Pool Testing Algorithm for Estimating Prevalence with Imperfect Test

Nyongesa L. Kennedy
Department of Mathematics,
Masinde Muliro University of Science and Technology.

Abstract
In recent years, pooling algorithms have gained applications in epidemiological studies. Pooling is cost effective especially if the disease prevalence is low. In this study, estimation of disease prevalence in the presence of inspection errors is discussed. It is observed that the model becomes more efficient as the pool sizes increases. The CI of the estimator have been constructed.

AMS subject classification:
Keywords: Likelihood, Prevalence, Pool.

1. Introduction
The testing of pooled samples of biological specimens for the presence of disease has a long history, beginning with Dorfman (1943). The basic idea is to pool together biological samples (e.g. urine, sera, or plasma) from a number of individuals and then test the pooled sample rather than test individual samples for evidence of disease. Epidemiological studies that use pooling have two objectives. The first objective is to screen a large population in order to identify those individuals with the disease (Dorfman, 1943). The second objective is to estimate the prevalence of the disease in the population (Thompson, 1962 and Sobel and Elashoff, 1975). For either objective, pooling is more cost effective than individual testing especially when the underlying testing disease prevalence is low, because if a pool tests negative it implies that none of the individuals that comprise the pool has the disease and thus it is not necessary to test each individual that comprise this pool.

In recent years, there has been renewed interest in pooling studies of biological specimens because of its application in HIV/Aids epidemiology (Kline et al., 1989, Monzon
et al., 1992). The procedure has potential in the application of HIV/AIDS testing because disease prevalence is estimated without necessarily identifying the diseased individuals as this may lead to denial of insurance, work permit or psychological torture (Gastwirth and Hammick, 1989). For example, Johnson et al., (1991) studied the cost effectiveness of pooling algorithm for the first objective of identifying diseased individuals. The algorithm involves dividing positive pools into halves, testing each of the two new smaller pools and continue subdividing positive pools. Litvak et al., (1994) extended this work by considering pooling algorithms when there are test errors (imperfect sensitivity and specificity) and showed that some of these algorithms can decrease the error rates of the screening procedures (the false positive and false negative) compared to individual testing. Nyongesa (2004) generalized and extended Monzon et al., (1992) model; of which it was observed that repeated testing increases the sensitivity and specificity of the group testing algorithm. Maheswaran et al., (2008) have computed statistical measures in their proposed block testing strategy.

Recent studies have focused on the second objective of estimating prevalence (Litvak et al., 1994 and Hardwick et al., 1998). These studies were generalized by Xie et al., (2001) who introduced blockers and synergism during early stages of drug discovery. Brookmeyer (1999) discussed the generalization of Thompson’s (1962) work on the development of estimators for single-stage pooling studies. Also, Tu et al., (1995) extended the work of Thompson (1962) by introducing errors in the procedure. In this paper we introduce the error component in Thompson’s estimation model and estimate the prevalence in the presence of errors, as test kits are not always hundred percent accurate as assumed by Thompson (1962), Brookmeyer (1999) among others. There is a large literature on the treatment of confidence interval (CI) of the prevalence rate using pool-testing schemes for instance see Hepworth (1996, 2004, and 2005). The most recent paper on this subject is by Hepworth and Watson (2009) that investigated the behavior of the bias of the maximum likelihood estimator of the prevalence in group-testing when pools of different sizes are used.

2. Single-Stage Pool Testing Model

In this section, we discuss our proposed testing scheme. First, suppose we have a population of size \( N \) say pool the population into \( n \) pool samples (all of equal sizes) say \( k \), then test each pool separately for the characteristic. If the pooled sample tests negative, it is dropped from further investigation. On the other hand, if it tests positive, individual testing is carried out. The procedure under discussion is a single-stage model as presented by Dorfman (1943) and used by Thompson (1962) to estimate the prevalence of a disease. The following notation will be useful in our discussion: the following indicator functions;

\[
T_h = \begin{cases} 
1, & \text{if the } h^{th} \text{ pool test is positive on the screening} \\
0, & \text{otherwise.}
\end{cases}
\]
The notations are adopted from Gastwirth and Hammick (1989). This indicator function will be utilized in the derivation of core probabilities for example, the probability that a $h^{th}$ pool of size $h$ is positive is $p(D_h=1) = (1 - p)^k$ and test result is positive is $p(T_h = 1)$. The sensitivity is $pr(T_h = 1|D_h = 1)$ and specificity $P(T_h = 0|D_h = 0)$. In the next section, we estimate the prevalence based on the described model.

2.1. Estimation of Prevalence in a Single-Stage Pool Testing Scheme

Consider a population of size $N$, to be investigated for some attributes or trait. The idea here is to pool the population $N$ into $n$ pools each of size $k$. Perform a test on each $n$ constructed pools. Notice that in practice, we observe the $T_h's$, i.e the readings on the test kits. For simplicity, let $X_1$ be the number of pools that test positive on the test kit and $X_2$, the number of pools that test negative on the test kit and, here $X_2 = n_1 - X_1$. Therefore, $X_1$ and $X_2$ are discrete random variables, representing the number of pools classified positive and negative respectively. Clearly, $X_1$ can take the values 0, 1, 2, ..., $n_1$ and hence assumed to be binomially distributed for ease of mathematical tractability in this study.

Let $p$ be a prevalence rate of the attribute or trait under investigation, that is to be estimated, then $q = 1 - p$, is the probability that an individual is negative. Our interest in this study is to consider the pool testing problem when tests in use are imperfect and extend the procedure into hierarchical estimation. For a single stage model with imperfect tests, we have a probability of interest, that is $\pi_1(p)$, the probability of classifying a pool as positive by the test or simply a probability of a positive reading on the test kit. To obtain $\pi_1(p)$, we apply the law of total probability on $T_h's$ to obtain

$$\pi_1(p) = (1 - q^{k_1})\eta + q^{k_1}(1 - \phi)$$

where $\eta$ and $\phi$ are the sensitivity and specificity of the test kits in use. By sensitivity, we mean the probability of correctly classifying a positive pool while specificity means the probability of correctly classifying a negative pool. It is assumed that $\eta$ and $\phi$ remain the same for the entire experiment. It follows that the number of pools that test positive on the test, $X_1$ follow binomial distribution i.e., $X_1 \sim binomial(n_1, \pi_1(p))$, where the probability $\pi_1(p)$ for a positive batch depends on the prevalence $p$, sensitivity, specificity, and pool size $k_1$. For a perfect test $\eta = \phi = 1$ implying that $\pi_1(p) = 1 - (1 - p)^{k_1}$. Clearly, $p \in [0, 1]$, thus

$$1 - \phi \leq \eta$$

Equation (2) shows that $\phi_1(p)$ is a continuous function of $p$, bounded above by $\eta$ and below by $1 - \phi$. Back to our main discussion, that is, $X_1 \sim binomial(n_1, \pi_1(p))$, then the likelihood function is

$$L(p|x, \eta, \phi) = \binom{n_1}{x_1} (\pi_1(p))^{x_1} (1 - \pi_1(p))^{n_1-x_1}$$

(3)
or simply written as

\[ L(p|x, \eta, \phi) \propto (\pi_1(p))^x (1 - \pi_1(p))^n_{1-x} \]

where the constant of proportionality is \( \frac{n_1}{x_1} \). Thus, the log-likelihood is

\[ \log L(.) \propto x \log \pi_1(p) + (n_1 - x_1) \log(1 - \pi_1(p)) \]

Maximizing the log-likelihood with respect to \( q \), we obtain

\[ \hat{p} = 1 - \left[ \frac{(X_1 - \eta n_1)}{n_1(1 - \phi - \eta)} \right]^{k_1} \]

But \( \phi + \eta - 1 > 0 \) (Tu et al., 1995); hence (5) becomes

\[ \hat{p} = 1 - \left[ \frac{(\eta n_1 - X_1)}{n_1(\phi + \eta - 1)} \right]^{\frac{1}{k_1}} \]

Now, assuming that tests have perfect sensitivity and specificity i.e., \( \eta = \phi = 1 \), becomes

\[ \hat{p} = 1 - \left( \frac{X_2}{n_1} \right) \]

a common result obtained by both Thompson (1962) and Brookmeyer (1999), thus our model generalizes the procedures into an imperfect inspection model, which is realistic in practice. The mean of the estimator \( \hat{p} \) is

\[ E(\hat{p}) = 1 - \sum_{i=0}^{n_1} \left[ n_1(\eta - 1) + i \right] \frac{1}{\binom{n_1}{i}} \left( \frac{1 - \pi_1(p)}{\pi_1(p)} \right)^i \]

Setting \( k_1 = 1 \) in (7), we obtain \( E(\hat{p}) = p \) implying that \( \hat{p} \) is an unbiased estimator of \( p \) for \( k_1 = 1 \). For \( k_1 > 1 \), and upon using Jensen’s inequality (Billingsley, 1995 p 276), with the estimate of \( p \) given by (5), we have for \( k_1 \geq 1 \) and \( \eta > \frac{X_1}{n_1} \),

\[ E\left[ \frac{n_1 \eta - X_1}{n_1(\eta + \phi - 1)} \right]^{\frac{1}{k_1}} \leq \left[ \frac{n_1 \eta - E(X_1)}{n_1(\eta + \phi - 1)} \right]^{\frac{1}{k_1}} \]

In practice \( \eta \) is close to one (see Kline et al., 1989) but from pool testing literature, the procedure is only worthwhile if the prevalence is small (Dorfman, 1943). i.e., \( \frac{X_1}{n_1} \to 0 \). Hence \( \eta \approx 1 \) and \( \frac{X_1}{n_1} \to 0 \) implying that \( \eta > \frac{X_1}{n_1} \) in practice, thus the above inequality holds and with \( \eta = 1 \) i.e., in the absence of test errors

\[ E(\hat{p}) \geq p, \]
with equality holding when \( k_1 = 1 \). As observed in (8) for \( k_1 > 1 \), \( \hat{p} \) overestimate \( p \) (see Nyongesa, (2011)). The computation of the variance of the estimate in (5) is obtained from \( \text{var}(\hat{p}) = E[(\hat{p} - E(\hat{p}))^2] \) and after a few tabulation, we obtain the MSE as

\[
MSE(\hat{p}) = \sum_{i=0}^{n_1} \left[ \frac{n_1(\eta - 1) + i}{n_1(\phi + \eta - 1)} \right]^2 \frac{i}{k_1} \pi_1(p)^{n_1-i}(1 - \pi_1(\hat{p}))^i
\]

(10)

where \( i = X \) in this case if we set \( \eta = \phi = 1 \) and \( k_1 = 1 \) in the above equation, we have

\[
MSE(\hat{p}) = \frac{(1 - p)p}{n_1},
\]

(11)

the variance of the estimator in the case where individual testing with perfect tests are performed. Suppose \( \eta = \phi = 1 \), Equation (9) yields

\[
MSE(\hat{p}) = \frac{E[i^{2/k_1}]}{n_1^{2/k_1}} - [1 - E(\hat{p})]^2,
\]

(12)

similarly, \( i = X_1 \) in particular consider \( E(i^{2/k_1}) \), if \( k_1 = 1 \), we get a result similar to (10). Also from (5), \( \hat{p} \) is a strong consistent estimate of \( p \) i.e.,

\[
pr\{ \lim_{n_1 \to \infty} (\hat{p} - p) = 0 \} = 1
\]

by the asymptotic property of \( \hat{p} \). Thus, for \( p = 0 \) and \( \phi < 1 \), then \( pr\{ \hat{p} = 0 \} \longrightarrow 0.5 \) as \( n_1 \longrightarrow 1 \) while for \( p = 1 \) and \( \eta < 1 \), we have \( pr(\hat{p} = 1) = 0 \) as \( n_1 \longrightarrow 1 \). These results will be further evaluated by the computation of moments of \( \hat{p} \) below. Next, we compute the asymptotic variance of \( \hat{p} \) from

\[
\left\{ -E\left[ \frac{\partial^2}{\partial p^2} \log L(p|\eta, \phi, x) \right] \right\}^{-1}
\]

(13)

From (13), we get

\[
\text{var}(\hat{p}) = \frac{(1 - p)^2 \pi_1(p)(1 - \pi_1(p))(1 - p)^{-2k_1}}{n_1 k_1^2 \pi_1(p)^{n_1}(\eta + \phi - 1)^2}
\]

(14)

Upon setting \( \phi = \eta = 1 \), the asymptotic variance obtained in (14) reduces to

\[
\text{var}(\hat{p}) = \frac{1 - (1 - p)^{k_1}}{n_1 k_1^2 (1 - p)^{k_1-2}}
\]

(15)

a result also obtained by Thompson (1962) and others. Similarly, the asymptotic variance (13) can also be derived by the Delta method, Lehmann and Casella, (1998 p 58).
Equation (14) provides an estimate of the asymptotic variance of $\hat{p}$ in the perfect inspection model of Dorfman (1943). Further, as pointed out before the derivation of (14), we need to investigate the asymptotic behavior of our estimator $\hat{p}$ and this will be achieved by investigating the properties of the moments of $\hat{p}$. The next discussion provides an investigation on the behavior of $\hat{p}$.

We investigate the behavior of our estimator by considering some specified parameters. For $p = 0$ and $\phi < 1$, expanding (5) about $1 - \phi$ by Taylor’s series expansion, we get

$$E(\hat{p}^2) = \frac{\phi(1 - \phi)}{k_1^2 n_1 (\eta + \phi - 1)^2} + O(n_1^{-2})$$

(16)

Also, applying the combination of Jensen’s inequality and Chebychev’s inequality on (5) yields

$$E(1 - \hat{p}^2) \leq \left(\frac{1}{\eta + \phi - 1}\right)^{\frac{1}{2}} \left(\frac{\phi(1 - \phi)}{n_1}\right)^{\frac{1}{2}}$$

(17)

To investigate the bias, we compute $E(\hat{p} - p)$ by Taylor’s series, we have

$$E(\hat{p} - p) = \frac{(k_1 - 1) \text{var}(\hat{\pi}_1(p))}{2k_1^2 n_1 (\eta + \phi)^2} + O(n_1^{-2})$$

(18)

where $\hat{\pi}_1(p) = [1 - (1 - \hat{p})^{k_1}]\eta + (1 - \hat{p})^{k_1}(1 - \phi)$. This shows that $\hat{p}$ is upwardly biased as is with pooling in general. Therefore, the procedure is only recommended when $p$ is small otherwise individual testing is preferable. Now, applying the inequality $E|x| \leq \{E(X^2)\}^{\frac{1}{2}}$ to (15) and (16), we get

$$E(\hat{p}) \leq \frac{\{\phi(1 - \phi)\}^{\frac{1}{2}}}{k_1 n_1 (\eta + \phi - 1)} + O(n_1^{-2})$$

(19)

and

$$E(1 - \hat{p}) \leq \left(\frac{1}{(\eta + \phi - 1)}\right)^{\frac{1}{2}} \left(\frac{\phi(1 - \phi)}{n_1}\right)^{\frac{1}{2}}$$

(20)

respectively. A point estimate such as the one in (5) is not very satisfactory since it is almost certain that the derived results will not be exactly equal to the unknown true prevalence rate. To overcome this difficulty, the prevalence rate may be estimated in terms of a confidence interval rather than as a single-value, and one can state with certain degree of confidence that the computed interval contains the unknown prevalence rate.

Having derived the formula for the estimator $\hat{p}$ and its variance, the confidence interval for the prevalence rate may be deduced with ease if a normal distribution of
the estimate is assumed. Asymptomatically, \( \hat{p} \) is unbiased, normally distributed and efficient; that is, for fixed \( k_1 \) and \( n_1 \to 1 \), i.e., as the number of pools increases to \( \infty \),

\[
\sqrt{n_1}(\hat{p} - p) \overset{D}{\to} \text{normal} \left( 0, \frac{(1 - p)^2 \pi_1(p)(1 - \pi_1(p))(1 - p)^{-2k_1}}{k_1^2(\phi + \eta - 1)^2} \right)
\]  

(21)

In practice, some prior value \( p \), say \( p_1 \), can be used to determine the group size. However, the distribution of the estimate \( \hat{p} \) is likely to be skewed and it is not known how fast this distribution approaches normal as the number of pools increases. Investigation can be carried out by Monte Carlo simulation in this case; consequently the confidence interval derived on the basis of the distribution of \( \hat{p} \) is likely to be accurate (cf. Chiang and Reeves, 1962). The 100\((1 - \alpha)\)% confidence interval for \( \hat{p} \) may be computed from

\[
\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{(1 - p)^2 \pi_1(p)(1 - \pi_1(p))(1 - p)^{-2k_1}}{n_1k_1^2(\phi + \eta - 1)^2}}
\]  

(22)

where, \( Z \sim \text{normal}(0, 1) \). Equation (21) provides the version of confidence interval proposed by Chiang and Reeves, (1962) when tests are imperfect.

3. Conclusion

We have constructed a prevalence estimator of pool testing strategy and discussed its properties. To help our discussion we compute the ARE’s for various groups as we compare individual-testing with pool-testing.

<table>
<thead>
<tr>
<th>Pool size</th>
<th>( p = 0.01, \eta = 0.99 )</th>
<th>( p = 0.01, \eta = 0.99 )</th>
<th>( p = 0.01, \eta = 0.99, \phi = 0.99 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 32 )</td>
<td>1.59</td>
<td>2.21</td>
<td>3.56</td>
</tr>
<tr>
<td>( 16 )</td>
<td>1.69</td>
<td>2.28</td>
<td>3.37</td>
</tr>
<tr>
<td>( 8 )</td>
<td>1.67</td>
<td>2.15</td>
<td>2.79</td>
</tr>
<tr>
<td>( 4 )</td>
<td>1.55</td>
<td>1.85</td>
<td>2.01</td>
</tr>
</tbody>
</table>

Clearly, the model is more efficient in situations where test kits have low sensitivity and specificity.

References


