Biostatistics Department Technical Report

BST2006-001

Estimation of Prevalence by Pool Screening With Equal Sized Pools and a Negative Binomial Sampling Model

Charles R. Katholi, Ph.D.
Emeritus Professor

Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
ckatholi@uab.edu

The screening of pools of insects to make estimates of the prevalence of infection of some disease in a vector species is becoming more and more common due to the increasing sensitivity and specificity of PCR methods. In this approach, the investigator collects "pools" or "groups" of the vector species. These pools are then tested using an assay method such as PCR and the pool is evaluated as positive or negative depending on whether the disease of interest is found in the pool. An early use of this approach was testing groups of men drafted to serve in the military for syphilis. In this case, the method was use to provide screening at reduced cost since if a pool was found to be negative it was certain that none of the men in the pool had the disease.This approach is particularly appropriate in the case where the prevalence is low and screening pools allows one to check a large number of insects with a smaller amount of labor than would be required to test each individual insect by dissection or some other protocol. The appropriate statistical model to use in conjunction with the pool screening approach depends very much on the way that the sampling and testing are done. Many investigators have considered the case of binomial sampling [1-5]. For testing individual insects (i.e., pools of size 1) George and Elston [6] recommended geometric sampling when the probability of an event was small. They gave confidence intervals for the prevalence based on this model. They did not, however, investigate the statistical properties of the estimator. Lui [7] extended their work on the confidence interval by considering negative binomial sampling and showed that as the number of successes required increased, the width of the confidence interval decreased. Lui also did not discuss point estimators, their statistical properties nor did he investigate the statistical properties of his confidence intervals. In this report we investigate these sampling models when an investigator collects and tests pools until some pre-determined number of positive pools is observed. We shall consider point and interval estimators obtained by both classical and Bayesian methods and investigate their statistical properties.

In what follows we shall denote the size of the pools collected by N, the prevalence of infection by p, the number of positive results to be observed before quitting by r, and the number of times the experiment is carried out by m. If the prevalence of infection is p, then the probability that a pool of size N tests negative is given by $(1-p)^N$ and the probability that a pool is positive is positive is $[1-(1-p)^N]$. If we let Y be the number of negative pools observed prior to getting r positive pools, then Y has a negative binomial distribution and we take this as the probability model upon which to base our calculations. If $Y_1, Y_2, \cdots, Y_m$ are the results of m such experiments we shall often denote them as a vector $\underline{Y}^T = (Y_1, \cdots, Y_m)$. Following normal practice we shall generally denote random variables by capital letters and their realizations by small letters.

Classical Methods: We begin by finding the maximum likelihood estimator of p given that Y has the negative binomial distribution,

$$f(y \mid p) = P(Y = y \mid p) = \binom{y+r-1}{y}\left[1-(1-p)^N\right]^r\left[(1-p)^N\right]^y, y = 0,1,2,\cdots; 0 < p < 1$$

Given the results of m replications of the sampling procedure, $Y_1, Y_2, \cdots, Y_m$ the likelihood function is given by

$$l(p \mid \underline{Y}) = \prod_{j=1}^{m} \binom{Y_j + r - 1}{Y_j} \left[1-(1-p)^N\right]^r \left[(1-p)^N\right]^{Y_j}$$

If we define $L(p \mid \underline{Y}) = \ln l(p \mid \underline{Y})$ we have

$$L(p \mid \underline{Y}) = mr \ln\left[1-(1-p)^N\right] + TN \ln\left[1-p\right] + \sum_{j=1}^{m} \binom{Y_j + r - 1}{Y_j} \text{ where } T = \sum_{j=1}^{m} Y_j$$

Recall that since the $Y_1, Y_2, \cdots, Y_m$ are i.i.d negative binomial with parameters r and p, then T is negative binomial with parameters mr and p. Following the usual procedure we take the derivative of the log likelihood with respect to p, set it equal to zero and solve the resulting equation. We shall show that the solution is a maximum and that it is unique.

$$\frac{\partial L}{\partial p} = \frac{mrN(1-p)^{N-1}}{\left[1-(1-p)^N\right]} - \frac{NT}{(1-p)} = \left(\frac{N}{(1-p)}\right)\left(\frac{mr(1-p)^N}{\left[1-(1-p)^N\right]} - T\right)$$

Since $\left(\frac{N}{(1-p)}\right)$ is positive for 0<p<1 we need only consider the right hand factor when this is set to zero. Setting the right hand term to zero and solving for p yields $\hat{p} = 1 - \left(\frac{T}{T+mr}\right)^{1/N}$. Examination of $u(p) = \left(\frac{mr(1-p)^N}{\left[1-(1-p)^N\right]} - T\right)$ reveals that as p approaches zero the u(p) tends to $+\infty$. Similarly, as p approaches 1, the u(p) approaches (-T). Because the function is continuous on (0, 1) and changes sign in the interval it follows from the intermediate value theorem that there is at least one root of the equation in the interval (0, 1). Next note that it is easily demonstrated that the left hand term in the expression for u(p) is strictly monotone decreasing so there is only one solution on (0, 1). Finally,

$$\frac{\partial^2 L}{\partial p^2} = -\frac{N}{(1-p)^2}\left\{\frac{mr(N-1)(1-p)^N}{\left[1-(1-p)^N\right]} + \frac{mrN(1-p)^{2N}}{\left[1-(1-p)^N\right]^2} + T\right\} < 0$$

for all 0 < p < 1 and so $\hat{p}$ is the unique maximum likelihood estimator (MLE) of p.

Next we consider the asymptotic properties of the MLE. It is commonly the custom to assume the consistency and asymptotic normality of maximum likelihood estimators as given when certain (generally unspecified) regularity conditions are met. Most basic texts in statistics do not say what these conditions are. They can be found, for example in Singer and Sen [10] , Serfling [11] and Furgeson [12]. The exact assumptions

differ somewhat among the authors and some can be difficult to establish in practice. Wald [13] established the strong consistency of the MLE under very weak conditions compared to those given in general. On the other hand, when the MLE is available in closed form as it is here, it is often possible to establish these properties directly and that approach is taken here.

In what follows we will denote the MLE by $\hat{p}_m$ to emphasize that it is the estimator based m replications of the sampling process. We begin by proving the following result.

Theorem 0: $\hat{p}_m$ is a strongly consistent estimator; that is, $\hat{p}_m \xrightarrow{a.s.} p$ as $m \to \infty$.

proof: We have previously shown that $\hat{p}_m = 1 - \left(\frac{T}{T+mr}\right)^{1/N} = 1 - \left(\frac{\frac{1}{m}T}{\frac{1}{m}T + r}\right)^{1/N}, T = \sum_{i=1}^{m} Y_i$.

Let $X_m = \frac{1}{m}T = \frac{1}{m}\sum_{i=1}^{m} Y_i$ and note that $E(Y_i) = \frac{r\left[(1-p)^N\right]}{\left[1-(1-p)^N\right]} = E(X_m)$, then by

Khintchine's strong law of large numbers $X_m \xrightarrow{a.s} \frac{r(1-p)^N}{1-(1-p)^N} = c$ and

$W_m = X_m - c \xrightarrow{a.s} 0$. With this notation we have that $\hat{p}_m = 1 - \left(\frac{X_m}{X_m + r}\right)^{1/N}$ and we expand the right hand side of this expression in a Taylor series about c. Thus we can write that

$$\hat{p}_m = p - \frac{r}{N}\left[h_m c + (1-h_m)X_m\right]^{\frac{1}{N}-1}\left[h_m c + (1-h_m)X_m + r\right]^{-\frac{1}{N}-1}\left(X_m - c\right) =$$

$$p - \frac{r}{N}\left[X_m - h_m W_m\right]^{\frac{1}{N}-1}\left[X_m + r - h_m W_m\right]^{-\frac{1}{N}-1}\left(W_m\right)$$

for some $h_m, 0 < h_m < 1$. Define the vector random variable $\underline{V}_m = \begin{pmatrix} X_m \\ W_m \end{pmatrix}$ and note that

$\underline{V}_m \xrightarrow{a.s.} \underline{v} = \begin{pmatrix} c \\ 0 \end{pmatrix}$ as $m \to \infty$. Also define the function

$$g(\underline{V}_m) = -\frac{rW_m}{N}\left[X_m - h_m W_m\right]^{\frac{1}{N}-1}\left[X_m + r - h_m W_m\right]^{-\frac{1}{N}-1}$$

$g(\underline{V}_m)$ is clearly a continuous function of its arguments and hence is a Borel function. Hence by general convergence results for Borel functions we have

$$g(\underline{V}_m) \xrightarrow{a.s.} g(\underline{v}) = 0.$$

This completes the proof.

We are also interested to asymptotic behavior of the moments of $\hat{p}_m$ .To this end, consider the function $f(t)=1-\left(\frac{t}{t+mr}\right)^{\frac{1}{N}}=1-t^{\frac{1}{N}}\left(t+mr\right)^{-\frac{1}{N}}$ and note that $E(T)=\frac{mr(1-p)^N}{\left[1-(1-p)^N\right]}$. In order to expand $f(t)$ about $t_0=E(T)$ we need the derivatives of $f(t)$. To this end we note that $f(t)$ can be written in terms of the product of two functions; $u(t)=t^{\frac{1}{N}}$ and $v(t)=\left(t+mr\right)^{-\frac{1}{N}}$. That is, $f(t)=1-u(t)v(t)$. We shall also utilize Leibnitz's rule for taking the s-th derivative of a product of two functions. That is,

$$\frac{d^s(uv)}{dt^s}=\sum_{j=0}^{s}\binom{s}{j}\left(\frac{d^{s-j}u}{dt^{s-j}}\right)\left(\frac{d^jv}{dt^j}\right) \qquad (1)$$

It is straight forward to show that

$$\left.\frac{d^ku(t)}{dt^k}\right|_{t=t_0=E(T)}=\left[\prod_{j=0}^{k-1}\left(\frac{1}{n}-j\right)\right]\left(mr\right)^{\frac{1}{N}-k}\left[\left(1-p\right)^N\right]^{\frac{1}{N}-k}\left[1-(1-p)^N\right]^{k-\frac{1}{N}}$$

and that

$$\left.\frac{d^kv(t)}{dt^k}\right|_{t=t_0}=\left(-1\right)^k\left[\prod_{j=0}^{k-1}\left(\frac{1}{N}+j\right)\right]\left(mr\right)^{-\left(\frac{1}{N}+k\right)}\left[1-(1-p)^N\right]^{\frac{1}{N}+k}$$

Similarly, we have that

$$u(t)\big|_{t=E(T)}=(mr)^{\frac{1}{N}}\left[(1-p)^N\right]^{\frac{1}{N}}\left[1-(1-p)^N\right]^{-\frac{1}{N}}$$

and

$$v(t)\big|_{t=E(T)}=(mr)^{-\frac{1}{N}}\left[1-(1-p)^N\right]^{\frac{1}{N}}$$

If we make the rule that $\prod_{j=0}^{-1}\left(\frac{1}{N}\pm j\right)=1$ then for $0\le j\le s$ we can write that

$$\left.\frac{d^{s-j}u}{dt^{s-j}}\frac{d^jv}{dt^j}\right|_{t=E(T)}=(-1)^j\left(\frac{1}{mr}\right)^s\left[(1-p)^N\right]^{\frac{1}{N}-s}\left[1-(1-p)^N\right]^s\left\{\left[\prod_{i=0}^{s-j-1}\left(\frac{1}{N}-i\right)\right]\left[\prod_{i=0}^{j-1}\left(\frac{1}{N}+i\right)\right]\left[(1-p)^N\right]^j\right\}$$

This is used in conjunction with Leibnitz's rule, equation (1), to calculate all needed derivatives. As it happens, examination of the first few of these reveals a pattern and so the general k-th derivative can be expressed in closed form. To this end, we consider several cases:

Case i: (s = 1)

$$\frac{d(uv)}{dt}=\binom{1}{0}v\frac{du}{dt}+\binom{1}{0}u\frac{dv}{dt}=\frac{1}{N}\left(\frac{1}{mr}\right)\left\{\left[(1-p)^N\right]^{\frac{1}{N}-1}\left[1-(1-p)^N\right]-\left[(1-p)^N\right]^{\frac{1}{N}}\left[1-(1-p)^N\right]\right\}$$

or

$$\frac{d(uv)}{dt}=\frac{1}{N}\left(\frac{1}{mr}\right)\left[(1-p)^N\right]^{\frac{1}{N}-1}\left[1-(1-p)^N\right]^2 \qquad (2)$$

Case ii: (s = 2)

$$\frac{d^2(uv)}{dt^2}=\binom{2}{0}v\frac{d^2u}{dt^2}+\binom{2}{1}\frac{du}{dt}\frac{dv}{dt}+\binom{2}{2}u\frac{d^2v}{dt^2}$$

which after plugging in the parts and gathering terms is

$$\frac{d^2(uv)}{dt^2}=\frac{1}{N}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{1}{N}-2}\left[1-(1-p)^N\right]^2\left\{\left(\frac{1}{N}-1\right)-\frac{2}{N}(1-p)^N+\left(\frac{1}{N}+1\right)\left[(1-p)^N\right]^2\right\}$$

But the term in braces on the right hand side of the equation is divisible by $\left[1-(1-p)^N\right]$ leading to the final result

$$\frac{d^2(uv)}{dt^2}=\frac{-1}{N}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{1}{N}-2}\left[1-(1-p)^N\right]^3\left\{\left(1-\frac{1}{N}\right)+\left(\frac{1}{N}+1\right)(1-p)^N\right\} \qquad (3)$$

Examination of several more derivatives (s = 3,4 and 5) lead to the general result,

$$\frac{d^k(uv)}{dt^k}=\frac{(-1)^{k+1}}{N}\left(\frac{1}{mr}\right)^k\left[(1-p)^N\right]^{\frac{1}{N}-k}\left[1-(1-p)^N\right]^{k+1}\left\{\sum_{l=0}^{k-1}\binom{k-1}{l}\left[\prod_{j=1}^{k-1-l}\left(j-\frac{1}{N}\right)\right]\left[\prod_{i=1}^{l}\left(i+\frac{1}{N}\right)\right]\left[(1-p)^N\right]^l\right\}$$

where we have made the rule that $\prod_{j=1}^{0}\left(j\pm\frac{1}{N}\right)=1$. Having developed this formula, it is now possible to expand the equation for the MLE in a Taylor series about E(T). Thus we have

$$\hat{p} = 1 - \left(\frac{T}{T + mr} \right)^{1/N} = 1 - u(t)\,v(t)\big|_{t=E(T)} - \left.\frac{d(uv)}{dt}\right|_{t=E(T)} \left(T - E(T)\right) - \left.\frac{d^2(uv)}{dt^2}\right|_{t=E(T)} \frac{\left(T - E(T)\right)^2}{2!} - \cdots$$

That is,

$$\begin{aligned}
\hat{p} = p &- \frac{1}{N}\left(\frac{1}{mr}\right)\left[(1-p)^N\right]^{\frac{1}{N}-1}\left[1-(1-p)^N\right]^2 (T - E(T)) + \\
&\frac{1}{N}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{1}{N}-2}\left[1-(1-p)^N\right]^3 g_2(p)\frac{(T-E(T))^2}{2!} - \\
&\frac{1}{N}\left(\frac{1}{mr}\right)^3\left[(1-p)^N\right]^{\frac{1}{N}-3}\left[1-(1-p)^N\right]^4 g_3(p)\frac{(T-E(T))^3}{3!} + \\
&\frac{1}{N}\left(\frac{1}{mr}\right)^4\left[(1-p)^N\right]^{\frac{1}{N}-4}\left[1-(1-p)^N\right]^5 g_4(p)\frac{(T-E(T))^4}{4!} - \cdots
\end{aligned} \tag{4}$$

where

$$g_k(p) = \sum_{l=0}^{k-1}\binom{k-1}{l}\left[\prod_{j=1}^{k-1-l}\left(j - \frac{1}{N}\right)\right]\left[\prod_{i=1}^{l}\left(i + \frac{1}{N}\right)\right]\left[(1-p)^N\right]^l \tag{5}$$

Noting that each term in this expansion has a factor of the form

$$\left(\frac{1}{m}\right)^k \left(T - E(T)\right)^k = \left(\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - E(Y_i)\right)\right)^k$$

we see that taking the expected value of $\hat{p}$ involves finding the expected value of averages of i.i.d. random variables. Clearly, by the way we constructed the series,

$$E\left(\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - E(Y_i)\right)\right) = \frac{1}{m}E\left(T - E(T)\right) = 0$$

Similarly,

$$E\left(\frac{1}{m}\sum_{i=1}^{m}(Y_i - E(Y_i))\right)^2 = \frac{1}{m}Var(Y) = \frac{r(1-p)^N}{m\left[1-(1-p)^N\right]^2}$$

while

$$E\left(\frac{1}{m}\sum_{i=1}^{m}(Y_i - E(Y_i))\right)^3 = \frac{1}{m^2}E\left(\left(Y - E(Y)\right)^3\right) = \frac{r(1-p)^N\left[1+(1-p)^N\right]}{m^2\left[1-(1-p)^N\right]^3}$$

and finally

$$E\left(\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - E(Y_i)\right)\right)^4 = \frac{3}{m^2}\left(Var(Y_i)\right)^2 + \bigcirc\left(\frac{1}{m^3}\right) = \frac{3r^2\left[(1-p)^N\right]^2}{m^2\left[1-(1-p)^N\right]^4} + \bigcirc\left(\frac{1}{m^3}\right)$$

If we now take the expectation term-wise in the series for $\hat{p}$ we obtain,

$$E(\hat{p}) = p + \frac{1}{N}\left(\frac{1-p}{mr}\right)\left(\frac{1-(1-p)^N}{2!(1-p)^N}\right)g_2(p) \quad -$$
$$\frac{1}{N}\left(\frac{1}{mr}\right)^2 \frac{(1-p)\left[1-(1-p)^N\right]\left[1+(1-p)^N\right]}{3!\left[(1-p)^N\right]^2} g_3(p) \quad +$$
$$\frac{3}{N}\left(\frac{1}{mr}\right)^2 \frac{(1-p)\left[1-(1-p)^N\right]}{4!\left[(1-p)^N\right]^2} g_4(p) \; + \quad \bigcirc\left(\frac{1}{m^3}\right)$$

or gathering terms in powers of $1/mr$ yields

Theorem 1: The first few terms in the expansion for $E(\hat{p})$ in powers of 1/m are

$$E(\hat{p}) = p + \frac{1}{N}\left(\frac{1-p}{mr}\right)\left(\frac{1-(1-p)^N}{(1-p)^N}\right)\frac{g_2(p)}{2!} \quad +$$
$$\frac{1}{N}\left(\frac{1}{mr}\right)^2 \frac{(1-p)\left[1-(1-p)^N\right]}{\left[(1-p)^N\right]^2}\left\{\frac{3g_4(p)}{4!} - \frac{\left[1+(1-p)^N\right]g_3(p)}{3!}\right\} \quad +$$
$$+ \quad \bigcirc\left(\frac{1}{m^3}\right)$$

where the $g_k(p)$ are defined as in equation (5).

We note that both $\frac{g_2(p)}{2!}$ and $\left\{\frac{3g_4(p)}{4!} - \frac{\left[1+(1-p)^N\right]g_3(p)}{3!}\right\}$ are positive and so the Maximum likelihood Estimator is upwardly biased. From the expansion in equation (4) we can also find the first few terms in the expansion for $(\hat{p}-p)^2$, take the expectation with respect to T as above an obtain the second order approximation to the Mean Square Error for $\hat{p}$. To this end, moving p to the left hand side of equation (4) and squaring both sides yields,

$$(\hat{p}-p)^2 = \frac{1}{N^2}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{2}{N}-2}\left[1-(1-p)^N\right]^4\left(T-E(T)\right)^2 \quad -$$
$$\frac{2}{N^2}\left(\frac{1}{mr}\right)^3\left[(1-p)^N\right]^{\frac{2}{N}-3}\left[1-(1-p)^N\right]^5 g_2(p)\frac{\left(T-E(T)\right)^3}{2!} \quad +$$
$$\frac{2}{N^2}\left(\frac{1}{mr}\right)^4\left[(1-p)^N\right]^{\frac{2}{N}-4}\left[1-(1-p)^N\right]^6 g_3(p)\frac{\left(T-E(T)\right)^4}{4!} \quad -$$
$$\frac{2}{N^2}\left(\frac{1}{mr}\right)^5\left[(1-p)^N\right]^{\frac{2}{N}-5}\left[1-(1-p)^N\right]^7 g_4(p)\frac{\left(T-E(T)\right)^5}{5!} \quad +\cdots+$$
$$\frac{1}{N^2}\left(\frac{1}{mr}\right)^4\left[(1-p)^N\right]^{\frac{2}{N}-4}\left[1-(1-p)^N\right]^6 g_2^2(p)\frac{\left(T-E(T)\right)^4}{(2!)^2} \quad -\cdots$$

Next taking the expectation on both sides of the equation and gathering terms in powers of $1/mr$ yields,

Theorem 2: The first few terms in the expansion for the Mean Square Error in powers of 1/m are,

$$E(\hat{p}-p)^2 = \frac{1}{N^2}\left(\frac{1}{mr}\right)\left[(1-p)^N\right]^{\frac{2}{N}-1}\left[1-(1-p)^N\right]^2 \quad +$$
$$\frac{1}{N^2}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{2}{N}-2}\left[1-(1-p)^N\right]^2\left\{\frac{2g_3(p)}{4!}+\frac{g_2^2(p)}{(2!)^2}-\left[1+(1-p)^N\right]g_2(p)\right\} \quad +$$
$$O\left(\frac{1}{(rm)^3}\right)$$

By combining the expansions from Theorems 1 and 2 we can find the first few terms of the expansion for the variance of $\hat{p}$. Carrying out this process leads to the following theorem.

Theorem 3: The first few terms in the expansion for the variance of $\hat{p}$ are

$$Var(\hat{p}) = \frac{1}{N^2}\left(\frac{1}{mr}\right)\left[(1-p)^N\right]^{\frac{2}{N}-1}\left[1-(1-p)^N\right]^2 \quad +$$
$$\frac{1}{N^2}\left(\frac{1}{mr}\right)^2\left[(1-p)^N\right]^{\frac{2}{N}-2}\left[1-(1-p)^N\right]^2\left\{\frac{2g_3(p)}{4!}-\left[1+(1-p)^N\right]g_2(p)\right\} \quad +$$
$$O\left(\frac{1}{(rm)^3}\right)$$

Note that as we would expect, the first term in this expansion is just the reciprocal of the Fischer information.

We now want to consider the asymptotic distribution of $\hat{p}$. We shall prove the following theorem:

Theorem 4: $$\left\{\frac{(\hat{p}-p)}{\frac{(1-p)(1-(1-p)^N)}{N\sqrt{rm}(1-p)^{\frac{N}{2}}}}\right\} \xrightarrow{D} N(0,1)$$

proof: We shall again utilize a truncated Taylor expansion to demonstrate this result. In a manner analogous to equation (4) we can obtain

$$\hat{p}-p = -\frac{1}{N}\left(\frac{1}{r}\right)\left[(1-p)^N\right]^{\frac{1}{N}-1}\left[1-(1-p)^N\right]^2 W_m \quad +$$
$$\frac{1}{N}\left[X_m - h_m W_m\right]^{\frac{1}{N}-2}\left[X_m + r - h_m W_m\right]^{-\frac{1}{N}-2}\left[2r(X_m - h_m W_m)+\left(1-\frac{1}{N}\right)r^2\right]\frac{W_m^2}{2!}$$

where as before $X_m = \frac{1}{m}T = \frac{1}{m}\sum_{i=1}^{m} Y_i$ and $W_m = X_m - c = \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \frac{r(1-p)^N}{(1-(1-p)^N)}\right)$

Next we note that the first term can be written as

$$\left\{\frac{(1-p)\left[1-(1-p)^N\right]}{N\sqrt{r}\sqrt{m}(1-p)^{\frac{N}{2}}}\right\}\left\{\frac{\sqrt{m}\left[1-(1-p)^N\right]}{\sqrt{r}(1-p)^{\frac{N}{2}}}\right\}$$

We note that the term in the left hand bracket is just the square root of the first term in the expansion for the variance given in Theorem 3 while the term in the right hand bracket is just the reciprocal of the square root of the variance of W_m. Hence we can write that,

$$\left\{\frac{(\hat{p}-p)}{\frac{(1-p)(1-(1-p)^N)}{N\sqrt{rm}(1-p)^{\frac{N}{2}}}}\right\}=-\left\{\frac{W_m}{\frac{\sqrt{r}(1-p)^{\frac{N}{2}}}{\sqrt{m}\left[1-(1-p)^N\right]}}\right\}+$$

$$\frac{\sqrt{r}(1-p)^{\frac{N}{2}}}{2!(1-p)(1-(1-p)^N)}\left[X_m-h_mW_m\right]^{\frac{1}{N}-2}\left[X_m+r-h_mW_m\right]^{-\frac{1}{N}-2}\times \qquad (6)$$

$$\left[2r(X_m-h_mW_m)+\left(1-\frac{1}{N}\right)r^2\right]\sqrt{m}W_m^2$$

We have noted previously that $X_m \xrightarrow{a.s} \frac{r(1-p)^N}{1-(1-p)^N}$ and that $W_m \xrightarrow{a.s.} 0$. Since almost sure convergence implies convergence in probability each of these converges to its respective limit in probability as well. Finally, consider the quantity $\sqrt{m}W_m^2$. We shall show that this converges in probability to 0 as m tends to infinity. To this end,

$P(\sqrt{m}W_m^2>\varepsilon)<\frac{E(\sqrt{m}W_m^2)}{\varepsilon}$ by Chebyshev's inequality and

$$E(\sqrt{m}W_m^2)=\sqrt{m}E(W_m^2)=\sqrt{m}Var(W_m)=\frac{r((1-p)^N}{\varepsilon\sqrt{m}\left[1-(1-p)^N\right]^2}$$

so that $P(\sqrt{m}W_m^2>\varepsilon)=0$ as $m\to\infty$; that is, $\sqrt{m}W_m^2$ converges to zero in probability. An argument like to one used in Theorem 0 now shows that the second term on the right side of the equal sign in equation (6) converges to zero in probability. By the central limit theorem the first term (in brackets) on the right hand side converges in distribution to a N(0,1) and by symmetry of the normal so does the negative of this random variable. Finally application of Slutsky's theorem completes the proof.

To summarize, we have shown that a unique maximum likelihood estimator exists and that it has all the usual properties we associate with an MLE. We have also obtained the first few terms of asymptotic expansions for the Bias and Variance of the MLE. One

practical matter needs to be noted. That is, the MLE is well defined if the experiment is carried out only once. That is, in case the investigator collects and tests pools until r positive pools are found, the MLE is just $\hat{p} = 1 - \left(\frac{Y}{Y+r} \right)^{\frac{1}{N}}$ where Y is the number of negative pools tested prior to obtaining r positive pools. On the other hand, it is not reasonable in this case to invoke any of the asymptotic results noted above. It is particularly worthy of note that in the event of a very rare event, it might be of practical interest to set r = 1. For the rare event case, it is also important to exercise care in the computation of $\hat{p}$. This is because Y will be large compared to r and so the ratio $Y/(Y+r)$ will approach 1. The ratio raised to the 1/N power is even closer to 1 and so there will be excessive cancellation leading to loss of precision in the computation of $\hat{p}$. This problem can be solved by noting that $\left(\frac{Y}{Y+r} \right)^{\frac{1}{N}} = \left(1 - \frac{r}{Y+r} \right)^{\frac{1}{N}} = e^{\frac{1}{N}\ln\left(1-\frac{r}{Y+r}\right)}$ and using the McClaurin expansion

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^k}{k} - \cdots, -1 < x < 1$$

with $x = \frac{r}{Y+r}$ to calculate the small negative quantity $\xi = \frac{1}{N}\ln\left(1 - \frac{r}{Y+r}\right)$. Then the McClaurin expansion $1 - e^{\xi} = -(\xi + \frac{\xi^2}{2!} + \frac{\xi^3}{3!} + \cdots)$ is used to calculate $\hat{p}$ without loss of precision due to cancellation.

To get a sense of the asymptotic behavior of the MLE, we note that the key factor in each term of the expansions given in Theorems 2 and 3 is the quantity rm. Table I shows the effect of rm on the bias for several values of p when the pool size N = 50. From the table it is clear that they key number from an asymptotic point of view is the number of positive cells observed (rm).

It can also be noted here that the results of Theorems 2 and 3 still hold if the experiment is changed so that at each site the investigator collects specimens until $r_i, i = 1, 2, \cdots, m$ are observed. That is, if the sites are indexed as $i = 1, 2, \cdots, m$, the investigator observes r_i positive pools at the i-th site. In this case the log likelihood function is

$$L(p \mid \underline{Y}) = \left(\sum_{j=1}^{m} r_j \right) \ln\left[1 - (1-p)^N\right] + TN \ln\left[1 - p\right] + \sum_{j=1}^{m} \binom{Y_j + r - 1}{Y_j}, T = \sum_{j=1}^{m} Y_j$$

and the maximum likelihood estimator is $\hat{p} = 1 - \left(\frac{T}{T + \sum_{j=1}^{m} r_j} \right)^{1/N}$. If $\sum_{j=1}^{m} r_j$ is written as $mr_m^*, r_m^* = \frac{1}{m} \sum_{j=1}^{m} r_j$ then the expansions of theorems 2 and 3 are the same but with r replaced by r_m^* and so the bias and means square error depend asymptotically on powers of $\sum_{j=1}^{m} r_j$ which grows without bound as m tends to infinity.

p	**rm**	$E(\hat{p})$	**Bias**
1/1000 = 0.001	1	0.05125	0.05025
	2	0.00425	0.00325
	3	0.00161	0.00061
	4	0.00134	0.00034
	5	0.00125	0.00025
	10	0.00111	0.00011
	15	0.00107	0.00007
	20	0.00105	0.00005
	25	0.00104	0.00004
1/10,000=0.0001	1	0.00546	0.00536
	2	0.00024	0.00013
	3	0.00015	0.00005
	4	0.00013	0.00003
	5	0.00012	0.00002
	10	0.00011	0.00001
	15	0.000107	0.000007
	20	0.000105	0.000005
	25	0.000104	0.000004

Table I: Behavior of the MLE with respect to bias as the quantity rm increases given p and pools of size N = 50.

We next consider confidence intervals produced from this modeling approach. As mentioned previously, George and Elston [6] considered a confidence interval for the case rm = 1 while Lui [7] considered the case r > 1 while m = 1. The extension of their results to the screening of pools is very easy. As we have previously observed, if we have collected specimens at m sites until r positive pools are observed at a site, and if we denote the number of negative pools collected at site i by Y_i then the distribution of

$T = \sum_{i=1}^{m} Y_i$ is a negative binomial with parameters $q = \left[1-(1-p)^N\right]$ and rm. Given that the investigator has observed t total failures, the classical confidence intervals are then given by solving the equations

$$\left[1-(1-p)^N\right]^{rm} \sum_{y=0}^{t} \binom{y+rm-1}{y} \left[(1-p)^N\right]^y = \alpha/2 \qquad (8)$$

$$\left[1-(1-p)^N\right]^{rm} \sum_{y=t}^{\infty} \binom{y+rm-1}{y} \left[(1-p)^N\right]^y = \alpha/2 \qquad (9)$$

A number of investigators [7-9] have shown that these sums can be replaced with equivalent binomial sums and hence that the sums can be found from the incomplete beta function. In particular, following Patil [9] let

$$w(y, \left[1-(1-p)^N\right], rm) = \binom{y+rm-1}{y} \left[1-(1-p)^N\right]^{rm} \left[(1-p)^N\right]^y$$

and $W(t, \left[1-(1-p)^N\right], rm) = \sum_{y=0}^{t} w(y, \left[1-(1-p)^N\right], rm)$ then

$$W(t, \left[1-(1-p)^N\right], rm) = I_{\left[1-(1-p)^N\right]}(rm, t+1)$$

where

$$I_{\left[1-(1-p)^N\right]}(rm, t+1) = \frac{1}{B(rm, t+1)} \int_0^{\left[1-(1-p)^N\right]} u^{rm-1} (1-u)^{(t+1)-1} du$$

is the incomplete beta function with parameters rm and t + 1. Lui then uses the known relationship between the beta distribution and the F-distribution to obtain closed form equations for the end points of the confidence interval based on critical values from the F-distribution. From a computational point of view, it is no easier to calculate the quantile values from the F-distribution than it is to calculate those from the beta distribution and so in our study we have used the appropriate quantiles from the incomplete beta function. The confidence intervals should be exact (ie, for any value of p, $P(p_l < p < p_u) \geq 1-\alpha$ and infimum with respect to p of such probabilities is exactly $1-\alpha$). Table II shows the coverage probabilities of the intervals for two values of the unknown parameter p and for a number of values of rm.

p	rm	Coverage probability
1/1000 = 0.001	2	0.95355
	3	0.95255
	4	0.95135
	5	0.95159
	10	0.95125
	15	0.95122
	20	0.95089
	25	0.95045
1/10,000 = 0.0001	2	0.95091
	3	0.95039
	4	0.95018
	5	0.95025
	10	0.95013
	15	0.95002
	20	0.95009
	25	0.95006

Table II: Coverage probabilities for the confidence intervals when the true values of the unknown parameter p are as given in the table. For this table, the pool size is N = 50.

In Table III we give results about coverage probabilities for the case N = 50, rm = 10 and p taking on values in the interval $\left(\frac{1}{10,000}, \frac{1}{100}\right)$.

$p \times 10^4$	Coverage Probability
0.450	0.95006
0.750	0.95006
0.961	0.95007
1.230	0.95009
1.585	0.95022
2.000	0.95003
2.610	0.95045
3.500	0.95035
4.310	0.95024
5.500	0.95086
7.100	0.95078
9.000	0.95067
10.000	0.95125

Table III. Coverage probabilities for 95% confidence intervals for $\frac{1}{10,000} \leq p \leq \frac{1}{1000}$ when the pool size is N = 50 and rm = 10.

In Table III we see the typical variation of the coverage probability with changes in p one expects with discrete distributions. These intervals are exact and slightly

conservative, but not nearly as conservative as the Binomial confidence intervals, for example.

Bayesian Methods: We now take a look at the use of the Bayesian method to find point estimates and confidence intervals for this sampling situation. The new decision which must be made here is the choice of the prior distribution. We shall assume initially that we have no prior experience upon which to base a decision concerning a prior distribution. For that reason, we shall utilize the Jeffreys prior for our analysis. It is not difficult to show that for our negative binomial model, this prior, g(p), is such that

$$g(p) \propto \frac{(1-p)^{n/2}}{(1-p)\left[1-(1-p)^{N}\right]}$$

Combining this with the likelihood

$$K\left[1-(1-p)^{N}\right]\left[(1-p)^{N}\right]^{T}, T=\sum_{j=1}^{m} Y_j, K=\prod_{i=1}^{m}\binom{Y_i+r-1}{Y_i}$$

leads to the posterior distribution,

$$\pi(p)=\frac{n}{B(rm,t+\frac{1}{2})}\left[1-(1-p)^{N}\right]^{rm-1}\left[(1-p)^{N}\right]^{t+\frac{1}{2}-\frac{1}{n}}, 0<p<1$$

where t is the value of the random variable T actually observed. Although the classical concept of point estimate is not part of the Bayes approach, it is none the less possible to consider the mode of the distribution as analogous to a point estimate. In this case, as long as $rm \geq 2$, $\pi(p)$ has a single maximum on 0 < p < 1 at the point,

$$p_b = 1-\left(1-\frac{rm-1}{t+rm-\frac{1}{2}-\frac{1}{n}}\right)^{\frac{1}{N}}$$

We noted earlier that the maximum likelihood estimator, $\hat{p}$ is upwardly biased. It requires only simple algebra to show that $p_b < \hat{p}$ suggesting that it might be a less biased estimator. This possibility was investigated numerically and as is shown in Table IV, p_b is far less biased than $\hat{p}$. This would suggest that if bias in the point estimator is important to the investigator, p_b is a better choice for the point estimator particularly when rm is small.

p	**rm**	$E(\hat{p})$	$E(p_b)$
1/1000 = 0.001	2	4.25×10^{-3}	9.8450×10^{-4}
	3	1.61×10^{-3}	1.0004×10^{-3}
	5	1.25×10^{-3}	1.0002×10^{-3}
	10	1.11×10^{-3}	1.0001×10^{-3}
1/10,000 = 0.0001	2	2.24×10^{-4}	9.9860×10^{-5}
	3	1.50×10^{-4}	1.00005×10^{-4}
	5	1.25×10^{-4}	1.00002×10^{-4}
	10	1.11×10^{-4}	1.00001×10^{-4}

Table IV: Comparison of the Expected value of the MLE, $\hat{p}$, to the expected value of the mode of the Bayesian posterior distribution, p_b. Calculations shown are for pool sizes of N = 50.

Equal tail area $(1-\alpha)\%$ credibility intervals are also easily calculated given the posterior distribution, $\pi(p)$. To this end we must find values p_l and p_u such that,

$$\int_0^{p_l} \pi(p)dp = \alpha/2 \text{ and } \int_0^{p_u} \pi(p)dp = 1-\alpha/2$$

That is, for p_l we must solve the equation,

$$\int_0^{p_l} \frac{N\left[1-(1-p)^N\right]^{r-1}\left[(1-p)^N\right]^{y+\frac{1}{2}-\frac{1}{n}}}{B(r, y+\frac{1}{2})}dp = \int_0^{\left[1-(1-p_l)^N\right]} \frac{u^{r-1}(1-u)^{y+\frac{1}{2}-1}}{B(r, y+\frac{1}{2})}du = \frac{\alpha}{2}$$

Note that if we let $u_l = \left[1-(1-p_l)^N\right]$ the right hand integral can be solved for u_l be means of the Beta distribution quantile function. p_l is then easily calculated from u_l by simple algebra. Computation of p_u is accomplished in the same manner. Again, coverage probability is not a Bayesian concept but for comparison purposes Table V gives the calculated coverage for the 95% Bayesian credibility for a range of values of p when N = 50 and rm = 3.

From Table V it is clear that the credibility is not exact in the sense described above. On the other hand, it appears to be reasonably close to the nominal level across all values of p considered and so should be quite useful. Before using this in a situation where you have prior knowledge which leads you to believe that the random variable p is in a particular interval, a simulation study for a moderate number of values of p in this interval will be valuable in assessing results.

$p \times 10^4$	Coverage Probability
0.961	0.94984
1.234	0.95019
1.585	0.95010
2.035	0.94979
2.613	0.95049
3.355	0.94992
4.307	0.94925
5.531	0.95098
7.102	0.95131
9.119	0.94769
11.709	0.94712
15.034	0.95200
19.305	0.95480
24.788	0.95024

Table V: Coverage probabilities for the 95% credibility interval for p in the approximate range $\left(\frac{1}{10,000}, \frac{1}{100}\right)$. The actual values of p were chosen equally spaced on a logarithmic scale. For this table, the pool size N = 50 and rm = 3.

General Conclusions: This method of sampling has been suggested as a reasonable approach to take when the population prevalence is believed to be very small. We have investigated the Maximum likelihood approach for finding a point estimate, shown that it is strongly consistent and that it is asymptotically normally distributed. We have found asymptotic expansions for the bias and mean square error and have investigated the bias of the estimator numerically. We have also considered exact confidence intervals analogous to the Clopper-Pearson intervals for the Binomial Distribution. Finally we have considered a Bayesian analysis based on the noninformative Jeffreys prior. We have found that the MLE is upwardly biased and severely so when the number of positive pools required prior to stopping the sampling is small (r = 1 or 2). The confidence intervals are slightly conservative but not nearly as conservative as the Clopper-Pearson intervals for the Binomial sampling model. We have also seen that the mode of the posterior distribution, viewed as a point estimate, is nearly unbiased even when r is small. The Bayesian credibility intervals, although not exact, have nice coverage properties from a Frequentist point of view.

References:

(1). Thompson ,K.H., "Estimation of the proportion of vectors in a natural population of insects insects". *Biometrics 18*: 568-578 (1962).

(2). Chiang ,C.L., Reeves ,W.C., " Statistical estimation of virus infection rates in mosquito vector populations", *Am J Hyg 75*: 377-391 (1962).

(3). Katholi, C.R., Toé L., Merriweather ,A., Unnasch, T.R., "Determining the prevalence of *Onchocerca volvulus* infection in vector populations by PCR screening of pools of black flies. *J Infect Dis 172*: 1414-1417 (1995).

(4). Barker, J.T., "Statistical Estimators of Infection Potential Based on PCR Pool Screening With Unequal Pool Sizes", *Biostatistics*. Birmingham: University of Alabama at Birmingham, (2000).

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

(6). George, V.T. and Elston, R.C., "Confidence intervals based on the first occurance of an event", *Statistics in Medicine*, Vol. 12, pp 685-690 (1993).

(7). Lui, K., "Confidence limits for the population prevalence rate based on the negative binomial distribution", *Statistics in Medicine*, Vol. 14, pp 1471-1477 (1995).

(8). Morris, K. W., "A note on direct and inverse binomial sampling", *Biometrika*, Vol. 50, N0. 3/4, pp 544-545, (1963).

(9). Patil, G. P., "On the evaluation of the Negative Binomial Distribution with examples", *Technometrics*, Vol. 2, No. 4, pp 501-505 (1960).

(10). Singer, J and Sen, P.K., 'Large Sample Methods in Statistics", Chapman & Hall, New York, (1993).

(11). Serfling, "Approximation Theorems of Mathematical Statistics", Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York, (1980).

(12). Ferguson, T.S., "A Course in Large Sample Theory", Chapman and Hall, New York (1996).

(13). Wald, A., "Note on Consistency of the Maximum Likelihood Estimator", Annals of Mathematical Statistics, Vol 20, No. 4, pp 595-601 (1949).