: The domain of a Beta random variable is restricted to lie between 0 and 1, which coincides with the range of p . The shape of the Beta distribution is sufficiently flexible (its curve can be U-or inverse U-shaped or flat).

We are proposing this Bernoulli-Beta instead of a Binomial-Beta model, because $T = \sum_{i=1}^M X_i$ (the number of positive pools and random variable in the Binomial-Beta model) is not a sufficient statistic for p in the unequal pool size case (Gao, 2010). A future expansion to the unequal pool size case and potential comparisons of results between the two cases are most easily accomplished if the basic model is identical (i.e. both are based on X instead of one on T and the other on X).

We find the unconditional distribution of each Bernoulli trial as:

$$f_x(x; n, \alpha, \beta) = \begin{cases} \frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} & \text{when } x=0, \text{ the probability of a negative pool} \\ 1 - \frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} & \text{when } x=1, \text{ the probability of a positive pool} \end{cases} \quad (3)$$

Next we derive the Maximum Likelihood Estimates of α and β . The likelihood function has the form:

$$L(\alpha, \beta | x_i) = \prod_{i=1}^M \left\{ \left[\frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} \right]^{1-x_i} \left[1 - \frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} \right]^{x_i} \right\} \quad (4)$$

We are left with one equation for the two parameters to be estimated:

$$\sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^k}{1-e^k} \right] = 0, \text{ where} \quad (5)$$

$$k = \ln[\Gamma(\beta+n)] + \ln[\Gamma(\alpha+\beta)] - \ln[\Gamma(\alpha+\beta+n)] - \ln[\Gamma(\beta)],$$

and where $\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx$.

We impose the constraint $\alpha = 1$. This choice does not bias the Beta distribution towards or away from 0. We explore this issue in some detail in section 4.

Theorem I (MLE of β)

Let $X_1, X_2, \dots, X_M$ be Bernoulli random samples that follow the hierarchical model defined in equation (2) where M denotes the number of pools of equal size n . Let $\alpha = 1$. Then the probability mass function for testing of each pool is,

$$f(x; n, \beta) = \begin{cases} \frac{B(1, \beta+n)}{B(1, \beta)} = \frac{\beta}{\beta+n} & \text{when } x=0 \\ 1 - \frac{B(1, \beta+n)}{B(1, \beta)} = 1 - \frac{\beta}{\beta+n} = \frac{n}{\beta+n} & \text{when } x=1 \end{cases}$$

Furthermore, the maximum likelihood estimator for β is given by:

$$\hat{\beta} = \frac{Mn}{T} - n, \quad (6)$$

$$\text{where } T = \sum_{i=1}^m X_i.$$

Proof—See Appendix A.

We will be looking at developing inferences based on the properties of $\hat{\beta}$ and MLEs to construct confidence intervals and test of hypothesis in the future.

Remark: Since T , the number of positive pools, is a sufficient statistic in the equal pool size case for p , no information is lost by using T instead of X .

Note that $\hat{\beta} = \frac{Mn}{T} - n$ is the same estimator as obtained by Bilder and Tebbs (2005) in their empirical Bayes approach using the method of moments and taking $\alpha = 1$ in their Beta prior. However, their approach is based on a completely different philosophy (i.e. Bayesian) and as a consequence has to be evaluated and interpreted differently.

3. PROPERTIES OF NEW ESTIMATORS

3.1. Properties of $\hat{\beta}$

In all asymptotic results stated below the pool size n is fixed and the number of pools M is the variable going to infinity.

Theorem II (Consistency of $\hat{\beta}$)—Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed Bernoulli $\left(\frac{n}{\beta+n}\right)$ random variables where $EX_i = \mu = \frac{n}{\beta+n}$ and $Var X_i = \frac{\beta n}{(\beta+n)^2} < \infty$. Let $\bar{T}_m = \frac{1}{m} \sum_{i=1}^m X_i$. Then $\hat{\beta}$ is a consistent estimator of β .

Proof: By the strong law of large numbers $\bar{T}_m$ converges almost surely to μ . Almost sure convergence implies convergence in probability, which implies consistency. Since $\bar{T}_m$

converges in probability, then $\hat{\beta} = f(\bar{T}_m) = \frac{mn}{m\bar{T}_m} - n = \frac{mn}{\bar{T}_m} - n$ converges in probability to $f[E(T)] = \beta$ by the continuous mapping theorem. Hence $\hat{\beta}$ is a consistent estimator for β .

Remark: All other estimators derived as continuous functions of $\hat{\beta}$ are also consistent by the continuous mapping theorem.

Next, we assess the characteristics of the estimator $\hat{\beta}$ and estimators pertaining to p based on $\hat{\beta}$. We consider the usual measures of merit of an estimator, such as Bias, Variance, and

Mean Squared Error. First we determine $E(\hat{\beta}) = mnE\left(\frac{1}{\bar{T}}\right) - n$. To find $E\left(\frac{1}{\bar{T}}\right)$ we use

the Taylor series expansion. We define $f(T) = \frac{1}{T}$ and let T be a random variable such that $E(T) = \mu_T$. We expand $f(T)$ about μ_T .

Theorem III—Under the same assumptions as stated in Theorem I we obtained the following results:

Result A (Bias):

$$E(\hat{\beta}) = \beta + \frac{(\beta+n)\beta}{(mn)} + \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2} + \frac{(\beta+n)\beta(6\beta^2+6\beta n+n^2)}{(mn)^3} + \frac{(\beta+n)\beta(24\beta^3+36\beta^2 n+14\beta n^2+n^3)}{(mn)^4} + O\left[\frac{1}{(mn)^5}\right] \quad (7)$$

Proof: See Appendix B1.

Result B (Mean Squared Error):

$$E(\hat{\beta} - \beta)^2 = \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(7\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2\beta(38\beta^2+28\beta n+3n^2)}{(mn)^3} + O\left[\frac{1}{(mn)^4}\right] \quad (8)$$

Proof: See Appendix B2.

Result C (Variance): We find an estimate of the Variance by subtracting the expression for (Bias)² from (8).

$$\text{Var}(\hat{\beta}) \approx \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(6\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2\beta(34\beta^2+26\beta n+3n^2)}{(mn)^3} \quad (8)$$

Proof: See Appendix B3.

We observe that $\hat{\beta}$ on average overestimates β . However, as $m \rightarrow \infty$, $E(\hat{\beta}) \rightarrow \beta$. Hence $\hat{\beta}$ is asymptotically unbiased.

Furthermore, we note that (7) has the form

$$\beta + \frac{a_1(\beta)}{m} + \frac{a_2(\beta)}{m^2} + \dots \quad (10)$$

Hence the extended Jackknife can be applied to obtain less biased estimates (see Schucany, Gray and Owen, 1971). We note that the linear bias term in (7) can be also obtained by applying the method described in Gart (1991). Gart's method has been applied to the maximum likelihood estimator of prevalence p by Hepworth and Watson (2009). An alternative to the Jackknife for finding a bias corrected estimate $\tilde{\beta}$ is to subtract the bias terms up to the desired order of magnitude from the estimate $\hat{\beta}$

$$\tilde{\beta} = \hat{\beta} - \text{Bias}(\hat{\beta}). \quad (11)$$

3.2. Properties of the “Credibility interval” for p

So far we have explored properties of the intermediate variable $\hat{\beta}$. Our actual interest focuses on p , the infection prevalence.

Theorem IV—Under the same assumptions as stated in Theorem I and given the MLE $\hat{\beta}$, the expected value of p is

$$\widehat{E(p)} = \hat{\beta} \int_0^1 p (1-p)^{\hat{\beta}-1} dp = \hat{\beta} B(2, \hat{\beta}) = \frac{1}{1+\hat{\beta}} \quad (12)$$

Proof: See Appendix C1.

Using the results of Theorem IV, we can construct an estimate of a $(1 - \alpha)100\%$ *credibility interval* for p . Define a $(1 - \alpha)100\%$ credibility interval as the area between two quantile points under the estimated prevalence distribution requiring equal tail areas. This *credibility interval* is conceptually different from a standard frequentist *confidence interval* and also from a Bayesian *credible interval*. In the standard frequentist approach confidence intervals are calculated by some algorithm, and the resulting probability (e.g. 99%) is the percent of time given a very large number of sets of data that the algorithm yields an interval containing the true value of the parameter (see Rodriguez-Perez *et al.*, 2006). The Bayesian credible interval is a probability statement about the quantity of interest based on prior knowledge and the currently observed data.

The expected values of the lower and upper bounds for the credibility interval for p given $\hat{\beta}$

denoted by $p_L = p_L(\hat{\beta})$, $p_H = p_H(\hat{\beta})$ such that $\int_{p_L}^{p_H} \hat{\beta} (1-p)^{\hat{\beta}-1} dp = 1 - \alpha$ were used to find an estimate of the true $(1 - \alpha)100\%$ credibility interval. The method is described as follows.

Given the general expressions $p_L = 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$, $p_H = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$, we expanded about $\hat{\beta}$ and took the expectation to get

$$\begin{aligned} E_{\hat{\beta}}(p_L) &= 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(1 - \frac{\alpha}{2}\right)}{\hat{\beta}^2} Bias \\ &- \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\hat{\beta}^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(1 - \frac{\alpha}{2}\right)}{\hat{\beta}^3} \right] \frac{1}{2} MSE + \dots \end{aligned} \quad (13)$$

$$\begin{aligned} E_{\hat{\beta}}(p_H) &= 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\hat{\beta}^2} Bias \\ &- \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\hat{\beta}^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\hat{\beta}^3} \right] \frac{1}{2} MSE + \dots \end{aligned} \quad (14)$$

Dropping the remaining terms and plugging in only the linear term for the Bias and MSE of $\hat{\beta}$ (from equations (7) and (8)) we obtain

$$\begin{aligned} E_{\hat{\beta}}(p_L) &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(1 - \frac{\alpha}{2}\right) \frac{\left(1 + \frac{n}{\hat{\beta}}\right)}{mn} \\ &- \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{2\hat{\beta}^3} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(1 - \frac{\alpha}{2}\right)}{\hat{\beta}^2} \right] \frac{(\hat{\beta} + n)^2}{mn} \end{aligned} \quad (15)$$

$$\begin{aligned} E_{\hat{\beta}}(p_H) &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right) \frac{\left(1 + \frac{n}{\hat{\beta}}\right)}{mn} \\ &- \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{2\hat{\beta}^3} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\hat{\beta}^2} \right] \frac{(\hat{\beta} + n)^2}{mn} \end{aligned} \quad (16)$$

For more details of the above results, see Appendix C2. We note that the expressions for $E_\beta(p_L)$, $E_\beta(p_H)$ consist of the true p_L , p_H plus/minus some correction term. Those correction

terms are almost negligible because $\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \approx 1$, $\ln\left(\frac{\alpha}{2}\right)$ is small and β is large. Hence the credibility interval for p is nearly unbiased on average.

3.3. Properties of $P(\widehat{p} \leq p_0)$ as an Estimator of $P(p \leq p_0)$

To decide when to end a treatment program and move to the surveillance stage, a very low infection prevalence level has to be achieved. The probability of being below an established or postulated threshold value for no recrudescence of the disease to occur is

$$P(\widehat{p} \leq p_0) = \int_0^{p_0} \widehat{\beta} (1-p)^{\widehat{\beta}-1} dp = - (1-p)^{\widehat{\beta}} \Big|_0^{p_0} = 1 - (1-p_0)^{\widehat{\beta}} \quad (17)$$

Using a Taylor series expansion about β we found:

$$E[P(\widehat{p} \leq p_0)] = P(p \leq p_0) - (1-p_0)^\beta \times \left\{ \ln(1-p_0) \text{Bias}(\widehat{\beta}) + \frac{[\ln(1-p_0)]^2}{2!} \text{MSE}(\widehat{\beta}) + \dots \right\}. \quad (18)$$

For details of the derivation of equation (18), see Appendix D. This result shows that, on average, $P(\widehat{p} \leq p_0)$ is an underestimate of $P(p \leq p_0)$. However, for small p_0 , which is the case we are interested in, $\ln(1-p_0)$ is close to 0 and $P(\widehat{p} \leq p_0)$ is nearly unbiased. Note also that $(1-p_0)^\beta \rightarrow 0$ as β gets large.

3.4. Comparison with Usual Pool Screening Model

To compare the estimator of prevalence given by $\widehat{E(p)} = \frac{1}{1+\widehat{\beta}}$ from the hierarchical model

with the regular pool screen estimate $\widehat{p} = 1 - \left(1 - \frac{T}{m}\right)^{\frac{1}{n}}$ (as derived for example in Chiang and Reeves, 1962 or Thompson, 1962) we found expansions for both estimators and derived expressions of the absolute and relative difference. The expansions take the following form:

$$\begin{aligned} \widehat{E(p)} &= \frac{T}{mn} + \frac{1}{n} \left(\frac{n-1}{n}\right) \left(\frac{T}{m}\right)^2 + \frac{1}{n} \left(\frac{n-1}{n}\right)^2 \left(\frac{T}{m}\right)^3 + \dots \\ \widehat{p} &= \frac{T}{mn} + \frac{1}{n} \left(\frac{n-1}{n}\right) \left(\frac{1}{2}\right) \left(\frac{T}{m}\right)^2 + \frac{1}{n} \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \left(\frac{1}{6}\right) \left(\frac{T}{m}\right)^3 + \dots \end{aligned} \quad (19)$$

Thus we see that $\widehat{p}$ and $\widehat{E(p)}$ agree in the first term and the absolute difference is given by

$$\widehat{p} - \widehat{E(p)} = -\left(\frac{1}{2}\right) \frac{1}{n} \left(\frac{n-1}{n}\right) \left(\frac{T}{m}\right)^2 - \frac{1}{n} \left(\frac{n-1}{n}\right) \left(\frac{5}{6}\right) \left(\frac{n-4}{n}\right) \left(\frac{T}{m}\right)^3 - \dots \quad (20)$$

The relative difference $\left| \frac{\widehat{p} - \widehat{E(p)}}{\widehat{E(p)}} \right|$ is $O\left(\frac{T}{m}\right)$ when m is large. By Theorem II $\frac{T}{m}$ converges almost surely to $\frac{n}{\beta+n}$ as m gets large. When p is small, β is large and $\frac{n}{\beta+n}$ is small. Thus if the number of pools is large the two estimators are practically the same. The comparison

considers only the point estimate of prevalence. The hierarchical model will be still preferable if $P(\widehat{p} \leq p_0)$ is of interest.

4. CHOICE OF ALPHA

To explore the impact of the choice of α on the shape of the prevalence distribution obtained by the proposed Bernoulli-Beta hierarchical model, we determined $\hat{\beta}$ numerically for a given value of α (Fortran executable is available upon request) and plotted the prevalence distribution (Beta(α, β)) for different values of α and the corresponding $\hat{\beta}$ assuming $M = 100$, $n = 25$ and $T = 2$. Figures 1 through 3 illustrate the influence of the choice of α for the two cases $\alpha < 1$ and $\alpha > 1$.

Figure 1 displays the cumulative density functions (cdfs) of the prevalence for $\alpha_1 = 0.5$, $\alpha_2 = 1.0$, and $\alpha_3 = 1.5$ and corresponding $\hat{\beta}$. We notice a steep increase in the prevalence distribution for $\alpha_1 = 0.5$. For this choice of α more of the distributional mass is concentrated near 0 compared to the other α values considered, indicating a bias towards 0 for $\alpha < 1$. We can also see that the cdf for $\alpha_3 = 1.5$ starts out below the cdf for $\alpha_2 = 1.0$. The latter indicates a bias away from 0 when α is greater than 1.

Figure 2 displays the cdfs for a range of α values less than 1 ($\alpha = 1$ is included as a reference). The closer to zero α is chosen the more mass of the density is concentrated near zero. Hence for $\alpha < 1$ we observe a bias towards zero. This bias gets larger the smaller α becomes.

Figure 3 depicts the cdfs for $\alpha > 1$; the cdf for $\alpha = 1$ is included as a reference. We notice that the curves “shift” to the right in the area closest to 0 as α increases; hence we observe a bias away from 0 for $\alpha > 1$. Figures 1 through 3 justify our choice of $\alpha = 1$, as the value that introduces neither a bias towards zero nor a bias away from zero.

5. CONCLUSION

This paper proposes a frequentist hierarchical Bernoulli-Beta model for the estimation of disease prevalence in the equal pool size case and investigates the properties of the estimators. The model allows for different values of prevalence (beyond sampling error) across a large sampling area by estimating a prevalence distribution instead of a point estimate of prevalence and permits the estimation of $P(p = p_0)$. Due to a lack of sufficient independent information in the likelihood function it was necessary to preset one of the parameters of the Beta (α, β) distribution. An investigation was undertaken in the choice of the parameter α in the Beta (α, β) prevalence distribution. Setting $\alpha = 1$ proved to be the unbiased option. Given $\alpha = 1$ the maximum likelihood estimator for β was found.

We have shown for n fixed and M approaching infinity that the intermediate estimator ($\hat{\beta}$) and the estimators of ultimate interest ($p_L, p_H, P(\widehat{p} \leq p_0)$) have reasonable statistical properties using standard frequentist criteria – first and foremost they are consistent estimators. The estimators $\hat{\beta}, p_L, p_H$, and $P(\widehat{p} \leq p_0)$ have minimal bias, especially in the case of p near zero.

A feature that sets our model apart from classical frequentist estimation methods based on pool screening is its ability to determine the probability of the infection prevalence being below a certain threshold value – a question frequently asked by entomologists.

A preliminary study has shown, that there exists a simple functional relationship between the α and β parameters in the Bernoulli-Beta hierarchical model. The evaluation of the

impact of other choices of α on the prevalence estimates through a simulation study appears worthwhile. In another extension, one could impose an ancillary condition when maximizing the likelihood function, which might result in the ability to uniquely estimate both parameters. The ancillary condition would be aimed at optimizing some statistical property (for instance to minimize the bias).

Acknowledgments

Support: Thomas Birkner is supported by the National Institute of Neurological Disorders and Stroke training grant (T32NS054584).

Appendix A

Maximum Likelihood Estimate of β

Considering the testing of each pool as a Bernoulli trial we can write out the following likelihood function (assuming independence between pools):

$$L(\alpha, \beta | x_i) = \prod_{i=1}^M \left\{ \left[\frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} \right]^{1-x_i} \left[1 - \frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} \right]^{x_i} \right\}$$

$$\text{Let } k = \ln \left[\frac{B(\alpha, \beta+n)}{B(\alpha, \beta)} \right]$$

$$\ln L = l = \sum_{i=1}^M \left[(1-x_i)k + x_i \ln(1-e^k) \right]$$

$$\frac{dl}{d\alpha} = \sum_{i=1}^M \left[(1-x_i) \frac{dk}{d\alpha} + x_i \frac{1}{1-e^k} (-e^k) \frac{dk}{d\alpha} \right] = \frac{dk}{d\alpha} \sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^k}{1-e^k} \right]$$

$$\frac{dl}{d\beta} = \sum_{i=1}^M \left[(1-x_i) \frac{dk}{d\beta} + x_i \frac{1}{1-e^k} (-e^k) \frac{dk}{d\beta} \right] = \frac{dk}{d\beta} \sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^k}{1-e^k} \right]$$

By properties of log and using the fact that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$$k = \ln [\Gamma(\beta+n)] + \ln [\Gamma(\alpha+\beta)] - \ln [\Gamma(\alpha+\beta+n)] - \ln [\Gamma(\beta)]$$

Recall the definition of the digamma function: $\frac{d}{d\alpha} \ln \Gamma(\alpha) = \Psi(\alpha)$

Applying the chain rule: $\frac{d}{d\alpha} \ln \Gamma[\zeta(\alpha)] = \Psi(\zeta) \frac{d\zeta}{d\alpha}$, we get

$$\begin{aligned} \frac{dk}{d\alpha} &= \Psi(\alpha+\beta) - \Psi(\alpha+\beta+n) \\ \frac{dk}{d\beta} &= \Psi(\beta+n) + \Psi(\alpha+\beta) - \Psi(\alpha+\beta+n) - \Psi(\beta) \end{aligned}$$

Using the result: $\Psi(\alpha+\beta+n) = \frac{1}{\alpha+\beta} + \frac{1}{(\alpha+\beta+1)} + \dots + \frac{1}{(\alpha+\beta+n-1)} + \Psi(\alpha+\beta)$

$$\text{We get } \frac{dk}{d\alpha} = - \sum_{j=0}^{n-1} \frac{1}{\alpha+\beta+j}$$

For $\alpha, \beta > 0$ this is always negative (not zero)

Similarly, we get $\frac{dk}{d\beta} = \sum_{j=0}^{n-1} \frac{\alpha}{(\beta+j)(\alpha+\beta+j)}$, which is > 0 for $\alpha, \beta > 0$ and finite.

Since $\frac{dk}{d\alpha}$ and $\frac{dk}{d\beta}$ cannot be zero, $\sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^k}{1-e^k} \right] \stackrel{!}{=} 0$

Now we are left with two unknowns α, β but only one equation. Hence we impose the constraint $\alpha = 1$ (do not bias towards or away from 0, see section 4 of paper)

For $\alpha = 1$ and using the fact that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, for $\alpha > 0$, we get $k = \ln\left(\frac{\beta}{\beta+n}\right)$
Therefore,

$$\begin{aligned} \sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^k}{1-e^k} \right] &= \sum_{i=1}^M \left[(1-x_i) + x_i \frac{-e^{\ln\left(\frac{\beta}{\beta+n}\right)}}{1-e^{\ln\left(\frac{\beta}{\beta+n}\right)}} \right] \\ &= \sum_{i=1}^M \left[(1-x_i) - x_i \frac{\beta}{n} \right] = M - T - \frac{\beta}{n} T \stackrel{!}{=} 0 \end{aligned}$$

Where $T = \sum_{i=1}^M x_i$, the number of positive pools. Solve for β , we get $\hat{\beta} = \frac{Mn}{T} - n$

Appendix B1

Bias

We found $\hat{\beta} = \frac{mn}{T} - n$ for $\alpha = 1$ by finding the maximum of the log-likelihood function. We

want to determine $E\left(\frac{1}{T}\right) = nmE\left(\frac{1}{T}\right) - n$. To find $E\left(\frac{1}{T}\right)$ we use the Taylor series

expansion. We define $f(T) = \frac{1}{T}$ and let T be a random variable such that $E(T) = \mu_T$. We shall expand $f(T)$ about μ_T .

Recall that the Taylor series expansion for $f(x)$ about $x = x_0$ is

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{(x-x_0)^2}{2!} + f'''(x_0)\frac{(x-x_0)^3}{3!} + \dots + f^{(n)}(x_0)\frac{(x-x_0)^n}{n!}$$

Note that $E(X) = \frac{n}{\beta+n}$ (for each Bernoulli trial), so that $E(T) = \sum_{j=1}^m E(X_j) = \frac{mn}{\beta+n} = \mu_T$

Applying the Taylor series expansion for $f(T) = \frac{1}{T}$ about $t = \mu_T$ we get:

$$\begin{aligned} f(T) &= \frac{1}{\mu_T} - \left(\frac{1}{\mu_T}\right)^2 (T - \mu_T) + \left(\frac{1}{\mu_T}\right)^3 (T - \mu_T)^2 - \left(\frac{1}{\mu_T}\right)^4 (T - \mu_T)^3 + \left(\frac{1}{\mu_T}\right)^5 (T - \mu_T)^4 \\ &\quad - \left(\frac{1}{\mu_T}\right)^6 (T - \mu_T)^5 + \left(\frac{1}{\mu_T}\right)^7 (T - \mu_T)^6 - \left(\frac{1}{\mu_T}\right)^8 (T - \mu_T)^7 + \left(\frac{1}{\mu_T}\right)^9 (T - \mu_T)^8 \dots \end{aligned}$$

Plugging in $\mu_T = \frac{mn}{\beta+n}$ and taking the expectation we obtain:

$$E(f(T)) = \frac{\beta+n}{mn} - \left(\frac{\beta+n}{mn}\right)^2 E\left(T - \frac{mn}{\beta+n}\right) + \left(\frac{\beta+n}{mn}\right)^3 E\left(T - \frac{mn}{\beta+n}\right)^2 - \left(\frac{\beta+n}{mn}\right)^4 E\left(T - \frac{mn}{\beta+n}\right)^3 \\ + \left(\frac{\beta+n}{mn}\right)^5 E\left(T - \frac{mn}{\beta+n}\right)^4 - \left(\frac{\beta+n}{mn}\right)^6 E\left(T - \frac{mn}{\beta+n}\right)^5 + \left(\frac{\beta+n}{mn}\right)^7 E\left(T - \frac{mn}{\beta+n}\right)^6 \\ - \left(\frac{\beta+n}{mn}\right)^8 E\left(T - \frac{mn}{\beta+n}\right)^7 + \left(\frac{\beta+n}{mn}\right)^9 E\left(T - \frac{mn}{\beta+n}\right)^8 \dots$$

We notice the second term equals zero. We find the remaining central moments using the following moment generating function (with the help of Maple 13):

$$M_{T - \left(\frac{mn}{\beta+n}\right)}(t) = e^{-\left(\frac{mnt}{\beta+n}\right)} \left(\frac{\beta+ne^t}{\beta+n}\right)^m$$

$$\begin{aligned} \text{We find: } E\left(T - \frac{mn}{\beta+n}\right)^2 &= \frac{mn\beta}{(\beta+n)^2} \\ E\left(T - \frac{mn}{\beta+n}\right)^3 &= \frac{mn\beta(\beta-n)}{(\beta+n)^3} \\ E\left(T - \frac{mn}{\beta+n}\right)^4 &= \frac{mn\beta(\beta^2-4\beta n+n^2)+(mn\beta)^2 3}{(\beta+n)^4} \\ E\left(T - \frac{mn}{\beta+n}\right)^5 &= \frac{mn\beta(\beta^3-11\beta^2 n+11\beta n^2-n^3)+(mn\beta)^2(10\beta-10n)}{(\beta+n)^5} \\ E\left(T - \frac{mn}{\beta+n}\right)^6 &= \frac{mn\beta(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4)}{(\beta+n)^6} \\ &\quad + \frac{(mn\beta)^2(25\beta^2-80n\beta+25n^2)+(mn\beta)^3 15}{(\beta+n)^6} \\ E\left(T - \frac{mn}{\beta+n}\right)^7 &= \frac{mn\beta(\beta^5-57n\beta^4+302n^2\beta^3-302n^3\beta^2+57n^4\beta-n^5)}{(\beta+n)^7} \\ &\quad + \frac{(mn\beta)^2(56\beta^3-406n\beta^2+406n^2\beta-56n^3)+(mn\beta)^3(105\beta-105n)}{(\beta+n)^7} \\ E\left(T - \frac{mn}{\beta+n}\right)^8 &= \frac{mn\beta(\beta^6-120n\beta^5+1191n^2\beta^4-2416n^3\beta^3+1191n^4\beta^2-120n^5\beta+n^6)}{(\beta+n)^8} \\ &\quad + \frac{(mn\beta)^2(-1680n\beta^3+3710n^2\beta^2-1680n^3\beta+119n^4+119\beta^4)}{(\beta+n)^8} \\ &\quad + \frac{(mn\beta)^3(490\beta^2-1400n\beta+490n^2)+(mn\beta)^4 105}{(\beta+n)^8} \end{aligned}$$

Inserting these moments into $E[f(T)]$, simplifying and ordering powers we obtain:

$$\begin{aligned}
E(f(T)) = & \frac{\beta+n}{mn} + \frac{(\beta+n)\beta}{(mn)^2} - \frac{(\beta+n)\beta(\beta-n)}{(mn)^3} + \frac{(\beta+n)(\beta)^2 3}{(mn)^3} \\
& + \frac{(\beta+n)\beta(\beta^2-4\beta n+n^2)}{(mn)^4} - \frac{(\beta+n)(\beta)^2(10\beta-10n)}{(mn)^4} + \frac{(\beta+n)(\beta)^3(15)}{(mn)^4} \\
& - \frac{(\beta+n)\beta(\beta^3-11\beta^2 n+11\beta n^2-n^3)}{(mn)^5} + \frac{(\beta+n)(\beta)^2(25\beta^2-80\beta n+25n^2)}{(mn)^5} \\
& - \frac{(\beta+n)(\beta)^3(105\beta-105n)}{(mn)^5} + \frac{(\beta+n)(\beta)^4 105}{(mn)^5} \\
& + \frac{(\beta+n)\beta(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4)}{(mn)^6} - \frac{(\beta+n)(\beta)^2(56\beta^3-406n\beta^2+406n^2\beta-56n^3)}{(mn)^6} \\
& + \frac{(\beta+n)(\beta)^3(490\beta^2-1400n\beta+490n^2)}{(mn)^6} \\
& - \frac{(\beta+n)\beta(\beta^5-57n\beta^4+302n^2\beta^3-302n^3\beta^2+57n^4\beta-n^5)}{(mn)^7} \\
& + \frac{(\beta+n)(\beta)^2(-1680n\beta^3+3710n^2\beta^2-1680n^3\beta+119n^4+119\beta^4)}{(mn)^7} \\
& + \frac{(\beta+n)\beta(\beta^6-120n\beta^5+1191n^2\beta^4-2416n^3\beta^3+1191n^4\beta^2-120n^5\beta+n^6)}{(mn)^8}
\end{aligned}$$

Simplifying again, we get:

$$\begin{aligned}
E(f(T)) = & \frac{\beta+n}{mn} + \frac{(\beta+n)\beta}{(mn)^2} + \frac{(\beta+n)\beta(2\beta+n)}{(mn)^3} + \frac{(\beta+n)\beta(6\beta^2+6\beta n+n^2)}{(mn)^4} \\
& + \frac{(\beta+n)\beta(24\beta^3+36\beta^2 n+14\beta n^2+n^3)}{(mn)^5} + O\left[\frac{1}{(mn)^6}\right]
\end{aligned}$$

$$\text{Recall } E(\hat{\beta}) = nmE\left(\frac{1}{T}\right) - n$$

$$\begin{aligned}
E(\hat{\beta}) = & \beta + \frac{(\beta+n)\beta}{(mn)} + \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2} + \frac{(\beta+n)\beta(6\beta^2+6\beta n+n^2)}{(mn)^3} \\
& + \frac{(\beta+n)\beta(24\beta^3+36\beta^2 n+14\beta n^2+n^3)}{(mn)^4} + O\left[\frac{1}{(mn)^5}\right]
\end{aligned}$$

We observe that as $m \rightarrow \infty$ $E(\hat{\beta}) \rightarrow \beta$. Hence $\hat{\beta}$ is asymptotically unbiased.

Appendix B2

Mean Squared Error

$$\text{Recall } \text{MSE} = E_{\theta}(W - \theta)^2 = \text{Var}_{\theta}W + (E_{\theta}W - \theta)^2 = \text{Var}_{\theta}W + (\text{Bias}_{\theta}W)^2$$

First we need the expression for $\hat{\beta}$, above we only found $E(\hat{\beta})$. We have the MLE for

$$\alpha=1: \hat{\beta} = \frac{mn}{T} - n$$

We also have the expansion for $\frac{1}{T}$, plugging in $\mu_T = \frac{mn}{\beta+n}$ in this expansion we obtain:

$$\begin{aligned}
\hat{\beta} &= mn \left[\frac{\beta+n}{mn} - \left(\frac{\beta+n}{mn} \right)^2 \left(T - \frac{mn}{\beta+n} \right) + \left(\frac{\beta+n}{mn} \right)^3 \left(T - \frac{mn}{\beta+n} \right)^2 - \left(\frac{\beta+n}{mn} \right)^4 \left(T - \frac{mn}{\beta+n} \right)^3 \right. \\
&\quad \left. + \left(\frac{\beta+n}{mn} \right)^5 \left(T - \frac{mn}{\beta+n} \right)^4 - \dots \right] - n \\
\hat{\beta} &= \beta - \frac{(\beta+n)^2}{mn} \left(T - \frac{mn}{\beta+n} \right) + \frac{(\beta+n)^3}{(mn)^2} \left(T - \frac{mn}{\beta+n} \right)^2 - \frac{(\beta+n)^4}{(mn)^3} \left(T - \frac{mn}{\beta+n} \right)^3 \\
&\quad + \frac{(\beta+n)^5}{(mn)^4} \left(T - \frac{mn}{\beta+n} \right)^4 - \frac{(\beta+n)^6}{(mn)^5} \left(T - \frac{mn}{\beta+n} \right)^5 + \dots \\
\hat{\beta} - \beta &= -\frac{(\beta+n)^2}{mn} \left(T - \frac{mn}{\beta+n} \right) + \frac{(\beta+n)^3}{(mn)^2} \left(T - \frac{mn}{\beta+n} \right)^2 - \frac{(\beta+n)^4}{(mn)^3} \left(T - \frac{mn}{\beta+n} \right)^3 \\
&\quad + \frac{(\beta+n)^5}{(mn)^4} \left(T - \frac{mn}{\beta+n} \right)^4 - \frac{(\beta+n)^6}{(mn)^5} \left(T - \frac{mn}{\beta+n} \right)^5 + \dots
\end{aligned}$$

Squaring above expression is somewhat daunting, but there is the Cauchy product (product of two power series):

$$\text{Given } \sum_{i=0}^{\infty} a_i x^i, \sum_{j=0}^{\infty} b_j x^j \text{ then } (\Sigma a_i)(\Sigma b_j) = \Sigma c_i, \text{ where } c_n = \sum_{j=0}^n a_j b_{n-j}$$

$$\text{And so when } a_j = b_j, \forall j \text{ we have } c_n = \sum_{j=0}^n a_j a_{n-j} = a_0 a_n + \sum_{j=1}^{n-1} a_j a_{n-j} + a_n a_0$$

$$\text{When } a_0 = 0 \text{ we have } \sum_{j=1}^{n-1} a_j a_{n-j}, \text{ for } n \geq 2$$

Note in this case $c_0 = c_1 = 0$, since $c_1 = 2a_0a_1 = 2 \cdot 0 \cdot a_1 = 0$

$$\begin{aligned}
c_2 &= a_1 a_1 = a_1^2 & a_1 &= -\frac{(\beta+n)^2}{mn} \\
c_3 &= a_1 a_2 + a_2 a_1 = 2a_1 a_2 & a_2 &= \frac{(\beta+n)^3}{(mn)^2} \\
c_4 &= a_1 a_3 + a_2 a_2 + a_3 a_1 = 2a_1 a_3 + a_2^2 & a_3 &= -\frac{(\beta+n)^4}{(mn)^3} \\
c_5 &= a_1 a_4 + a_2 a_3 + a_3 a_2 + a_4 a_1 = 2a_1 a_4 + 2a_2 a_3 & a_4 &= \frac{(\beta+n)^5}{(mn)^4} \\
c_6 &= a_1 a_5 + a_2 a_4 + a_3 a_3 + a_4 a_2 + a_5 a_1 = 2a_1 a_5 + 2a_2 a_4 + a_3^2 & a_5 &= -\frac{(\beta+n)^6}{(mn)^5}
\end{aligned}$$

$$\begin{aligned}
c_2 &= a_1^2 = \frac{(\beta+n)^4}{(mn)^2} \\
c_3 &= 2a_1 a_2 = -2 \frac{(\beta+n)^5}{(mn)^3} \\
c_4 &= 2a_1 a_3 + a_2^2 = 3 \frac{(\beta+n)^6}{(mn)^4} \\
c_5 &= 2a_1 a_4 + 2a_2 a_3 = -4 \frac{(\beta+n)^7}{(mn)^5} \\
c_6 &= 2a_1 a_5 + 2a_2 a_4 + a_3^2 = 5 \frac{(\beta+n)^8}{(mn)^6}
\end{aligned}$$

$$\begin{aligned}
 (\hat{\beta} - \beta)^2 &= c_2(T - \mu_T)^2 + c_3(T - \mu_T)^3 + c_4(T - \mu_T)^4 + \dots \\
 &= \frac{(\beta+n)^4}{(mn)^2}(T - \mu_T)^2 - 2\frac{(\beta+n)^5}{(mn)^3}(T - \mu_T)^3 + 3\frac{(\beta+n)^6}{(mn)^4}(T - \mu_T)^4 \\
 &\quad - 4\frac{(\beta+n)^7}{(mn)^5}(T - \mu_T)^5 + 5\frac{(\beta+n)^8}{(mn)^6}(T - \mu_T)^6 - \dots
 \end{aligned}$$

Taking expectation on both sides:

$$\begin{aligned}
 E(\hat{\beta} - \beta)^2 &= MSE = \frac{(\beta+n)^4}{(mn)^2}E(T - \mu_T)^2 - 2\frac{(\beta+n)^5}{(mn)^3}E(T - \mu_T)^3 \\
 &\quad + 3\frac{(\beta+n)^6}{(mn)^4}E(T - \mu_T)^4 - 4\frac{(\beta+n)^7}{(mn)^5}E(T - \mu_T)^5 \\
 &\quad + 5\frac{(\beta+n)^8}{(mn)^6}E(T - \mu_T)^6 - \dots
 \end{aligned}$$

Taking above piece by piece:

$$\begin{aligned}
 c_2E(T - \mu_T)^2 &= \frac{(\beta+n)^2\beta}{mn} \\
 c_3E(T - \mu_T)^3 &= -\frac{2(\beta+n)^2\beta(\beta-n)}{(mn)^2} \\
 c_4E(T - \mu_T)^4 &= \frac{3(\beta+n)^2\beta(\beta^2-4\beta n+n^2)}{(mn)^3} + \frac{9(\beta+n)^2(\beta)^2}{(mn)^2} \\
 c_5E(T - \mu_T)^5 &= -4\frac{(\beta+n)^2\beta(\beta^3-11\beta^2n+11\beta n^2-n^3)}{(mn)^4} - 4\frac{(\beta+n)^2(\beta)^2(10\beta-10n)}{(mn)^3} \\
 c_6E(T - \mu_T)^6 &= 5\frac{(\beta+n)^2\beta(\beta^4-26\beta n\beta^3+66n^2\beta^2-26n^3\beta+n^4)}{(mn)^5} \\
 &\quad + 5\frac{(\beta+n)^2(\beta)^2(25\beta^2-80\beta n+25n^2)}{(mn)^4} + \frac{75(\beta+n)^2(\beta)^3}{(mn)^3}
 \end{aligned}$$

Putting those pieces together up to $\frac{1}{(mn)^3}$

$$\begin{aligned}
 &E(\hat{\beta} - \beta)^2 \\
 &= \frac{(\beta+n)^2\beta}{mn} - \frac{2(\beta+n)^2\beta(\beta-n)}{(mn)^2} + \frac{9(\beta+n)^2(\beta)^2}{(mn)^2} + \frac{3(\beta+n)^2\beta(\beta^2-4\beta n+n^2)}{(mn)^3} \\
 &\quad - 4\frac{(\beta+n)^2(\beta)^2(10\beta-10n)}{(mn)^3} + \frac{75(\beta+n)^2(\beta)^3}{(mn)^3}
 \end{aligned}$$

$$\begin{aligned}
 &E(\hat{\beta} - \beta)^2 \\
 &= \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(7\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2\beta(38\beta^2+28\beta n+3n^2)}{(mn)^3} + O\left(\frac{1}{(mn)^4}\right)
 \end{aligned}$$

Appendix B3

Variance

Considering only terms up to $\frac{1}{(mn)^3}$ we obtain an estimate of the Variance:

$$\begin{aligned}
 (\text{MSE} &= \text{Var} + \text{Bias}^2 \Rightarrow \text{Var} = \text{MSE} - \text{Bias}^2) \\
 \text{Var}(\hat{\beta}) &\approx \frac{(\beta+n)^2 \beta}{mn} + \frac{(\beta+n)^2 \beta (7\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2 \beta (38\beta^2+28\beta n+3n^2)}{(mn)^3} - \frac{(\beta+n)^2 \beta^2}{(mn)^2} \\
 &- \frac{2\beta^2 (\beta+n)^2 (2\beta+n)}{(mn)^3} \\
 &= \frac{(\beta+n)^2 \beta}{mn} + \frac{(\beta+n)^2 \beta (6\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2 \beta (34\beta^2+26\beta n+3n^2)}{(mn)^3}
 \end{aligned}$$

We see that the difference in Var and MSE is only in 2nd order term and above.

Computation of Bias²

Another application of the Cauchy product: $c_n = \sum_{j=0}^n a_j a_{n-j} = a_0 a_n + \sum_{j=1}^{n-1} a_j a_{n-j} + a_n a_0$

$$(\text{Bias})^2 = \left[\frac{(\beta+n)\beta}{(mn)} + \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2} + \frac{(\beta+n)\beta(6\beta^2+6\beta n+n^2)}{(mn)^3} + \frac{(\beta+n)\beta(24\beta^3+36\beta^2 n+14\beta n^2+n^3)}{(mn)^4} + \dots \right]^2$$

$$\begin{aligned}
 a_0 &= 0 \\
 a_1 &= \frac{(\beta+n)\beta}{(mn)} \\
 a_2 &= \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2} \\
 a_3 &= \frac{(\beta+n)\beta(6\beta^2+6\beta n+n^2)}{(mn)^3} \\
 a_4 &= \frac{(\beta+n)\beta(24\beta^3+36\beta^2 n+14\beta n^2+n^3)}{(mn)^4}
 \end{aligned}$$

$$\begin{aligned}
 c_1 &= \sum_{j=1}^0 a_j a_{1-j} = 0 \\
 c_2 &= \sum_{j=1}^1 a_j a_{2-j} = a_1^2 = \frac{(\beta+n)^2 \beta^2}{(mn)^2} \\
 c_3 &= \sum_{j=1}^2 a_j a_{3-j} = a_1 a_2 + a_2 a_1 = 2a_1 a_2 = \frac{2\beta^2 (\beta+n)^2 (2\beta+n)}{(mn)^3} \\
 c_4 &= \sum_{j=1}^3 a_j a_{4-j} = a_1 a_3 + a_2 a_2 + a_3 a_1 = 2a_1 a_3 + a_2^2 = \frac{\beta^2 (\beta+n)^2 (16\beta^2+16\beta n+3n^2)}{(mn)^4} \\
 (\text{Bias})^2 &\approx \frac{(\beta+n)^2 \beta^2}{(mn)^2} + \frac{2\beta^2 (\beta+n)^2 (2\beta+n)}{(mn)^3} + \frac{\beta^2 (\beta+n)^2 (16\beta^2+16\beta n+3n^2)}{(mn)^4}
 \end{aligned}$$

Appendix C1

Estimated expected value of prevalence

$$\widehat{E(p)} = \int_0^1 p \frac{p^{\alpha-1} (1-p)^{\hat{\beta}-1}}{B(\alpha, \hat{\beta})} dp$$

For $\alpha = 1$

$$\begin{aligned}
E(p) &= \frac{1}{B(1, \hat{\beta})} \int_0^1 p(1-p)^{\hat{\beta}-1} dp = \hat{\beta} \int_0^1 p(1-p)^{\hat{\beta}-1} dp \\
&= \hat{\beta} \int_0^1 p^{2-1} (1-p)^{\hat{\beta}-1} dp = \hat{\beta}^* B(2, \hat{\beta}) \\
&\quad \left[B(1, \hat{\beta}) = \frac{\Gamma(1)\Gamma(\hat{\beta})}{\Gamma(1+\hat{\beta})} = \frac{\Gamma(\hat{\beta})}{\hat{\beta}^* \Gamma(\hat{\beta})} = \frac{1}{\hat{\beta}} \right] \\
E(p) &= \hat{\beta}^* B(2, \hat{\beta}) = \hat{\beta} \frac{\Gamma(2)\Gamma(\hat{\beta})}{\Gamma(2+\hat{\beta})} = \frac{\hat{\beta}^* (\hat{\beta}-1)!}{(\hat{\beta}+1)!} = \frac{\hat{\beta}!}{(\hat{\beta}+1)!} = \frac{1}{\hat{\beta}+1}
\end{aligned}$$

Appendix C2

Credibility interval

Expressions for lower and upper bound for p

$$\begin{aligned}
p &\sim \frac{p^{\alpha-1} (1-p)^{\hat{\beta}-1}}{B(\alpha, \hat{\beta})}, \text{ for } \alpha = 1 \text{ this simplifies to } \hat{\beta}(1-p)^{\hat{\beta}-1} \\
\int_{p_L}^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} dp &= 1 - \alpha \\
\int_0^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} dp &= - (1-p)^{\hat{\beta}} \Big|_0^{p_H} = 1 - (1-p_H)^{\hat{\beta}} \\
\int_{p_H}^1 \hat{\beta}(1-p)^{\hat{\beta}-1} dp &= 1 - \int_0^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = 1 - \left[1 - (1-p_H)^{\hat{\beta}} \right] = (1-p_H)^{\hat{\beta}} \stackrel{!}{=} \frac{\alpha}{2} \\
\Rightarrow (1-p_H) &= \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \Rightarrow p_H = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \\
\int_0^{p_L} \hat{\beta}(1-p)^{\hat{\beta}-1} dp &= - (1-p)^{\hat{\beta}} \Big|_0^{p_L} = 1 - (1-p_L)^{\hat{\beta}} \stackrel{!}{=} \frac{\alpha}{2} \\
1 - \frac{\alpha}{2} &= (1-p_L)^{\hat{\beta}} \Rightarrow \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} = (1-p_L) \Rightarrow p_L = 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}
\end{aligned}$$

The upper bound of the credibility interval: $p_H = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$

We will expand $\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$ about β using the Taylor series expansion

$$\begin{aligned}
\frac{d}{d\beta} \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} &= - \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} \\
\frac{d^2}{d\beta^2} \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} &= \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \\
f(\hat{\beta}) &\approx \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} - \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} (\hat{\beta} - \beta) + \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right] \frac{(\hat{\beta} - \beta)^2}{2!} \\
p_H &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} (\hat{\beta} - \beta) - \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right] \frac{(\hat{\beta} - \beta)^2}{2!} \\
E_{\beta}(p_H) &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} E(\hat{\beta} - \beta) - \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} E(\hat{\beta} - \beta)^2 \\
E_{\beta}(p_H) &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} Bias - \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} MSE
\end{aligned}$$

We found bias and MSE earlier:

$$\begin{aligned} \text{Bias}(\hat{\beta}) &= \frac{(\beta+n)\beta}{mn} + \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2} + \dots \\ \text{MSE}(\hat{\beta}) &= \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(7\beta+2n)}{(mn)^2} + \dots \end{aligned}$$

Plugging in expressions for Bias and MSE (only first term each):

$$\begin{aligned} E_{\beta}(p_H) &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} \frac{(\beta+n)\beta}{mn} - \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} \frac{(\beta+n)^2\beta}{mn} \\ E_{\beta}(p_H) &\approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right) \frac{(1+\frac{n}{\beta})}{mn} - \left[\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(\frac{\alpha}{2}\right)\right]^2}{2\beta^3} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} \right] \frac{(\beta+n)^2}{mn} \end{aligned}$$

We see that the expression for $E_{\beta}(p_H)$ consists of the true p_H plus/minus some correction

term, where $\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \sim 1$ and $\ln\left(\frac{\alpha}{2}\right)$ is small

Lower bound of credibility interval: $p_L = 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}}$

$f(\beta) = \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}}$, expand about β

$$\begin{aligned} \frac{d}{d\beta} \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} &= - \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} \\ \frac{d^2}{d\beta^2} \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} &= \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \\ f(\hat{\beta}) &\approx \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} - \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} (\hat{\beta} - \beta) \\ &\quad + \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \right] \frac{(\hat{\beta} - \beta)^2}{2!} \\ p_L &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} (\hat{\beta} - \beta) \\ &\quad - \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \right] \frac{(\hat{\beta} - \beta)^2}{2!} \\ E_{\beta}(p_L) &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} E(\hat{\beta} - \beta) \\ &\quad - \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} E(\hat{\beta} - \beta)^2 \\ E_{\beta}(p_L) &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} \text{Bias} \\ &\quad - \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} \text{MSE} \end{aligned}$$

Plugging in expressions for Bias and MSE (only first term each):

$$\begin{aligned}
E_{\beta}(p_L) &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right) (\beta+n) \beta}{\beta^2 mn} \\
&\quad - \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3} \right] \frac{1}{2} \frac{(\beta+n)^2 \beta}{mn} \\
E_{\beta}(p_L) &\approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right) \frac{\left(1 + \frac{n}{\beta}\right)}{mn} \\
&\quad - \left[\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left[\ln\left(1 - \frac{\alpha}{2}\right)\right]^2}{2\beta^3} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} \right] \frac{(\beta+n)^2}{mn}
\end{aligned}$$

Appendix D

Threshold Probability

Investigate the properties of $P(p \leq p_0) = \int_0^{p_0} \hat{\beta}(1-p)^{\hat{\beta}-1} dp$ as an estimator of $P(\widehat{p} \leq p_0)$

$$P(\widehat{p} \leq p_0) = \int_0^{p_0} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = - (1-p)^{\hat{\beta}} \Big|_0^{p_0} = 1 - (1-p_0)^{\hat{\beta}}.$$

Note that the true value is β . So we expand $(1-p_0)^{\hat{\beta}}$ about β

$$\begin{aligned}
f(\hat{\beta}) &= (1-p_0)^{\hat{\beta}} = e^{\hat{\beta} \ln(1-p_0)} \\
\frac{df}{d\hat{\beta}} &= e^{\beta \ln(1-p_0)} \ln(1-p_0) = (1-p_0)^{\beta} \ln(1-p_0) \\
\frac{d^2 f}{d\hat{\beta}^2} &= e^{\beta \ln(1-p_0)} [\ln(1-p_0)]^2 = (1-p_0)^{\beta} [\ln(1-p_0)]^2 \\
&\vdots \\
\frac{d^k f}{d\hat{\beta}^k} &= e^{\beta \ln(1-p_0)} [\ln(1-p_0)]^k = (1-p_0)^{\beta} [\ln(1-p_0)]^k \\
f(\hat{\beta}) &= (1-p_0)^{\beta} + (1-p_0)^{\beta} \ln(1-p_0) (\hat{\beta} - \beta) + (1-p_0)^{\beta} [\ln(1-p_0)]^2 \frac{(\hat{\beta} - \beta)^2}{2!} \\
&\quad + (1-p_0)^{\beta} [\ln(1-p_0)]^3 \frac{(\hat{\beta} - \beta)^3}{3!} + \dots \\
P(\widehat{p} \leq p_0) &= 1 - (1-p_0)^{\hat{\beta}} = 1 - (1-p_0)^{\beta} \ln(1-p_0) (\hat{\beta} - \beta) \\
&\quad - (1-p_0)^{\beta} [\ln(1-p_0)]^2 \frac{(\hat{\beta} - \beta)^2}{2!} - (1-p_0)^{\beta} [\ln(1-p_0)]^3 \frac{(\hat{\beta} - \beta)^3}{3!} - \dots
\end{aligned}$$

$$\text{We know } \int_0^{p_0} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = - (1-p)^{\hat{\beta}} \Big|_0^{p_0} = 1 - (1-p_0)^{\hat{\beta}} \Rightarrow P(p \leq p_0) = 1 - (1-p_0)^{\beta}$$

Hence

$$\begin{aligned}
P(\widehat{p} \leq p_0) &= P(p \leq p_0) - (1-p_0)^{\beta} \left\{ \ln(1-p_0) (\hat{\beta} - \beta) + [\ln(1-p_0)]^2 \frac{(\hat{\beta} - \beta)^2}{2!} + [\ln(1-p_0)]^3 \frac{(\hat{\beta} - \beta)^3}{3!} - \dots \right\} \\
E[P(\widehat{p} \leq p_0)] &= P(p \leq p_0) - (1-p_0)^{\beta} \left\{ \ln(1-p_0) E(\hat{\beta} - \beta) + [\ln(1-p_0)]^2 \frac{E(\hat{\beta} - \beta)^2}{2!} + [\ln(1-p_0)]^3 \frac{E(\hat{\beta} - \beta)^3}{3!} - \dots \right\} \\
E[P(\widehat{p} \leq p_0)] &= P(p \leq p_0) - (1-p_0)^{\beta} \times \left\{ \ln(1-p_0) \text{Bias}(\hat{\beta}) + \frac{[\ln(1-p_0)]^2}{2!} \text{MSE}(\hat{\beta}) + \dots \right\}.
\end{aligned}$$

References

- Amazigo, U.; Noma, M.; Bump, J.; Benton, B.; Liese, B.; Yameogo, HZ.; Seketeli, A. Onchocerciasis. In: Jamison, DT.; Feachurn, RG.; Makgoba, MW.; Bos, ER.; Baingana, FK.; Hofman, KJ.; Rogoet, KO., editors. Disease and Mortality in Sub-Saharan Africa. Second Edition. World Bank; 2006.
- Bilder CR, Tebbs JM. Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biometrical Journal*. 2005; 47:502–516. [PubMed: 16161807]
- Chaubey YP, Li W. Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binomial Probability with Sample Compositing. *Journal of Official Statistics*. 1995; 11:379–390.
- Chen P, Tebbs JM, Bilder CR. Group Testing Regression Models with Fixed and Random Effects. *Biometrics*. 2009; 65:1270–1278. [PubMed: 19210734]
- Chiang CL, Reeves WC. Statistical estimation of virus infection rates in mosquito vector populations. *American Journal of Hygiene*. 1962; 75:377–391. [PubMed: 13878878]
- Chick SE. Bayesian Models for Limiting Dilution Assay and Group Test Data. *Biometrics*. 1996; 52:1055–1062.
- Dorfman R. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*. 1943; 14:436–440.
- Gao, H. Hypothesis testing in unequal sized pool screening. University of Alabama at Birmingham; Birmingham: 2010.
- Garner FC, Stapanian MA, Yfantis EA, Williams LR. Probability estimation with sample compositing techniques. *Journal of Official Statistics*. 1989; 5:365–374.
- Gart, JJ. An application of score methodology: confidence intervals and tests of fit for one-hit curves. In: Rao, CR.; Chakraborty, R., editors. *Handbook of Statistics*. Vol. 8. Elsevier; 1991.
- Gastwirth JL, Johnson WO. Screening with cost-effective quality control: potential applications to HIV and drug testing. *Journal of the American Statistical Association*. 1994; 89:972–981.
- Hepworth G. Exact Confidence Intervals for Proportions Estimated by Group Testing. *Biometrics*. 1996; 52:1134–1146.
- 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. *Journal of the Royal Statistical Society-Series C Applied Statistics*. 2009; 58:105–121.
- Katholi CR, Toe L, Merriweather A, Unnasch TR. Determining the prevalence of *Onchocerca volvulus* infection in vector populations by polymerase chain reaction screening of pools of black flies. *Journal of Infectious Diseases*. 1995; 172:1414–1417. [PubMed: 7594692]
- Messam, LLMcV; Branscum, AJ.; Collins, MT.; Gardner, IA. Frequentist and Bayesian approaches to prevalence estimation using examples from John's disease. *Animal Health Research Review*. 2008; 9:1–23.
- Rodoni BC, Hepworth G, Richardson C, Moran JR. The use of a sequential batch testing procedure and ELISA to determine the incidence of five viruses in Victorian cut-flower Sim carnations. *Australian Journal of Agricultural Research*. 1994; 45:223–230.
- Rodriguez-Perez MA, Katholi CR, Hassan HK, Unnasch TR. Large-scale entomologic assessment of *Onchocerca volvulus* transmission by poolscreen PCR in Mexico. *American Journal of Tropical Medicine and Hygiene*. 2006; 74:1026–1033. [PubMed: 16760515]
- Schucany WR, Gray HL, Owen DB. On Bias Reduction in Estimation. *Journal of the American Statistical Association*. 1971; 66:524–533.
- Swallow WH. Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology*. 1985; 75:882–889.
- Tebbs JM, Bilder CR, Moser BK. An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics - Theory and Methods*. 2003; 32:983–995.
- Tebbs JM, Bilder CR. 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.

- Thompson KH. Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics*. 1962; 18:568–578.
- Tu XM, 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; 82:287–297.
- 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]

Figure 1.

Prevalence distributions (cdfs) for different combinations of α and $\hat{\beta}$ ($M = 100$, $n = 25$ and $T = 2$)

Figure 2.

Prevalence distributions (cdfs) for different combinations of $\alpha \leq 1$ and $\hat{\beta}$. ($M = 100$, $n = 25$ and $T = 2$)

Figure 3. Prevalence distributions (cdfs) for different combinations of $\alpha \geq 1$ and $\hat{\beta}$ ($M = 100$, $n = 25$ and $T = 2$)

Table 1

Summary of the position of the frequentist hierarchical model approach in relation to other possible methods for estimating prevalence. ($\propto$ stands for proportional)

Frequentist Approach		Bayesian Approach	
Classical	Hierarchical	Empirical	Classical
Prevalence is parameter (constant)	Prevalence is random variable	Prevalence is random variable	Prevalence is random variable
<ul style="list-style-type: none"> Find point estimate of prevalence using MLE and confidence interval (CI) 	<ul style="list-style-type: none"> Choose functional form of prevalence distribution a priori Find MLE's for parameters of prevalence distribution By plugging in those estimates obtain estimated prevalence distribution 	<ul style="list-style-type: none"> Prevalence distribution is determined by: posterior $\propto$ likelihood*prior Determine prior density based on historical data or data at hand 	<ul style="list-style-type: none"> Prevalence distribution is determined by: posterior $\propto$ likelihood*prior Specify prior density based on existing knowledge or beliefs
<ul style="list-style-type: none"> Obtain new point estimate and CI as new data becomes available disregarding previous data 	<ul style="list-style-type: none"> Obtain new prevalence distribution as new data becomes available disregarding previous data 	<ul style="list-style-type: none"> Update posterior as new data becomes available 	<ul style="list-style-type: none"> Update posterior as new data becomes available
<ul style="list-style-type: none"> Statistics are evaluated using frequentist criteria (such as Bias and MSE) 	<ul style="list-style-type: none"> Statistics are evaluated using frequentist criteria (such as Bias and MSE) 	<ul style="list-style-type: none"> Evaluate posterior distribution 	<ul style="list-style-type: none"> Evaluate posterior distribution