2011

Hierarchical and Bayesian Approaches for Estimating Prevalence Based on Pool Screening

Thomas Birkner
University of Alabama at Birmingham

HIERARCHICAL AND BAYESIAN APPROACHES FOR ESTIMATING PREVALENCE BASED ON POOL SCREENING

by

THOMAS BIRKNER

INMACULADA B. ABAN, COMMITTEE CHAIR
CHARLES R. KATHOLI, CO-CHAIR
SADEEP SHRESTHA
HEMANT K. TIWARI
O DALE WILLIAMS

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2011

HIERARCHICAL AND BAYESIAN APPROACHES FOR ESTIMATING PREVALENCE BASED ON POOL SCREENING

THOMAS BIRKNER

BIOSTATISTICS

ABSTRACT

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.

In paper 1, we present a frequentist Bernoulli-Beta hierarchical model to relax the constant prevalence assumption underlying the traditional frequentist prevalence estimation approach. This assumption is called into question when sampling from a large geographic area. Using the hierarchical model an investigator can determine the probability of the prevalence being below a pre-specified threshold value, a value at which no reemergence of the disease is expected. 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. An investigation into the least biased choice of the α parameter in the Beta$\left(\alpha,\beta\right)$ prevalence distribution leads to the choice of $\alpha = 1$.

In paper 2, we propose and evaluate the performance of a sequential Bayesian approach for the case that zero positive pools are observed in a particular year. Such observations become more likely the longer an elimination program is in place. A Bayesian approach can incorporate results from previous years and will provide a more sensible prevalence estimate compared to the estimate of zero from the traditional approach. Through simulation, we investigate the amount of data (number of years for which pool screen results are available) required such that the type of objective prior chosen does not make a significant difference with respect to the prevalence estimate. We also evaluate the accuracy of the estimates and propose three strategies to improve the performance of this Bayesian estimation approach.

In the last paper we make the case for the Bayesian estimation approach when the elimination programs are close to succeeding by presenting and comparing numerical results calculated from real data using different approaches.

Keywords: Pool Screening, Bernoulli-Beta, sequential Bayes, objective prior

DEDICATION

This research is dedicated to my parents, Waltraut and Konrad Birkner, and my wife, Elizabeth Birkner.

ACKNOWLEDGMENTS

I want to thank Dr. Charles Katholi for introducing me to the pool screening topic and for providing constant guidance throughout this research. His office door was always open and our conversations encompassed everything from his bike ride through Germany way back then to the lemurs he and his wife visited. I want to equally thank Dr. Inmaculada “Chichi” Aban for her talent to keep the research on track, her detailed and superfast reviews and her flexibility in accommodating to another unscheduled meeting. I am indebted as well to Drs. O Dale Williams, Sadeep Shrestha and Hemant Tiwari for providing helpful feedback on the manuscripts. Last but not least I am very thankful for the financial support throughout my graduate studies and dissertation research from the National Heart, Lung and Blood Institute.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

Table *Page*

EVALUATION OF A FREQUENTIST HIERARCHICAL MODEL TO ESTIMATE PREVALENCE WHEN SAMPLING FROM A LARGE GEOGRAPHIC AREA USING POOL SCREENING

BAYESIAN APPROACHES FOR ESTIMATING PREVALENCE BASED ON POOL SCREENING WHEN OBSERVING ZERO POSITIVE POOLS

THE CASE FOR A BAYESIAN APPROACH TO MONITORING

LIST OF FIGURES

Figure *Page*

EVALUATION OF A FREQUENTIST HIERARCHICAL MODEL TO ESTIMATE PREVALENCE WHEN SAMPLING FROM A LARGE GEOGRAPHIC AREA USING POOL SCREENING

BAYESIAN APPROACHES FOR ESTIMATING PREVALENCE BASED ON POOL SCREENING WHEN OBSERVING ZERO POSITIVE POOLS

THE CASE FOR A BAYESIAN APPROACH TO MONITORING

INTRODUCTION

When to Use Pool Screening

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, for instance, positive versus negative, infected versus non-infected, polluted versus not polluted. The method is currently applied in various fields including environmental studies, safety of blood products, disease screening in humans, animals or plants. In an environmental application, soil samples from different sampling locations within a larger area where contamination is suspected are combined and tested in order to see whether they surpass a certain threshold value. Blood samples from different blood donors are tested together for certain diseases. A pool of a number of black flies is homogenized and tested for the DNA of a parasite. Leaves of different plants are grouped and tested for a plant virus. The purpose of group testing is either to detect each positive or "defective" member in the sample or to estimate the prevalence of those members in the population. If the goal is detecting each "defective" member, a re-testing design with decreasing group sizes is usually employed. 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 application is the disease prevalence estimation in arthropod vector populations such as mosquitoes or black flies. These vectors transmit viral and parasitic diseases (e.g. West Nile virus, St. Louis encephalitis, Onchocerciasis). As a concrete example 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 prevalence 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 stops further transmission of the disease from the treated human.

In order to evaluate the progress of control and elimination programs, pools of black flies are collected at different locations within a broader area. Before the advent of Polymerase Chain Reaction (PCR) methodology in the 1990s (for PCR method in black flies see: Katholi *et al.*, 1995), each fly would have to be dissected and evaluated separately for the presence of microfilariae under a microscope. This method provides an unbiased estimate of the disease prevalence. However, when the disease prevalence is very small this method is inefficient, e.g. 1000 flies might have to be evaluated before the first infected one is found. If 10 pools consisting of 100 flies each are formed, then only 10 PCR assays will have to be performed and one of these pools will test positive. A positive pool implies that it contains at least one infected insect. PCR cannot determine the number of infected insects in a pool.

Unresolved and Ignored Statistical Issues in Pool Screening

Non-constant prevalence

Traditional frequentist approaches in prevalence estimation assume 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 across a large geographic area which could also lead to locally different values of the prevalence.
3. There are different vector sub-species exhibiting varying efficiencies in transmitting the parasite.

Hence, instead of treating the prevalence as a constant it might be more realistic to view it as a random variable having a distribution. This distribution has to be estimated given the data, which consist of pool size (n), number of pools (m) and number of positive pools (t). By using a hierarchical model, the non-constant prevalence assumption can be incorporated within the frequentist framework. The advantage of estimating a distribution for the prevalence over a mere point estimate and/or confidence interval lies in the ability to make statements about the probability of observing any specific prevalence value; it will be possible to state the probability that the prevalence is below a certain threshold value. This threshold could be established as the value where further disease transmission ceases. Practitioners in the field appear to still prefer Frequentist over Bayesian methods. The hierarchical model developed and evaluated in paper 1 is a way of attaining some of the advantages otherwise only provided in the Bayesian framework.

Zero positive pools

When control and elimination programs have been in place for awhile and have led to a decrease in infection prevalence in the vector population, it is not uncommon in practice to encounter zero positive pools even after an extensive collection of vectors (for example see Yameogo *et al.* 1999, Guevara *et al.* 2003, Rodriguez-Perez *et al*. 2006). This does not imply that the prevalence is truly zero, but that the prevalence has fallen to a level where zero positive pools are not unlikely to be obtained. The question then arises: what is a good estimate of the true infection prevalence?

Both the traditional maximum likelihood method as well the hierarchical model approach mentioned above fail in this situation. The Maximum Likelihood Estimate (MLE) for the prevalence is zero. The distribution in the hierarchical model approach is degenerate.

It is imperative that the success of the control/elimination program be evaluated reliably. It is crucial to know whether the parasite has been eliminated or at least suppressed to a level where no recurrence is expected. These facts are needed to decide whether to continue the mass drug administration or end the program. After a program ends post treatment monitoring will ensue and be challenged by very low prevalence numbers. The need for methods capable of handling the observation of zero positive pools, without providing unrealistic prevalence estimates is great. Bayesian approaches that incorporate knowledge of previous samplings from the same or similar geographic area into the current prevalence estimation might offer a solution. Several of such approaches are proposed and investigated in paper 2. A numerical comparison using OEPA data between the traditional and the proposed Bayesian estimation techniques is presented in paper 3.

Literature Review

Frequentist contributions

The pool screen approach dates back to the beginning of the 20th century (for examples see Watson (1936) and Dorfman (1943)). The following assumptions are made in the traditional model:

1. The disease prevalence (*p*) is constant over the entire sampling area.
2. The screening assay/test has perfect sensitivity and specificity.
3. The result of a test of pool of size n_i is $x_i = \{0,1\}$ with

$$f_{x_i}(x \mid n_i, p) = \begin{cases} (1-p)^{n_i} & \text{when x = 0} \\ 1-(1-p)^{n_i} & \text{when x = 1} \end{cases}$$

The prevalence (*p*) is estimated by Maximum Likelihood Estimation given the m data points $(n_i, x_i), i = 1,...,m.$

Widely cited as the original paper in the area of pool screening is Dorfman (1943). His objective was to develop an efficient procedure to detect “defective” members of large populations. Dorfman (1943) illustrated his approach with data gathered by the US Public Health Service with the purpose of “weed[ing] out all syphilitic men called up for induction” into the military. He derived expressions for the expected number of chemical analysis to be conducted and for the expected relative cost. He presented a table giving the optimum group size for different values of prevalence. Dorfman (1943) concluded that when the prevalence is sufficiently small then it is more economical to obtain a measure on a group than on the individual units.

The foundational work for prevalence estimation was done by Chiang and Reeves (1962) and Thompson (1962). Chiang and Reeves developed a method to estimate the viral infection rate in a mosquito population. Thompson was interested in the proportion of a leafhopper population capable of transmitting aster-yellows virus. Both authors found the MLE for *p* given equal pool sizes, using our notation, as

$$\hat{p} = 1 - \left(\frac{m - T}{m} \right)^{\frac{1}{n}}, T = \text{number of positive pools.}$$

Thompson (1962) noted that $\hat{p}$ is a biased estimator of p given any group size greater than 1. Considering the first order approximation of the bias of the MLE provided by Tu *et al*. (1995) or alternatively by Barker (2000) again in our notation:

$$E\left(\hat{p} - p\right) = \frac{\left(n-1\right)\left(1-p\right)^{1-n}\left(1-\left(1-p\right)^{n}\right)}{2mn^{2}},$$

we recognize that for $n > 1$ the MLE is on average an overestimate. For low values of p and n, even for a small number of pools, the bias in $\hat{p}$ will be small (Thompson 1962).

As *m* approaches infinity $\hat{p}$ is distributed asymptotically normal and converges in probability to *p* (Thompson 1962). The asymptotic variance of $\hat{p}$ was found as

$$\lim_{m \to \infty} V\left(\hat{p}\right) = \frac{1-\left(1-p\right)^{n}}{mn^{2}\left(1-p\right)^{n-2}} \text{ (Thompson, 1962).}$$

Chiang and Reeves (1962) derived an exact confidence interval $\left(p_L, p_U\right)$ for p from the confidence limits for π, the binomial probability that a pool will be positive:

$$p_L = 1 - \left[1 - \pi_1\right]^{\frac{1}{n}}$$
$$p_U = 1 - \left[1 - \pi_2\right]^{\frac{1}{n}}.$$

π_1 and π_2 are determined by the desired α level. These intervals can be viewed as Clopper-Pearson confidence intervals (compare to Katholi *et al*. 1995). Thompson (1962) presented the following approximate $100\left(1-\alpha\right)\%$ confidence interval for p:

$$\hat{p} \pm t_\alpha \left[\sum_{i=1}^{m} \left[\left(1-\hat{p}\right) - \left(\frac{i}{m}\right)^{\frac{1}{n}} \right]^2 \binom{m}{i} \left[\left(1-\hat{p}\right)^n \right]^i \left[1 - \left(1-\hat{p}\right)^n \right]^{m-i} \right]^{\frac{1}{2}}.$$

With respect to the determination of pool size, Chiang and Reeves (1962) suggested to choose *n* such that there are some positive and some negative pools among the *m* pools. Given an upper limit on *n* = 100, they derived the following:

$$n = \frac{\log\left(1/2\right)}{\log\left(1-p\right)}.$$

This follows from the demand that any pool have a 50% chance of being positive. The formula is only reasonable for $n \le 100,$ otherwise the suggested pool size is far too large when *p* is small. We notice that some prior knowledge about *p* is necessary to determine the pool size.

Thompson (1962) derived an approximate formula for the optimal pool size given a constant number of pools (m) by minimizing the asymptotic variance of $\hat{p},$ as

$$n = \frac{1.5936 - p}{p}.$$

Note that when p is small n becomes very large, for instance suppose p = 0.0001 then n required is approximately 15,935.

Many other contributions addressing the topic of pool screening have been made. Bhattacharyya *et al*. (1979) developed a finite population estimator for p. Walter *et al.* (1980) incorporated pools of different size in the estimation of p. Swallow (1985) provided extensive tables and graphs to illustrate bias, variance, and mean-squared-error (MSE) properties of $\hat{p}$. Swallow elaborated on the choice of the pool size n considering relative cost for the cases when the limiting factor is a) the number of pools (tests) or b) the total number of vectors. Swallow (1985) also provided an approximate confidence interval for p,

$\hat{p} \pm z\left[\widehat{V(\hat{p})}\right]^{\frac{1}{2}}$, with $\widehat{V(\hat{p})} = \left[1-(1-\hat{p})^n\right] \Big/ \left[mn^2(1-\hat{p})^{n-2}\right]$ and z being standard normal.

Burrows (1986) presented an alternative to the MLE estimator for p with smaller bias and MSE:

$$\tilde{p} = 1 - \left[(m-T+a)/(m+b)\right]^{\frac{1}{n}}, a = b = (n-1)/2n .$$

The value of a and b was chosen, so as to eliminate the dominant term of the bias when expanded as a power series. Katholi *et al.* (1995) provided approximate confidence interval formulae for p based on the F distribution. Barker (2000) derived the following expansion for $E(\hat{p})$:

$$
\begin{aligned}
E\left(\hat{p}\right) = p + \frac{1}{m}\left[\frac{n-1}{2!n^2}\left\{\left(w-w^2\right)-\left(\frac{1}{n}-2\right)\left(w^2-w^3\right)+\left(\frac{1}{n}-2\right)\left(\frac{1}{n}-3\right)\frac{1}{4}\left(w^3-w^4\right)+\ldots\right\}\right]+ \\
\frac{1}{m^2}\left[-\frac{\left(n-1\right)}{n}\left(\frac{1}{n}-2\right)\frac{1}{3!}\left\{\left(w-3w^2+3w^3\right)+\left(\frac{1}{n}-3\right)\frac{1}{4}\left(7w^2-18w^3+11w^4\right)+\ldots\right\}\right]+ \\
\frac{1}{m^3}\left[\frac{\left(n-1\right)}{n}\left(\frac{1}{n}-2\right)\left(\frac{1}{n}-3\right)\frac{1}{4!}\left\{\left(w-7w^2+12w^3-6w^4\right)+\ldots\right\}\right]
\end{aligned}
$$

where $w = 1-\left(1-p\right)^n$.

This expansion can be used to evaluate bias and mean squared error of *p*.

Hepworth (1996) as well as Tebbs and Bilder (2004) provided a review of different interval estimators for p. Hepworth (1996) investigated exact confidence intervals in the unequal group size case. In particular he compared “Sterne Intervals”, in which outcomes are ordered according to their probability of occurrence to intervals where outcomes are ordered by the magnitudes of their MLE’s. The rejection region with significance level α for the Sterne interval “includes all the least probable outcomes such that their combined probability is no greater than α” (Hepworth 1996). The Sterne method can produce empty intervals and disjoint intervals. Hepworth (1996) concluded that those issues are avoided when outcomes are ordered by their MLE’s.

Tebbs and Bilder (2004) evaluated the Wald, the Thompson, a variance-stabilizing, the Clopper-Pearson, the Blaker, the Mid-P and the Wilson Score interval in terms of coverage probability and mean length. They recommended the use of the Wilson or Blaker intervals.

Katholi and Unnasch (2006) drew attention to different sampling models (binomial, negative binomial and hypergeometric) and their underlying assumptions.

Tebbs and McCann (2007) examined large-sample, likelihood-based hypothesis tests in the context of stratified group testing. They concluded that likelihood ratio tests are the most appropriate, especially when the sample sizes n_i are small.

Hepworth and Watson (2009) investigated techniques to correct for the bias in the MLE for pools of different size. They compared Burrows's method (Burrows 1987) to a more general bias adjustment described by Gart (see Gart, 1991). Burrows's method performs well in one-stage procedures, as does Gart's in fixed multistage procedures. For sequential procedures, a numerical correction was proposed (Hepworth and Watson, 2009).

Gao (2010) proposed an exact two-sided hypothesis test procedure based on the number of positive pools in the unequal pool size case. He presented "modified versions of the likelihood-ratio, Wald's and Score tests where simulated quantiles are used instead of the quantiles based on the standard asymptotic distribution to obtain the rejection region for each test" (Gao 2010). Gao (2010) also investigated the likelihood ratio test procedure for the one-sided hypothesis test for unequal pool sizes.

Bayesian contributions

Before we describe some of the Bayesian contributions to the field of pool screening, we will outline some of the major differences between the Frequentist and Bayesian methods. In Bayesian statistics there are no parameters (constants). Everything we measure is regarded as a random variable, which follows a probability density function. Bayesians, for instance, would assign a probability to the speed of light being of a certain value. For Frequentists the speed of light is one specific number with probability equal to 1.

Frequentists assign a probability to an estimate (statistic) of a parameter by means of deriving or assuming a sampling distribution. Bayesians consider parameters as random variables about which probability statements can be made. Bayesians incorporate previous knowledge or beliefs (prior) into their derivation of the probability distribution (posterior). Frequentists consider only the data at hand by means of the likelihood function. Bayesian credibility intervals are different from frequentist confidence intervals in that they are expressing true probabilities of the random variable taking a value within the interval. Frequentist confidence intervals (CI) are meaningful only when considered in the context of repeating an experiment a large number of times. There we state our confidence that the CI will contain the parameter value N*(1-α) times out of the total N times the experiment is performed. For one particular experiment there is no way of knowing whether the CI contains the true value of the parameter.

Bayesian approaches to prevalence estimation start to appear in the literature in the early 1990s. Whereas much of the literature considers three random variables – prevalence, sensitivity, and specificity – we will focus on the developments with respect to prevalence, since the tests used in our application (arthropod vector control) are assumed to have perfect sensitivity and specificity.

Boswell *et al.* (1992) in an environmental application studied an empirical Bayes procedure to predict the prevalence rate and the corresponding composite sample size. They were interested in a cost-efficient approach to classify samples as polluted or not polluted. Boswell *et al.* (1992) used composite samples initially in each sampling stage, but performed separate tests for each individual sample from composite samples that were polluted. The experimental unit in their classification approach is the individual sample, while our experimental unit is the composite sample or pool itself. They chose the conjugate beta$\left(\alpha,\beta\right)$ prior on p and hence used the standard Beta-binomial model. There procedure works as follows:

1. Choose an initial beta prior on p with parameter α_1 and β_1.
2. First sampling stage (a. test composite samples, b. test individual units where needed). Result: X_1 individual samples out of n_1 individual samples are classified as polluted.
3. The posterior distribution of p given X_1 is a beta distribution with parameters $\alpha_2=\alpha_1+X_1$ and $\beta_2=\beta_1+n_1-X_1$. The expected value of $p_2=\dfrac{\alpha_2}{\alpha_2+\beta_2}$
4. Second sampling stage (a. test composite samples, b. test individual units where needed). Proceed as before and move to next sampling stage.

Chaubey and Li (1995) compared the maximum likelihood estimator (MLE) of a binomial probability based on sample compositing with a Bayes estimator. First they derived the Bayes estimator using a Beta$\left(\alpha,\beta\right)$ prior for p, the population proportion, for the equal pool size case:

$$\hat{p}=E\left(p\,|\,T\right)=\frac{\sum_{j=0}^{T}\binom{T}{j}(-1)^{j}\,\mathrm{B}\left(\alpha+1,nj+nm-nT+\beta\right)}{\sum_{j=0}^{T}\binom{T}{j}(-1)^{j}\,\mathrm{B}\left(\alpha,nj+nm-nT+\beta\right)},$$

where *T* is the number of positive pools, *n* is the pool size and *m* is the number of pools.

Secondly they evaluated the estimators using Bayesian (Bayes relative efficiency) and Frequentist (relative bias and relative efficiency) criteria. They concluded that the MLE is inferior to the Bayes estimator under all three criteria.

Chick (1996) in a paper titled “Bayesian Models for Limiting Dilution Assay and Group Test Data” expanded the Bayesian prevalence estimation approach to the unequal pool size case choosing also a Beta prior. He compared posterior probability distributions for three different sets of α and β $(\alpha = \beta = 1; \alpha = 2, \beta = 3; \alpha = 10, \beta = 15)$. The three posterior distributions obtained were not very different from each other due to the relatively large amount of data available. Chick (1996) conjectured that this would be the case for all Beta priors with small values (<50) of $\alpha + \beta$.

Tebbs *et al.* (2003) developed an empirical Bayes procedure to estimate *p* using a $\text{Beta}(1, \beta)$ prior distribution. Their estimate of *p* is the mean of the empirical posterior $f_{P|T}\left(p \mid t, \hat{\beta}\right) = \dfrac{f_{T,P}\left(t, p \mid \hat{\beta}\right)}{f_T\left(t \mid \hat{\beta}\right)}$. They found $\hat{p}_{eb} = 1 - \dfrac{\Gamma\left(m + \hat{\beta}/n + 1\right)\Gamma\left(m - t + \hat{\beta}/n + 1/n\right)}{\Gamma\left(m - t + \hat{\beta}/n\right)\Gamma\left(m + \hat{\beta}/n + 1 + 1/n\right)}$.

Tebbs *et al.* (2003) point out that for $T = 0$ (no positive pools) or $T = m$ (all positive pools) $f_T\left(t \mid \hat{\beta}\right)$ cannot be maximized and no $\hat{p}_{eb}$ can be computed.

They showed that their empirical Bayes estimator outperforms the traditional maximum likelihood estimator with respect to relative bias and relative efficiency for small group sizes and small *p*. Tebbs *et al.* (2003) also derived an empirical credible interval for *p*:

$$p_L = 1 - \left(\mathrm{B}_{1-\alpha/2; m-t+\hat{\beta}/n, t+1}\right)^{1/s}, \; p_U = 1 - \left(\mathrm{B}_{\alpha/2; m-t+\hat{\beta}/n, t+1}\right)^{1/s},$$

where $\mathrm{B}_{y;a,b}$ denotes the γ quantile of the two parameter Beta distribution.

Bilder and Tebbs (2005) adapted the empirical Bayes estimator from above to fit the multiple-vector-transfer designs method. This method entails moving several vectors, for instance leafhoppers, from a diseased plant to a healthy plant. Researchers are interested in whether the healthy plant develops disease symptoms, indicating that at least one of the vectors carried the disease, for instance a plant virus.

One issue we have with the approach used in both papers (2003 and 2005) is the fact that the *currently* observed data are used to estimate the hyperparameter β. The data are then used again together with $\hat{\beta}$ to estimate p. This amounts to "double dipping" into the data. Tebbs *et al.* (2003) and Bilder and Tebbs (2005) argue that their approach is an improvement over (arbitrarily) choosing the values of model hyperparameters a priori. We will follow up on this by suggesting a) the use of a noninformative prior in the beginning and b) the use of historical data as it becomes available (see papers 2 and 3).

Messam *et al*. (2008) in an animal health research application suggested the following mixture prior distribution to account for the possibility that p is truly zero:

$$\begin{aligned}&p \sim \text{Beta}\left(\alpha,\beta\right) \text{ with probability} = \lambda \\ &p = 0 \text{ with probability} = 1-\lambda,\end{aligned}$$

where λ denotes the probability that the herd is infected. Here the herd represents the sampling universe from which a number of animals is randomly selected and pooled into groups. The test is performed on the groups and not the individual animals. The determination of λ appears to be the Achilles heel of this approach.

EVALUATION OF A FREQUENTIST HIERARCHICAL MODEL TO ESTIMATE PREVALENCE WHEN SAMPLING FROM A LARGE GEOGRAPHIC AREA USING POOL SCREENING

by

THOMAS BIRKNER, INMACULADA B. ABAN, CHARLES R. KATHOLI

In preparation for Communications in Statistics – Theory and Methods

Format adapted for dissertation

1. INTRODUCTION

Pool screening, also known as group testing or composite sampling, is a method that combines individual units into several groups, or pools of a certain size. The pools are then tested in place of the individual units. 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). 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)). In an environmental application, soil samples from different sampling locations within a larger area where contamination is suspected are combined and tested in order to see whether they surpass a particular threshold value. Blood samples from different blood donors are tested together for certain diseases. A pool of a number of black flies is homogenized and tested for the DNA of a parasite. Leaves of different plants are grouped and tested for a plant virus. 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. If the goal is detecting each defective member, a sequential/re-testing design with decreasing group sizes is usually employed. The rationale behind the use of pool screening versus the testing of each individual unit 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 disease 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 we will 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 from the treated human.

In order to evaluate the progress of control and elimination programs, pools of insects are collected at different locations within a broader area. Before the advent of Polymerase Chain Reaction (PCR) methodology in the 1990s (for PCR method in black flies see: Katholi *et al.* (1995)), each insect would have to be dissected and evaluated separately for the presence of the parasite under a microscope. This method provides an unbiased estimate of the disease prevalence. However, when the disease prevalence is very small this method is inefficient, e.g. 1000 insects might have to be evaluated before the first infected one is found. If 10 pools consisting of 100 insects each are formed, then only 10 PCR assays will have to be performed and one of these pools will test positive. A positive pool implies that it contains at least one infected insect. This method cannot determine the number of infected insects in a pool.

Traditional frequentist approaches in prevalence estimation assume 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. 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 might 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). In a classical Bayesian approach a prior distribution for the prevalence is chosen and then updated by the observed data through the likelihood function. Inferences are based on the posterior distribution, which can be updated as more data becomes available. Practitioners in the field appear to still prefer the frequentist method and are reluctant to adopt the Bayesian paradigm. The frequentist method does not involve a prior density and inferences are based on the maximum likelihood estimate (MLE) of the prevalence. By using a hierarchical model, the non-constant prevalence assumption can be incorporated within the frequentist framework. The hierarchical model uses the data at hand and maximum likelihood estimates 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 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. There is no updating; previous data is disregarded. 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 make statements about the probability of observing any specific prevalence value; it will be possible to state the probability that the prevalence is below a pre-specified threshold value. This threshold could be established as the value where further disease transmission ceases.

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
- Find point estimate of prevalence using MLE and confidence interval (CI)	- 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	- Prevalence distribution is determined by: posterior $\propto$ likelihood*prior - Determine prior density based on historical data or data at hand	- Prevalence distribution is determined by: posterior $\propto$ likelihood*prior - Specify prior density based on existing knowledge or beliefs
- Obtain new point estimate and CI as new data becomes available disregarding previous data	- Obtain new prevalence distribution as new data becomes available disregarding previous data	- Update posterior as new data becomes available	- Update posterior as new data becomes available
- Statistics are evaluated using frequentist criteria (such as Bias and MSE)	- Statistics are evaluated using frequentist criteria (such as Bias and MSE)	- Evaluate posterior distribution	- Evaluate posterior distribution

The objective of this paper is to derive and evaluate a frequentist hierarchical model - which accounts for different prevalence values within a broader geographic area - for the estimation of infection prevalence. This model will allow answering the question about the probability of the prevalence being 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 (which assumes constant prevalence and is not equipped to answer the threshold probability question) without "going Bayesian".

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 $\left(X=0\right)$ or positive $\left(X=1\right)$ is distributed as Bernoulli $\left(1-\left(1-p\right)^{n}\right)$, where p is the probability of an insect being infected (carrier of virus or parasite).

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

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

$$\begin{aligned} X \mid P &\sim Bernoulli\left(1-\left(1-p\right)^{n}\right)=\begin{cases}\left(1-p\right)^{n} \text{ when x = 0} \\ 1-\left(1-p\right)^{n} \text{ when x = 1}\end{cases} \\ P &\sim Beta\left(\alpha,\beta\right)=\frac{p^{\alpha-1}\left(1-p\right)^{\beta-1}}{\mathrm{B}\left(\alpha,\beta\right)} \end{aligned} \quad (2)$$

The $Beta\left(\alpha,\beta\right)$ pdf is defined as

$$f\left(p\middle|\alpha,\beta\right)=\frac{1}{\mathrm{B}\left(\alpha,\beta\right)}p^{\alpha-1}\left(1-p\right)^{\beta-1},0<p<1,\alpha>0,\beta>0,$$

and $\mathrm{B}(\alpha,\beta)$ denotes the beta function, $\mathrm{B}(\alpha,\beta)=\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,dp$.

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:

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

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

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

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

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

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

and where $\Gamma\left(\lambda\right)$ denotes the gamma function, $\Gamma\left(\lambda\right)=\int_0^\infty x^{\lambda-1}e^{-x}dx.$

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

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

$$\hat{\beta}=\frac{Mn}{T}-n, \tag{6}$$

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

Proof: See Appendix A.

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 $\hat{\beta}$

Theorem II (Consistency of $\hat{\beta}$): Let $X_1, X_2 ...$ 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 $VarX_i = \frac{\beta n}{(\beta+n)^2} < \infty$. Let $\overline{T}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$. Then $\hat{\beta}$ is a consistent estimator of β .

Proof: By the strong law of large numbers $\overline{T}_m$ converges almost surely to μ . Almost sure convergence implies convergence in probability, which implies consistency. Since $\overline{T}_m$ converges in probability, then $\hat{\beta} = f\left(\overline{T}_m\right) = \frac{mn}{m\overline{T}_m} - n = \frac{mn}{T_m} - n$ converges in probability to $f\left(E\left(\overline{T}\right)\right) = \beta$ 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 $E\left(\hat{\beta}\right)=nmE\left(\frac{1}{T}\right)-n$. To find $E\left(\frac{1}{T}\right)$ we use the Taylor series expansion. We define $f\left(T\right)=\frac{1}{T}$ and let T be a random variable such that $E\left(T\right)=\mu_T$. We expand $f\left(T\right)$ about μ_T.

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

Result A (Bias)

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

Proof: See Appendix B1.

Result B (Mean Squared Error)

$$\begin{aligned} E(\widehat{\beta}-\beta)^2=\frac{(\beta+n)^2\beta}{mn}+\frac{\left(\beta+n\right)^2\beta\left(7\beta+2n\right)}{\left(mn\right)^2}+\frac{\left(\beta+n\right)^2\beta\left(38\beta^2+28\beta n+3n^2\right)}{\left(mn\right)^3} \\ +O\left(\frac{1}{\left(mn\right)^4}\right) \end{aligned} \tag{8}$$

Proof: See Appendix B2.

Result C (Variance)

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

$$Var\left(\hat{\beta}\right) \approx \frac{(\beta+n)^2\beta}{mn} + \frac{\left(\beta+n\right)^2 \beta\left(6\beta+2n\right)}{\left(mn\right)^2} + \frac{\left(\beta+n\right)^2 \beta\left(34\beta^2+26\beta n+3n^2\right)}{\left(mn\right)^3} \quad (9)$$

Proof: See Appendix B3.

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

Furthermore, we note that (7) has the form

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

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 $\hat{p}$ by Hepworth and Watson (2009). An alternative to the Jackknife for finding a bias corrected estimate $\breve{\beta}$ is to subtract the bias terms up to the desired order of magnitude from the estimate $\hat{\beta}$

$$\breve{\beta} = \hat{\beta} - Bias\left(\hat{\beta}\right). \quad (11)$$

3.2. Properties of "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

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

Proof: See Appendix C1.

Using the results of Theorem IV, we can construct an estimate of a (1 – α)100% credibility interval for p. Define a (1 – α)100% credibility interval as the area between two quantile points under the estimated prevalence distribution requiring equal tail areas. The expected values of the lower and upper bounds for the credibility interval for p given $\hat{\beta}$ denoted by $p_L = p_L\left(\hat{\beta}\right), p_H = p_H\left(\hat{\beta}\right)$ 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 – α)100% credibility interval. The method is described as follows. Given the general expressions $p_L = 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}$, $p_H = 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}$, we expanded about β and took the expectation to get

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

$$E_\beta\left(p_H\right) = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} Bias - \left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4} + \frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^3} \right) \frac{1}{2} MSE + \ldots \tag{14}$$

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

$$E_\beta\left(p_L\right) \approx 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right) \frac{\left(1 + \frac{n}{\beta}\right)}{mn} - \left(\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left(\ln\left(1 - \frac{\alpha}{2}\right)\right)^2}{2\beta^3} + \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2} \right) \frac{\left(\beta + n\right)^2}{mn} \tag{15}$$

$$E_\beta\left(p_H\right) \approx 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} + \left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right) \frac{\left(1 + \frac{n}{\beta}\right)}{mn} - \left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{2\beta^3} + \frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(\frac{\alpha}{2}\right)}{\beta^2} \right) \frac{\left(\beta + n\right)^2}{mn} \tag{16}$$

For more details of the above results, see Appendix C2. We note that the expressions for $E_\beta\left(p_L\right), E_\beta\left(p_H\right)$ 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 $\widehat{P\left(p \le p_0\right)}$ as an Estimator of $P\left(p \le p_0\right)$

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

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

Using a Taylor series expansion about β we found:

$$E\left[\widehat{P\left(p \le p_0\right)}\right] = P\left(p \le p_0\right) - \left(1-p_0\right)^{\beta}\left[\ln\left(1-p_0\right) Bias\left(\hat{\beta}\right) + \frac{\left[\ln\left(1-p_0\right)\right]^2}{2!} MSE\left(\hat{\beta}\right) + \ldots\right] \quad (18)$$

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

3.4. Comparison With Usual Pool Screening Model

To compare the estimator of prevalence given by $\widehat{E(p)} = \frac{1}{1+\hat{\beta}}$ from the hierarchical model with the regular pool screen estimate $\hat{p} = 1 - \left(1 - \frac{T}{m}\right)^{\frac{1}{n}}$ (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:

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

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

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

The relative difference $\left|\frac{\hat{p} - \widehat{E(p)}}{\widehat{E(p)}}\right|$ is $O\left(\frac{T}{m}\right)$ when m is large. Thus if the number of pools is large the two estimators are practically the same. This result is independent of the size of p. The comparison considers only the point estimate of prevalence. The hierarchical model will be still preferable if $\widehat{P(p \le p_0)}$ 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 pre-valence distribution (Beta$\left(\alpha,\hat{\beta}\right)$) 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.

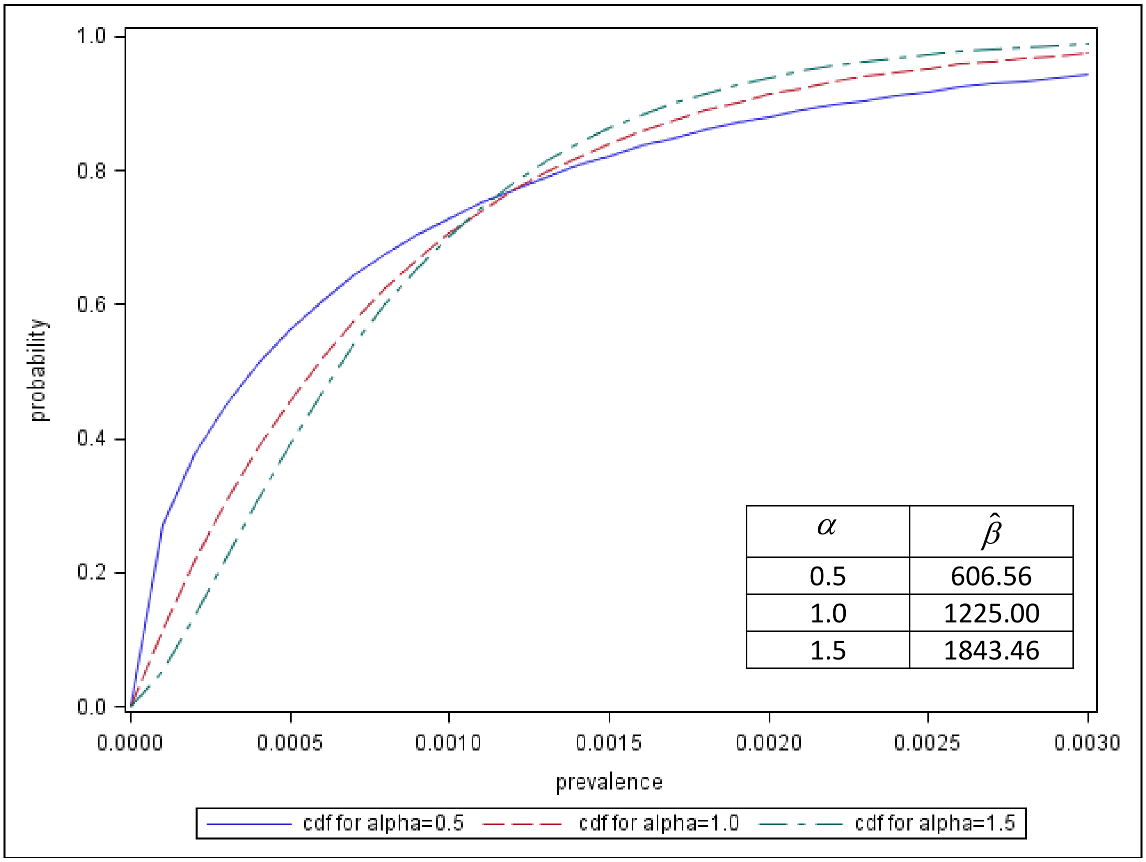
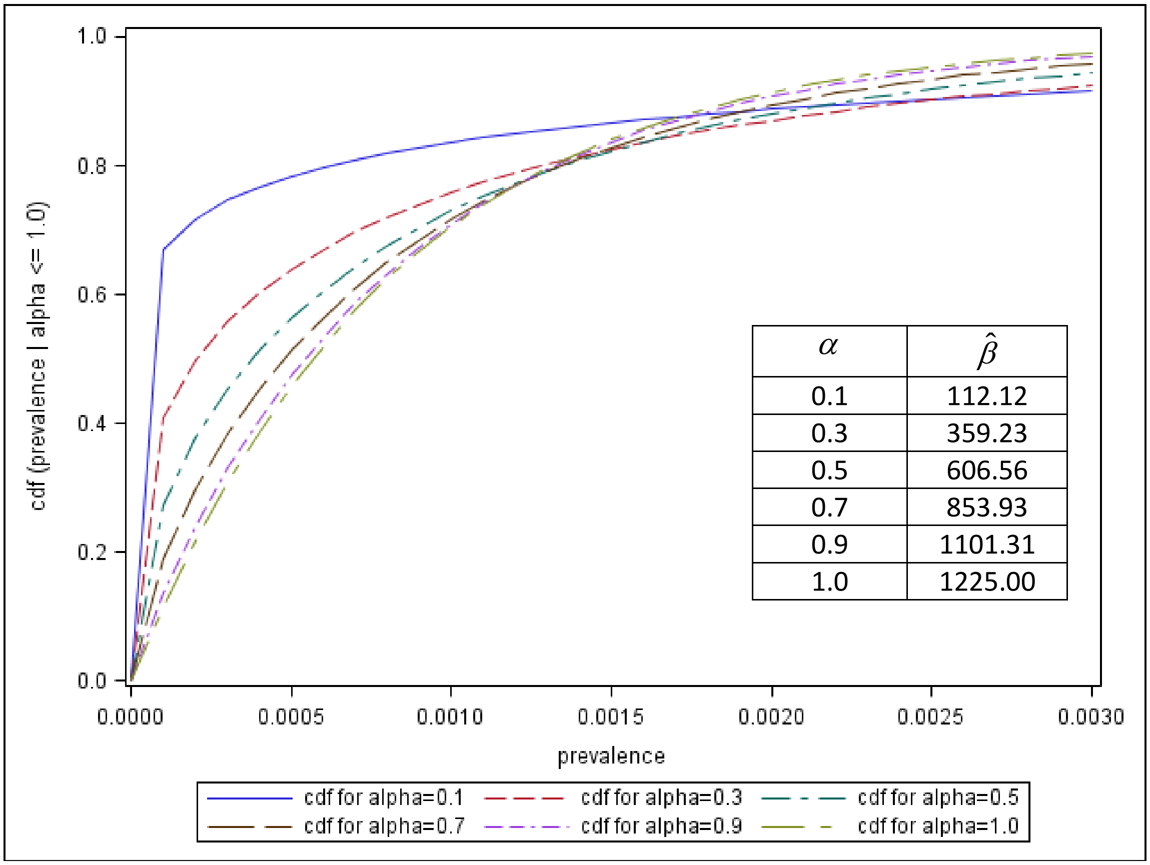


Figure 1. Prevalence distributions (cdfs) for different combinations of α and $\hat{\beta}$**.** $\left(M = 100, n = 25 \text{ and } T = 2\right)$

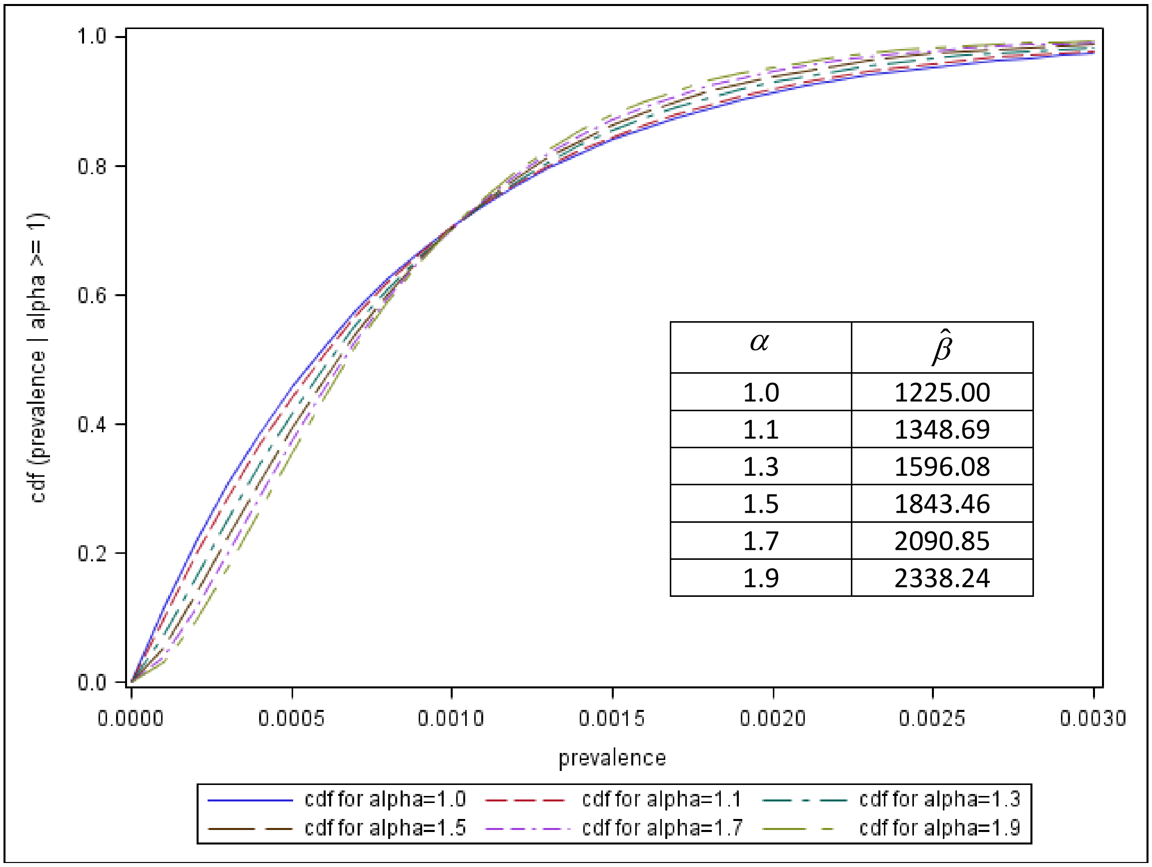
Figure 2 displays the cdfs for a range of α values less than 1 ($\alpha = 1$ is included as a reference). The closer to zero α is chosen the more mass of the density is concentrated near zero. Hence for $\alpha < 1$ we observe a bias towards zero. This bias becomes larger the smaller α becomes.



α	$\hat{\beta}$
0.1	112.12
0.3	359.23
0.5	606.56
0.7	853.93
0.9	1101.31
1.0	1225.00

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

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 $\alpha > 1$.



α	$\hat{\beta}$
1.0	1225.00
1.1	1348.69
1.3	1596.08
1.5	1843.46
1.7	2090.85
1.9	2338.24

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

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 affords the estimation of $P\left(p \le p_0\right)$ without "going Bayesian". Due to a lack of sufficient independent information in the likelihood function it was necessary to preset one of the parameters of the Beta$\left(\alpha,\beta\right)$ distribution. An investigation was undertaken in the choice of the parameter α in the Beta$\left(\alpha,\beta\right)$ 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 that the intermediate estimator $\left(\hat{\beta}\right)$ and the estimators of ultimate interest $\left(p_L, p_H, \widehat{P\left(p \le p_0\right)}\right)$ have reasonable statistical properties using standard frequentist criteria – first and foremost they are consistent estimators. The estimators $\hat{\beta}, p_L, p_H,$ and $\widehat{P\left(p \le p_0\right)}$ 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.

REFERENCES

Amazigo, U. , Noma, M., Bump, J., et al. (2006). Onchocerciasis. In Disease and Mortality in Sub-Saharan Africa. (eds. D.T. Jamison, R. G. Feachurn, M. W. Makgoba, et al.), Second Edition.

Bilder, C. R., and Tebbs, J. M. (2005). Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biom J* **47**, 502-516.

Chaubey, Y. P., and Li, W. (1995). Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binomial Probability with Sample Compositing. *Journal of Official Statistics* **11**, 379-390.

Chiang, C. L., and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. *American Journal of Hygiene* **75**, 377-391.

Chick, S. E. (1996). Bayesian Models for Limiting Dilution Assay and Group Test Data. *Biometrics* **52**, 1055-1062.

Dorfman, R. (1943). The Detection of Defective Members of Large Populations. The Annals of Mathematical Statistics 14, 436-440.

Gao, H. (2010). Hypothesis testing in unequal sized pool screening. University of Alabama at Birmingham, Birmingham.

Garner, F.C., Stapanian, M.A., Yfantis, E.A. and Williams, L.R. (1989): Probability estimation with sample compositing techniques. Journal of Official Statistics 5, 365-374.

Gart, J.J. (1991). An application of score methodology: confidence intervals and tests of fit for one-hit curves. In Handbook of Statistics (eds. C.R. Rao and R. Chakraborty), vol. 8, pp. 395-406. Amsterdam: Elsevier.

Gastwirth, J.L. and Johnson, W.O. (1994). Screening with cost-effective quality control: potential applications to HIV and drug testing. Journal of the American Statistical Association 89, 972-981.

Hepworth, G., and Watson, R. (2009). Debiased estimation of proportions in group testing. Journal of the Royal Statistical Society-Series C Applied Statistics 58, 105-121.

Katholi, C. R., Toe, L., Merriweather, A., and Unnasch, T. R. (1995). Determining the prevalence of Onchocerca volvulus infection in vector populations by polymerase chain reaction screening of pools of black flies. J Infect Dis 172, 1414-1417.

Messam, L. L. McV., Branscum, A. J., Collins, M. T., and Gardner, I. A. (2008). Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease. Anim Health Res Rev 9, 1-23.

Rodoni, B. C., Hepworth, G., Richardson, C., and Moran, J. R. (1994). 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 **45**, 223-230.

Schucany, W.R., Gray, H.L., Owen, D.B. (1971). On Bias Reduction in Estimation. *Journal of the American Statistical Association* **66**, 524-533.

Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology* **75**, 882-889.

Tebbs, J. M., Bilder, C. R., and Moser, B. K. (2003). An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics - Theory and Methods* **32**, 983-995.

Thompson, K. H. (1962). Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics* **18**, 568-578.

Tu, X.M., Litvak, E., and Pagano, M. (1995). On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika* **82**, 287-297.

BAYESIAN APPROACHES FOR ESTIMATING PREVALENCE BASED ON POOL SCREENING WHEN OBSERVING ZERO POSITIVE POOLS

by

THOMAS BIRKNER, INMACULADA B. ABAN, CHARLES R. KATHOLI

In preparation for Computational Statistics & Data Analysis

Format adapted for dissertation

1. INTRODUCTION

1.1. Pool Screening

Pool screening, also known as group testing or composite sampling, is a method that combines individual items into several groups, or pools of a known size. The pools are then tested in place of the individual items. The result is binary, for instance, positive versus negative, infected versus non-infected, polluted versus not polluted. 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 the disease prevalence estimation in arthropod vector populations such as mosquitoes or black flies. These vectors transmit viral and parasitic diseases (e.g. West Nile virus, St. Louis encephalitis, Onchocerciasis, Malaria, Filariasis). As a concrete example we will 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 from the treated human.

In order to evaluate the progress of control and elimination programs, pools of insects are collected at different locations within a broader area. The pools are tested using a Polymerase Chain Reaction (PCR) method (for PCR in black flies see: Katholi *et al.*, 1995). A positive result implies that the pool contains at least one infected insect. PCR cannot determine the number of infected insects in a pool. The number of positive pools is then used to estimate the infection prevalence in the vector population.

1.2. Zero Positive Pools

When control or elimination programs have been in place for awhile and have led to a decrease in infection prevalence in the vector population, it is not uncommon in practice to encounter zero positive pools even after an extensive collection of vectors (for example see Yameogo *et al.,* 1999, Guevara *et al.,* 2003, Rodriguez-Perez *et al.,* 2006). This does not imply that the prevalence is truly zero, but that the prevalence has fallen to a level where zero positive pools are likely to be obtained. Katholi and Unnasch (2006) provide a simple formula to calculate the probability of detecting an infected insect when screening different numbers of pools given a particular infection rate. The question then arises: what is a good estimate of the true infection prevalence? Both the traditional maximum likelihood method (see for instance: Chiang and Reeves, 1962) as well as the hierarchical model approach by Birkner, Aban and Katholi (2011) fail in this situation. The Maximum Likelihood Estimate (MLE) for the prevalence is zero. The distribution in the hierarchical model approach is degenerate.

It is crucial to know whether the infection prevalence has been decreased to a level where no recurrence is expected. The prevalence estimate will inform the decision whether to continue the mass drug administration or end the program. After a program ends, post treatment monitoring will ensue and be challenged by very low prevalence numbers. The need for methods capable of handling the observation of zero positive pools, without providing unrealistic prevalence estimates is great. There exists no viable frequentist approach. Bayesian approaches that incorporate knowledge of previous samplings from the same or similar geographic area into the current prevalence estimation can offer a solution.

Vector control programs are evaluated periodically. A natural way to include historical data is provided by the Bayesian approach in the form of a prior distribution – the posterior distribution is then proportional to the likelihood times the prior distribution. We believe that practitioners have not embraced a Bayesian prevalence estimation method in the past due to the difficulty of specifying a prior. The use of an objective (versus subjectively elicited) prior does not require the specification of prior parameters. A Bayesian approach incorporating an objective prior will be able to handle samples where none of the pools are infected.

1.3. Bayesian Contributions

In a classical Bayesian approach a prior distribution for the prevalence is chosen and then updated by the observed data through the likelihood function. Inferences are based on the posterior distribution, which can be updated as more data become available.

Bayesian approaches to prevalence estimation start to appear in the literature in the early 1990s. Whereas much of the literature considers three random variables – prevalence, sensitivity, and specificity – we will focus on the developments with respect to prevalence, since the tests used in our application (arthropod vector control) are assumed to have perfect sensitivity and specificity (for supporting evidence see Katholi *et al.*, 1995)

Chaubey and Li (1995) compared the maximum likelihood estimator (MLE) of a binomial probability based on sample compositing with a Bayes estimator. First they derived the Bayes estimator using a beta$\left(\alpha,\beta\right)$ prior for p, the population proportion, for the equal pool size case:

$$\hat{p} = E\left(p \mid T\right) = \frac{\sum_{j=0}^{T} \binom{T}{j} \left(-1\right)^{j} \mathrm{B}\left(\alpha+1, nj+nm-nT+\beta\right)}{\sum_{j=0}^{T} \binom{T}{j} \left(-1\right)^{j} \mathrm{B}\left(\alpha, nj+nm-nT+\beta\right)},$$

where *T* is the number of positive pools, *n* is the pool size and *m* is the number of pools.

Here we define $\binom{T}{j} = \frac{T!}{j!(T-j)!}$,

the beta$\left(\alpha,\beta\right)$ pdf as $f\left(p\middle|\alpha,\beta\right) = \frac{1}{\mathrm{B}\left(\alpha,\beta\right)} p^{\alpha-1}\left(1-p\right)^{\beta-1}, 0 < p < 1, \alpha > 0, \beta > 0$,

and the beta function, $\mathrm{B}\left(\alpha,\beta\right) = \int_{0}^{1} p^{\alpha-1}\left(1-p\right)^{\beta-1} dp$.

Secondly they evaluated the estimators using Bayesian (Bayes relative efficiency) and Frequentist (relative bias and relative efficiency) criteria. They concluded that the MLE is inferior to the Bayes estimator under all three criteria.

Chick (1996) in a paper titled “Bayesian Models for Limiting Dilution Assay and Group Test Data” expanded the Bayesian prevalence estimation approach to the unequal pool size case choosing also a beta prior. He compared posterior probability distributions for three different sets of α and β$\left(\alpha=\beta=1;\alpha=2,\beta=3;\alpha=10,\beta=15\right)$.The three posterior distributions obtained were not very different from each other due to the relatively large amount of data available. Chick (1996) conjectured that this would be the case for all beta priors with small values (<50) of $\alpha+\beta$.

Tebbs *et al.* (2003) developed an empirical Bayes procedure to estimate *p* using a beta$\left(1,\beta\right)$prior distribution. Their estimate of *p* is the mean of the empirical posterior

$f_{P|T}\left(p\,|\,t,\hat{\beta}\right)=\dfrac{f_{T,P}\left(t,p\,|\,\hat{\beta}\right)}{f_T\left(t\,|\,\hat{\beta}\right)}$. They found

$$\hat{p}_{eb}=1-\frac{\Gamma\left(m+\hat{\beta}/n+1\right)\Gamma\left(m-t+\hat{\beta}/n+1/n\right)}{\Gamma\left(m-t+\hat{\beta}/n\right)\Gamma\left(m+\hat{\beta}/n+1+1/n\right)},$$

where $\Gamma\left(\alpha\right)$ defines the gamma function, $\Gamma\left(\alpha\right)=\int\limits_0^\infty x^{\alpha-1}e^{-x}dx.$

Tebbs *et al.* (2003) point out that for $T=0$ (no positive pools) or $T=m$ (all positive pools) $f_T\left(t\,|\,\hat{\beta}\right)$cannot be maximized and no $\hat{p}_{eb}$ can be computed. They showed that their empirical Bayes estimator outperforms the traditional maximum likelihood estimator with respect to relative bias and relative efficiency for small group sizes and small *p*. Tebbs *et al.* (2003) also derived an empirical credible interval for *p*:

$$p_L=1-\left(\mathrm{B}_{1-\alpha/2;m-t+\hat{\beta}/n,t+1}\right)^{1/s},\ p_U=1-\left(\mathrm{B}_{\alpha/2;m-t+\hat{\beta}/n,t+1}\right)^{1/s},$$

where $\mathrm{B}_{y;a,b}$ denotes the γ quantile of the two parameter beta distribution.

Bilder and Tebbs (2005) adapted the empirical Bayes estimator from above to fit the multiple-vector-transfer designs method. This method entails moving several vectors, for instance leafhoppers, from a diseased plant to a healthy plant. Researchers are interested in whether the healthy plant develops disease symptoms, indicating that at least one of the vectors carried the disease, for instance a plant virus.

The approach proposed in both papers (2003 and 2005) uses the *currently* observed data to estimate the hyperparameter β . The data are then used again together with $\hat{\beta}$ to estimate p . This amounts to “double dipping” into the data. Tebbs *et al.* (2003) and Bilder and Tebbs (2005) argue that their approach is an improvement over (arbitrarily) choosing the values of model hyperparameters a priori.

1.4. Goal and Outline of the Paper

The objective of this paper is to:

1) Develop a sequential Bayes algorithm for prevalence estimation from the beginning of a control program to the end and beyond (post treatment surveillance). This algorithm will be based on updating the prior and posterior distribution as new information (new pool screening results) becomes available. We will consider two objective priors: a) Bayes/Laplace and b) Jeffreys’ prior.

2) Demonstrate that the proposed sequential Bayes procedure produces a sensible estimate of the prevalence even if zero positive pools are observed in the current sampling. This situation is encountered most often after several years of treatment and in the post treatment surveillance phase.

3) Suggest strategies to minimize inherent weaknesses of the sequential Bayes approach.

We are interested in the number of years (amount of positive pools) required such that the difference between the posteriors based on either of the two priors becomes practically insignificant. This is important, because given this condition is met, the choice between the objective priors considered here no longer matters. We also want to evaluate the impact of the priors on the accuracy of the resulting prevalence estimators, in particular, how quickly the posterior distributions change when the underlying true prevalence changes (due to the treatment effect or reintroduction of the parasite).

The outline of the paper is as follows: In section 2 we will derive two objective priors (Bayes/Laplace, Jeffreys') and present analytical expressions of the posterior distributions in the sequential Bayes framework. In section 3 we will discuss simulation results to gauge the amount of data needed to minimize the impact of a particular prior choice. Simulation results pertaining to the accuracy of the Bayesian estimates and to three strategies to reduce the "inertia" embedded in the posteriors due to the sequential nature of our approach are presented as well.

2. SEQUENTIAL BAYES

The idea behind the sequential Bayes method is to carry forward the prevalence information gathered for the previous year(s) to the current estimation of the prevalence distribution. At year 1, we assume that we know very little about the prevalence distribution, and hence, use an objective (least informative prior), which in our case would amount to either the Bayes/Laplace or Jeffreys' prior. Starting at year 2 we use the posterior distribution from the previous year as prior in the preceding year (see schematic below, where $\propto$ stands for proportional).

year 1: $posterior_1\left(p \mid n,m,t\right) \propto \left[1-\left(1-p\right)^n\right]^{t_1}\left[\left(1-p\right)^n\right]^{m_1-t_1} *$ objective prior

year 2: $posterior_2\left(p \mid n,m,t\right) \propto \left[1-\left(1-p\right)^n\right]^{t_2}\left[\left(1-p\right)^n\right]^{m_2-t_2} * posterior_1\left(p \mid n,m,t\right)$

year 3: $posterior_3\left(p \mid n,m,t\right) \propto \left[1-\left(1-p\right)^n\right]^{t_3}\left[\left(1-p\right)^n\right]^{m_3-t_3} * posterior_2\left(p \mid n,m,t\right)$

$\vdots$

2.1. Objective Priors

A prior distribution is supposed to represent knowledge about parameters before the results of an experiment are known. There are two things to keep in mind: A) One can never be in a state of complete ignorance; and B) “knowing little a priori” can only have meaning relative to the information provided by an experiment (Box and Tiao (1992), 25). “Thus, the main issue is how to select a prior which provides little information relative to what is expected to be provided by the intended experiment” (Box and Tiao (1992), 25). The need to elicit a subjective prior distribution for the prevalence at the beginning or some other stage of a treatment program constitutes a major obstacle to the application of a Bayesian prevalence estimation approach. Using an objective (least informative) prior eliminates this step and also the potential for specifying a poor prior distribution. Our choice among available objective priors was guided by finding relatively simple, well accepted and computationally convenient ones. We selected the Bayes/Laplace (uniform distribution on the number of positive pools) and Jeffreys’ prior. A catalog of non-informative priors and their properties is provided by Yang and Berger (1998).

2.1.1. Bayes/Laplace Prior

We shall consider the finding of an objective prior analogous to the “flat” or Uniform prior advocated by Bayes and Laplace for the Binomial sampling model. The discussion is based on the Stigler (1982) interpretation of Thomas Bayes’s, “An Essay Towards Solving a Problem in the Doctrine of Chance”, published posthumously in 1764. Stigler argues that Bayes did not base his choice of the Uniform prior on an appeal to the “Principle of Insufficient Reason”. Rather, the justification is based on choosing a prior such that the marginal distribution of the binomial random variable given this prior is a discrete uniform. That is, prior to having collected any data no particular value for the number of successes in n trials is any more likely than any other. In modern terms, if one considers the $Beta(\alpha,\beta)$ family of distributions as a reasonable candidate for characterization of the uncertainty, Bayes’s argument leads to the choice $\alpha=\beta=1$ which is the *Uniform*(0,1) distribution.

We recall that in the pool screening model (equal pool sizes), *m* pools of size *n* are screened and the Bernoulli random variable X is 0 when a pool found to be negative and 1 when a pool is positive. Thus, for each pool,

$$X_j \sim \left[1-(1-p)^n\right]^{X_j}\left[(1-p)^n\right]^{1-X_j},0<p<1$$

If we define $T=\sum_{j=1}^{m} X_j$ then T is a $Binomial(m,\theta)$ random variable with parameter, $\theta=1-(1-p)^n$. Hence T has the probability mass function given by:

$$f(T;m|\theta)=\binom{m}{T}\theta^T(1-\theta)^{m-T}=\binom{m}{T}\left(1-(1-p)^n\right)^T\left((1-p)^n\right)^{m-T}$$

Following Bayes' argument as given by Stigler, we look for a prior in the natural conjugate prior family (which is the Beta family of distributions),

$$f_p\left(p \middle| n,\alpha,\beta\right)=\frac{\theta(p)^{\alpha-1}\left[1-\theta(p)\right]^{\beta-1}}{\mathrm{B}(\alpha,\beta)}=\frac{\theta(p)^{\alpha-1}\left[1-\theta(p)\right]^{\beta-1}}{\int_0^1\theta(p)^{\alpha-1}\left[1-\theta(p)\right]^{\beta-1}dp},$$

where $\theta(p)$ highlights the fact that θ is a function of *p*.

Using this we can obtain the marginal distribution of T :

$$\begin{aligned}f_T\left(t \middle| \alpha,\beta\right)&=\binom{m}{T}\left(\frac{1}{\mathrm{B}\left(\alpha,\beta+\frac{1}{n}-1\right)}\right)\int_0^1[\theta]^{\alpha+T-1}[1-\theta]^{m-T+\beta+\frac{1}{n}-1-1}d\theta\\&=\binom{m}{T}\left(\frac{\mathrm{B}\left(\alpha+T,m-T+\beta+\frac{1}{n}-1\right)}{\mathrm{B}\left(\alpha,\beta+\frac{1}{n}-1\right)}\right),\alpha,\beta>0\end{aligned}$$

Result: In particular the distribution function $f_T\left(t \middle| \alpha,\beta\right)$ has value $1/(m+1)$ for all $t\in\{0,1,2,...,m\}$ when $\alpha=1$ and $\beta=1+\frac{(n-1)}{n}$.

The objective Bayes prior for the equal pool size pool screening model is obtained as

$$f_p(p)=\frac{n*1*\left[(1-p)^n\right]^{\frac{n-1}{n}}}{\mathrm{B}(1,1)}=\frac{n(1-p)^{n-1}}{\mathrm{B}(1,1)}=n(1-p)^{n-1}$$

For details and a proof of the result above see appendix E1.

2.1.2. Jeffreys' Prior

Many interpreters of Bayes have interpreted him as suggesting a uniform prior distribution in cases when we know little (or nothing) about the parameter before the experiment is conducted (see Stigler (1982)). This interpretation opened the door for the following criticism: if the distribution of a continuous parameter θ were uniform, then the distribution of $\log\theta,\ \theta^{-1}$ or some other transformation of θ would not be uniform (Box and Tiao (1992), 24). Thus, the application of Bayes theorem to different transformations of θ would lead to inconsistent posterior distributions (and inferences) from the same data. To overcome this issue, Jeffreys (1946) developed a rule which produces a non-informative prior that is invariant under transformation. Jeffreys defined the objective prior density as $p(\theta)\propto\left[J(\theta)\right]^{\frac{1}{2}}$, where $J(\theta)$ is the Fisher information for θ:

$J(\theta)=E\left[\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2|\theta\right]=-E\left[\frac{d^2\log p(y|\theta)}{d\theta^2}|\theta\right]$ (Gelman *et al*, 2004, p. 63; Jeffreys 1946).

The Jeffreys' prior in our setting is proportional to $\sqrt{J(p)}=\frac{\sqrt{m}n(1-p)^{\frac{n-2}{2}}}{\left[1-(1-p)^n\right]^{\frac{1}{2}}}$.

For details consult appendix E2. Using Jeffreys' rule the inferences drawn will not differ whether we use a prior on p (the probability of one insect being infected and the population prevalence) or θ (a function of p and the probability of a pool containing at least one infected insect). For another application but using a non-sequential approach of the Jeffreys' prior in the analysis of entomologic data, see Rodriguez-Perez *et al.* (2006).

2.2. Posteriors

In this section we present the posterior distributions and expressions for the two main quantities of interest: the expected value (the prevalence estimate) and the probability of being below a certain threshold value. The algebra of the sequential Bayes approach is simple since we chose conjugate priors. A prior is conjugate when it is of the same functional form as the likelihood, which in turn produces a posterior of the same functional form. The derivation of the posteriors beyond year 1 amounts to updating the exponents by simple addition or subtraction of the number of positive pools (t) and number of pools tested (m) for the current year. A sketch of the algebra for the Bayes/Laplace prior is provided in appendix F.

2.2.1. The Bayes/Laplace Prior Posterior Distribution

Let $f_{BL}(p)$ be the posterior distribution given the Bayes/Laplace prior. Then

$$\begin{aligned} f_{BL}(p) &\propto \left[1-(1-p)^n\right]^T \left[(1-p)^n\right]^{m-T} n(1-p)^{n-1} \\ &\propto \left[1-(1-p)^n\right]^T \left[(1-p)^{n(m+1-T)-1}\right] n. \end{aligned}$$

The fully normalized posterior can be expressed as follows:

$$f_{BL}(p) = \frac{n\left[1-(1-p)^n\right]^T \left[(1-p)^{n(m+1-T)-1}\right]}{\mathrm{B}\left[T+1, m+1-T\right]}. \; (1)$$

The expected value is derived as:

$$E_{BL}\left(p\right)=1-\frac{\Gamma\left(m+2\right)\Gamma\left(m+1-T+\frac{1}{n}\right)}{\Gamma\left(m+2+\frac{1}{n}\right)\Gamma\left(m+1-T\right)}.(2)$$

An expression for the threshold probability is provided next:

$$P\left(p\le p_0\right)=\int_0^{p_0} f_{BL}\left(p\right)dp=\int_0^{p_0}\frac{n\left[1-\left(1-p\right)^n\right]^T\left[\left(1-p\right)^{n\left(m+1-T\right)-1}\right]}{\mathrm{B}\left[T+1,m+1-T\right]}dp.(3)$$

For derivation, see appendix G. The formulas (1) through (3) above are directly applicable after the number of positive pools has been determined for the first year of pool screening. For subsequent years replace $T=t$ with $t_{total}=\sum_{year=1}^{current\ year} t_{year}$ and m with $m_{total}=\sum_{year=1}^{current\ year} m_{year}$, where $\sum$ stands for summation.

2.2.2. Jeffreys' Prior Posterior Distribution

We now present the posterior distribution for *p* employing the Jeffreys' prior, denoted by $f_J(p)$, as derived in Barker (2000):

$$f_J\left(p\right)=\frac{n}{\mathrm{B}\left(m-T+\frac{1}{2},T+\frac{1}{2}\right)}\left(1-p\right)^{n\left(m-T+\frac{1}{2}\right)-1}\left[1-\left(1-p\right)^n\right]^{\left(T+\frac{1}{2}\right)-1}.(4)$$

The expected value of p, given in Barker (2000), is:

$$E_J\left(p\right)=1-\frac{\Gamma\left(m-T+\frac{1}{2}+\frac{1}{n}\right)\Gamma\left(m+1\right)}{\Gamma\left(m-T+\frac{1}{2}\right)\Gamma\left(m+1+\frac{1}{n}\right)}.(5)$$

An expression for the threshold probability follows next:

$$P\left(p \le p_0\right) = \int_0^{p_0} f_J\left(p\right)dp = \int_0^{p_0} \frac{n}{\mathrm{B}\left(m-T+\frac{1}{2},T+\frac{1}{2}\right)}\left(1-p\right)^{n\left(m-T+\frac{1}{2}\right)-1}\left[1-\left(1-p\right)^n\right]^{\left(T+\frac{1}{2}\right)-1}dp. \tag{6}$$

To generalize (4) through (6) replace $T = t$ with t_{total} and m with m_{total}, each the summation over all years in which pool screening has been conducted.

3. SIMULATIONS

3.1. Objective

The objective of the simulation study presented here is to determine:

a) How many years of data are needed such that the difference between the posteriors derived by either using the Bayes/Laplace or Jeffreys' prior is not of any practical concern?

b) Do we observe a strong memory/inertia effect (defined in section 3.4.)?

  b1) How well do the sequential Bayes estimators (expected values from posteriors) match the simulated ("true") prevalence?

  b2) Are the threshold probabilities $P\left(p \le p_0\right)$ read off the posterior distribution meaningful? In other words: Are they indicating the right time when to stop treatment and shift to surveillance only?

c) If we observe an inertia effect, how can we reduce it?

c1) How many years of low prevalence (later years) have to be included to overcome the inertia?

c2) How strong is the effect of omitting a certain number of higher prevalence years (early years) with respect to alleviating the inertia? In other words: When switching to the sequential Bayes prevalence estimation approach at some point during a treatment program how many years of past data do we want to include?

d) How does the Bayesian method given the Bayes/Laplace or Jeffreys' prior compare to the Maximum Likelihood approach when used in a non-sequential manner?

3.2. Simulation parameters

We are considering the following setup, which is meant to mimic the real world situation for Onchocerciasis and Filariasis elimination programs at this time:

Number of pools: 300

Pool size (number of vectors per pool): 25

Infection prevalence: Linear decrease from year 1 (p=1/1000) to year 20 (p=1/10000). This linear decrease is simulating the effect of continuous treatment.

The prevalence at year 1 is set markedly lower compared to the typical observed prevalence at the start of a treatment program in the past. The focus here is prevalence estimation after several decades of successful treatment. We suggest that the sequential Bayes method should replace the traditional Maximum Likelihood estimation approach in low prevalence scenarios because it can handle the observation of zero positive pools (late stage treatment phase or surveillance phase) without producing an implausible prevalence estimate.

3.3. Evaluation of Simulation Error (Number of Replications Required)

To ensure meaningful simulations we need to consider the between-simulation variability. This quantity is known as simulation error or Monte Carlo error (MCE). We follow Koehler, Brown and Haneuse (2009) and let ϕ denote some target quantity of interest and $\hat{\phi}_R$ denote the Monte Carlo estimate of ϕ from a simulation with R replicates. The Monte Carlo error is defined as:

$$MCE\left(\hat{\phi}_R\right)=\sqrt{Var\left[\hat{\phi}_R\right]}.(7)$$

The target quantity ϕ in our application is the expected value of the infection prevalence based on either the Bayes/Laplace or Jeffreys' prior and a simulated treatment course of 20 years. We considered 10, 100, 500, 1000, 10000, and 100,000 replications. **Figure 1** displays the averaged expected values (Bayes/Laplace prior) and simulation errors for 6 different replication sizes. We notice that both curves are not changing anymore for $R \geq 500$. The Monte Carlo Error decreases below 2/10000 at year 5 and further decreases to a value below 1/10000 around year 10. The simulation error is small relative to the prevalence estimate. We decided to use 100,000 replications in our simulations, since the computations are not very intensive.

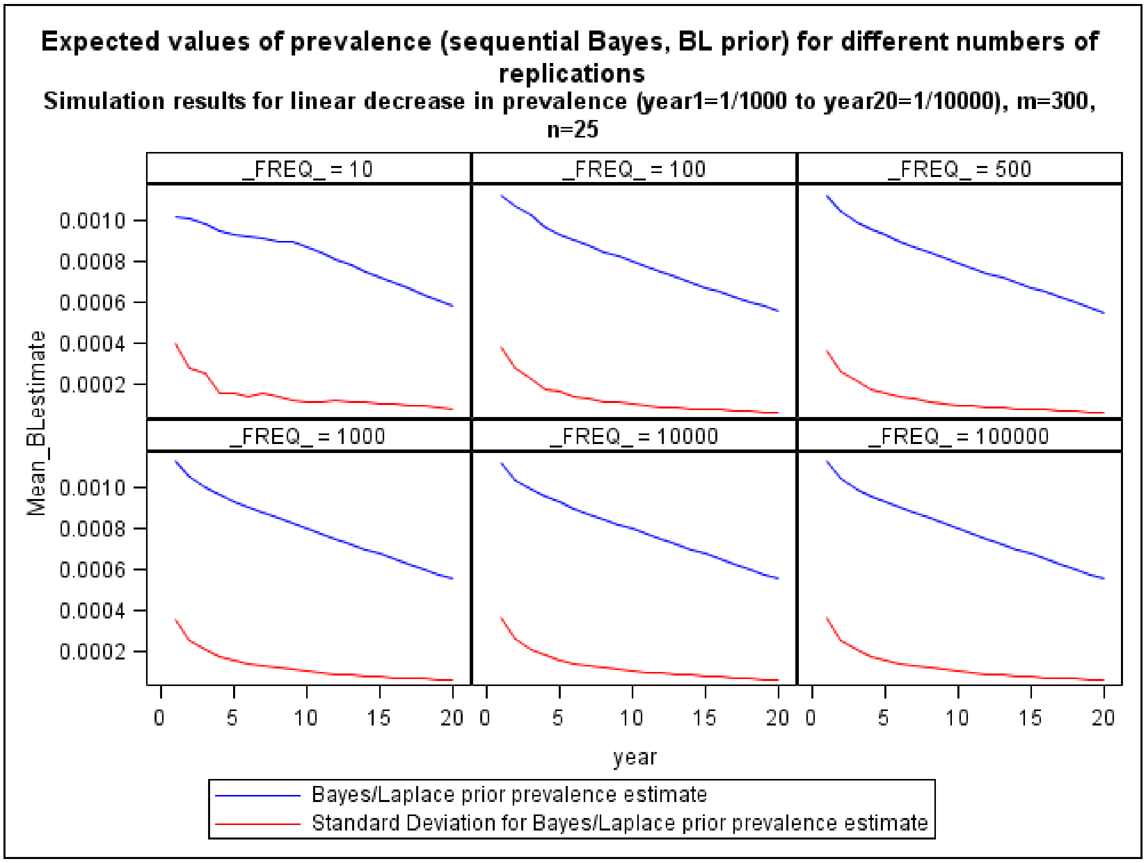


Figure 1. Prevalence estimator (using Bayes/Laplace prior) and standard deviation (simulation error) shown for six replication sizes.

3.4. Simulation of 20 years with Linear Decrease in Prevalence

[m=300, n=25, years=20 (p_{year1}=1/1000, p_{year20}=1/10000), p_0=5/10000]

The expected values (means of posterior distributions employing either the Bayes/Laplace or Jeffreys' prior) averaged over 100,000 replications for each year in the 20 year sequence and associated standard deviations (simulation error) are depicted in **Figure 2**. We observe that the initial small difference between the expected values becomes practically insignificant around year 5 and completely vanishes around year 10. The magnitude of the estimators declines in accordance with the simulated decrease in prevalence, but not to the full extent of the underlying (simulated) prevalence. We simulated a decline in prevalence from 1/1000 to 1/10000, but the decline in the estimators is much slower. The expected values at year 20 are greater than 5/10000, which is an overestimate of the true prevalence. This indicates that the sequentially estimated posterior distributions retain the higher prevalence of the earlier years for too long – a phenomenon we term “inertia”.

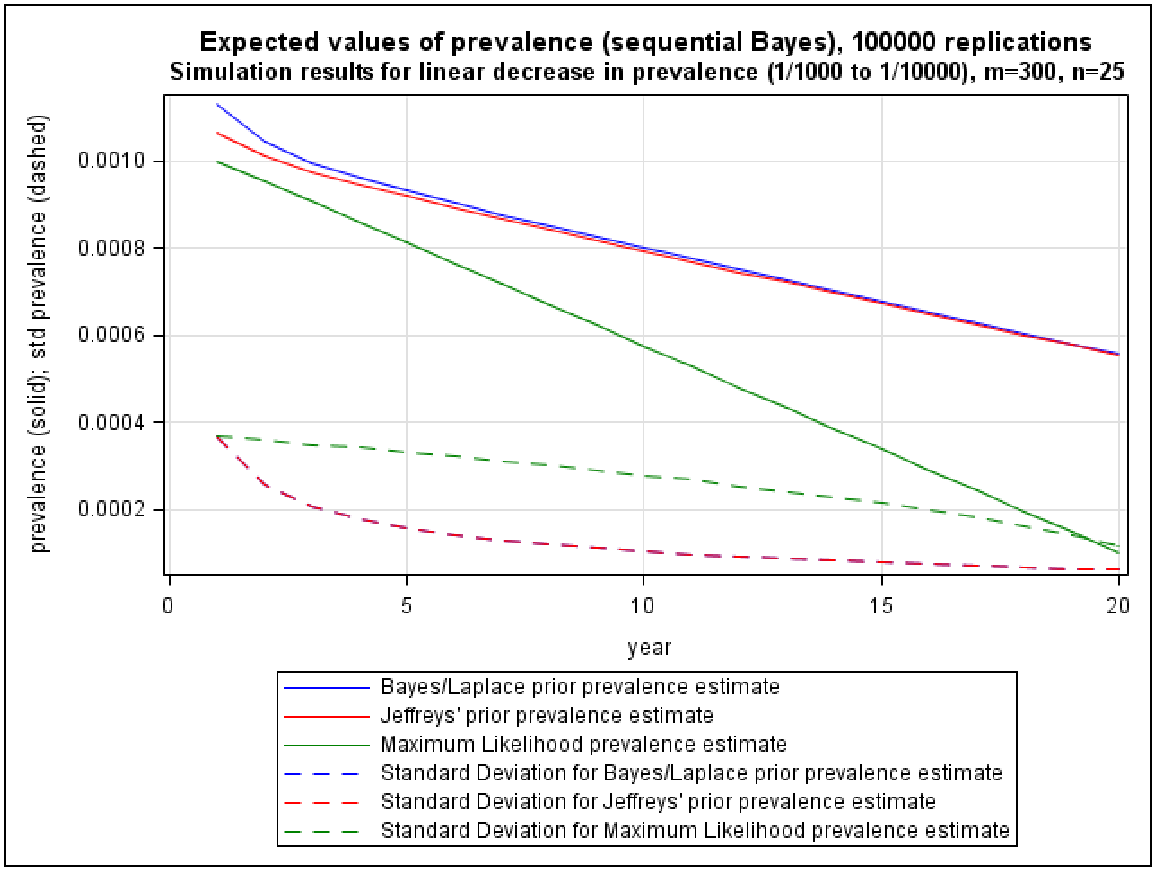


Figure 2. Prevalence estimators using Bayes/Laplace or Jeffreys’ prior, MLE as reference, and standard deviations (simulation error) shown for 100,000 replications of 20 year sequence.

It appears the MLE is superior to the Bayesian Estimators. The principal weakness of the traditional MLE approach though is the implausible prevalence estimate of 0 for years when no positive pools are observed (none of the pools contained at least one infected insect). For our simulation, where the range of the underlying prevalence ranges from 1/1000 to 1/10000, this is the case 157,968 times out of 2,000,000 (7.9%). The most extreme proportion of the MLE=0 is recorded at year 20 where the underlying prevalence is the lowest at 1/10000. Close to half of the time (or with probability=1/2) the MLE estimate equals zero (see **Figure 3**).

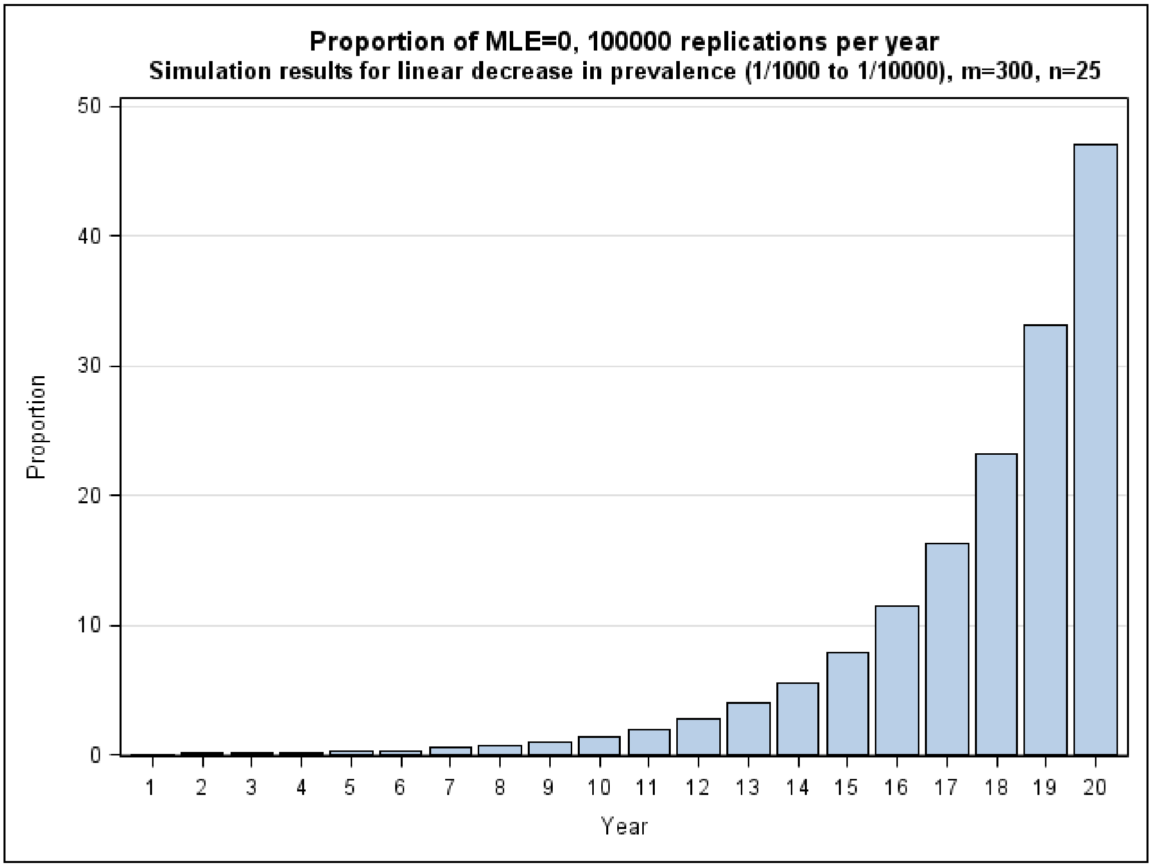


Figure 3. Proportion (%) for which MLE = 0, shown for 100,000 replications of 20 year sequence.

Even more important in practical terms than prevalence estimates are probability statements with regard to the prevalence being below a particular threshold value. It is believed that when the prevalence has been pushed below such threshold value the disease (parasite) has been eliminated and will not reemerge. If the $P\left(p \le p_0\right)$ is high then treatment can be stopped and the post treatment surveillance phase can begin. For Onchoceriasis and Filariasis such proposed threshold values are currently in the neighborhood of $p_0 = 1/10000$ and $p_0 = 25/10000$ respectively. **Figure 4** depicts the median over the 100,000 replications of $P\left(p \le 5/10000\right)$ for each year in the 20 year sequence. The U-shaped curves with a wide flat middle portion indicate again the inertia of the posteriors. At year 10 the underlying prevalence falls below the threshold value of 5 in 10000. We observe an increasing slope in the following years, but the small magnitudes of the actual probabilities still reflect the inertia $\left(p \le 0.2 \text{ for all years}\right)$. We chose the median statistic because the sampling distribution of $P\left(p \le 5/10000\right)$ for either prior revealed skewness. Box and whisker plots for each year (not shown) revealed how extreme values shift the position of the mean upwards, such that the mean is not a good measure for the center of the sampling distribution.

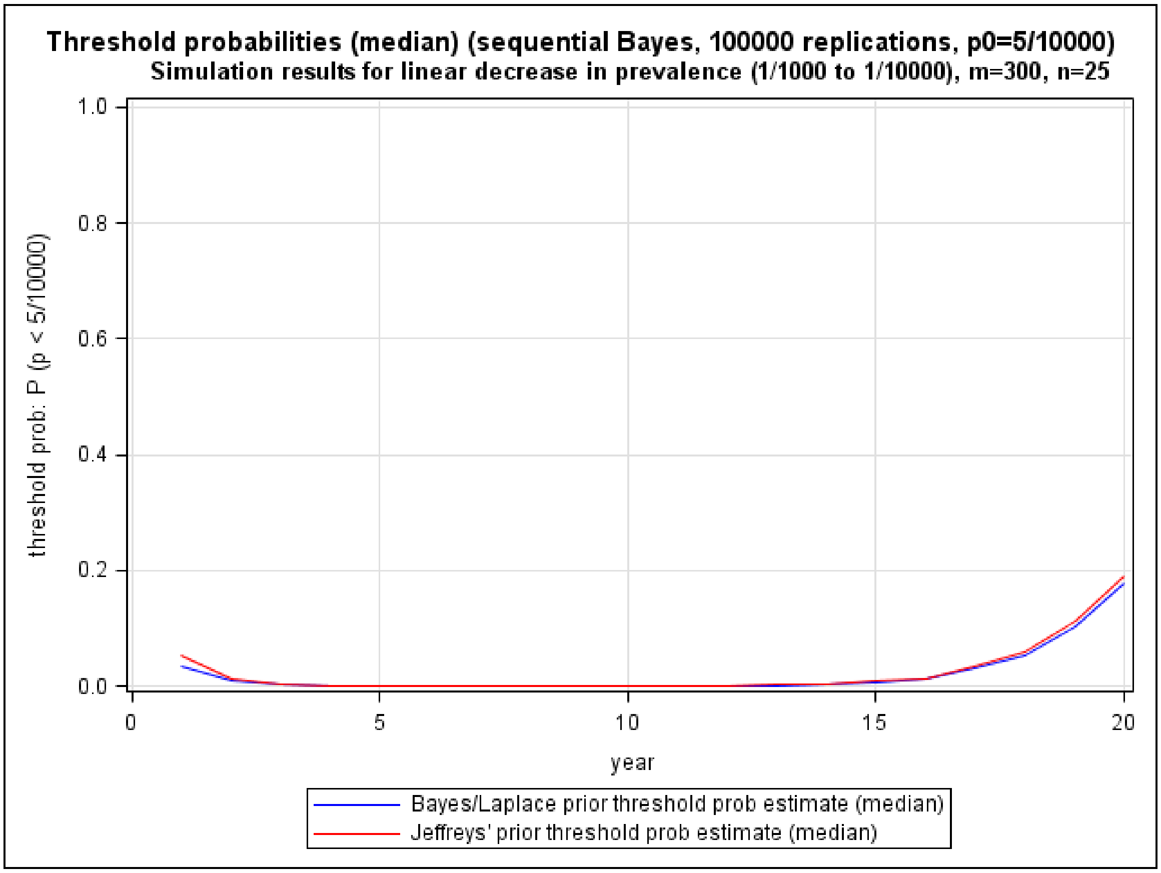


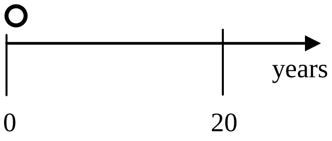
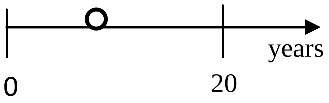
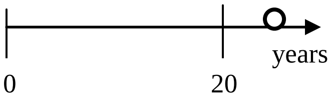
Figure 4. Threshold probabilities $P\left(p \leq 5/10000\right)$ median of 100,000 replications per year.

3.5. Strategies to Overcome Inertia

Given the observation of inertia above, we wanted to assess possible strategies to overcome the inertia effect. Strategy 1) extends the data collection for a number of years, while Strategy 2) omits a number of early (in our case high prevalence) years. Under Strategy 3) we apply the Bayesian approach to the results of each year independent of earlier results (non sequential application). The applicability of these strategies depends on the time the sequential Bayes approach is implemented relative to the progression of the treatment and pool screen programs. **Table 1** summarizes the potential situations and indicates the feasibility of strategies 1), 2) or 3) for each. We view the [sequential] Bayes approach as a - “one fits all” – strategy to handle those different implementation time points and ignore the possibility of deriving a subjectively elicited prior after a few years of pool screening results have been obtained.

Table 1

Applicability of suggested strategies to reduce inertia given different scenarios when the sequential Bayes estimation approach is implemented.

Three situations/ time points for use of sequential Bayes approach (○ represents start of sequential Bayes approach)	**Feasibility of using strategies 1, 2 or 3**		
	Strategy 1 (extend program duration, add results)	**Strategy 2** (exclude results from early years)	**Strategy 3** (non-sequential)
I. prospectively • pool screen program about to begin • no historical pool screening data available 	yes	no	yes
II. mixture of prospectively and retrospectively • pool screen program has been in place for at least one year • some historical data available 	yes	yes*	yes
III. retrospectively • pool screen program ended • only historical data available 	no	yes**	yes

*include only the five most recent years of historical results; ** exclude x number (e.g. x=5) of early years

3.5.1. Strategy 1: Add five years with constant low prevalence

[m=300, n=25, years=25 (p_{year1}=1/1000, p_{year20}=1/10000, p_{year21}-p_{year25}=1/10000), p_0=5/10000]

Under this strategy, we investigate the impact of including additional years/pool screen results. The rationale behind this strategy is to allow for some extra time for the posterior distribution to adjust to the actual prevalence level. When adding five years of constant low prevalence (year 20 through year 25: p=1/10000) we observe a decline in the value of the estimators, but what we are not seeing is a quick downward adjustment (**Figure 5**). The slopes of the curves depicting the estimators longitudinally hardly change. The value of the estimators decreases from 0.00055 (year 20) to 0.00046 (year 25) - not very close to the underlying value of 1/10000 at year 25. Including five years of constant low prevalence in the later stages of a treatment program does not have a significant effect on overcoming the embedded inertia with respect to the expected values.

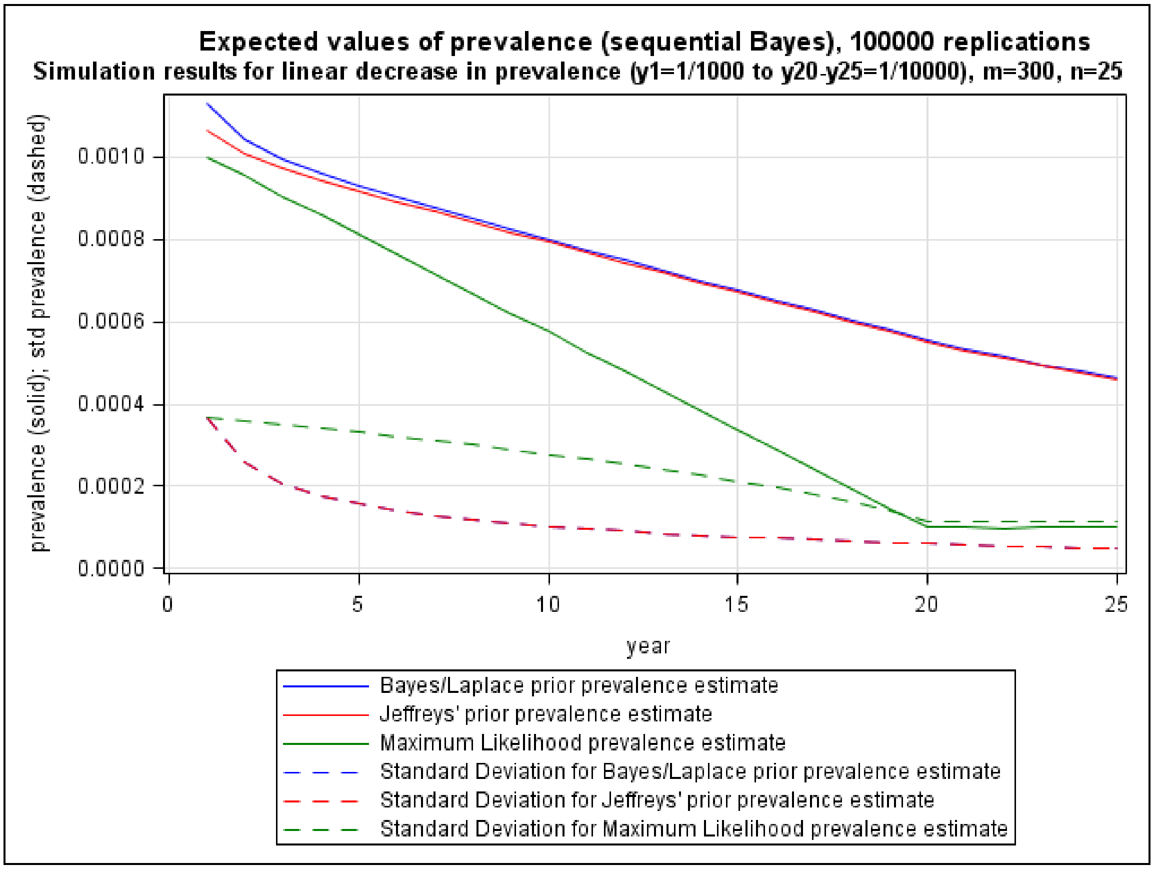


Figure 5. Prevalence estimators using Bayes/Laplace or Jeffreys' prior, MLE as reference, and standard deviations (simulation error) shown for 100,000 replications of 25 year sequence, prevalence year 20 through year 25 set at 1/10000.

There is a markedly stronger effect of adding those years with respect to the threshold probabilities. Considering the median of the threshold probabilities demonstrates a strong impact of the inclusion of five low prevalence years in the computations (**Figure 6**). The probability of the prevalence being below the threshold value of 5/10000 increases form approximately 0.18 (year 20) to 0.79 (year 25).

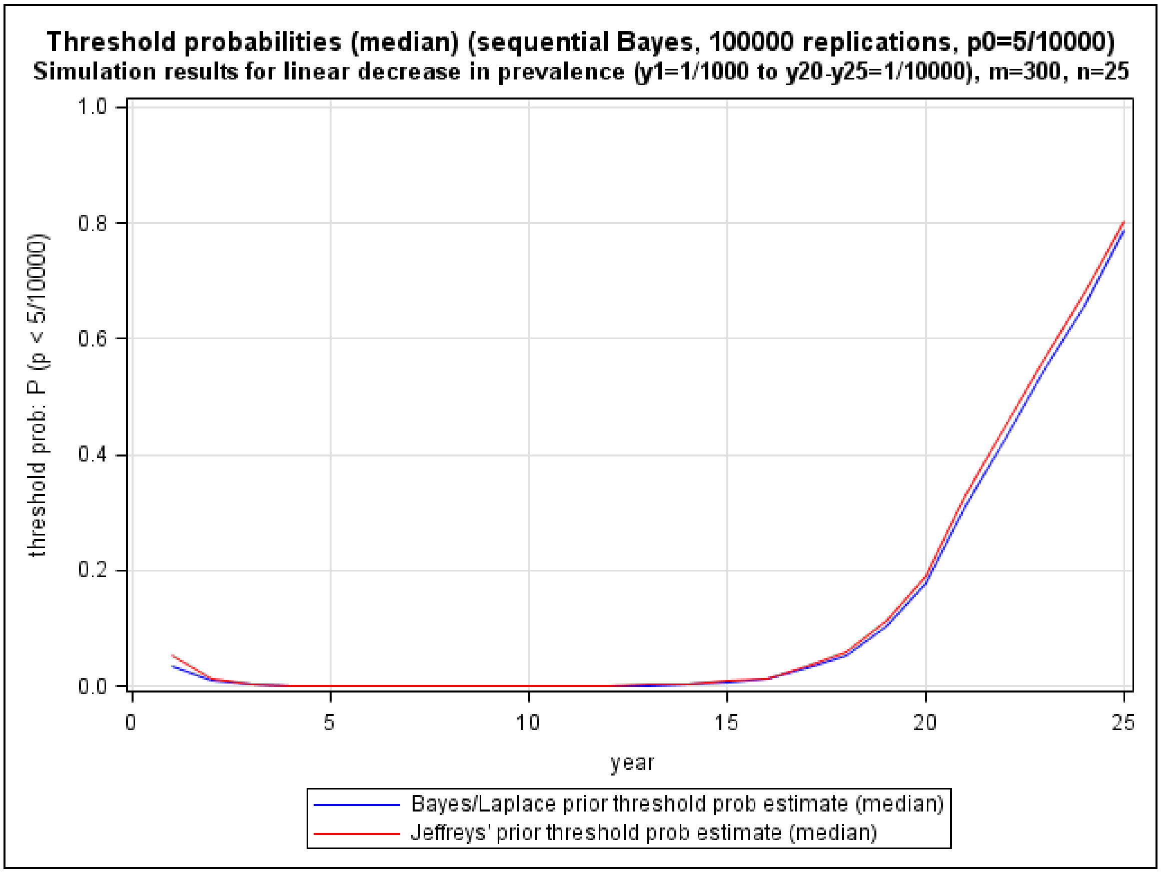


Figure 6. Threshold probabilities $P\left(p \le 5/10000\right)$ median of 100,000 replications per year, prevalence year 20 through year 25 set at 1/10000.

3.5.2. Strategy 2: Omit first five years of higher prevalence

[m=300, n=25, years=20 (p_{year6}=7.6/10000, p_{year20}=1/10000), p_0=5/10000]

Judging the inclusion of more low prevalence years at the end (strategy 1)) a partial success in overcoming the inertia of our sequential Bayes approach we wanted to assess the impact of leaving out a number of early high prevalence years. This question is of practical relevance when the sequential Bayes estimation approach is applied to a treatment program in progress, where pool screening results are already available for a number of years. How many of those early, probably higher prevalence years do we want to include? It seems reasonable to go back about five years, since as we have shown earlier that differences between the Bayes/Laplace and Jeffreys' prior are very small after the sequential evaluation of five years of data. Using too much historical data appears to increase chances of "overloading the memory" of the posteriors with higher prevalence values and so worsening the inertia. For the sake of argument, assume a treatment program has been going on for 10 years. We disregard the results from the first five years, include the results from years 6 through 10, and go forward another 10 years. We observe that the averaged expected values of prevalence decline as before following the simulated linear decrease in prevalence (**Figure 7**). The size of the expected values at year 6 (which is the first year included in computations under strategy 2)) is of course much smaller compared to the expected values at year 1 (the start year under strategy 1)). The estimators take a value of approximately 0.00044 at year 20 – a clear improvement over strategy 1), where the prevalence estimates are 0.00046 at year 25.

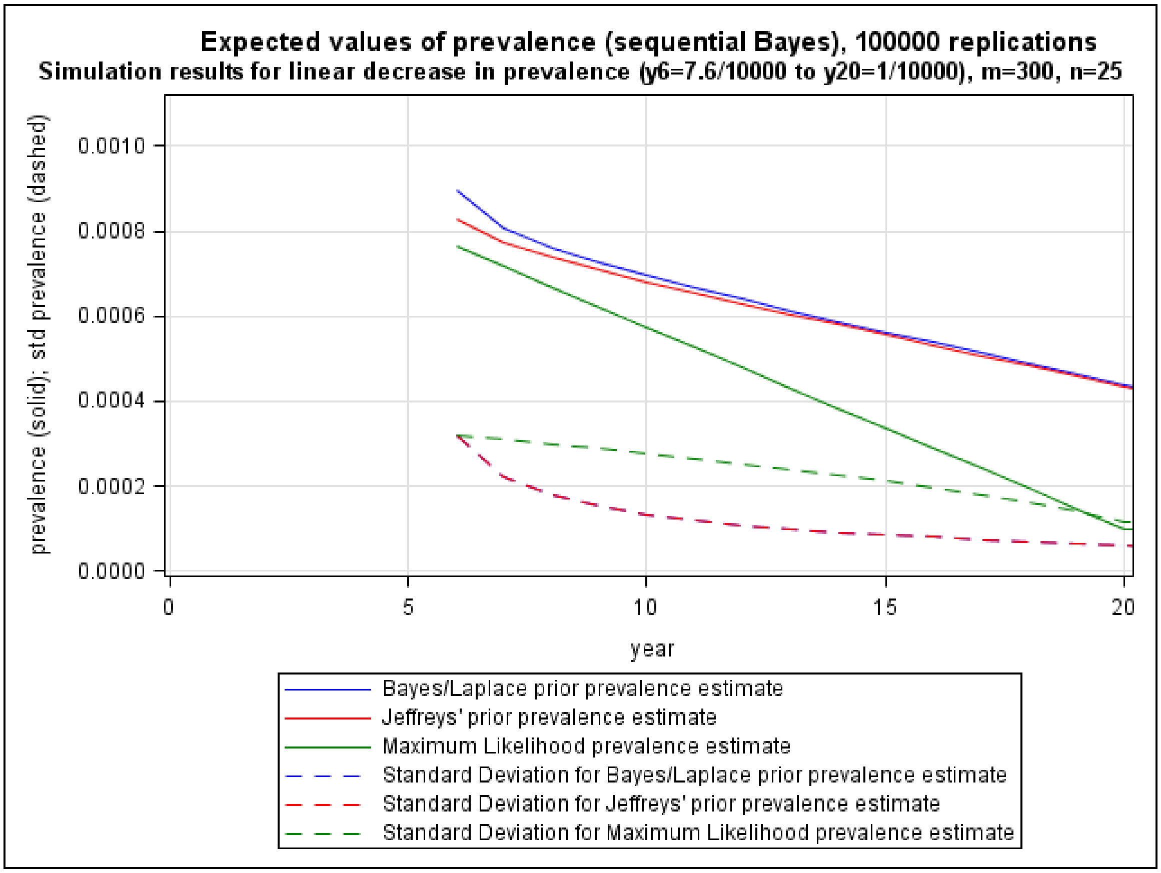


Figure 7. Prevalence estimators using Bayes/Laplace or Jeffreys' prior, MLE as reference, and standard deviations (simulation error) shown for 100,000 replications of 15 year sequence, omitted first five (higher prevalence) years.

The effect of disregarding the first five years is seen even more when considering the threshold probabilities (**Figure 8**), whereas in the original setup (section 3.4.) the probability of $p \leq 5/10000$ at year 20 was less than 0.2 it is now approximately 0.85. The inclusion of five low prevalence years (year 21 through 25) under strategy 1) results in threshold probabilities of approximately 0.8 at year 25 - a value that is already succeeded at year 20 under strategy 2).

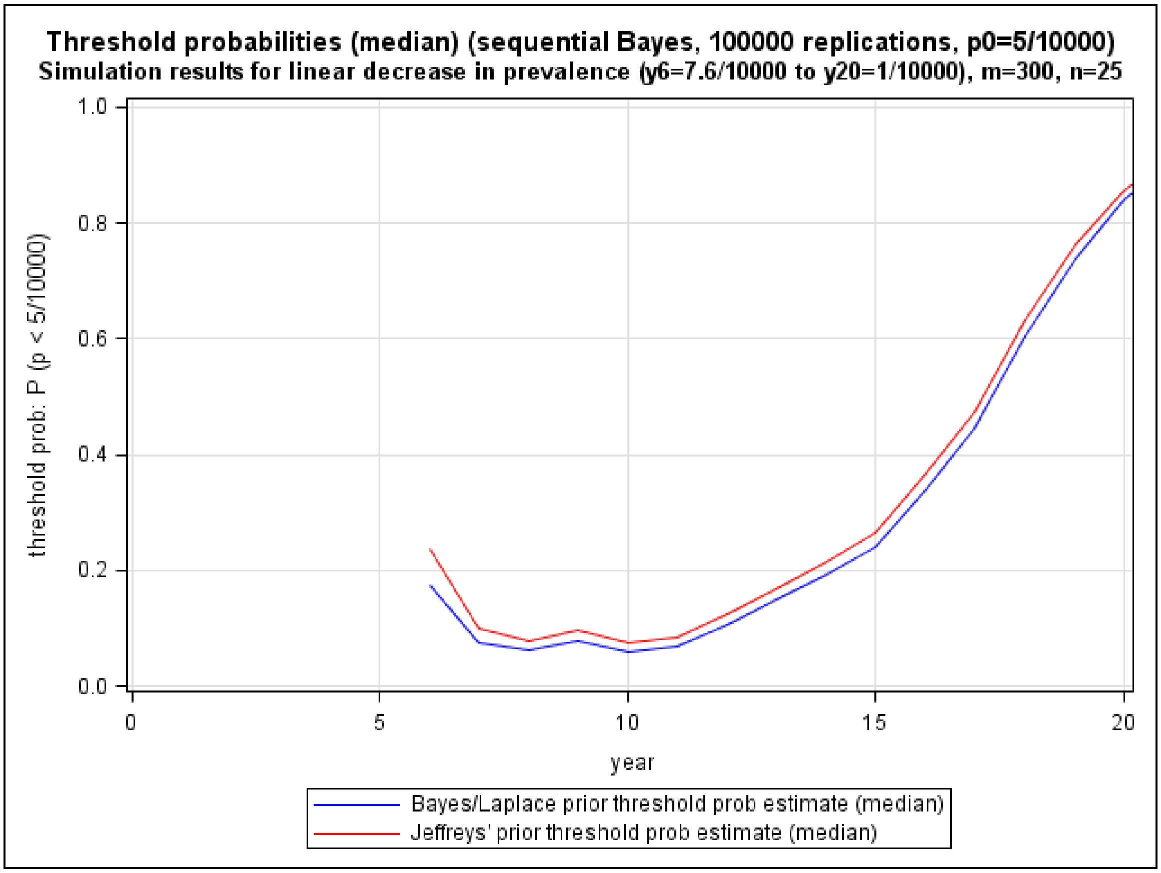


Figure 8. Threshold probabilities $P(p \leq 5/10000)$ median of 100,000 replications per year, omitted first five (higher prevalence) years.

3.6. Strategy 3: Evaluation of Objective Priors when used Non-Sequentially

Given the inertia observed in the posteriors when applying the sequential Bayes method to several years of declining prevalence, we wanted to explore how the approach compares to the MLE when applied to data of a single year (strategy 3). The setup in this case simplifies to: posterior $\propto$ likelihood *objective prior. We found that the prevalence estimates obtained by either using the Bayes/Laplace or Jeffreys' prior track very well with the MLE (**Figure 9**). The estimates are slightly larger than the MLE on average (Bayes/Laplace consistently greater than Jeffreys' estimator), but offer the compensating advantage of non-zero prevalence estimates in cases where all pools are negative for infection.

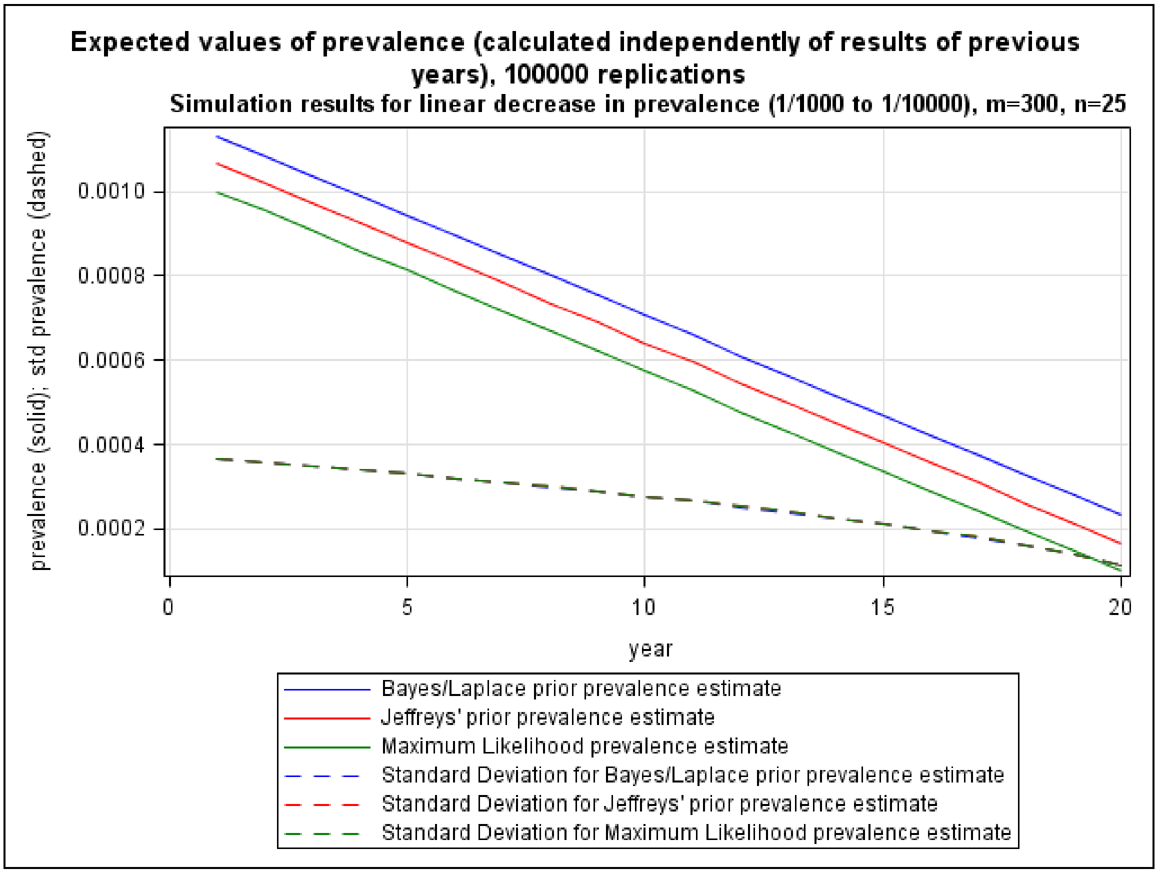


Figure 9. Prevalence estimators using Bayes/Laplace or Jeffreys' prior, MLE as reference, and standard deviations (simulation error) shown for 100,000 replications calculated for each year not using information from previous year(s).

This approach also allows to calculate $P(p \leq p_0)$. The medians of the threshold probabilities over the 100,000 replications are depicted in **Figure 10**. The probabilities are increasing as the underlying prevalence decreases from 1/1000 to 1/10000, with the probabilities calculated using the Jeffreys' prior tracking slightly above the Bayes/Laplace probabilities. The probability of $p \leq 5/10000$ reaches values around 0.9 at year 20.

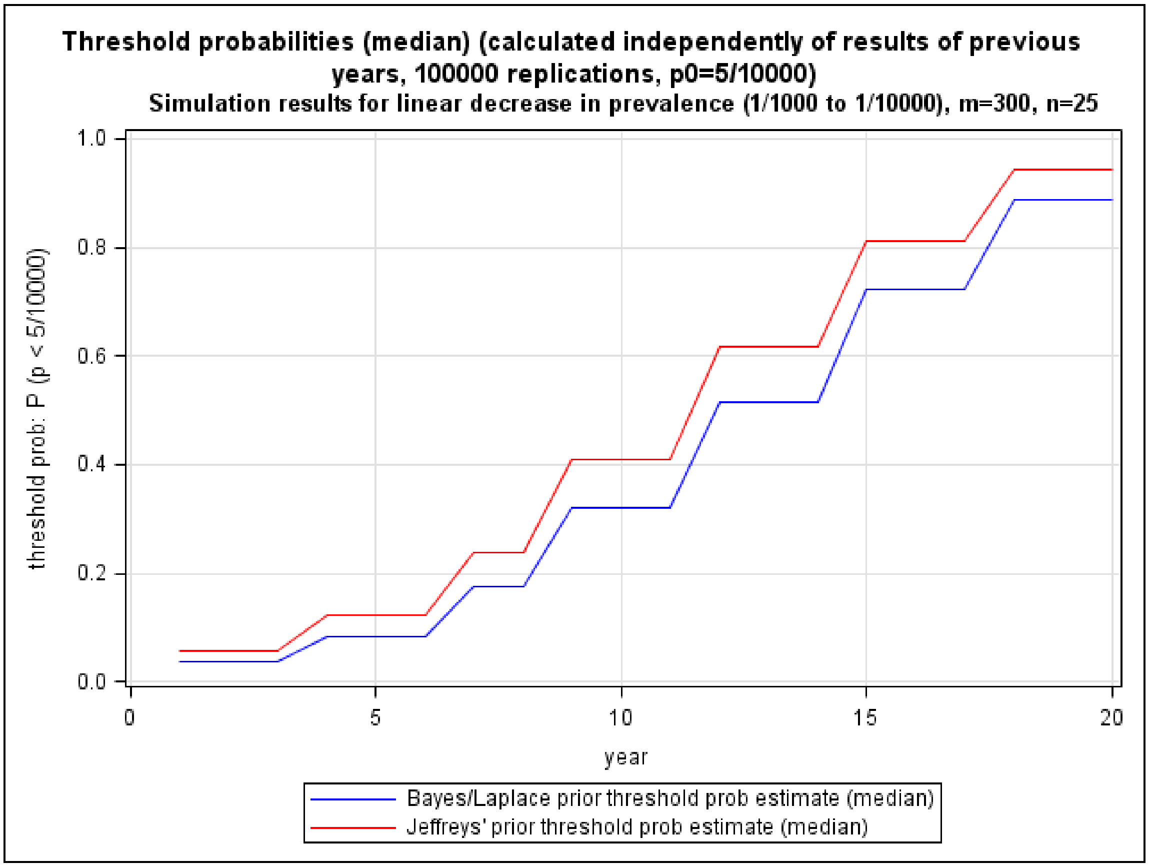


Figure 10. Threshold probabilities $P\left(p \leq 5/10000\right)$ median of 100,000 replications per year, calculated for each year not using information from previous year(s).

4. CONCLUSIONS

We propose a sequential Bayes method, using an objective prior (Bayes/ Laplace or Jeffreys') in the first year, and the posterior distribution of the previous year as prior for all subsequent years.

Including results from previous years makes the observation of zero positive pools for a particular year a non-issue (versus the prevalence estimate of zero using the Maximum Likelihood Estimator). An appropriate treatment of zero positive pools becomes more important as the probability of such an event increases, which is the case after years of successful treatment or in the post treatment surveillance stage.

Since we are estimating the prevalence distribution, we can calculate $P\left(p \leq p_0\right)$, an important determinant for ending the treatment phase and initiating the surveillance only phase.

The choice between the Bayes/Laplace and Jeffreys' prior is inconsequential after about five years of data have been included (given our setup of m=300, n=25, p_{year1}=1/1000 with linear decrease to p_{year20}=1/10000). The posterior distributions become indistinguishable after 5 to 10 years of data have been incorporated [objective a)].

Earlier higher prevalence values are a strong determinant of the posterior distribution longer than anticipated (an observation we termed "inertia") [objective b)]. This affects the performance of the expected values and threshold probabilities. The expected values tend to overestimate the simulated prevalence. The differences are small (less than 5/10000 for each year), but nonetheless critical when considering the potential of ending a treatment program too early and see the infection rebound. The computed threshold probabilities for our declining prevalence scenario are smaller than they should be [objectives b1) and b2)].

We attribute these observations to the inertia resulting from the carry over effect of earlier pool screening results. To reduce the effect of this inertia, we considered 3 strategies.

1) Continue the pool screening beyond the typical 15 to 20 years of prevalence estimation, to allow the posteriors to adjust fully to the changed prevalence. This approach works, but the adjustment process is very slow. There are many extra years required to overcome the "memory" of the posterior distribution [objective c1)].

2) Omit data of the first few years of treatment in the sequential Bayes computations to avoid “overloading the memory” of the posterior distribution. In the situation where the treatment program has begun in the past and the sequential Bayes approach is now to be used going forward, it should be sufficient and beneficial not to include all pool screen results beginning at year 1, but only incorporate the most recent 5 year span and go forward from there. This strategy appears to be better suited then strategy 1) to mediate the inertia in a practical way because it does not require additional screenings and also has a stronger impact on reducing the inertia [objective c2)].

3) When the approach is applied to data from one year only (posterior $\propto$ likelihood *objective prior) it compares very well to the traditional Maximum Likelihood approach. This strategy provides (as do strategies 1) and 2)) a non-zero estimate in the case zero positive pools are observed, and the possibility to calculate $P\left(p \leq p_0\right)$ [objective d)]. The results shown for all three strategies are averaged over the 100,000 replications and the estimates derived under strategy 3) for one particular sequence of years will tend to be noisier than the estimates under the *sequential* strategies 1) and 2).

We showed that the sequential Bayes approach has the inherent property of inertia (slow adjustment to changing underlying prevalence). If over the years a change in the true prevalence occurs, this property is a clear drawback of the method. If the prevalence remains constant however, the inertia has no negative impact on the accuracy of the prevalence estimates. In the latter case additional years of results only refine the estimate and reduce its variance. For general use of the sequential Bayes approach we recommend that it is accompanied by at least one of the other two estimation approaches (MLE or annual Bayes [strategy 3]) to detect potential changes in prevalence quickly and to be able to gauge the amount of inertia. Another option is to use the MLE or annual Bayes methods until the infection prevalence has been reduced to a very low level at which little further change is expected. At that time the sequential Bayes approach can be implemented avoiding its weakness (inertia) and utilizing its strengths (nonzero estimate even if 0 positive pools are observed in one particular year).

In the future, we would like to explore the use of a sliding window, for example apply the sequential Bayesian approach to five years worth of pool screening results at a time. The idea is to limit the impact of the inertia by limiting the number of years for which prior information can be carried forward. Other extensions we are considering are modifications to allow for unequal pool sizes and the use of a conjugate subjectively elicited prior.

REFERENCES

Amazigo, U. , Noma, M., Bump, J., et al. (2006). Onchocerciasis. In Disease and Mortality in Sub-Saharan Africa. (eds. D.T. Jamison, R. G. Feachurn, M. W. Makgoba, et al.), Second Edition.

Barker, J. T. (2000). Statistical estimators of infection potential based on PCR pool screening with unequal pool sizes. University of Alabama at Birmingham, Birmingham.

Bayes, T. (1764). An Essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London* **53**, 370-418.

Bilder, C. R., and Tebbs, J. M. (2005). Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biom J* **47**, 502-516.

Birkner, T., Aban, I.B., Katholi, C. R. (2011). Evaluation of a Frequentist Hierarchical Model to estimate prevalence when sampling from a large geographic area using pool screening (unpublished manuscript).

Box, G. E. P. and Tiao, G. C. (1992): Bayesian Inference in Statistical Analysis. John Wiley and Sons.

Chaubey, Y. P., and Li, W. (1995). Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binomial Probability with Sample Compositing. Journal of Official Statistics 11, 379-390.

Chiang, C. L., and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. American Journal of Hygiene 75, 377-391.

Chick, S. E. (1996). Bayesian Models for Limiting Dilution Assay and Group Test Data. Biometrics 52, 1055-1062.

Gao, H. (2010). Hypothesis testing in unequal sized pool screening. University of Alabama at Birmingham, Birmingham.

Gelman, A. et al (2004). Bayesian Data Analysis. 2nd edition, Chapman &Hall/CRC.

Guevara, A. G., Vieira, J. C., Lilley, B. G., et al. (2003). Entomological evaluation by pool screen polymerase chain reaction of Onchocerca volvulus transmission in Ecuador following mass Mectizan distribution. Am J Trop Med Hyg 68, 222-227.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London-Series A Mathematical and Physical Sciences 186, 453-461.

Katholi, C. R., Toe, L., Merriweather, A., and Unnasch, T. R. (1995). Determining the prevalence of Onchocerca volvulus infection in vector populations by polymerase chain reaction screening of pools of black flies. J Infect Dis 172, 1414-1417.

Katholi, C. R., and Unnasch, T. R. (2006). Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches. Am J Trop Med Hyg 74, 779-785.

Koehler, E., Brown, E., and Haneuse, S. J.-P. A. (2009). On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. The American Statistician 63, 155-162.

Rodriguez-Perez, M. A., Katholi, C. R., Hassan, H. K., and Unnasch, T. R. (2006). Large-scale entomologic assessment of Onchocerca volvulus transmission by poolscreen PCR in Mexico. Am J Trop Med Hyg 74, 1026-1033.

Stigler, S. M. (1982). Thomas Bayes's Bayesian Inference. Journal of the Royal Statistical Society-Series A Statistics in Society 145, 250-258.

Tebbs, J. M., Bilder, C. R., and Moser, B. K. (2003). An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. Communications in Statistics - Theory and Methods 32, 983-995.

Yameogo, L., Toe, L., Hougard, J. M., Boatin, B. A., and Unnasch, T. R. (1999). Pool screen polymerase chain reaction for estimating the prevalence of Onchocerca volvulus infection in Simulium damnosum sensu lato: results of a field trial in an area subject to successful vector control. Am J Trop Med Hyg 60, 124-128.

Yang, R. and Berger, J. O. (1998). A Catalog of Noninformative Priors. [unpublished manuscript]. http://www.stats.org.uk/priors/noninformative/YangBerger1998.pdf, accessed: 03/11/2011

THE CASE FOR A BAYESIAN APPROACH TO MONITORING

by

THOMAS BIRKNER, INMACULADA B. ABAN, CHARLES R. KATHOLI

In preparation for American Journal of Tropical Medicine and Hygiene

Format adapted for dissertation

1. INTRODUCTION

Onchocerciasis or River Blindness is a leading cause of skin disease and blindness in Africa and to lesser degree in the Americas. Embryonic microfilariae of the parasitic worm *Onchocerca volvulus* enter the human through the bite of a black fly (species Simulium) taking a blood meal. These microfilariae develop into worms inhabiting subcutaneous nodules. Female worms produce microfilariae, which swarm underneath the epidermis and may enter the eye. During subsequent bites transmission to other humans may occur. Ivermectin (Mectizan) is a safe drug that kills the microfilariae, but not the adult worms. Therefore it has to be given repeatedly over several years. The population at risk in the Americas is about 500,000. Onchocerciasis was initially endemic in 13 regions (foci) in six countries: Brazil, Colombia, Ecuador, Guatemala, Mexico and Venezuela. The "Onchocerciasis Elimination Program for the Americas" (OEPA) was launched in 1992 as a regional partnership that includes the governments of the endemic countries, the Pan American Health Organization, The Carter Center, Lions Clubs International, the US Centers for Disease Control and Prevention, the Bill and Melinda Gates Foundation, several universities, and the Mectizan Donation Program. OEPA's goal is the elimination of the disease in the Americas. In 2009, seven foci still required mass mectizan administration; transmission was interrupted in the other six foci. For an overview of the program we recommend Sauerbrey (2008), Blanks *et al.* (1998) and OEPA's annual reports published in World Health Organization's "Weekly Epidemiological Record".

To evaluate the progress of the elimination programs periodic screens for infection are conducted in the human and in the vector population in several communities within each region. For the latter, insects are caught and tested in groups using a Polymerase Chain Reaction (PCR) method. The outcome for each group is either positive (at least one fly was infected) or negative (none of the flies were infected). Given the number of pools tested, the number of insects per pool and the number of pools that tested positive, an infection prevalence estimate can be calculated. Several methods based on different assumptions and with different strengths and weaknesses are available to perform those calculations (see for instance: Chiang and Reeves, 1962; Tebbs, Bilder, Moser, 2003; Birkner, Aban, Katholi, 2011 [1] and [2]). The predominant approach so far has been Maximum Likelihood Estimation (MLE). The success of the elimination programs in reducing the infection prevalence to a very small value poses challenges for this method. This paper will outline the assumptions, advantages and drawbacks of several available methods (MLE, frequentist hierarchical model, empirical Bayes and classical Bayesian) by real data examples. The SAS code to perform the calculations is available from the authors. The second and third section outlines the different approaches divided into frequentist and Bayesian techniques, in the fourth section we present numerical results to compare the estimates resulting from the different methods. We end with an illustration of prior influence when the amount of data is small.

2. THE FREQUENTIST APPROACH

2.1. Maximum Likelihood Estimation

The current Pool Screening protocol is based on the *classical (frequentist) approach* to statistical inference. In this approach the quantity of interest (prevalence of infection, p) is a parameter in the model and is assumed to be fixed, but unknown. A sample of flies is collected, grouped into pools and tested. The results of these tests and the size of pools are the data used to estimate the prevalence of infection. Dorfman (1943), Chiang and Reeves (1962) and Thompson (1962) provide information regarding this standard method. The approach works well until we find no positive pools. The infection prevalence estimate in this case is zero, which appears to be an implausible value for the true prevalence especially when considering that the data stems from one sample only. One-sided confidence bounds [0, U] can be found, but the performance of these intervals can be poor (Tebbs and Bilder, 2004). A supposedly 95% confidence interval may actually have only 80% coverage, meaning that the random interval [0,U] covers the parameter only with 80% probability instead of the assumed 95%. Whenever the estimate falls on the boundary of the allowable range of the parameter ($0 \leq p \leq 1$) the frequentist approach gives poor results.

2.2. Estimation based on Hierarchical Model

In some applications, the assumption that the infection rate is a constant is not realistic. A way to address this issue is to specify a distribution for the prevalence (p). This introduces a lower level in the probability model. The value of p calculated at this level (assuming the parameters of the prevalence distribution are known) is passed on to the higher level, and used in the calculation of the probability of observing an infected pool. Usually the parameters are not known and the observed number of infected pools is used to estimate them. Due to the multi-level structure those models are called hierarchical models. Casella and Berger (2002, p. 162-168) provide a cursory overview of classic frequentist hierarchical models. A reasonable distribution for p would be the Beta distribution with parameters α=1 and β=unknown (Birkner, Aban, Katholi 2011 [1]). The larger the value for β the smaller becomes the value for p. The samples of flies are collected and processed in the same way. Given the number of pools, the pool size and the number of infected pools one can calculate an estimate for β. Given $\hat{\beta}$, one can obtain the "expected value" of the prevalence distribution, which would be analogous to the MLE. The interval between the appropriate upper and lower "critical points" of the Beta distribution with α=1 and β=$\hat{\beta}$ would be a credibility interval for p. The term 'credibility' interval is used to differentiate this interval from a Bayesian 'credible' interval. Unfortunately, this approach also has problems when there are no positive pools. The estimate of the parameter β becomes infinite. This implies that p has a distribution with all of its mass at zero.

3. THE BAYESIAN APPROACH

Another estimation approach that handles a non-constant infection prevalence is the Bayesian approach. The variable (here, the infection prevalence) has a distribution, known as the prior distribution, which incorporates available information and data before the initial experiment is conducted. After conducting the experiment, one can update the prior probability model by incorporating the information inherent in the collected data to obtain what is known as the "posterior" distribution. All inferences are made using this posterior model. Once the posterior distribution is found then all probability statements can be calculated. For example, one can calculate a credible interval and make a statement like: “Based on prior beliefs and the current data the probability that p is between p_1 and p_2 is 99%”. One can also find the probability that p is less than some transmission threshold value p_0.

Given these probability statements made on the basis of the posterior distribution, the investigator can make decisions based on the risk level with which he/she is comfortable.

3.1. Reasons to Consider the Bayesian Approach

A key assumption of the classical approach is that the quantity of interest is an unknown constant. If sampling is taking place over a large geographic area, then the assumption of a homogeneous population with the infection rate a constant may be a bit questionable. The infection prevalence is treated as a variable in the Bayesian approach – a variable having a distribution.

An added advantage to using the general Bayesian approach is that it can handle the case when no positive pools are observed. For any reasonable choice of prior distribution, the resulting posterior distribution will be well defined. It will not be a degenerate distribution with all its mass at a single point.

3.2. What about the Choice of the Prior?

Many investigators are uncomfortable with the Bayesian approach because of the need to specify a prior distribution. How does one choose a prior? Below are some priors that we will consider in this paper.

3.2.1. Empirical Bayes

A prior can be specified using historical data or even the data at hand. The latter approach is problematic because it uses the same data for the prior specification and also in the calculation of the posterior distribution. Using historical data (when available) to find parameter estimates of the prior distribution is the better approach. This requires that the current prevalence and the historical prevalence be considered a random sample from a common distribution. Since we are interested in the situation of very small prevalence values, the effect of further treatment (such as the mass drug administration) is small and the condition is satisfied for all practical purposes.

3.2.2. Objective priors

The uncertainty of specifying a prior can be alleviated by an “Objective Bayes” approach utilizing an objective (least informative) prior. There are several priors, which can be used in the case of the Pool screening model. We consider two: the Bayes/Laplace prior and the Jeffreys’ prior. Our experimentation shows that they lead to similar conclusions. If longitudinal data is available then the posterior distribution calculated from the previous sample can be used as prior distribution for the next round of estimation. This scenario we term “sequential Bayes” (see Birkner, Aban, Katholi 2011 [2]).

3.2.3. Subjectively elicited priors

On the other hand, the investigator has the option to incorporate his/her prior experience and beliefs into the analysis. This prior experience can be used to put a value on the parameters of a prior distribution. For example, suppose that based on years of measurement the investigator believes that the prevalence is a certain value, say p_0. The parameter of the prior can be chosen so that the expected value of the prior distribution is p_0.

4. ILLUSTRATION USING OEPA DATA

4. 1. About the Data

The data for all numerical examples that follow are real data collected by the OEPA[1]. For the following example, we are using data from Mexico from the year 2004. Community level pool screening results for two foci (Oaxaca and Southern Chiapas) are displayed in **Table 1**. Transmission has been interrupted in Oaxaca since 2008 (WHO, 2010 and Rodriguez-Perez *et al.*, 2010) and transmission is classified as suppressed in Southern Chiapas (WHO, 2010 and Rodriguez-Perez *et al.*, 2008). The pool size (number of flies per pool) in all examples is 50.

[1] Data files provided by Thomas R. Unnasch.

Table 1

Pool Screen results Mexico 2004.

Region	Community	Number	
		Pools	Infected pools
Oaxaca	La Esperanza	82	0
	Santa Maria La Chichina	28	0
	Santigo Lalopa	63	0
		173	0
Southern Chiapas	1 de Mayo	68	0
	Ampliacion Las Malvinas	27	0
	Estrella Roja	31	0
	Jose Maria Morelos	86	7
	Las Golondrinas	66	0
	Las Nubes	62	1
	Nueva Costa Rica	54	0
	Nueva Reforma Agraria	67	0
		461	8

The classical frequentist estimates obtained by the Maximum Likelihood method and the estimates from the frequentist hierarchical model are shown in **Table 2**.

Table 2

Prevalence estimates from traditional and hierarchical model approaches for Mexico 2004 data.

Region	Traditional frequentist estimate (95% confidence interval)	Frequentist hierarchical model estimate (95% credibility interval) P (p≤1/2000)
Oaxaca	0 (0, 3.46)*	0 nd** 1***
Southern Chiapas	3.50 (1.51, 6.90)	3.53 (0.09, 13.02) 0.757
Both	2.54 (1.10, 5.00)	2.56 (0.06, 9.42) 0.859

* calculated for all of α = 0.05 in upper tail; ** nd = not defined for point mass distribution; *** distribution reduces to point mass of 1 at 0 and by right continuity property: P ($p \leq p_0$) = 1 for any p_0 ($0 \leq p_0 \leq 1$); point and interval estimates expressed per 10,000 flies.

From Table 2 we notice that both approaches give a point estimate of 0 when 0 positive pools are observed for a particular year (Oaxaca focus). We also see that the point estimates derived using either method are very close (Southern Chiapas focus). The credibility interval from the hierarchical model is wider compared to the standard method confidence interval, reflecting the distribution of prevalence values at the community level. Only the hierarchical model approach allows calculating the probabilities of being below a specified prevalence threshold.

To overcome the limitation of the small number of pools at the community level we compare the estimators also at the foci level by combining data from the Oaxaca and Southern Chiapas regions. Within the hierarchical model context this amounts to allowing for the prevalence to vary between the two regions instead between the communities within a region. We reach the same conclusions as above – nearly identical prevalence point estimates and a wider credibility interval.

4.2. Empirical Bayes

For comparison purposes, it is worth considering the Empirical Bayes approach suggested by Tebbs, Bilder and Moser (2003). The prior distribution in their analysis has the form of the Beta density with α=1 and β=unknown. Under this method the data is used twice - first to calculate the unknown parameter in the prior distribution and secondly in the likelihood function.

A more typical empirical Bayes approach is to use historical data to calculate the parameters of the prior distribution. For the example below, data from Mexico from the year 2000 is utilized to calculate the α and β parameters of a three parameter Beta distribution, with the third parameter being set equal to the pool size. The historical data (8 communities) allow to obtain two pieces of information (mean and variance), which is required to estimate the two parameters.

Table 3 summarizes the estimates for the two empirical Bayes approaches. The standard method estimates from Table 2 are included for comparison.

Table 3

Empirical Bayes Estimates for Mexico 2004 data.

	Traditional frequentist estimate	**Empirical Bayes Estimate with prior based on**	
Region	**(95% confidence interval)**	**Current sample** **(95% credible interval)** **P ($p \leq 1/2000$)**	**Historical data** **(95% credible interval)** **P ($p \leq 1/2000$)**
Oaxaca	0 (0, 3.46)*	0 nd** 1***	0.39 (0.00004, 2.27) 0.999
Southern Chiapas	3.50 (1.51, 6.90)	3.50 (1.60, 6.13) 0.893	3.55 (1.57, 6.33) 0.878
Both	2.54 (1.10, 5.00)	2.54 (1.16, 4.45) 0.992	2.60 (1.15, 4.63) 0.987

* calculated for all of $\alpha = 0.05$ in upper tail; ** nd = not defined for point mass distribution; *** posterior reduces to point mass of 1 at 0 and by right continuity property: $P(p \leq p_0) = 1$ for any p_0 ($0 \leq p_0 \leq 1$); point and interval estimates expressed per 10,000 flies.

A great degree of agreement between the point as well as interval estimates is apparent. The empirical Bayes estimates are almost identical to the standard method estimates.

It is important to note that the one sample Empirical Bayes approach suffers from the same problem as the standard method when there are no positive pools. In this case, the estimate of the parameter in the prior distribution is infinity and so the method fails. The empirical Bayes approach utilizing historical data avoids this issue as long as the historic data contain a few communities whose pool screen results were positive.

4.3. Objective Bayes

Table 4 contains the prevalence estimates, 95% credible intervals and threshold probabilities for the two objective priors chosen and the standard estimates for comparison.

Table 4

Objective Bayes Estimates for Mexico 2004 data.

Region	**Traditional frequentist estimate** **(95% confidence interval)**	**Bayes/Laplace prior Estimate** **(95% credible interval)** **P (p≤1/2000)**	**Jeffreys' prior Estimate** **(95% credible interval)** **P (p≤ 1/2000)**
Oaxaca	0 (0, 3.46)*	1.15 (0.03, 4.24) 0.987	0.58 (0.0005, 2.90) 0.997
Southern Chiapas	3.50 (1.51, 6.90)	3.93 (1.80, 6.88) 0.806	3.72 (1.65, 6.60) 0.846
Both	2.54 (1.10, 5.00)	2.85 (1.30, 5.00) 0.975	2.70 (1.20, 4.79) 0.983

* calculated for all of α = 0.05 in upper tail; point and interval estimates expressed per 10,000 flies.

As expected the point estimates of infection prevalence are not 0 for the Bayesian estimators even if no positive pools are observed (Oaxaca focus). The lower bounds of the Bayesian credible intervals are greater than 0 as well. The credible intervals themselves agree to a greater degree and are also more similar to the confidence intervals in the case that some pools test positive. The Bayes/Laplace and Jeffreys' prior prevalence estimates are close to each other and only slightly above the MLE (Southern Chiapas focus). The probability of the prevalence being below 1 in 2000 flies is approximately 0.99 in Oaxaca and greater 0.8 in Southern Chiapas. Those probability estimates cannot be obtained from the standard approach.

4.4. Sequential Bayes

Since prevalence estimates are needed repeatedly over the course of a treatment program more pool screening results become available as time goes on. A natural approach of deriving a prior would be a new specification of the prior parameters each time more data becomes available. Instead of estimating say α and β in the case of the 3 parameter Beta prior repeatedly we proceeded to apply the previous posterior distribution as the prior distribution for the next round of testing.

Table 5 displays data from two communities (San Miguel and El Tigre) in Ecuador (see Vieira *et al*. 2007 for background information regarding the elimination program in Ecuador). Data was collected every second month from November 1995 through November 1996.

Table 5

Pool screening results from two communities in Ecuador from 1996.

	San Miguel Number		**El Tigre Number**	
	Pools	**Infected Pools**	**Pools**	**Infected Pools**
Nov95	6	0	3	0
Jan96	18	2	18	3
Mar96	31	2	25	10
May96	26	12	28	4
Jul96	34	2	26	0
Sep96	5	1	1	0
Nov96	5	0	1	0

Figure 1 displays the prevalence estimates for the Bayes/Laplace and Jeffreys' prior when the posterior distribution is successively updated with data from the next month in which samples were collected. The objective priors are used in the calculations for the November 1995 estimate (assuming nothing or little is known about the infection prevalence at this point). Thereafter the posterior distribution estimated based on the previous round of testing is employed as the prior distribution for the next round of testing (for example: the posterior from November 1995 is the prior for January 1996). Figure 1 also displays the estimates from the classical Maximum Likelihood estimation approach. In order to allow for a fair comparison with the sequential Bayes approach, sequential ML estimates based on successively adding results from the bi-monthly tests were calculated and are also shown in Figure 1.

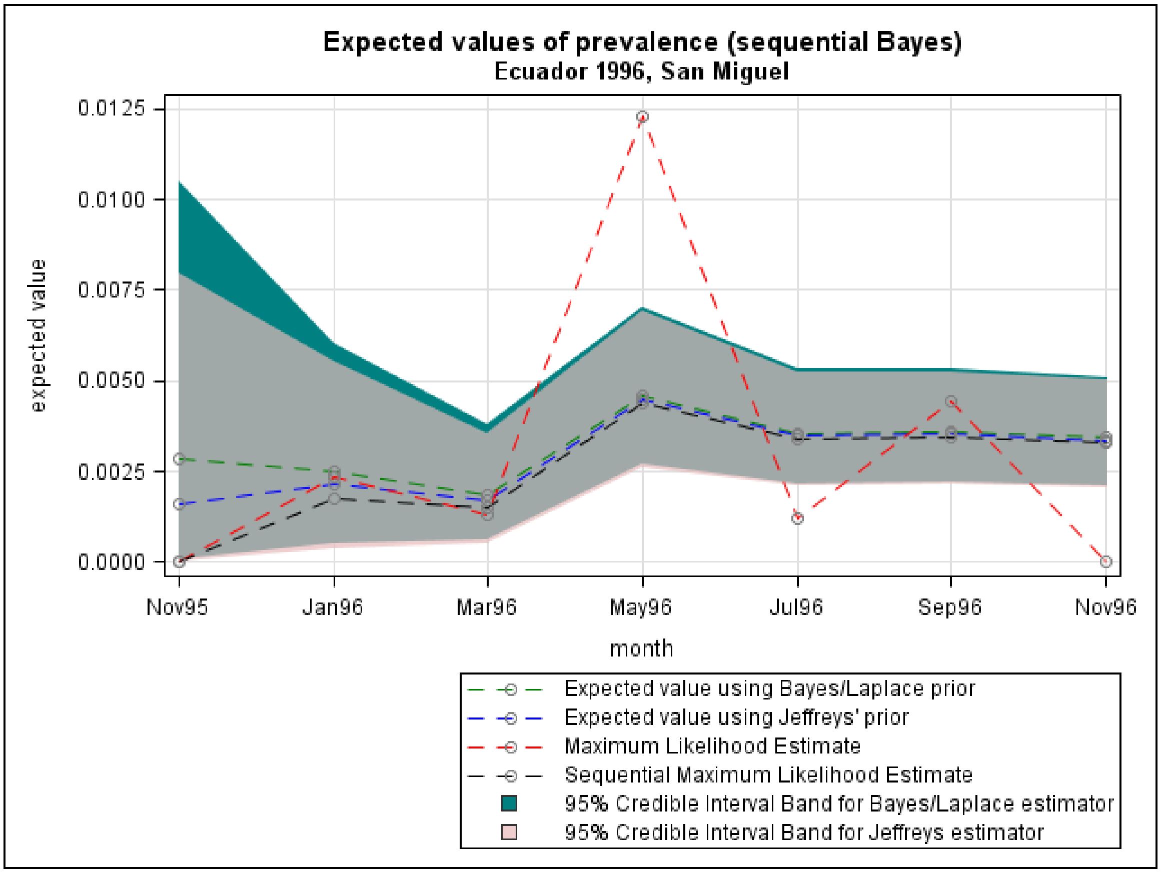


Figure 1. Sequential Bayes estimates compared to MLE and sequential MLE based on data from San Miguel.

We observe that the sequential Bayes/Laplace and Jeffreys estimates are not zero in November 1995 despite observing zero positive pools. This is clearly an effect of the priors. The amount of data (6 negative pools of 50 flies each) that month is not great enough to dominate the priors and push the expected value to 0. Overall the Bayesian estimates are changing in a less erratic manner compared to the classical MLE. They are comparable to the sequential MLE approach, the only difference being that they never take a value of zero as the sequential MLE does in November 1995 (all pools tested negative in the first month data was collected). After all seven months of data are included the three sequential prevalence estimators are very close in value (E_B/L= 0.00344, E_J=0.00337, E_MLE_Seq=0.00329).

A benefit of the Bayesian approach is that the original choice of the prior distribution "washes out" over successive data collections. Differences between the Bayes/Laplace and Jeffreys' prior estimates based on the Ecuador 1996 data have essentially vanished as the fourth batch of data (May 1996) is included in the estimation (see Figure 1). This observation has also been made for simulated data in Birkner, Aban, Katholi (2011) [2]. Ideally the sequential approach should be applied to data from consecutive years where results from more recent years is given a greater weight compared to earlier results. Without any weighting the estimates are slow to adjust to changes in the true prevalence as shown in Birkner, Aban and Katholi (2011) [2].

4.5. The Prior matters when the Amount of Data is Small

Even a least informative or objective prior carries some information. The prior can be pictured as additional data. In general, if an adequate number of pools is tested then those results will dominate/override the information provided by the prior. **Figures 2** and **3** show the difference between sufficient information to overwhelm the prior and the case where the prior has a great impact on determining the estimate. Here we naively employed the objective Bayes approach to each bi-monthly dataset separately (naively, because of the small number of flies tested per month). We observe that the Bayes estimates track well with the MLE as long as the number of pools is greater 15. Differences emerge for small pool sizes and in the case that zero positive pools are observed. The estimates for September and November 1996 for El Tigre display the greatest divergence (Figure 3). For each of those months only one pool was tested and found to contain no infected flies. The MLE estimate is 0 and the estimate based on the Bayes/Laplace prior approximately 0.01 with the Jeffreys' estimate slightly below. Given an adequate sample size the difference between the classical frequentist and Bayesian approaches is negligible.

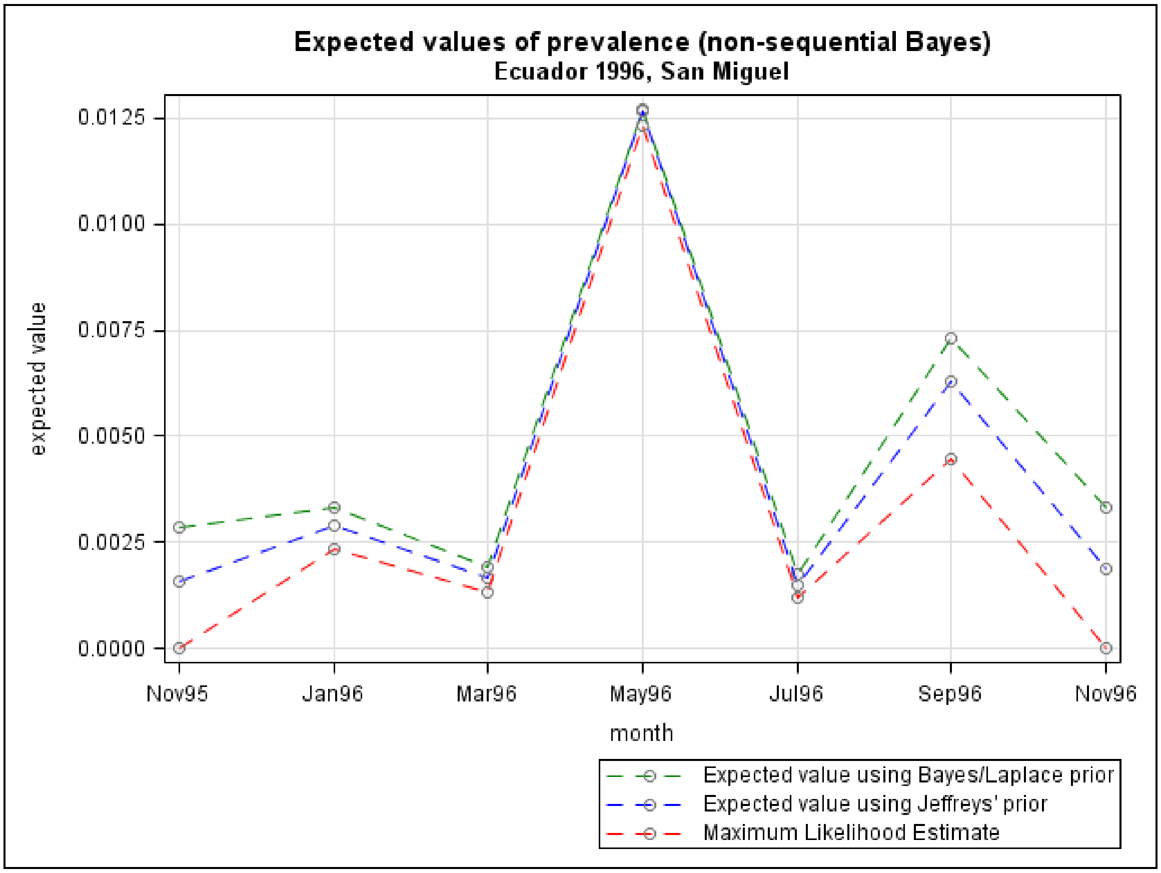


Figure 2. Objective Bayes estimators compared to MLE (San Miguel, Ecuador 1996).

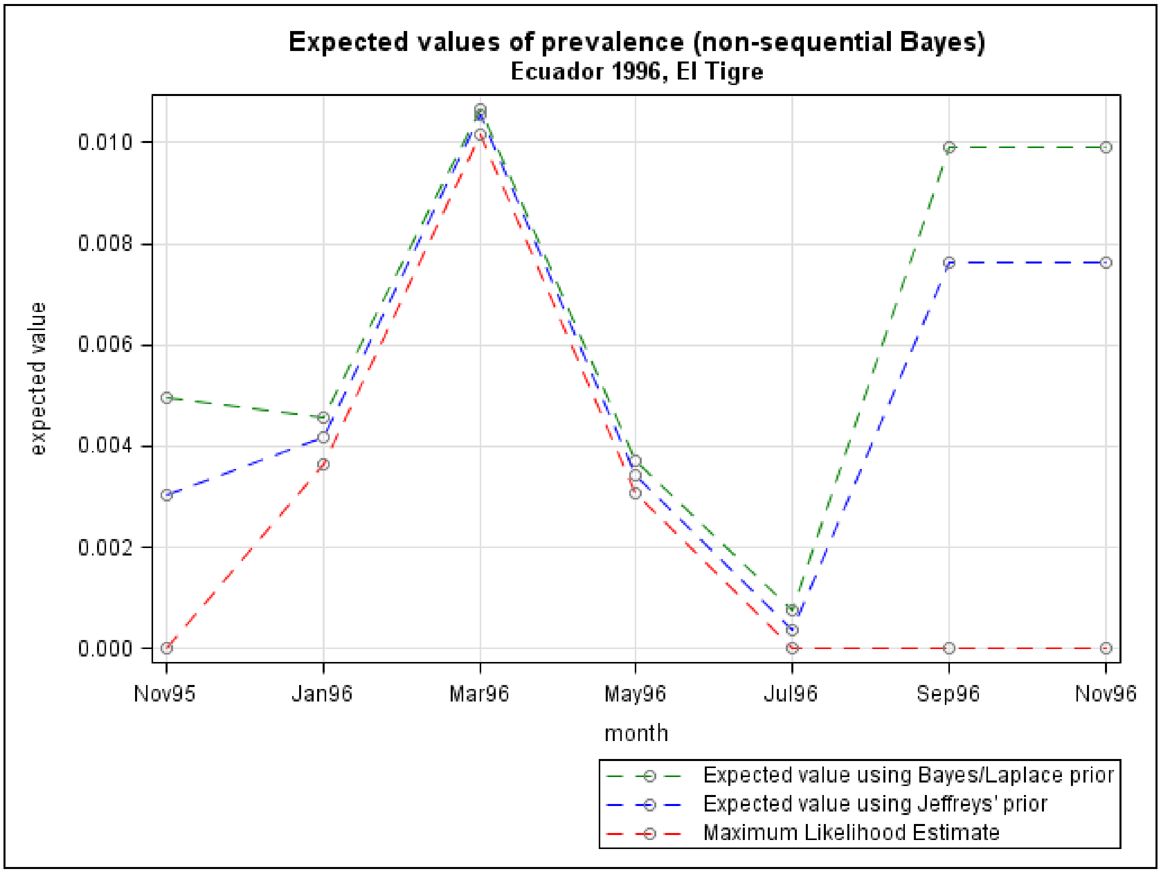


Figure 3. Objective Bayes estimators compared to MLE (El Tigre, Ecuador 1996).

5. CONCLUSIONS

Since the intent of OEPA is to eliminate River Blindness in the Americas, it is inevitable that at some point the investigator will begin to find no positive pools. In this case, the frequentist approach begins to perform poorly (in the sense that it produces an unrealistic point estimate of 0 and coverage probabilities at times below the nominal level and at other times excessively conservative). The Bayesian approach with various prior specifications (empirical Bayes utilizing historical data, objective priors, sequential Bayes) does not lead to this problem. Moreover, the approach naturally allows for different prevalence values within a broader sampling region and also provides the possibility of probability statements, such as P ($p \leq 1/2000$). The Bayesian point and interval estimates are strikingly close to the frequentist ones provided that we observe a few positive pools. In this case the choice between the frequentist and Bayesian approaches is inconsequential. Bayesian estimation techniques are preferable when zero infected pools are observed.

REFERENCES

Birkner, T., Aban, I. B., Katholi, C. R. (2011) [1]. Evaluation of a Frequentist Hierarchical Model to estimate prevalence when sampling from a large geographic area using pool screening (unpublished manuscript).

Birkner, T., Aban, I. B., Katholi, C. R. (2011) [2]. Bayesian approaches for estimating prevalence based on pool screening when observing zero positive pools (unpublished manuscript).

Blanks, J., Richards, F., Beltran, F., Collins, R., Alvarez, E., Zea Flores, G., Bauler, B., Cedillos, R., Heisler, M., Brandling-Bennett, D., Baldwin, W., Bayona, M, Klein, R., Jacox, M. (1998). The Onchocerciasis Elimination Program for the Americas: a history of partnership. *Pan Am J Public Health* **3**, 367-374.

Casella, G. and Berger, R. L. (2002). Statistical Inference. 2nd edition. Duxbury, Pacific Grove.

Chiang, C. L., and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. *American Journal of Hygiene* **75**, 377-391.

Dorfman, R. (1943). The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics* **14**, 436-440.

Rodriguez-Perez, M. A., Lutzow-Steiner, M. A., Segura-Cabrera, A. et al. (2008). Rapid suppression of Onchocerca volvulus transmission in two communities of the Southern Chiapas focus, Mexico, achieved by quarterly treatments with Mectizan. *Am J Trop Med Hyg*, **79**:239-244.

Rodriguez-Perez , M. A., Unnasch, T. R., Dominguez-Vazquez, A. et al. (2010). Interruption of transmission of Onchocerca volvulus in the Oaxaca focus, Mexico. *Am J Trop Med Hyg*. **83**, 21-27.

Sauerbrey, M. (2008). The Onchocerciasis Elimination Program for the Americas (OEPA). *Annals of Tropical Medicine & Parasitology* **102**, Supplement No. 1, 25-29.

Tebbs, J.M., Bilder, C.R., and Moser, B.K. (2003). An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics – Theory and Methods* **32**, 983-995.

Tebbs, J.M. and Bilder, C.R. (2004). Confidence Interval Procedures for the Probability of Disease Transmission in Multiple-Vector-Transfer Designs. *Journal of Agricultural, Biological, and Environmental Statistics* **9**, 75-90.

Thompson, K. H. (1962). Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics* **18**, 568-578.

Vieira, J. C., Cooper, P. J., Lovato, R. et al. (2007). Impact of long-term treatment of onchocerciasis with ivermectin in Ecuador: potential for elimination of infection. *BMC Med*. **5:9**.

WHO (2010). Report from the 2009 InterAmerican Conference on Onchocerciasis: progress towards eliminating river blindness in the Region of the Americas. *Weekly epidemiological record* **33**, 321-327.

CONCLUSION

Summary

Two statistical issues have been recognized when estimating the infection prevalence of viral or parasitic diseases in a vector population based on pooled samples. One issue is the assumption of the traditional estimation method, that the infection prevalence is constant throughout a possible large sampling region. The treatment of the infection prevalence as unknown and fixed parameter in the traditional frequentist approach ignores good reasons for a varying prevalence across a region, such as differences in vector habitat (for instance, the existence of fast flowing streams), the existence and distribution of highly and less effective vectors, and differences in the implementation of the treatment program.

The second issue is a consequence of the success of the treatment programs in reducing the infection prevalence in the vector population from between one and five percent to levels close to zero. It becomes very likely to observe only non-infected pools when the prevalence is that low. The point estimate obtained by the traditional method in this case is zero, a value that most likely does not equal the true prevalence.

To address the first issue we proposed a Bernoulli-Beta hierarchical model, which assumes a Beta(α, β) distribution for the prevalence. The prevalence is treated as random variable instead of being a fixed parameter. Due to limited information in the likelihood we imposed the constraint α=1 and found the Maximum Likelihood estimator for β.

The choice of α=1 is justified by showing that the resulting prevalence distribution is neither biased towards nor away from zero. Given the values of α and $\hat{\beta}$ we derived the new prevalence estimate $\widehat{E(p)}$, an interval estimator and an estimate of the probability of the prevalence being below a specified value. This threshold probability cannot be computed under the traditional model, but is a value frequently inquired about by program staff to aid them in deciding when to end a treatment program. We found that all estimators perform well in terms of the usual measures of merit for frequentist estimators, such as Variance, Bias and Mean Squared Error. We also showed that they are consistent estimators. Given these findings we are confident that the proposed Bernoulli-Beta model can replace the traditional model whenever the assumption of constant prevalence across the region is in doubt.

As a solution to the second issue we proposed and investigated a sequential Bayesian approach, which incorporates all available pool screening results (historical and current) for a particular region. Under the Bayesian paradigm all information gathered before the current sampling is contained in the prior distribution, which is updated by the current data to form the posterior distribution. The prevalence estimate is derived from the posterior distribution and hence depends not only on the current data (for instance, data containing no positive pools)—but also on the prior distribution. The use of an objective prior in the first year and of the previous posterior distribution as prior in the next year for all subsequent years in our approach resolves the difficult question of prior choice. The conjugate property of the two objective priors considered produces posteriors that are Beta distributions for any number of pool screening results. Any software containing the Beta distribution function can be used to calculate percentiles and credible intervals.

A simulation study showed that the choice between the Bayes/Laplace and Jeffreys' prior is inconsequential after results of five to eight years have been incorporated. The study revealed an inertia problem, in the sense that the posterior distributions are slow to adjust to changes in prevalence. The sequential Bayes estimates will over- or underestimate the true prevalence depending on the underlying trend.

Three strategies to reduce the inertia were considered: (1) continue the pool screening efforts for some additional years, (2) omit early years, and (3) use a non-sequential approach. We found that strategy (1) has only a small effect on the prevalence estimates, strategy (2) results in a similar small correction, but without the need for additional data, and strategy (3) produces estimates just slightly above the traditional ones on average. All three strategies exhibit a strong impact on the threshold probabilities, resulting in more realistic values. Differences in the estimators due to the prior choice disappear in the sequential approach, but are maintained in the non-sequential formulation. Within the sequential framework we recommend the use of strategy (2), provided that at least five years worth of results are still available after the omission of a number of early results. Further research is needed exploring the use of a subjective or empirical Bayes prior in the non-sequential framework.

The third paper uses data from the "Onchocerciasis Elimination Program for the Americas" to make numerical comparisons between the traditional, hierarchical and non-sequential Bayesian type approaches. Point and interval estimates, as well as the threshold probability P ($p \leq 5/10000$) were calculated. Empirical Bayes estimators using either only the current data or current and historical data were also considered. The point and interval estimators are fairly similar across all approaches as long as the sample contains some positive pools. If this is not the case then only the Bayesian type approaches produce a non-zero point estimate and non-zero lower bound for the interval estimate.

By using longitudinal data collected bi-monthly over the duration of one year the differences between the traditional estimate and the sequential Bayes estimates are demonstrated. The accumulation of information in the sequential approach produces more consistent estimates compared to the fluctuating traditional estimate. Calculating the traditional estimate in a sequential mode resolves that difference and produces an estimate for the whole year very similar to the sequential Bayes one. Another point that is illustrated graphically is that even objective priors contain some information resulting in non-zero prevalence estimates even if no positive pools are observed. As expected the influence of the prior gets stronger as the number of pools in the sample gets smaller.

Limitations and Future Research

The assumption of perfect sensitivity and specificity of the test simplifies the evaluation of the proposed methods but is not necessarily true. Generally tests do not achieve such levels. Incorporating measurement error in the simulations would allow a more realistic comparison between the Bayesian and traditional methods.

The methods proposed are applicable when all pools are of equal size. However, due to a lack of knowledge about the underlying prevalence investigators sometimes purposefully design a study where different pool sizes are used sequentially (for instance see: Hepworth, 1996). Also, the use of unequal sized pools might be desirable to avoid excessive handling of the vectors. Instead of forming equal sized pools, one might want to test whatever number of insects is collected within the specified sampling time. The extension of the proposed methods for unequal pool sizes will increase their usefulness.

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).

Other modifications to reduce the inertia in the sequential Bayes method need to be considered. One idea is the use of a sliding window approach – include for instance five years worth of data in the computations and then start over with next five years and so on. This will reduce the impact of earlier data while still incorporating information beyond the current sample. Another idea is to introduce some weighting scheme to increase the impact of the most recent results.

The simulations assumed the availability of pool screen data for every year from the exact same communities within a region over a 20 year span. In reality entomological evaluations are not undertaken on a yearly basis and the communities included can vary to a degree. It is important to investigate the behavior of the sequential Bayes estimator given such data and to incorporate mechanisms to handle unequal temporal distances between samples.

In the third paper we calculated empirical Bayes estimates using historical data to specify the prior. We went beyond approaches already available (for instance, see Tebbs, Bilder, Moser 2003) by specifying a three parameter Beta prior distribution, where the parameters are $\alpha, \beta, \gamma = n$. This prior is conjugate for the Binomial likelihood resulting in a Beta posterior distribution. The performance of this Bayesian approach needs to be evaluated as possible alternative to the proposed sequential model for longitudinal data. The development of an easy to use algorithm for the entomologist (for instance: R program) to specify the prior parameters is another priority.

GENERAL LIST OF REFERENCES

Amazigo, U., Noma, M., Bump, J., Benton, B., Liese, B., Yameogo, H. Z., and Seketeli, A. (2006). Onchocerciasis. In Disease and Mortality in Sub-Saharan Africa. (eds. D.T. Jamison, R. G. Feachurn, M. W. Makgoba, E. R. Bos, F. K. Baingana, K. J. Hofman, and K. O. Rogoet), Second Edition. Washington (DC): World Bank.

Barker, J. T. (2000). Statistical estimators of infection potential based on PCR pool screening with unequal pool sizes. University of Alabama at Birmingham, Birmingham.

Bhattacharyya, G. K., Karandinos, M. G., and DeFoliart, G. R. (1979). Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies. *Am J Epidemiol* **109**, 124-131.

Bilder, C. R., and Tebbs, J. M. (2005). Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biom J* **47**, 502-516.

Boswell, M. T., Gore, S. D., Patil, G. P., and Taillie, C. (1992). A Bayesian approach to classifying samples as polluted or not polluted. *Technical Report 92-0801*.

Burrows, P. M. (1987). Improved Estimation of Pathogen Transmission Rates by Group Testing. *Phytopathology* **77**, 363-365.

Chaubey, Y. P., and Li, W. (1995). Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binomial Probability with Sample Compositing. *Journal of Official Statistics* **11**, 379-390.

Chiang, C. L., and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. *American Journal of Hygiene* **75**, 377-391.

Chick, S. E. (1996). Bayesian Models for Limiting Dilution Assay and Group Test Data. *Biometrics* **52**, 1055-1062

Dorfman, R. (1943). The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics* **14**, 436-440.

Gao, H. (2010). Hypothesis testing in unequal sized pool screening. University of Alabama at Birmingham, Birmingham.

Gart, J.J. (1991). An application of score methodology: confidence intervals and tests of fit for one-hit curves. In Handbook of Statistics (eds. C.R. Rao and R. Chakraborty), vol. 8, pp. 395-406. Amsterdam: Elsevier.

Guevara, A. G., Vieira, J. C., Lilley, B. G., *et al.* (2003). Entomological evaluation by pool screen polymerase chain reaction of Onchocerca volvulus transmission in Ecuador following mass Mectizan distribution. *Am J Trop Med Hyg* **68**, 222-227.

Hepworth, G. (1996). Exact Confidence Intervals for Proportions Estimated by Group Testing. *Biometrics* **52**, 1134-1146.

Hepworth, G., and Watson, R. (2009). Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society-Series C Applied Statistics* **58**, 105-121.

Katholi, C. R., Toe, L., Merriweather, A., and Unnasch, T. R. (1995). Determining the prevalence of Onchocerca volvulus infection in vector populations by polymerase chain reaction screening of pools of black flies. *J Infect Dis* **172**, 1414-1417.

Katholi, C. R., and Unnasch, T. R. (2006). Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches. *Am J Trop Med Hyg* **74**, 779-785.

McV. Messam, L. L., Branscum, A. J., Collins, M. T., and Gardner, I. A. (2008). Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease. *Anim Health Res Rev* **9**, 1-23.

Rodriguez-Perez, M. A., Katholi, C. R., Hassan, H. K., and Unnasch, T. R. (2006). Large-scale entomologic assessment of Onchocerca volvulus transmission by poolscreen PCR in Mexico. *Am J Trop Med Hyg* **74**, 1026-1033.

Swallow, W. H. (1985). Group Testing for Estimating Infection Rates and Probabilities of Disease Transmission. *Phytopathology* **75**, 882-889.

Tebbs, J. M., and Bilder, C. R. (2004). Confidence interval procedures for the probability of disease transmission in Multiple-Vector-Transfer Designs. *Journal of Agricultural, Biological, and Environmental Statistics* **9**, 75-90.

Tebbs, J. M., Bilder, C. R., and Moser, B. K. (2003). An Empirical Bayes Group-Testing Approach to Estimating Small Proportions. *Communications in Statistics - Theory and Methods* **32**, 983-995.

Tebbs, J.M. and McCann, M.H. (2007). Large-Sample Hypothesis Tests for Stratified Group-Testing Data. *Journal of Agricultural, Biological and Environmental Statistics* **12**, 534-551.

Tu, X.M., Litvak, E., and Pagano, M. (1995). On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika* **82**, 287-297.

Thompson, K. H. (1962). Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics* **18**, 568-578.

Walter, S. D., Hildreth, S. W., and Beaty, B. J. (1980). Estimation of infection rates in population of organisms using pools of variable size. *Am J Epidemiol* **112**, 124-128.

Watson, M. (1936). Factors Affecting the Amount of Infection Obtained by Aphis Transmission of the virus Hy. III. *Philosophical Transactions of the Royal Society of London*, Series B, **226**, 457–489.

Yameogo, L., Toe, L., Hougard, J. M., Boatin, B. A., and Unnasch, T. R. (1999). Pool screen polymerase chain reaction for estimating the prevalence of Onchocerca volvulus infection in Simulium damnosum sensu lato: results of a field trial in an area subject to successful vector control. *Am J Trop Med Hyg* **60**, 124-128.

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):

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

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

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

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

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

Goal: evaluate $\frac{dk}{d\alpha}$ and $\frac{dk}{d\beta}$, start by simplifying k

Recall: $\mathrm{B}(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Hence $k=\ln\left\{\dfrac{\dfrac{\Gamma(\alpha)\Gamma(\beta+n)}{\Gamma(\alpha+\beta+n)}}{\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}\right\}=\ln\left\{\dfrac{\Gamma(\beta+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\beta)}\right\}$

By properties of log:

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

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

Using a fact introducing the digamma function: $\dfrac{d}{d\alpha}\ln\Gamma(\alpha)=\Psi(\alpha)$

plus an application of the chain rule: $\dfrac{d}{d\alpha}\ln\Gamma(\zeta(\alpha))=\Psi(\zeta)\dfrac{d\zeta}{d\alpha}$

We obtain:

$$\frac{dk}{d\alpha} = \Psi\left(\alpha+\beta\right) - \Psi\left(\alpha+\beta+n\right)$$

$$\frac{dk}{d\beta} = \Psi\left(\beta+n\right) + \Psi\left(\alpha+\beta\right) - \Psi\left(\alpha+\beta+n\right) - \Psi\left(\beta\right)$$

Using another result:

$$\Psi\left(\alpha+\beta+n\right) = \frac{1}{\alpha+\beta} + \frac{1}{\left(\alpha+\beta+1\right)} + \ldots + \frac{1}{\left(\alpha+\beta+n-1\right)} + \Psi\left(\alpha+\beta\right)$$

$$\begin{aligned}\frac{dk}{d\alpha} &= \Psi\left(\alpha+\beta\right) - \Psi\left(\alpha+\beta+n\right) \\ &= \Psi\left(\alpha+\beta\right) - \frac{1}{\alpha+\beta} - \frac{1}{\left(\alpha+\beta+1\right)} - \ldots - \frac{1}{\left(\alpha+\beta+n-1\right)} - \Psi\left(\alpha+\beta\right) \\ &= -\sum_{j=0}^{n-1} \frac{1}{\alpha+\beta+j}\end{aligned}$$

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

$$\frac{dk}{d\beta} = \Psi\left(\alpha+\beta\right) - \Psi\left(\alpha+\beta+n\right) + \Psi\left(\beta+n\right) - \Psi\left(\beta\right)$$

$$= -\sum_{j=0}^{n-1} \frac{1}{\alpha+\beta+j} + \frac{1}{\beta} + \frac{1}{\beta+1} + \ldots + \frac{1}{\beta+\left(n-1\right)} + \Psi\left(\beta\right) - \Psi\left(\beta\right)$$

$$= \sum_{j=0}^{n-1} \frac{1}{\beta+j} - \sum_{j=0}^{n-1} \frac{1}{\alpha+\beta+j} = \sum_{j=0}^{n-1} \left[\frac{\alpha+\beta+j-\beta-j}{\left(\beta+j\right)\left(\alpha+\beta+j\right)} \right] = \sum_{j=0}^{n-1} \frac{\alpha}{\left(\beta+j\right)\left(\alpha+\beta+j\right)}$$

, which is > 0 for $\alpha, \beta > 0$ and finite.

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

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)

We defined $k = \ln\left[\Gamma\left(\beta+n\right)\right] + \ln\left[\Gamma\left(\alpha+\beta\right)\right] - \ln\left[\Gamma\left(\alpha+\beta+n\right)\right] - \ln\left[\Gamma\left(\beta\right)\right]$

For $\alpha = 1$ and using the fact that $\Gamma(\alpha+1) = \alpha\Gamma\left(\alpha\right)$, for $\alpha > 0$

$$\begin{aligned}
k &= \ln\left[\Gamma\left(\beta+n\right)\right]+\ln\left[\Gamma\left(1+\beta\right)\right]-\ln\left[\Gamma\left(1+\beta+n\right)\right]-\ln\left[\Gamma\left(\beta\right)\right] \\
&= \ln\left[\Gamma\left(\beta+n\right)\right]+\ln\left[\beta*\Gamma\left(\beta\right)\right]-\ln\left[\left(\beta+n\right)\Gamma\left(\beta+n\right)\right]-\ln\left[\Gamma\left(\beta\right)\right] \\
&= \ln\left[\Gamma\left(\beta+n\right)\right]+\ln\beta+\ln\Gamma\left(\beta\right)-\ln\left(\beta+n\right)-\ln\Gamma\left(\beta+n\right)-\ln\left[\Gamma\left(\beta\right)\right] \\
&= \ln\beta-\ln\left(\beta+n\right)=\ln\left(\frac{\beta}{\beta+n}\right)
\end{aligned}$$

Plugging this back in:

$$\sum_{i=1}^{M}\left\{\left(1-x_i\right)+x_i\frac{-e^{\ln\left(\frac{\beta}{\beta+n}\right)}}{1-e^{\ln\left(\frac{\beta}{\beta+n}\right)}}\right\}=\sum_{i=1}^{M}\left\{\left(1-x_i\right)-x_i\frac{\frac{\beta}{\beta+n}}{1-\frac{\beta}{\beta+n}}\right\}$$

$$=\sum_{i=1}^{M}\left\{\left(1-x_i\right)-x_i\frac{\frac{\beta}{\beta+n}}{\frac{\beta+n-\beta}{\beta+n}}\right\}=\sum_{i=1}^{M}\left\{\left(1-x_i\right)-x_i\frac{\beta}{n}\right\}$$

$$=\sum_{i=1}^{M}1-\sum_{i=1}^{M}x_i-\frac{\beta}{n}\sum_{i=1}^{M}x_i=M-T-\frac{\beta}{n}T\overset{!}{=}0$$

Where $T=\sum_{i=1}^{M}x_i$, the number of positive pools.

Solve for $\hat{\beta}$

$$M - T = \frac{\beta}{n} T$$

$$\frac{M}{T} - 1 = \frac{\beta}{n}$$

$$\hat{\beta} = \frac{Mn}{T} - n$$

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 $E\left(\hat{\beta}\right) = nmE\left(\frac{1}{T}\right) - n$. To find $E\left(\frac{1}{T}\right)$ we use the Taylor series expansion. We define $f\left(T\right) = \frac{1}{T}$

and let T be a random variable such that $E\left(T\right) = \mu_T$. We shall expand $f\left(T\right)$ about μ_T.

Taylor series expansion:

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

We also calculated $E(X) = \frac{n}{\beta + n}$ (for each Bernoulli trial) and

$$E(T) = \sum_{j=1}^{m} E(X_j) = \frac{mn}{\beta + n} = \mu_T$$

Note:

$$\frac{\delta}{\delta\mu_T}\mu_T^{-1} = -1\mu_T^{-2} = -\left(\frac{1}{\mu_T}\right)^2$$

$$\frac{\delta}{\delta\mu_T}-1\mu_T^{-2} = (-1)(-2)\,\mu_T^{-3} = 2!\left(\frac{1}{\mu_T}\right)^3$$

$$\frac{\delta}{\delta\mu_T}2!\,\mu_T^{-3} = 2!(-3)\,\mu_T^{-4} = -3!\left(\frac{1}{\mu_T}\right)^4$$

$$\frac{\delta}{\delta\mu_T}-3!\,\mu_T^{-4} = -3!(-4)\,\mu_T^{-5} = 4!\left(\frac{1}{\mu_T}\right)^5$$

$$\frac{\delta}{\delta\mu_T}4!\,\mu_T^{-5} = 4!(-5)\,\mu_T^{-6} = -5!\left(\frac{1}{\mu_T}\right)^6$$

$$\frac{\delta}{\delta\mu_T}-5!\,\mu_T^{-6} = -5!(-6)\,\mu_T^{-7} = 6!\left(\frac{1}{\mu_T}\right)^7$$

$$\frac{\delta}{\delta\mu_T}6!\,\mu_T^{-7} = 6!(-7)\,\mu_T^{-8} = -7!\left(\frac{1}{\mu_T}\right)^8$$

$$\frac{\delta}{\delta\mu_T} - 7!\,\mu_T^{-8} = -7!(-8)\,\mu_T{}^{-9} = 8!\left(\frac{1}{\mu_T}\right)^9$$

$$\vdots$$

Next we rewrite the Taylor series expansion for our scenario:

$$f(T) = \frac{1}{\mu_T} - \left(\frac{1}{\mu_T}\right)^2 (T-\mu_T) + \left(\frac{1}{\mu_T}\right)^3 (T-\mu_T)^2 - \left(\frac{1}{\mu_T}\right)^4 (T-\mu_T)^3 + \left(\frac{1}{\mu_T}\right)^5 (T-\mu_T)^4 - \left(\frac{1}{\mu_T}\right)^6 (T-\mu_T)^5 +$$
$$\left(\frac{1}{\mu_T}\right)^7 (T-\mu_T)^6 - \left(\frac{1}{\mu_T}\right)^8 (T-\mu_T)^7 + \left(\frac{1}{\mu_T}\right)^9 (T-\mu_T)^8 \ldots$$

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

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

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):

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

We find:

$$E\left(T-\frac{mn}{\beta+n}\right)^{2}=\frac{mn\beta}{\left(\beta+n\right)^{2}}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^{3}=\frac{-mn\beta\left(n-\beta\right)}{\left(\beta+n\right)^{3}}=\frac{mn\beta\left(\beta-n\right)}{\left(\beta+n\right)^{3}}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^{4}=\frac{mn\beta\left(\beta^{2}-4\beta n+n^{2}+3mn\beta\right)}{\left(\beta+n\right)^{4}}=\frac{mn\beta\left(\beta^{2}-4\beta n+n^{2}\right)+\left(mn\beta\right)^{2}3}{\left(\beta+n\right)^{4}}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^{5}=\frac{mn\beta\left(10mn\beta^{2}-10mn^{2}\beta+\beta^{3}-11\beta^{2}n+11\beta n^{2}-n^{3}\right)}{\left(\beta+n\right)^{5}}=\frac{mn\beta\left(\beta^{3}-11\beta^{2}n+11\beta n^{2}-n^{3}\right)+\left(mn\beta\right)^{2}\left(10\beta-10n\right)}{\left(\beta+n\right)^{5}}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^6=\frac{mn\beta\left(15m^2n^2\beta^2+25mn\beta^3-80mn^2\beta^2+25mn^3\beta+\beta^4-26\beta^3n+66\beta^2n^2-26\beta n^3+n^4\right)}{\left(\beta+n\right)^6}$$

$$=\frac{mn\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)+\left(mn\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)+\left(mn\beta\right)^3\left(15\right)}{\left(\beta+n\right)^6}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^7=\frac{mn\beta\left(-56mn^4\beta-105m^2n^3\beta^2+105m^2n^2\beta^3+56mn\beta^4-406mn^2\beta^3+406mn^3\beta^2\right)}{\left(\beta+n\right)^7}$$

$$-\frac{57\beta^4n+302\beta^3n^2-302\beta^2n^3+57\beta n^4-n^5+\beta^5}{\left(\beta+n\right)^7}$$

$$=\frac{mn\beta\left(\beta^5-57n\beta^4+302n^2\beta^3-302n^3\beta^2+57n^4\beta-n^5\right)+\left(mn\beta\right)^2\left(56\beta^3-406n\beta^2+406n^2\beta-56n^3\right)}{\left(\beta+n\right)^7}$$

$$+\frac{\left(mn\beta\right)^3\left(105\beta-105n\right)}{\left(\beta+n\right)^7}$$

$$E\left(T-\frac{mn}{\beta+n}\right)^8=\frac{mn\beta\left(-1680mn^2\beta^4+3710mn^3\beta^3-1680mn^4\beta^2+119mn^5\beta+105m^3n^3\beta^3+490m^2n^2\beta^4-1400m^2n^3\beta^3\right)}{\left(\beta+n\right)^8}$$

$$+\frac{490m^2n^4\beta^2+119mn\beta^5-120\beta^5n+1191\beta^4n^2-2416\beta^3n^3+1191\beta^2n^4-120\beta n^5+n^6+\beta^6}{\left(\beta+n\right)^8}$$

$$=\frac{mn\beta\left(\beta^6-120n\beta^5+1191n^2\beta^4-2416n^3\beta^3+1191n^4\beta^2-120n^5\beta+n^6\right)}{\left(\beta+n\right)^8}$$

$$+\frac{\left(mn\beta\right)^2\left(-1680n\beta^3+3710n^2\beta^2-1680n^3\beta+119n^4+119\beta^4\right)+\left(mn\beta\right)^3\left(490\beta^2-1400n\beta+490n^2\right)+\left(mn\beta\right)^4105}{\left(\beta+n\right)^8}$$

Plugging these moments into $E\left(f\left(T\right)\right)$ we obtain:

$$
\begin{aligned}
E\left(f\left(T\right)\right) = {} & \frac{\beta+n}{mn} + \left(\frac{\beta+n}{mn}\right)^3 \frac{mn\beta}{\left(\beta+n\right)^2} - \left(\frac{\beta+n}{mn}\right)^4 \frac{mn\beta\left(\beta-n\right)}{\left(\beta+n\right)^3} + \left(\frac{\beta+n}{mn}\right)^5 \frac{mn\beta\left(\beta^2-4\beta n+n^2\right)+\left(mn\beta\right)^2 3}{\left(\beta+n\right)^4} \\
& -\left(\frac{\beta+n}{mn}\right)^6 \frac{mn\beta\left(\beta^3-11\beta^2 n+11\beta n^2-n^3\right)+\left(mn\beta\right)^2\left(10\beta-10n\right)}{\left(\beta+n\right)^5} \\
& +\left(\frac{\beta+n}{mn}\right)^7 \frac{mn\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)+\left(mn\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)+\left(mn\beta\right)^3\left(15\right)}{\left(\beta+n\right)^6} \\
& -\left(\frac{\beta+n}{mn}\right)^8 \frac{mn\beta\left(\beta^5-57n\beta^4+302n^2\beta^3-302n^3\beta^2+57n^4\beta-n^5\right)+\left(mn\beta\right)^2\left(56\beta^3-406n\beta^2+406n^2\beta-56n^3\right)+\left(mn\beta\right)^3\left(105\beta-105n\right)}{\left(\beta+n\right)^7} \\
& +\left(\frac{\beta+n}{mn}\right)^9 \left[\begin{array}{l} \dfrac{mn\beta\left(\beta^6-120n\beta^5+1191n^2\beta^4-2416n^3\beta^3+1191n^4\beta^2-120n^5\beta+n^6\right)}{\left(\beta+n\right)^8} \\ +\dfrac{\left(mn\beta\right)^2\left(-1680n\beta^3+3710n^2\beta^2-1680n^3\beta+119n^4+119\beta^4\right)+\left(mn\beta\right)^3\left(490\beta^2-1400n\beta+490n^2\right)+\left(mn\beta\right)^4 105}{\left(\beta+n\right)^8} \end{array}\right]
\end{aligned}
$$

$$
\begin{aligned}
E\left(f\left(T\right)\right) = {} & \frac{\beta+n}{mn} + \frac{\left(\beta+n\right)\beta}{\left(mn\right)^2} - \frac{\left(\beta+n\right)\beta\left(\beta-n\right)}{\left(mn\right)^3} + \left(\frac{\beta+n}{mn}\right)^5 \frac{mn\beta\left(\beta^2-4\beta n+n^2\right)}{\left(\beta+n\right)^4} + \left(\frac{\beta+n}{mn}\right)^5 \frac{\left(mn\beta\right)^2 3}{\left(\beta+n\right)^4} \\
& - \left(\frac{\beta+n}{mn}\right)^6 \frac{mn\beta\left(\beta^3-11\beta^2 n+11\beta n^2-n^3\right)}{\left(\beta+n\right)^5} - \left(\frac{\beta+n}{mn}\right)^6 \frac{\left(mn\beta\right)^2\left(10\beta-10n\right)}{\left(\beta+n\right)^5} \\
& + \left(\frac{\beta+n}{mn}\right)^7 \frac{mn\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)}{\left(\beta+n\right)^6} + \left(\frac{\beta+n}{mn}\right)^7 \frac{\left(mn\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)}{\left(\beta+n\right)^6} + \left(\frac{\beta+n}{mn}\right)^7 \frac{\left(mn\beta\right)^3\left(15\right)}{\left(\beta+n\right)^6} \\
& - \left(\frac{\beta+n}{mn}\right)^8 \frac{mn\beta\left(\beta^5-57n\beta^4+302n^2\beta^3-302n^3\beta^2+57n^4\beta-n^5\right)}{\left(\beta+n\right)^7} - \left(\frac{\beta+n}{mn}\right)^8 \frac{\left(mn\beta\right)^2\left(56\beta^3-406n\beta^2+406n^2\beta-56n^3\right)}{\left(\beta+n\right)^7} \\
& - \left(\frac{\beta+n}{mn}\right)^8 \frac{\left(mn\beta\right)^3\left(105\beta-105n\right)}{\left(\beta+n\right)^7} \\
& + \left(\frac{\beta+n}{mn}\right)^9 \frac{mn\beta\left(\beta^6-120n\beta^5+1191n^2\beta^4-2416n^3\beta^3+1191n^4\beta^2-120n^5\beta+n^6\right)}{\left(\beta+n\right)^8} \\
& + \left(\frac{\beta+n}{mn}\right)^9 \frac{\left(mn\beta\right)^2\left(-1680n\beta^3+3710n^2\beta^2-1680n^3\beta+119n^4+119\beta^4\right)}{\left(\beta+n\right)^8} \\
& + \left(\frac{\beta+n}{mn}\right)^9 \frac{\left(mn\beta\right)^3\left(490\beta^2-1400n\beta+490n^2\right)}{\left(\beta+n\right)^8} + \left(\frac{\beta+n}{mn}\right)^9 \frac{\left(mn\beta\right)^4 105}{\left(\beta+n\right)^8}
\end{aligned}
$$

Simplifying:

$$E\left(f\left(T\right)\right)=\frac{\beta+n}{mn}+\frac{\left(\beta+n\right)\beta}{\left(mn\right)^{2}}-\frac{\left(\beta+n\right)\beta\left(\beta-n\right)}{\left(mn\right)^{3}}+\frac{\left(\beta+n\right)\beta\left(\beta^{2}-4\beta n+n^{2}\right)}{\left(mn\right)^{4}}+\frac{\left(\beta+n\right)\left(\beta\right)^{2}3}{\left(mn\right)^{3}}$$

$$-\frac{\left(\beta+n\right)\beta\left(\beta^{3}-11\beta^{2}n+11\beta n^{2}-n^{3}\right)}{\left(mn\right)^{5}}-\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(10\beta-10n\right)}{\left(mn\right)^{4}}$$

$$+\frac{\left(\beta+n\right)\beta\left(\beta^{4}-26n\beta^{3}+66n^{2}\beta^{2}-26n^{3}\beta+n^{4}\right)}{\left(mn\right)^{6}}+\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(25\beta^{2}-80n\beta+25n^{2}\right)}{\left(mn\right)^{5}}+\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(15\right)}{\left(mn\right)^{4}}$$

$$-\frac{\left(\beta+n\right)\beta\left(\beta^{5}-57n\beta^{4}+302n^{2}\beta^{3}-302n^{3}\beta^{2}+57n^{4}\beta-n^{5}\right)}{\left(mn\right)^{7}}-\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(56\beta^{3}-406n\beta^{2}+406n^{2}\beta-56n^{3}\right)}{\left(mn\right)^{6}}$$

$$-\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(105\beta-105n\right)}{\left(mn\right)^{5}}+\frac{\left(\beta+n\right)\beta\left(\beta^{6}-120n\beta^{5}+1191n^{2}\beta^{4}-2416n^{3}\beta^{3}+1191n^{4}\beta^{2}-120n^{5}\beta+n^{6}\right)}{\left(mn\right)^{8}}$$

$$+\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(-1680n\beta^{3}+3710n^{2}\beta^{2}-1680n^{3}\beta+119n^{4}+119\beta^{4}\right)}{\left(mn\right)^{7}}$$

$$+\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(490\beta^{2}-1400n\beta+490n^{2}\right)}{\left(mn\right)^{6}}+\frac{\left(\beta+n\right)\left(\beta\right)^{4}105}{\left(mn\right)^{5}}$$

Ordering powers:

$$E\left(f\left(T\right)\right)=\frac{\beta+n}{mn}+\frac{\left(\beta+n\right)\beta}{\left(mn\right)^{2}}$$

$$-\frac{\left(\beta+n\right)\beta\left(\beta-n\right)}{\left(mn\right)^{3}}+\frac{\left(\beta+n\right)\left(\beta\right)^{2}3}{\left(mn\right)^{3}}$$

$$+\frac{\left(\beta+n\right)\beta\left(\beta^{2}-4\beta n+n^{2}\right)}{\left(mn\right)^{4}}-\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(10\beta-10n\right)}{\left(mn\right)^{4}}+\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(15\right)}{\left(mn\right)^{4}}$$

$$-\frac{\left(\beta+n\right)\beta\left(\beta^{3}-11\beta^{2}n+11\beta n^{2}-n^{3}\right)}{\left(mn\right)^{5}}+\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(25\beta^{2}-80n\beta+25n^{2}\right)}{\left(mn\right)^{5}}-\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(105\beta-105n\right)}{\left(mn\right)^{5}}+\frac{\left(\beta+n\right)\left(\beta\right)^{4}105}{\left(mn\right)^{5}}$$

$$+\frac{\left(\beta+n\right)\beta\left(\beta^{4}-26n\beta^{3}+66n^{2}\beta^{2}-26n^{3}\beta+n^{4}\right)}{\left(mn\right)^{6}}-\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(56\beta^{3}-406n\beta^{2}+406n^{2}\beta-56n^{3}\right)}{\left(mn\right)^{6}}+\frac{\left(\beta+n\right)\left(\beta\right)^{3}\left(490\beta^{2}-1400n\beta+490n^{2}\right)}{\left(mn\right)^{6}}$$

$$-\frac{\left(\beta+n\right)\beta\left(\beta^{5}-57n\beta^{4}+302n^{2}\beta^{3}-302n^{3}\beta^{2}+57n^{4}\beta-n^{5}\right)}{\left(mn\right)^{7}}+\frac{\left(\beta+n\right)\left(\beta\right)^{2}\left(-1680n\beta^{3}+3710n^{2}\beta^{2}-1680n^{3}\beta+119n^{4}+119\beta^{4}\right)}{\left(mn\right)^{7}}$$

$$+\frac{\left(\beta+n\right)\beta\left(\beta^{6}-120n\beta^{5}+1191n^{2}\beta^{4}-2416n^{3}\beta^{3}+1191n^{4}\beta^{2}-120n^{5}\beta+n^{6}\right)}{\left(mn\right)^{8}}$$

Combining terms:

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

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

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

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

APPENDIX B2

MEAN SQUARED ERROR

Recall MSE = $E_{\theta}\left(W-\theta\right)^{2} = Var_{\theta}W + \left(E_{\theta}W-\theta\right)^{2} = Var_{\theta}W + \left(Bias_{\theta}W\right)^{2}$

First we need the expression for $\hat{\beta}$, above we only found $E\left(\hat{\beta}\right)$. We have the MLE for $\alpha = 1: \hat{\beta} = \frac{mn}{T} - n$

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

$$\hat{\beta} = mn\left(\frac{\beta+n}{mn} - \left(\frac{\beta+n}{mn}\right)^{2}\left(T-\frac{mn}{\beta+n}\right) + \left(\frac{\beta+n}{mn}\right)^{3}\left(T-\frac{mn}{\beta+n}\right)^{2}\right.$$

$$\left. - \left(\frac{\beta+n}{mn}\right)^{4}\left(T-\frac{mn}{\beta+n}\right)^{3} + \left(\frac{\beta+n}{mn}\right)^{5}\left(T-\frac{mn}{\beta+n}\right)^{4} - \ldots\right) - n$$

$$\hat{\beta} = \beta - \frac{(\beta+n)^{2}}{mn}\left(T-\frac{mn}{\beta+n}\right) + \frac{(\beta+n)^{3}}{(mn)^{2}}\left(T-\frac{mn}{\beta+n}\right)^{2} - \frac{(\beta+n)^{4}}{(mn)^{3}}\left(T-\frac{mn}{\beta+n}\right)^{3} + \frac{(\beta+n)^{5}}{(mn)^{4}}\left(T-\frac{mn}{\beta+n}\right)^{4} - \frac{(\beta+n)^{6}}{(mn)^{5}}\left(T-\frac{mn}{\beta+n}\right)^{5} + \ldots$$

$$\hat{\beta} - \beta = -\frac{(\beta+n)^{2}}{mn}\left(T-\frac{mn}{\beta+n}\right) + \frac{(\beta+n)^{3}}{(mn)^{2}}\left(T-\frac{mn}{\beta+n}\right)^{2} - \frac{(\beta+n)^{4}}{(mn)^{3}}\left(T-\frac{mn}{\beta+n}\right)^{3} + \frac{(\beta+n)^{5}}{(mn)^{4}}\left(T-\frac{mn}{\beta+n}\right)^{4} - \frac{(\beta+n)^{6}}{(mn)^{5}}\left(T-\frac{mn}{\beta+n}\right)^{5} + \ldots$$

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

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

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

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

Note in this case $c_0 = c_1 = 0$, since $c_1 = 2a_0 a_1 = 2 * 0 * a_1 = 0$

$$c_2 = a_1a_1 = a_1^2$$

$$c_3 = a_1a_2 + a_2a_1 = 2a_1a_2$$

$$c_4 = a_1a_3 + a_2a_2 + a_3a_1 = 2a_1a_3 + a_2^2$$

$$c_5 = a_1a_4 + a_2a_3 + a_3a_2 + a_4a_1 = 2a_1a_4 + 2a_2a_3$$

$$c_6 = a_1a_5 + a_2a_4 + a_3a_3 + a_4a_2 + a_5a_1 = 2a_1a_5 + 2a_2a_4 + a_3^2$$

$$a_1 = -\frac{\left(\beta + n\right)^2}{mn}$$

$$a_2 = \frac{\left(\beta + n\right)^3}{\left(mn\right)^2}$$

$$a_3 = -\frac{\left(\beta + n\right)^4}{\left(mn\right)^3}$$

$$a_4 = \frac{\left(\beta + n\right)^5}{\left(mn\right)^4}$$

$$a_5 = -\frac{\left(\beta + n\right)^6}{\left(mn\right)^5}$$

$$c_2 = a_1^2 = \frac{(\beta + n)^4}{\left(mn\right)^2}$$

$$c_3 = 2a_1a_2 = 2\left(-\frac{\left(\beta + n\right)^2}{mn}\right)\left(\frac{\left(\beta + n\right)^3}{\left(mn\right)^2}\right) = -2\frac{\left(\beta + n\right)^5}{\left(mn\right)^3}$$

$$c_4 = 2a_1a_3 + a_2^2 = 2\left(-\frac{\left(\beta + n\right)^2}{mn}\right)\left(-\frac{\left(\beta + n\right)^4}{\left(mn\right)^3}\right) + \frac{\left(\beta + n\right)^6}{\left(mn\right)^4} = 2\frac{\left(\beta + n\right)^6}{\left(mn\right)^4} + \frac{\left(\beta + n\right)^6}{\left(mn\right)^4} = 3\frac{\left(\beta + n\right)^6}{\left(mn\right)^4}$$

$$c_5 = 2a_1a_4 + 2a_2a_3 = 2\left(-\frac{(\beta+n)^2}{mn}\right)\left(\frac{(\beta+n)^5}{(mn)^4}\right) + 2\left(\frac{(\beta+n)^3}{(mn)^2}\right)\left(-\frac{(\beta+n)^4}{(mn)^3}\right) = -2\frac{(\beta+n)^7}{(mn)^5} - 2\frac{(\beta+n)^7}{(mn)^5} = -4\frac{(\beta+n)^7}{(mn)^5}$$

$$c_6 = 2a_1a_5 + 2a_2a_4 + a_3^2 = 2\left(-\frac{(\beta+n)^2}{mn}\right)\left(-\frac{(\beta+n)^6}{(mn)^5}\right) + 2\left(\frac{(\beta+n)^3}{(mn)^2}\right)\left(\frac{(\beta+n)^5}{(mn)^4}\right) + \left(-\frac{(\beta+n)^4}{(mn)^3}\right)^2$$
$$= 2\frac{(\beta+n)^8}{(mn)^6} + 2\frac{(\beta+n)^8}{(mn)^6} + \frac{(\beta+n)^8}{(mn)^6} = 5\frac{(\beta+n)^8}{(mn)^6}$$

$$\left(\hat{\beta}-\beta\right)^2 = c_2\left(T-\mu_T\right)^2 + c_3\left(T-\mu_T\right)^3 + c_4\left(T-\mu_T\right)^4 + \ldots$$
$$= \frac{(\beta+n)^4}{(mn)^2}\left(T-\mu_T\right)^2 - 2\frac{(\beta+n)^5}{(mn)^3}\left(T-\mu_T\right)^3 + 3\frac{(\beta+n)^6}{(mn)^4}\left(T-\mu_T\right)^4 - 4\frac{(\beta+n)^7}{(mn)^5}\left(T-\mu_T\right)^5 + 5\frac{(\beta+n)^8}{(mn)^6}\left(T-\mu_T\right)^6 - \ldots$$

Taking expectation on both sides:

$$E(\hat{\beta}-\beta)^2 = MSE = \frac{(\beta+n)^4}{(mn)^2}E(T-\mu_T)^2 - 2\frac{(\beta+n)^5}{(mn)^3}E(T-\mu_T)^3 + 3\frac{(\beta+n)^6}{(mn)^4}E(T-\mu_T)^4 - 4\frac{(\beta+n)^7}{(mn)^5}E(T-\mu_T)^5$$
$$+5\frac{(\beta+n)^8}{(mn)^6}E(T-\mu_T)^6 - \ldots$$

Taking above piece by piece:

$$c_2E(T-\mu_T)^2=\frac{(\beta+n)^4}{(mn)^2}\frac{mn\beta}{\left(\beta+n\right)^2}=\frac{(\beta+n)^2\beta}{mn}$$

$$c_3E(T-\mu_T)^3=-2\frac{\left(\beta+n\right)^5}{\left(mn\right)^3}\frac{mn\beta\left(\beta-n\right)}{\left(\beta+n\right)^3}=\frac{-2\left(\beta+n\right)^2\beta\left(\beta-n\right)}{\left(mn\right)^2}$$

$$c_4E(T-\mu_T)^4=3\frac{\left(\beta+n\right)^6}{\left(mn\right)^4}\frac{mn\beta\left(\beta^2-4\beta n+n^2\right)+\left(mn\beta\right)^2 3}{\left(\beta+n\right)^4}=\frac{3\left(\beta+n\right)^6}{\left(mn\right)^4}\frac{mn\beta\left(\beta^2-4\beta n+n^2\right)}{\left(\beta+n\right)^4}+\frac{3\left(\beta+n\right)^6}{\left(mn\right)^4}\frac{\left(mn\beta\right)^2 3}{\left(\beta+n\right)^4}$$

$$=\frac{3\left(\beta+n\right)^2\beta\left(\beta^2-4\beta n+n^2\right)}{\left(mn\right)^3}+\frac{9\left(\beta+n\right)^2\left(\beta\right)^2}{\left(mn\right)^2}$$

$$c_5E(T-\mu_T)^5=-4\frac{\left(\beta+n\right)^7}{\left(mn\right)^5}\frac{mn\beta\left(\beta^3-11\beta^2n+11\beta n^2-n^3\right)+\left(mn\beta\right)^2\left(10\beta-10n\right)}{\left(\beta+n\right)^5}$$

$$=-4\frac{\left(\beta+n\right)^7}{\left(mn\right)^5}\frac{mn\beta\left(\beta^3-11\beta^2n+11\beta n^2-n^3\right)}{\left(\beta+n\right)^5}-4\frac{\left(\beta+n\right)^7}{\left(mn\right)^5}\frac{\left(mn\beta\right)^2\left(10\beta-10n\right)}{\left(\beta+n\right)^5}$$

$$=-4\frac{\left(\beta+n\right)^2\beta\left(\beta^3-11\beta^2n+11\beta n^2-n^3\right)}{\left(mn\right)^4}-4\frac{\left(\beta+n\right)^2\left(\beta\right)^2\left(10\beta-10n\right)}{\left(mn\right)^3}$$

$$c_6 E(T-\mu_T)^6 = 5\frac{\left(\beta+n\right)^8}{\left(mn\right)^6}\frac{mn\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)+\left(mn\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)+\left(mn\beta\right)^3\left(15\right)}{\left(\beta+n\right)^6}$$

$$=5\frac{\left(\beta+n\right)^8}{\left(mn\right)^6}\frac{mn\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)}{\left(\beta+n\right)^6}+5\frac{\left(\beta+n\right)^8}{\left(mn\right)^6}\frac{\left(mn\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)}{\left(\beta+n\right)^6}+5\frac{\left(\beta+n\right)^8}{\left(mn\right)^6}\frac{\left(mn\beta\right)^3\left(15\right)}{\left(\beta+n\right)^6}$$

$$=5\frac{\left(\beta+n\right)^2\beta\left(\beta^4-26n\beta^3+66n^2\beta^2-26n^3\beta+n^4\right)}{\left(mn\right)^5}+5\frac{\left(\beta+n\right)^2\left(\beta\right)^2\left(25\beta^2-80n\beta+25n^2\right)}{\left(mn\right)^4}+\frac{75\left(\beta+n\right)^2\left(\beta\right)^3}{\left(mn\right)^3}$$

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

$$E(\widehat{\beta}-\beta)^2 = \frac{(\beta+n)^2\beta}{mn}-\frac{2\left(\beta+n\right)^2\beta\left(\beta-n\right)}{\left(mn\right)^2}+\frac{9\left(\beta+n\right)^2\left(\beta\right)^2}{\left(mn\right)^2}+\frac{3\left(\beta+n\right)^2\beta\left(\beta^2-4\beta n+n^2\right)}{\left(mn\right)^3}-4\frac{\left(\beta+n\right)^2\left(\beta\right)^2\left(10\beta-10n\right)}{\left(mn\right)^3}$$
$$+\frac{75\left(\beta+n\right)^2\left(\beta\right)^3}{\left(mn\right)^3}$$

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

APPENDIX B3

VARIANCE

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

(MSE = Var + Bias^2 => Var=MSE-Bias^2)

$$Var(\hat{\beta}) \approx \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(7\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2\beta(38\beta^2+28\beta n+3n^2)}{(mn)^3} - \frac{(\beta+n)^2\beta^2}{(mn)^2} - \frac{2\beta^2(\beta+n)^2(2\beta+n)}{(mn)^3}$$

$$Var\left(\hat{\beta}\right) \approx \frac{(\beta+n)^2\beta}{mn} + \frac{2(\beta+n)^2\beta(3\beta+n)}{(mn)^2} + \frac{(\beta+n)^2\beta(34\beta^2+26\beta n+3n^2)}{(mn)^3} = \frac{(\beta+n)^2\beta}{mn} + \frac{(\beta+n)^2\beta(6\beta+2n)}{(mn)^2} + \frac{(\beta+n)^2\beta(34\beta^2+26\beta n+3n^2)}{(mn)^3}$$

$$\left[= \frac{(\beta+n)^2\beta\left(m^2n^2+6\beta mn+2mn^2+34\beta^2+26\beta n+3n^2\right)}{(mn)^3} \right]$$

Difference in Var and MSE is only in 2nd order term and above

Computation of Bias^2

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

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

$$a_0 = 0$$

$$a_1 = \frac{(\beta+n)\beta}{(mn)}$$

$$a_2 = \frac{(\beta+n)\beta(2\beta+n)}{(mn)^2}$$

$$a_3 = \frac{(\beta+n)\beta\left(6\beta^2+6\beta n+n^2\right)}{(mn)^3}$$

$$a_4 = \frac{(\beta+n)\beta\left(24\beta^3+36\beta^2 n+14\beta n^2+n^3\right)}{(mn)^4}$$

$$c_1 = \sum_{j=1}^{0} a_j a_{1-j} = 0$$

$$c_2 = \sum_{j=1}^{1} a_j a_{2-j} = a_1^2 = \frac{(\beta + n)^2 \beta^2}{(mn)^2}$$

$$c_3 = \sum_{j=1}^{2} a_j a_{3-j} = a_1 a_2 + a_2 a_1 = 2a_1 a_2 = \frac{2\beta^2 (\beta + n)^2 (2\beta + n)}{(mn)^3}$$

$$c_4 = \sum_{j=1}^{3} a_j a_{4-j} = a_1 a_3 + a_2 a_2 + + a_3 a_1 = 2a_1 a_3 + a_2^2 = \ldots = \frac{\beta^2 (\beta + n)^2 \left(16\beta^2 + 16\beta n + 3n^2\right)}{(mn)^4}$$

$$(Bias)^2 \approx \frac{(\beta + n)^2 \beta^2}{(mn)^2} + \frac{2\beta^2 (\beta + n)^2 (2\beta + n)}{(mn)^3} + \frac{\beta^2 (\beta + n)^2 \left(16\beta^2 + 16\beta n + 3n^2\right)}{(mn)^4}$$

APPENDIX C1

ESTIMATED EXPECTED VALUE OF PREVALENCE

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

For $\alpha = 1$

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

$$\left[\mathrm{B}(1,\hat{\beta}) = \frac{\Gamma(1)\Gamma(\hat{\beta})}{\Gamma(1+\hat{\beta})} = \frac{\Gamma(\hat{\beta})}{\hat{\beta} * \Gamma(\hat{\beta})} = \frac{1}{\hat{\beta}}\right]$$

$$\widehat{E(p)} = \hat{\beta} * \mathrm{B}(2,\hat{\beta}) = \hat{\beta}\frac{\Gamma(2)\Gamma(\hat{\beta})}{\Gamma(2+\hat{\beta})} = \frac{\hat{\beta} * (\hat{\beta}-1)!}{(\hat{\beta}+1)!} = \frac{\hat{\beta}!}{(\hat{\beta}+1)!} = \frac{1}{\hat{\beta}+1}$$

APPENDIX C2

CREDIBILITY INTERVAL

Expressions for lower and upper bound for p

$p \sim \frac{p^{\alpha-1}(1-p)^{\hat{\beta}-1}}{\mathrm{B}\left(\alpha,\hat{\beta}\right)}$, for $\alpha = 1$ this simplifies to $\hat{\beta}(1-p)^{\hat{\beta}-1}$

$$\int_{p_L}^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} = 1-\alpha$$

$$\int_{0}^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = -\left(1-p\right)^{\hat{\beta}}\Big|_{0}^{p_H} = 1-(1-p_H)^{\hat{\beta}}$$

$$\int_{p_H}^{1} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = 1-\int_{0}^{p_H} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = 1-\left(1-(1-p_H)^{\hat{\beta}}\right) = (1-p_H)^{\hat{\beta}} \overset{!}{=} \frac{\alpha}{2}$$

$$\Rightarrow (1-p_H) = \left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}} \Rightarrow p_H = 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$$

$$\int_{0}^{p_L} \hat{\beta}(1-p)^{\hat{\beta}-1} dp = -\left(1-p\right)^{\hat{\beta}}\Big|_{0}^{p_L} = 1-(1-p_L)^{\hat{\beta}} \overset{!}{=} \frac{\alpha}{2}$$

$$1-\frac{\alpha}{2}=\left(1-p_L\right)^{\hat{\beta}}\Rightarrow\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}=\left(1-p_L\right)\Rightarrow p_L=1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$$

The upper bound of 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

Taylor series expansion:

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

$$\frac{d}{d\beta}\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}=-\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^2}$$

$$\frac{d^2}{d\beta^2}\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}=\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^3}$$

$$\frac{d^3}{d\beta^3}\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}=-\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^3}{\beta^6}-\frac{6\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^5}-\frac{6\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^4}$$

$$\frac{d^4}{d\beta^4}\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}=\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^4}{\beta^8}+\frac{12\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^3}{\beta^7}+\frac{36\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^6}+\frac{24\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^5}$$

$$f\left(\hat{\beta}\right)\approx\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}-\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^2}\left(\hat{\beta}-\beta\right)+\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{\left(\hat{\beta}-\beta\right)^2}{2!}$$

$$p_H \approx 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^2}\left(\hat{\beta}-\beta\right)-\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{\left(\hat{\beta}-\beta\right)^2}{2!}$$

$$E_\beta\left(p_H\right)\approx 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^2}E\left(\hat{\beta}-\beta\right)-\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{1}{2}E\left(\hat{\beta}-\beta\right)^2$$

$$E_\beta\left(p_H\right)\approx 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^2}Bias-\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{1}{2}MSE$$

We found bias and MSE earlier:

$$Bias\left(\hat{\beta}\right)=\frac{\left(\beta+n\right)\beta}{mn}+\frac{\left(\beta+n\right)\beta\left(2\beta+n\right)}{\left(mn\right)^2}+\ldots$$

$$MSE\left(\hat{\beta}\right)=\frac{\left(\beta+n\right)^{2}\beta}{mn}+\frac{\left(\beta+n\right)^{2}\beta\left(7\beta+2n\right)}{\left(mn\right)^{2}}+\ldots$$

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

$$E_{\beta}\left(p_{H}\right)\approx 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^{2}}\frac{\left(\beta+n\right)\beta}{mn}-\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^{2}}{\beta^{4}}+\frac{2\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^{3}}\right)\frac{1}{2}\frac{\left(\beta+n\right)^{2}\beta}{mn}$$

$$E_{\beta}\left(p_{H}\right)\approx 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)\frac{\left(1+\frac{n}{\beta}\right)}{mn}-\left(\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(\frac{\alpha}{2}\right)\right)^{2}}{2\beta^{3}}+\frac{\left(\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(\frac{\alpha}{2}\right)}{\beta^{2}}\right)\frac{\left(\beta+n\right)^{2}}{mn}$$

We see that the expression for $E_{\beta}\left(p_{H}\right)$ 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

$$p_L = 1 - \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\hat{\beta}}}$$

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

$$\frac{d}{d\beta}\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} = -\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2}$$

$$\frac{d^2}{d\beta^2}\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} = \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left(\ln\left(1 - \frac{\alpha}{2}\right)\right)^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3}$$

$$f\left(\hat{\beta}\right) \approx \left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} - \frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^2}\left(\hat{\beta} - \beta\right) + \left(\frac{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \left(\ln\left(1 - \frac{\alpha}{2}\right)\right)^2}{\beta^4} + \frac{2\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln\left(1 - \frac{\alpha}{2}\right)}{\beta^3}\right)\frac{\left(\hat{\beta} - \beta\right)^2}{2!}$$

$$p_L \approx 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^2}\left(\hat{\beta}-\beta\right)-\left(\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(1-\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{\left(\hat{\beta}-\beta\right)^2}{2!}$$

$$E_\beta\left(p_L\right)\approx 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^2}E\left(\hat{\beta}-\beta\right)-\left(\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(1-\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{1}{2}E\left(\hat{\beta}-\beta\right)^2$$

$$E_\beta\left(p_L\right)\approx 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^2}Bias-\left(\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln\left(1-\frac{\alpha}{2}\right)\right)^2}{\beta^4}+\frac{2\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\ln\left(1-\frac{\alpha}{2}\right)}{\beta^3}\right)\frac{1}{2}MSE$$

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

$$E_{\beta}\left(p_{L}\right) \approx 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln \left(1-\frac{\alpha}{2}\right)}{\beta^{2}} \frac{(\beta+n) \beta}{m n}-\left(\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln \left(1-\frac{\alpha}{2}\right)\right)^{2}}{\beta^{4}}+\frac{2\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln \left(1-\frac{\alpha}{2}\right)}{\beta^{3}}\right) \frac{1}{2} \frac{(\beta+n)^{2} \beta}{m n}$$

$$E_{\beta}\left(p_{L}\right) \approx 1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}+\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln \left(1-\frac{\alpha}{2}\right) \frac{\left(1+\frac{n}{\beta}\right)}{m n}-\left(\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}}\left(\ln \left(1-\frac{\alpha}{2}\right)\right)^{2}}{2 \beta^{3}}+\frac{\left(1-\frac{\alpha}{2}\right)^{\frac{1}{\beta}} \ln \left(1-\frac{\alpha}{2}\right)}{\beta^{2}}\right) \frac{(\beta+n)^{2}}{m n}$$

APPENDIX D

THRESHOLD PROBABILITY

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

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

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

Taylor series expansion:

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

$$f\left(\hat{\beta}\right) = \left(1-p_0\right)^{\hat{\beta}} = e^{\hat{\beta}\ln\left(1-p_0\right)}$$

$$\frac{df}{d\beta}=e^{\beta\ln(1-p_0)}\ln(1-p_0)=(1-p_0)^{\beta}\ln(1-p_0)$$

$$\frac{d^2 f}{d\beta^2}=e^{\beta\ln(1-p_0)}\left[\ln(1-p_0)\right]^2=(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^2$$

$$\ldots$$

$$\frac{d^k f}{d\beta^k}=e^{\beta\ln(1-p_0)}\left[\ln(1-p_0)\right]^k=(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^k$$

$$f\left(\hat{\beta}\right)=(1-p_0)^{\beta}+(1-p_0)^{\beta}\ln(1-p_0)\left(\hat{\beta}-\beta\right)+(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^2\frac{\left(\hat{\beta}-\beta\right)^2}{2!}+(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^3\frac{\left(\hat{\beta}-\beta\right)^3}{3!}+\ldots$$

$$\widehat{P(p\le p_0)}=1-(1-p_0)^{\beta}-(1-p_0)^{\beta}\ln(1-p_0)\left(\hat{\beta}-\beta\right)-(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^2\frac{\left(\hat{\beta}-\beta\right)^2}{2!}-(1-p_0)^{\beta}\left[\ln(1-p_0)\right]^3\frac{\left(\hat{\beta}-\beta\right)^3}{3!}-\ldots$$

We know

$$\int_0^{p_0}\hat{\beta}(1-p)^{\hat{\beta}-1}dp=-\left(1-p\right)^{\hat{\beta}}\Big|_0^{p_0}=1-(1-p_0)^{\hat{\beta}}\Rightarrow P(p\le p_0)=1-(1-p_0)^{\beta}$$

Hence $\widehat{P\left(p\le p_0\right)}=P\left(p\le p_0\right)-\left(1-p_0\right)^{\beta}\left[\ln\left(1-p_0\right)\left(\hat{\beta}-\beta\right)+\left[\ln\left(1-p_0\right)\right]^2\frac{\left(\hat{\beta}-\beta\right)^2}{2!}+\left[\ln\left(1-p_0\right)\right]^3\frac{\left(\hat{\beta}-\beta\right)^3}{3!}-\ldots\right]$

$$E\left[\widehat{P\left(p\le p_0\right)}\right]=P\left(p\le p_0\right)-\left(1-p_0\right)^{\beta}\left[\ln\left(1-p_0\right)E\left(\hat{\beta}-\beta\right)+\left[\ln\left(1-p_0\right)\right]^2\frac{E\left(\hat{\beta}-\beta\right)^2}{2!}+\left[\ln\left(1-p_0\right)\right]^3\frac{E\left(\hat{\beta}-\beta\right)^3}{3!}-\ldots\right]$$

$$E\left[\widehat{P\left(p\le p_0\right)}\right]=P\left(p\le p_0\right)-\left(1-p_0\right)^{\beta}\left[\ln\left(1-p_0\right)Bias\left(\hat{\beta}\right)+\frac{\left[\ln\left(1-p_0\right)\right]^2}{2!}MSE\left(\hat{\beta}\right)+\ldots\right]$$

APPENDIX E1

DERIVATION OF BAYES/LAPLACE PRIOR

If we define $T=\sum_{j=1}^{m} X_j$ then T is a $Binomial\left(m,\theta\right)$ random variable with parameter, $\theta = 1-(1-p)^n$:

$$T \sim \binom{m}{T}\theta^T\left(1-\theta\right)^{m-T} = \binom{m}{T}\left(1-\left(1-p\right)^n\right)^T\left(\left(1-p\right)^n\right)^{m-T}$$

Following Bayes's argument as given by Stigler, we look for a prior in the natural conjugate prior family,

$$f_p\left(p\middle|n,\alpha,\beta\right)=\frac{\theta\left(p\right)^{\alpha-1}\left[1-\theta\left(p\right)\right]^{\beta-1}}{\mathrm{B}\left(\alpha,\beta\right)}=\frac{\theta\left(p\right)^{\alpha-1}\left[1-\theta\left(p\right)\right]^{\beta-1}}{\int_0^1\theta\left(p\right)^{\alpha-1}\left[1-\theta\left(p\right)\right]^{\beta-1}dp}$$

A change of variable

$$\theta=1-(1-p)^n \rightarrow (1-p)^n=1-\theta \rightarrow 1-p=\left(1-\theta\right)^{\frac{1}{n}} \rightarrow p=1-\left(1-\theta\right)^{\frac{1}{n}} \rightarrow dp=\frac{1}{n}\left(1-\theta\right)^{\frac{1}{n}-1}d\theta$$

results in the following equation:

$$f_p\left(p\middle|n,\alpha,\beta\right)=\frac{\left[1-(1-p)^n\right]^{\alpha-1}\left[(1-p)^n\right]^{\beta-1}}{\frac{1}{n}\int_0^1\theta\left(p\right)^{\alpha-1}\left[1-\theta\left(p\right)\right]^{\beta-1+\frac{1}{n}-1}d\theta}=\frac{n\left[1-(1-p)^n\right]^{\alpha-1}\left[(1-p)^n\right]^{\beta-1}}{\mathrm{B}\left(\alpha,\beta+\frac{1}{n}-1\right)},0<p<1;\alpha,\beta>0$$

Prior to collection of the data, the joint distribution of *T* and *p* is

$$g\left(T,p\right)=f\left(T\middle|\theta\left(p\right)\right)*f\left(\theta\left(p\right)\right)$$
$$=\binom{m}{T}\left(\frac{n}{\mathrm{B}\left(\alpha,\beta+\tfrac{1}{n}-1\right)}\right)\left[1-(1-p)^{n}\right]^{\alpha+T-1}\left[(1-p)^{n}\right]^{m-T+\beta-1}$$

The marginal distribution of T is

$$f_{T}\left(t\middle|\alpha,\beta\right)=\binom{m}{T}\left(\frac{n}{\mathrm{B}\left(\alpha,\beta+\tfrac{1}{n}-1\right)}\right)\int_{0}^{1}\left[1-(1-p)^{n}\right]^{\alpha+T-1}\left[(1-p)^{n}\right]^{m-T+\beta-1}dp$$

A further change of variable $\left(dp=\frac{1}{n}(1-\theta)^{\frac{1}{n}-1}d\theta\right)$ yields:

$$f_{T}\left(t\middle|\alpha,\beta\right)=\binom{m}{T}\left(\frac{1}{\mathrm{B}\left(\alpha,\beta+\tfrac{1}{n}-1\right)}\right)\int_{0}^{1}[\theta]^{\alpha+T-1}[1-\theta]^{m-T+\beta+\frac{1}{n}-1-1}\,d\theta$$
$$=\binom{m}{T}\left(\frac{\mathrm{B}\left(\alpha+T,m-T+\beta+\tfrac{1}{n}-1\right)}{\mathrm{B}\left(\alpha,\beta+\tfrac{1}{n}-1\right)}\right),\alpha,\beta>0$$

Theorem 1: The distribution function $f_{T}\left(t\middle|\alpha,\beta\right)$ has value $1\big/(m+1)$ for all $t\in\{0,1,2,...,m\}$ when $\alpha=1$ and $\beta=1+\frac{(n-1)}{n}$.

Proof: Note that $f_{T}\left(t\middle|\alpha,\beta\right)$ can be written as,

$$f_T\left(t\middle|\alpha,\beta\right)=\frac{\Gamma\left(m+1\right)\Gamma\left(\alpha+t\right)\Gamma\left(m-t+\beta+\frac{1}{n}-1\right)\Gamma\left(\alpha+\beta+\frac{1}{n}-1\right)}{\Gamma\left(t+1\right)\Gamma\left(m-t+1\right)\Gamma\left(m+\alpha+\beta+\frac{1}{n}-1\right)\Gamma\left(\alpha\right)\Gamma\left(\beta+\frac{1}{n}-1\right)}$$

When $\alpha=1$ and $\beta=1+\frac{(n-1)}{n}$ above becomes,

$$f_T\left(t\middle|\alpha=1,\beta=1+\frac{(n-1)}{n}\right)=\frac{\Gamma\left(m+1\right)\Gamma\left(t+1\right)\Gamma\left(m-t+1\right)\Gamma\left(2\right)}{\Gamma\left(t+1\right)\Gamma\left(m-t+1\right)\Gamma\left(m+2\right)\Gamma\left(1\right)}=\frac{\Gamma\left(m+1\right)}{\Gamma\left(m+2\right)}=\frac{1}{m+1},\forall t$$

since $\Gamma\left(1\right)=\Gamma\left(2\right)=1$ and $\Gamma\left(m+2\right)=\left(m+1\right)\Gamma\left(m+1\right)$.

End of proof.

Plugging $\alpha=1$ and $\beta=1+\frac{(n-1)}{n}$ into $f_p\left(p\middle|n,\alpha,\beta\right)=\frac{n\left[1-(1-p)^n\right]^{\alpha-1}\left[(1-p)^n\right]^{\beta-1}}{\mathrm{B}\left(\alpha,\beta+\frac{1}{n}-1\right)}$ we obtain the objective Bayes prior for the equal pool size pool screening model as

$$f_p\left(p\right)=\frac{n*1*\left[\left(1-p\right)^n\right]^{\frac{n-1}{n}}}{\mathrm{B}\left(1,1\right)}=\frac{n\left(1-p\right)^{n-1}}{\mathrm{B}\left(1,1\right)}=n\left(1-p\right)^{n-1}$$

APPENDIX E2

DERIVATION OF JEFFREYS' PRIOR

Jeffreys’ prior for *p*, the infection prevalence, will be calculated next.

$$L\left(p\middle|T\right)=\binom{m}{T}\left(1-\left(1-p\right)^{n}\right)^{T}\left(\left(1-p\right)^{n}\right)^{m-T}$$, where T is the # of positive pools.

$$\ln L=l=\ln\binom{m}{T}+T\ln\left(1-\left(1-p\right)^{n}\right)+\left(m-T\right)n\ln\left(1-p\right)$$

$$\frac{dl}{dp}=\frac{T}{1-\left(1-p\right)^{n}}*n\left(1-p\right)^{n-1}-\left(m-T\right)n\frac{1}{\left(1-p\right)}$$

$$\frac{d^{2}l}{dp^{2}}=-\frac{T\left(n\left(1-p\right)^{n-1}\right)}{\left(1-\left(1-p\right)^{n}\right)^{2}}\left(n\left(1-p\right)^{n-1}\right)-n\left(n-1\right)\left(1-p\right)^{n-2}*\frac{T}{1-\left(1-p\right)^{n}}-\left(m-T\right)n\frac{1}{\left(1-p\right)^{2}}$$

$$=-\frac{Tn^{2}\left(1-p\right)^{2n-2}}{\left(1-\left(1-p\right)^{n}\right)^{2}}-\frac{n\left(n-1\right)\left(1-p\right)^{n-2}T}{1-\left(1-p\right)^{n}}-\frac{\left(m-T\right)n}{\left(1-p\right)^{2}}$$

$E\left(T\right)=m\left(1-\left(1-p\right)^{n}\right)=m-m\left(1-p\right)^{n}$ since T ~ Bin (m,θ)

$$J\left(p\right)=-E\left(\frac{d^2l}{dp^2}\right)=\frac{m\left(1-\left(1-p\right)^n\right)n^2\left(1-p\right)^{2n-2}}{\left(1-\left(1-p\right)^n\right)^2}+\frac{n\left(n-1\right)\left(1-p\right)^{n-2}m\left(1-\left(1-p\right)^n\right)}{1-\left(1-p\right)^n}+\frac{\left(m-m\left(1-\left(1-p\right)^n\right)\right)n}{\left(1-p\right)^2}$$

$$=\frac{mn^2\left(1-p\right)^{2n-2}}{1-\left(1-p\right)^n}+\frac{n\left(n-1\right)\left(1-p\right)^{n-2}m\left(1-\left(1-p\right)^n\right)}{1-\left(1-p\right)^n}+\frac{mn\left(1-p\right)^n}{\left(1-p\right)^2}$$

$$=\frac{mn^2\left(1-p\right)^{n-2}}{1-\left(1-p\right)^n}$$

Now the Jeffreys' prior is proportional to $\sqrt{J\left(p\right)}=\dfrac{\sqrt{m}n\left(1-p\right)^{\frac{n-2}{2}}}{\left[1-\left(1-p\right)^n\right]^{\frac{1}{2}}}$.

APPENDIX F

SEQUENTIAL BAYES APPROACH FOR BAYES/LAPLACE PRIOR

year 1:

$$posterior_1\left(p\,|\,n,m,t\right)\propto\left[1-\left(1-p\right)^n\right]^{t_1}\left[\left(1-p\right)^n\right]^{m_1-t_1}*n\left[\left(1-p\right)^n\right]^{1-\frac{1}{n}}$$

$$posterior_1\left(p\,|\,n,m,t\right)\propto n\left[1-\left(1-p\right)^n\right]^{t_1}\left[\left(1-p\right)^n\right]^{m_1-t_1+1-\frac{1}{n}}$$

$$Normalized: posterior_1\left(p\,|\,n,m,t\right)=\frac{n}{\mathrm{B}\left[t_1+1,m_1-t_1+1\right]}\left[1-\left(1-p\right)^n\right]^{t_1}\left[\left(1-p\right)^n\right]^{m_1-t_1+1-\frac{1}{n}}$$

year 2:

$$posterior_2\left(p\,|\,n,m,t\right)\propto\left[1-\left(1-p\right)^n\right]^{t_2}\left[\left(1-p\right)^n\right]^{m_2-t_2}*posterior_1\left(p\,|\,n,m,t\right)$$

$$posterior_2\left(p\,|\,n,m,t\right)\propto\left[1-\left(1-p\right)^n\right]^{t_2}\left[\left(1-p\right)^n\right]^{m_2-t_2}*\frac{n}{\mathrm{B}\left[t_1+1,m_1-t_1+1\right]}\left[1-\left(1-p\right)^n\right]^{t_1}\left[\left(1-p\right)^n\right]^{m_1-t_1+1-\frac{1}{n}}$$

$$posterior_2\left(p\,|\,n,m,t\right)\propto\frac{n}{\mathrm{B}\left[t_1+1,m_1-t_1+1\right]}\left[1-\left(1-p\right)^n\right]^{t_1+t_2}\left[\left(1-p\right)^n\right]^{m_1+m_2-t_1-t_2+1-\frac{1}{n}}$$

$$Normalized: posterior_2\left(p\,|\,n,m,t\right)\propto\frac{n}{\mathrm{B}\left[t_1+t_2+1,m_1+m_2-t_1-t_2+1\right]}\left[1-\left(1-p\right)^n\right]^{t_1+t_2}\left[\left(1-p\right)^n\right]^{m_1+m_2-t_1-t_2+1-\frac{1}{n}}$$

$\vdots$

APPENDIX G

BAYES/LAPLACE PRIOR POSTERIOR DISTRIBUTION AND EXPECTED VALUE

$$\begin{aligned} f_{BL}(p) &\propto \left[1-(1-p)^n\right]^T \left[(1-p)^n\right]^{m-T} n(1-p)^{n-1} \\ &\propto \left[1-(1-p)^n\right]^T \left[(1-p)^{n(m+1-T)-1}\right] n \end{aligned}$$

We must find the scaling constant for $f_{BL}(p)$. To this end, let

$$K = \int_0^1 n\left[1-(1-p)^n\right]^T \left[(1-p)^{n(m+1-T)-1}\right] dp$$

Making the change of variable $\theta = 1-(1-p)^n$ so that $(1-p) = (1-\theta)^{\frac{1}{n}}$ and $dp = \frac{1}{n}(1-\theta)^{\frac{1}{n}-1} d\theta$. We notice when $p = 0,\ \theta = 0$ and when $p = 1,\ \theta = 1$.

$$\begin{aligned} K &= \int_0^1 n\theta^T \left[(1-\theta)^{m+1-T-\frac{1}{n}}\right] \frac{1}{n}(1-\theta)^{\frac{1}{n}-1} d\theta \\ &= \int_0^1 \theta^T \left[(1-\theta)^{m-T}\right] d\theta \\ &= \mathrm{B}\left[T+1, m+1-T\right] \end{aligned}$$

Hence

$$f_{BL}(p) = \frac{n\left[1-(1-p)^n\right]^T \left[(1-p)^{n(m+1-T)-1}\right]}{\mathrm{B}\left[T+1, m+1-T\right]}$$

Next, we want to find $E_{BL}(p) = 1 - E_{BL}(1-p)$.

$$E_{BL}\left(1-p\right)=\int_{0}^{1}\frac{n\left[1-\left(1-p\right)^{n}\right]^{T}\left[\left(1-p\right)^{n\left(m+1-T\right)-1+1}\right]}{\mathrm{B}\left[T+1,m+1-T\right]}dp$$

Making the same change of variable we have

$$E_{BL}\left(1-p\right)=\int_{0}^{1}\frac{\theta^{T}\left[\left(1-\theta\right)^{m+1-T+\frac{1}{n}-1}\right]d\theta}{\mathrm{B}\left[T+1,m+1-T\right]}$$

$$=\frac{\mathrm{B}\left[T+1,m+1-T+\frac{1}{n}\right]}{\mathrm{B}\left[T+1,m+1-T\right]}=\frac{\dfrac{\Gamma\left(T+1\right)\Gamma\left(m+1-T+\frac{1}{n}\right)}{\Gamma\left(m+2+\frac{1}{n}\right)}}{\dfrac{\Gamma\left(T+1\right)\Gamma\left(m+1-T\right)}{\Gamma\left(m+2\right)}}$$

$$=\frac{\Gamma\left(m+2\right)\Gamma\left(m+1-T+\frac{1}{n}\right)}{\Gamma\left(m+2+\frac{1}{n}\right)\Gamma\left(m+1-T\right)}$$

Hence

$$E_{BL}\left(p\right)=1-\frac{\Gamma\left(m+2\right)\Gamma\left(m+1-T+\frac{1}{n}\right)}{\Gamma\left(m+2+\frac{1}{n}\right)\Gamma\left(m+1-T\right)}$$