Estimation of Multiple Traits in an M – Stage Group Testing Model

John L. Sirengo
Department of Mathematics
Kibabii University, Kenya

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

Shem Aywa
Department of Mathematics, Kibabii University, Kenya

This article is distributed under the Creative Commons by-nc-nd Attribution License.
Copyright © 2020 Hikari Ltd.

Abstract

Multi – stage group testing design involves testing groups for the presence or absence of traits in a population and sequentially subdividing the groups into sub – groups. The sub – groups to be tested at a particular stage are based on the information obtained from the previous stage. This paper developed an M – stage design for testing the presence of multiple traits in a finite population. The asymptotic variance of the model is also discussed. The model gains efficiency with each additional stage when the occurrence of traits is rare in the population.

Keywords: M – stage, asymptotic variance and additional stage
1. Introduction

The testing of pooled samples of biological specimens for a disease has a long history, beginning with Dorfman (1943) seminal work on identifying individuals with syphilis during the World War II as an economical method of testing blood samples. The basic idea is to divide the population into groups and a test is performed on each group rather testing each individual unit of the group for the presence of a trait. The main benefit of group testing procedure is that it reduces the number of tests if the prevalence rate is low. Hughes-Oliver and Swallow (1994) developed a multistage group testing model for testing the presence of a single trait in a population. Brookmeyer (1999) advanced on this by carrying out the analysis of multistage pooling studies of Biological specimens for estimating disease incidence and prevalence. Recently, Hughes-Oliver and Rosenberger (2000) proposed a two – stage algorithm for testing the presence of multiple traits. To this end multistage group testing procedure can be used to estimate the prevalence rate of a trait if it occurs. Therefore the purpose of this paper is to develop an M – stage model for testing the presence a multiple number of traits in a finite a population with the use of Hughes-Oliver and Rosenberger (2000) group testing procedure.

For simplicity, throughout this paper we shall assume that samples being pooled are independent and identically distributed. In addition, the tests are also independent of one another (cf. Nyongesa, 2004). The rest of the paper is arranged as follows: Construction of the model as proposed by Hughes-Oliver and Rosenberger (2000) in section 2 and the probability of classification is discussed in section 3. The likelihood function is presented in section 4 and maximum likelihood estimator of the prevalence is presented in Section 5. Derivation and computation of the asymptotic variance is in section 6 while Section 7 provides the conclusion to the study.

2. The Model

The population N under study is assumed herein as sufficient for the experiment to be considered. Firstly, the population N is split into $n_1$ homogeneous pools each of size $k_1$. The $n_1$ constructed pools are subjected to testing for the presence or absence of T – traits. Positive results indicate the presence of at least one of the T – traits and the negative reading indicates the absence of all the traits. The pools that tested positive at stage one are split into smaller sub groups of size $k_2 (k_2 < k_1)$ that forms pools for testing at stage two, in total we shall have $n_2$ pools each of size $k_2$ for testing in stage two. The pools that tests positive at stage two are further split into smaller pools of size $k_3 (k_3 < k_2)$ for testing in stage three and in total we have $n_3$ pools that are constructed in this stage. The procedure is repeated up to $m^{th}$ stage where at this stage $n_m$ sub pools of size $k_m (k_m < k_{m-1})$ are constructed for testing. The amalgamated M – stage group testing is shown in Figure 1 below.
Estimation of multiple traits in an M – stage group testing model

Population Size ($k_0$)

Stage 1
$n_1$ groups of size $k_1$ ($k_1 < k_0$)

Stage 2
$n_2$ groups of size $k_2$ ($k_2 < k_1$)

Stage h
$n_h$ groups of size $k_h$ ($k_h < k_{h-1}$)

Stage m
$n_m$ groups of size $k_m$ ($k_m < k_{m-1}$)

Figure 1: Generalized Hughes-Oliver and Rosenberger (2000) model
3. Probability of classifying an \(i^{th}\) Pool in the \(h^{th}\) Stage

Here we are interested in the \(h^{th}\) stage since \(h\) will be allowed to vary from 1 to \(m\) as discussed above. Our objective here is to construct the probability of positive reading at this stage. Notice that \(k_m \in k_{m-1} \in k_{3m-2} \in \cdots \in k_2 \in k_1\), this forms a filtration therefore we shall employ the theory of Martingale in constructing this probability (cf Billingsley 1995). The probability of classifying a \(j^{th}\) individual from an \(i^{th}\) pool in the \(h^{th}\) stage is obtained as follows: The \(j^{th}\) unit is subjected to testing for the presence of \(T – traits, the unit can test positive for at least one of the \(T – traits or negative for all the traits.\)

Let

\[
Y_{h,j}^{(t)} = \begin{cases} 1 & \text{if the } j^{th} \text{ individual tests positive of the } t^{th} \text{ trait, for } t = 1, 2, \ldots, T \\ 0 & \text{otherwise} \end{cases}
\]

since we have \(T – traits the vector of responses is

\[
\left( Y_{h,j}^{(1)}, \ldots, Y_{h,j}^{(T)} \right)
\]

For simplicity, we shall denote this vector by \(Y_{h,j}\)

That is

\[
Y_{h,j} = \left( Y_{h,j}^{(1)}, \ldots, Y_{h,j}^{(T)} \right)
\]

The probability of vector (1) is

\[
\Pr(Y_{h,j} = y_{h,j}) = \Pr(Y_{h,j}^{(1)} = y_{h,j}^{(1)}, \ldots, Y_{h,j}^{(T)} = y_{h,j}^{(T)})
\]

Upon assuming independence in the \(T – traits (Jacqueline and Rosenberger, 2000) (2) simplifies to

\[
P_r(Y_{h,j} = y_{h,j}) = P_r(Y_{h,j}^{(1)} = y_{h,j}^{(1)}) \ldots P_r(Y_{h,j}^{(T)} = y_{h,j}^{(T)})
\]

Note that a random variable \(Y_{h,j}^{(t)}\) is a Bernoulli random variable with probability of success \(1 - (1 - p_t)\) for \(t = 1, 2, \ldots, T\) (cf Dorfman, 1943). Thus (3) reduces to

\[
P_r(Y_{h,j} = y_{h,j}) = \prod_{t=1}^{T} (1 - (1 - p_t))^{y_{h,j}^{(t)}} (1 - P_t)^{1-y_{h,j}^{(t)}}.
\]

Notice that the above working was devoted to classifying a \(j^{th}\) individual from an \(i^{th}\) group in the \(h^{th}\) stage; next we compute the probability of classifying the \(i^{th}\) group itself. This will be (4) for the \(i^{th}\) pool scenario.

Let

\[
Y_{h,i}^{(t)} = \begin{cases} 1 & \text{if the } i^{th} \text{ group tests positive of the } t^{th} \text{ trait, for } t = 1, 2, \ldots, T \\ 0 & \text{otherwise} \end{cases}
\]
Estimation of multiple traits in an M–stage group testing model

Also define the vector for the T-traits as follows

\[ Y_{hi} = \left( Y_{hi}^{(1)}, Y_{hi}^{(2)}, \ldots, Y_{hi}^{(T)} \right) \]

Thus the probability of classifying the i\textsuperscript{th} group in the h\textsuperscript{th} stage is the probability of (5).

That is

\[ P_r(Y_{hi} = y_{hi}) = P_r\left( Y_{hi}^{(1)} = y_{hi}^{(1)}, \ldots, Y_{hi}^{(T)} = y_{hi}^{(T)} \right) \]

Upon assuming independence in the groups we get

\[ P_r(Y_{hi} = y_{hi}) = \prod_{i=1}^{T} P_r\left( Y_{hi}^{(i)} = y_{hi}^{(i)} \right) \]

Also we note that \( Y_{hi}^{(i)} \) is a Bernoulli random variable with probability of success \( 1 - (1 - p_e)^{k_h} \)
(cf Dorfman, 1943). Hence (7)

\[ \Pr(Y_{hi} = y_{hi}) = \prod_{i=1}^{T} \left( 1 - (1 - p_e)^{k_h} \right)^{y_{hi}^{(i)}} \left( (1 - p_e)^{k_h} \right)^{1 - y_{hi}^{(i)}} \]

The sub-groups used at the h\textsuperscript{th} stage comes from positive sub pools in stage h – 1.
The probability of interest that is the probability of classifying the i\textsuperscript{th} group as positive given that it comes from a positive sub-group in stage h – 1 is

\[ \Pr (Y_{hi}^{(i)} = y_{hi}^{(i)} | Y_{h-1;i} = y_{h-1;i}) \]

Reorganizing this conditional probability we have

\[ \Pr (Y_{hi}^{(i)} = y_{hi}^{(i)} | Y_{h-1;i} = y_{h-1;i}) = \frac{Pr (Y_{hi}^{(i)} = y_{hi}^{(i)}, Y_{h-1;i} = y_{h-1;i})}{Pr (Y_{h-1;i} = y_{h-1;i})} \]

Notice that \( k_h \in k_{h-1} \), this implies that

\[ \Pr (Y_{hi} = y_{hi}^{(i)} | Y_{h-1;i} = y_{h-1;i}) = \Pr (Y_{hi} = y_{hi}^{(i)}) | \Pr (Y_{h-1;i} = y_{h-1;i}) \]

We recall that the i\textsuperscript{th} group is positive if at least one of the units in the group is positive, hence
This is the probability of classifying an \(i^{th}\) pool in the \(h^{th}\) as positive. Equation (11) is of a truncated model. Working similarly, the probability that a pool tests negative at the \(h^{th}\) stage is

\[
\frac{((1-p_i)^{k_h})^{i-y_{hi}}}{(1-(1-p_i)^{k_h-y_{hi}})\cdot (1^{i-y_{hi}})}
\]

Equations (11) and (12) are vital in the formulation of an M-stage multiple traits estimation model as they are the probability of classifying a group as positive and negative respectively for the \(i^{th}\) trait in the \(h^{th}\) stage.

4. Likelihood Function

The likelihood function at this stage is anchored on Equations (11) and (12). Also in this stage there are \(n_h\) sub-groups to be tested for the presence of the \(t^{th}\) trait, if the response is \(Y_{hi}^{(t)}\), where

\(i = 1, 2, \ldots, n_h, t = 1, 2, \ldots, T\) and \(h = 1, 2, \ldots, m\).

Thus utilizing the indicator function \(Y_{hi}^{(t)}\) as proposed above the likelihood function at the \(h^{th}\) stage is

\[
L_h(p_t) \propto \prod_{i=1}^{n_h} \frac{1-(1-p_i)^{K_h-y_{hi}}}{1-(1-p_i)^{K_h-1-y_{hi}}},
\]

Where \(p_t = (p_1, p_2, \ldots, p_T)^t\)

Model (13) is a truncated Binomial model. Notice that \(h = 1, 2, \ldots, m\), in model (13) thus the M-stage likelihood function is

\[
L_m(p_t) \propto \prod_{h=1}^{m} \prod_{i=1}^{n_h} \frac{1-(1-p_i)^{K_h-y_{hi}}}{1-(1-p_i)^{K_h-1-y_{hi}}},
\]

Equation (14) holds with \((1-p_t)^{k_0} = 0\), this is true because at initial stage \(k_0\) is equal to the entire population which is large and \((1-p_t)^{k_0} \rightarrow 0\) as \(k_0 \rightarrow \infty\) where \(k_0 = N\).

Upon setting \(m = 1\) in (14) the model reduces to Hughes-Oliver and Rosenberger (2000) model.
5. Estimator of the Prevalence rate

In this section we determine the estimator of the constructed design (14) by using the maximum likelihood estimate (MLE) method. Mathematically given as

$$\hat{p}_t = \arg\min_{p_t} \sum_{h=1}^{m} \sum_{i=1}^{n_h} \sum_{t=1}^{T} (\cdot)$$

For simplicity we let $q_t = 1 - p_t$, hence

$$f(q_t) = \frac{\partial}{\partial q_t} \log L_m(\cdot) = \sum_{h=1}^{m} \sum_{i=1}^{n_h} \left( \frac{-f(0)^q_k \alpha q_t^{\alpha-1} q_t^{\beta-1} (1-q_t)^{\gamma-1}}{q_t^{\beta-1} (1-q_t)^{\gamma-1}} + \frac{2f(0)^q_k \alpha q_t^{\alpha-1} q_t^{\beta-1} (1-q_t)^{\gamma-1}}{1-q_t^{\beta-1}} \right)$$

The optimal $q_t$ can be obtained by Newton – Raphson iteration method. With

$$q_{t+1} = q_t - \frac{f(q_t)}{f'(q_t)}$$

where $f'(q_t)$ is the derivative of $f(q_t)$ and the iteration ceases if $|q_{t+1} - q_t| < \varepsilon$, for some arbitrary $\varepsilon$. Equation (16) can easily be implemented on a desktop. The estimator $\hat{q}_t$ obtained in (16) is the estimate of $q_t$ for $t = 1, 2, \ldots, T$.

6. Asymptotic Variance

For large sample size that is $N \to \infty$, Tebbs et al. (2003) showed that the asymptotic variance of an estimator is obtained by use of the Cramer – Rao lower bound method. Mathematically written as

$$\text{Var}(\hat{q}_t) = - \left[ E \left( \frac{\partial^2}{\partial q_t^2} \log L_m(\cdot) \right) \right]^{-1}$$

Upon utilizing (17) on (14) we get the asymptotic variance of the model as

$$\text{Var}(\hat{q}_t) = \frac{1}{\sum_{h=1}^{m} \sum_{i=1}^{n_h} \left( \frac{f(0)^q_k \alpha q_t^{\alpha-1} q_t^{\beta-1} (1-q_t)^{\gamma-1}}{1-(1-p_t)q_t^{\beta-1}} + \frac{2f(0)^q_k \alpha q_t^{\alpha-1} q_t^{\beta-1} (1-q_t)^{\gamma-1}}{1-(1-p_t)^{\beta-1}} \right)}$$

Now utilizing (18) we compute the asymptotic variance for various values of $p_t$, using halving method with parallel testing procedure. In the computation of the variance the study considered five stages with ten groups in each stage. The computed results are shown in Table 1.
Table 1: Simulated asymptotic variance for specified values of $p_t$ with $N = 640$ and $n_h = 10$ where $\text{var}(\hat{p}_t) \times 10^{-6}$

<table>
<thead>
<tr>
<th>$m$</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.670</td>
<td>30.206</td>
<td>39.043</td>
<td>45.232</td>
<td>49.535</td>
</tr>
<tr>
<td>2</td>
<td>8.8350</td>
<td>15.103</td>
<td>19.522</td>
<td>22.616</td>
<td>24.768</td>
</tr>
<tr>
<td>3</td>
<td>5.8900</td>
<td>10.069</td>
<td>13.014</td>
<td>15.077</td>
<td>16.512</td>
</tr>
</tbody>
</table>

Table 1 illustrates the asymptotic variance of the constructed estimator after each additional stage, assuming there were $N = 640$ individual units that were initially subdivided into ten groups at stage each of size $k_1 = 64$ individual units and, at each successive stage, the groups were divided into half using the halving method and tested using parallel testing scheme. From the table we observe that as the prevalence increases the asymptotic variance increases and in fact at stage one with $p_t = 0.01$ the asymptotic variance is $1.767 \times 10^{-5}$ and when $p_t = 0.02$ the asymptotic variance is $3.021 \times 10^{-5}$ which implies that with high prevalence rate the probability of detecting a trait in a population is very high. We also observe that the constructed estimator is more efficient when the occurrence of the traits in a population is rare.

It can also be observed that the estimator gains efficiency with each additional stage. This result concurs with the result of Swallow (1987) and many others who proposed group testing procedure when the occurrence of a trait is rare.

7. Conclusion

In this study we have developed an M – stage group testing procedure and constructed an estimator for estimating the occurrence of rare traits in a finite population. The property of the estimator such as the asymptotic variance was discussed. It has been observed that the estimator gains efficiency with each additional stage and optimal results are achieved when the occurrence of traits in a
population is rare. This concurs with Shallow (1985) who recommended that relatively small group sizes should be used to obtain optimal results.

References


Received: April 11, 2020; Published: April 24, 2020