IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 16, Issue 3 Ser. II (May – June 2020), PP 29-34
www.iosrjournals.org

An M – Stage Hierachical Group Testing Model for Estimating Multiple Traits in a Population

Sirengo L. John; Nyonges L. Kennedy; Away Shem
Department of Mathematics, Kibabii University

Abstract: *Monitoring the presence or absence of a trait in a large population one – at – a – time is tedious, uneconomical and bound to errors. The remedy is to group the population into homogeneous groups and test each group for the presence of a trait. Multi – stage group testing procedure involves testing groups for the presence or absence of trait in a population and sequentially subdividing the positive 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 proposed an M – stage hierarchical design for testing the presence of multiple traits in a finite population. The design improves the efficiency of the estimators as evident via the computation of asymptotic variance.*

Keywords: *M – stage, i^{th} group, h^{th} stage.*

Date of Submission: 12-05-2020 Date of Acceptance: 24-05-2020

I. Introduction

The testing of pooled samples of biological specimens for 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. 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 hierarchical 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: Design 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 discussion of the hierarchical asymptotic relative efficiency in section 7 and Section 8 provides the conclusion.

II. Model Construction

The population N understudy 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.

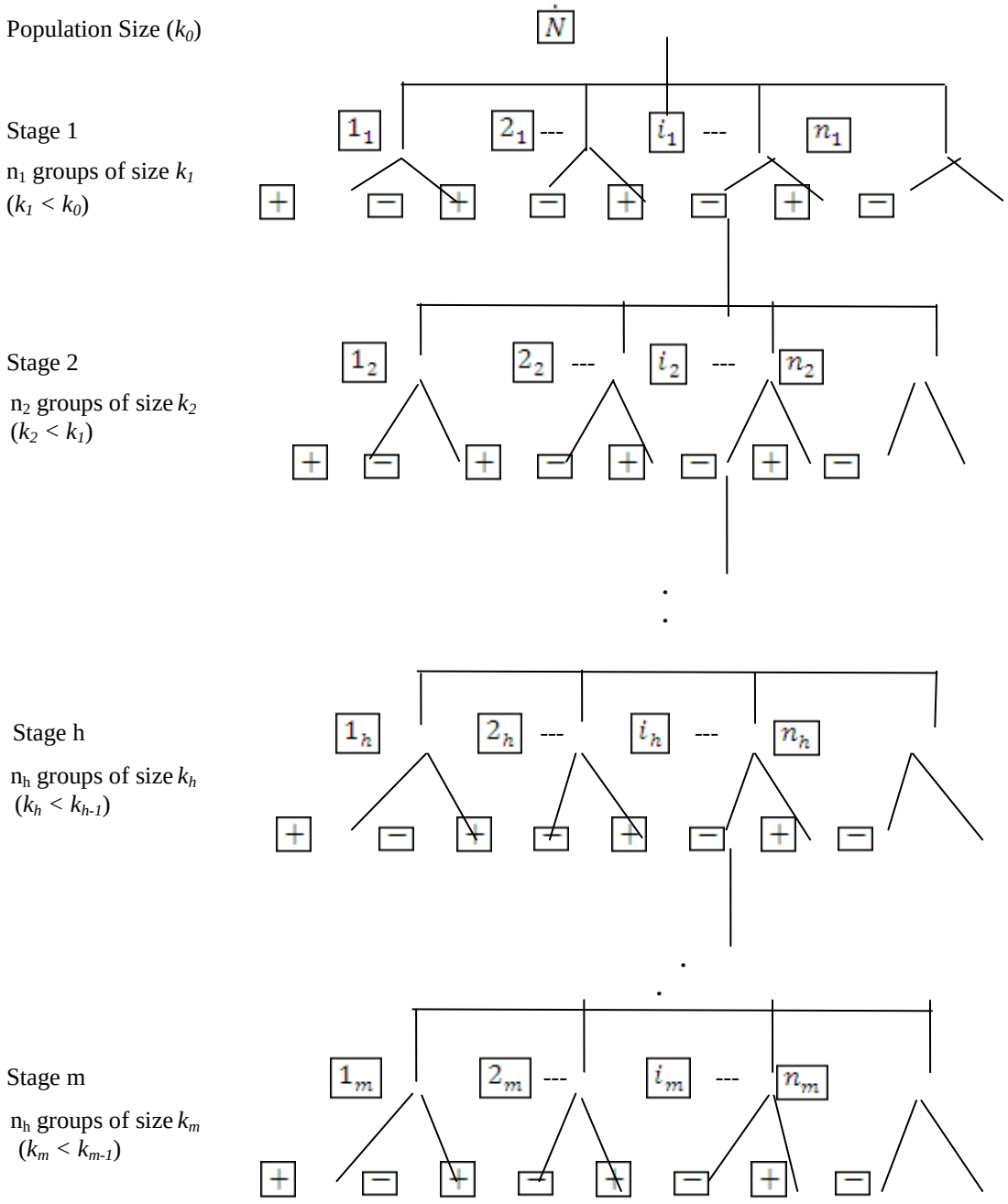

Figure 1: Generalized Hughes-Oliver and Rosenberger (2000) model.

III. Probability of classifying an i^{th} Group 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 $k_m \in k_{m-1} \in k_{m-2} \in \cdots \in k_2 \in k_1$, 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

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

since we have T – traits the vector of responses is

$\left(Y_{h.j}^{(1)}, Y_{h.j}^{(2)}, \ldots, Y_{h.j}^{(t)}, \ldots, Y_{h.j}^{(T)}\right)$.

For simplicity, we shall denote this vector by $Y_{h.j}$

That is

$$Y_{h.j} = \left(Y_{h.j}^{(1)}, Y_{h.j}^{(2)}, \ldots, Y_{h.j}^{(t)}, \ldots, Y_{h.j}^{(T)}\right). \quad (1)$$

The probability of vector (1) is

$$\Pr\left(Y_{h.j} = y_{h.j}\right) = \Pr\left(Y_{h.j}^{(1)} = y_{h.j}^{(1)} \ldots Y_{h.j}^{(T)} = y_{h.j}^{(T)}\right) \quad (2)$$

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

$$P_r\left(Y_{h.j} = y_{h.j}\right) = P_r\left(Y_{h.j}^{(1)} = y_{h.j}^{(1)}\right) \ldots P_r\left(Y_{h.j}^{(T)} = y_{h.j}^{(T)}\right) \quad (3)$$

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, . . . , T (cf Dorfman, 1943). Thus (3) reduces to

$$P_r\left(Y_{h.j} = y_{h.j}\right) = \prod_{t=1}^{T}\left(1 - (1 - P_t)\right)^{Y_{h.j}^{(t)}}(1 - P_t)^{1 - Y_{h.j}^{(t)}}. \quad (4)$$

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

$$Y_{hi.}^{(t)} = \begin{cases} 1 & if\ the\ i^{th} group\ tests\ positive\ of\ the\ t^{th} trait, for\ t = 1,2, \ldots, T \\ 0 & otherwise \end{cases}$$

Also define the vector for the T-traits as follows

$$Y_{hi.} = \left(Y_{hi.}^{(1)}, Y_{hi.}^{(2)}, \ldots., Y_{hi.}^{(t)}, \ldots, Y_{hi.}^{(T)}\right) \quad (5)$$

Thus the probability of classifying the ith group in the hth stage is the probability of (5).

That is

$$P_r(Y_{hi} = y_{hi.}) = P_r\left(Y_{hi}^{(1)} = y_{hi.}^{(1)}, \ldots, Y_{hi.}^{(T)} = y_{hi.}^{(T)}\right). \quad .6)$$

Upon assuming independence in the groups we get

$$P_r(Y_{hi.} = y_{hi.}) = \prod_{t=1}^{T} P_r\left(Y_{hi.}^{(t)} = y_{hi.}^{(t)}\right). \quad (7)$$

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

$$\Pr(Y_{hi.} = y_{hi.}) = \prod_{t=1}^{T}\left(1 - (1 - p_t)^{k_h}\right)^{Y_{hi.}^{t}}\left((1 - p_t)^{k_h}\right)^{1 - Y_{hi.}^{t}} \quad (8)$$

The sub-groups used at the h^{th} stage comes from positive sub pools in stage *h – 1*.

The probability of interest that is the probability of classifying the ith group as positive given that it comes from a positive sub-group in stage *h – 1* is

$$\Pr\left(Y_{hi.}^{t} = y_{hi.}^{t} \backslash Y_{h-1i.}^{t} = y_{h-1i}^{t}\right) \quad (9)$$

Reorganizing this conditional probability we have

$$\Pr\left(Y_{hi.}^{t} = y_{hi.}^{t} | Y_{h-1i.}^{t} = y_{h-1i}^{t}\right) = \frac{\Pr\left(Y_{hi}^{t} = y_{hi.}^{t},\ Y_{h-1i.}^{t} = y_{h-1i.}^{t}\right)}{\Pr\left(Y_{h-1i.}^{t} = y_{h-1i.}^{t}\right)}$$

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

$$\Pr\left(Y_{hi.}^{(t)} = y_{hi.}^{(t)} | Y_{h-1i.}^{(t)} = y_{h-1i.}^{(t)}\right) = \Pr\left(Y_{hi.}^{(t)} = y_{hi.}^{(t)}\right) | \Pr(Y_{h-1i.}^{(t)} = y_{h-1i.}^{(t)}) \quad (10)$$

We recall that the ith group is positive if at least one of the units in the group is positive, hence

$$\Pr(Y_{hi.} = y_{hi.} | Y_{h-1i.} = y_{h-1i.}) = \prod_{t=1}^{T} \frac{\left(1 - (1 - p_t)^{k_h}\right)^{Y_{hi.}^{(t)}}}{\left(1 - (1 - p_t)^{k_{h-1i.}}\right)^{Y_{h-1i.}^{(t)}}} \quad (11)$$

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

$$\frac{\left((1 - p_t)^{k_h}\right)^{1 - Y_{hi.}^{(t)}}}{\left(1 - (1 - p_t)^{k_{h-1i.}}\right)^{Y_{h-1i.}^{(t)}}} \quad (12)$$

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 t^{th} trait in the h^{th} stage.

IV. 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

$$L_h(p_t) \propto \prod_{i=1}^{n_h} \prod_{t=1}^{T} \frac{\left(1-(1-p_t)^{K_h}\right)^{Y_{hi}^{(t)}}}{\left(1-(1-p_t)^{K_{h-1}}\right)^{Y_{h-1i}^{(t)}}} \left(\frac{\left((1-p_t)^{K_h}\right)^{1-Y_{hi}^{(t)}}}{\left(1-(1-p_t)^{K_{h-1}}\right)^{1-Y_{h-1i}^{(t)}}} \right), \qquad (13)$$

Where $p_t = (p_1, p_2, \ldots, p_T)'$

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

$$L_m(p_t) \propto \prod_{h=1}^{m} \cdot \prod_{i=1}^{n_h} \prod_{t=1}^{T} \frac{\left(1-(1-p_t)^{K_h}\right)^{Y_{hi}^{(t)}}}{\left(1-(1-p_t)^{K_{h-1}}\right)^{Y_{h-1i}^{(t)}}} \left(\frac{\left((1-p_t)^{K_h}\right)^{1-Y_{hi}^{(t)}}}{\left(1-(1-p_t)^{K_{h-1}}\right)^{1-Y_{h-1i}^{(t)}}} \right) \qquad (14)$$

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 $(1-p_t)^{k_0} \to 0$ as $k_0 \to \infty$ where $k_0 = N$.

Upon setting $m = 1$ in (14) the model reduces to Hughes-Oliver and Rosenberger (2000) model.

V. Construction of the Estimator

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

$$\hat{p}_t = \underset{p_t}{argmin} \sum_{h=1}^{m} \sum_{i=1}^{n_h} \sum_{t=1}^{T} (\,.\,)$$

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

$$f(q_t) \overset{=}{D} \frac{\partial}{\partial q_t} \log lm(.) = \sum_{h=1}^{m} \sum_{i=1}^{n_h} \left(\frac{-Y_{hi}^{(t)} k_h q_t^{k_h-1} q_t^{k_h} + \left(1-Y_{hi}^{(t)}\right) k_h q_t^{k_h-1} \left(1-q_t^{k_h}\right)}{q_t^{k_h}\left(1-q_t^{k_h}\right)} + \frac{2Y_{h-1i}^{(t)} k_{h-1} q_t^{k_{h-1}-1}}{1-q_t^{k_{h-1}}} \right) \qquad (15)$$

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

With

$$q_{t+1} = q_t - \frac{f(q_t)}{f'(q_t)} \qquad (16)$$

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 ε. 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.

VI. 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

$$Var(\hat{q}_t) = -\left[E\left(\frac{\partial^2}{\partial q_t^2} \log \mathrm{L_m}(.) \right) \right]^{-1} \qquad (17)$$

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

$$Var(\hat{p}_t) = \frac{1}{\sum_{h=1}^{m} \sum_{i=1}^{n_h} \left(-\frac{k_h^2 (1-P_t)^{k_h^{-2}}}{1-(1-P_t)^{k_h}} + \frac{2K_{h-1}(1-P_t)^{k_{h-1}^{-2}} \left(k_{h-1} - \left(1-(1-P_t)^{k_{h-1}}\right)\right)}{1-(1-P_t)^{k_{h-1}}} \right)} \qquad (18)$$

In hierarchical modeling we only consider the asymptotic variance of the i^{th} group at the h^{th} stage *i.e* the group that tested positive at the previous stage. This procedure is illustrated diagrammatically in Figure 2 as previously discussed by Monzon et. al. (1992).

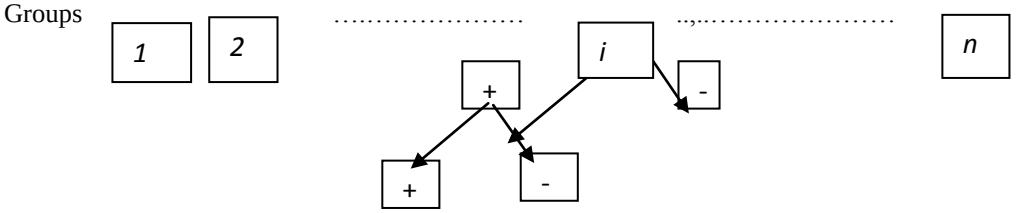

Figure 2: Monzon et. al. (1992) Group Testing Model

Upon invoking Figure 2 Equation (18) becomes

$$Var(\hat{p}_t) = \frac{1}{\sum_{h=1}^{m}\left(-\frac{k_h^2(1-P_t)^{k_h^{-2}}}{1-(1-P_t)^{k_h}} + \frac{2K_{h-1}(1-P_t)^{k_{h-1}^{-2}}(k_{h-1}-(1-(1-P_t)^{k_{h-1}}))}{1-(1-P_t)^{k_{h-1}}}\right)} \quad (19)$$

In the computation of the asymptotic variance we considered fives stages in the analysis in order to make an all inclusive conclusion. Table 1 gives the computed asymptotic variance of the estimator hierarchically.

	p_t				
m	0.01	0.02	0.03	0.04	0.05
1	0.00015	0.00050	0.00130	0.00356	0.01326
2	0.00009	0.00028	0.00061	0.00127	0.00260
3	0.00008	0.00022	0.00045	0.00084	0.00148
4	0.00007	0.00019	0.00039	0.00069	0.00116
5	0.00007	0.00018	0.00036	0.00063	0.00103

Table 1: Simulated asymptotic variance for the i^{th} group for specified values of p_t

Table 1 illustrates the asymptotic variance of the i^{th} group at the h^{th} stage for a range of values of p_t of the constructed estimator were N = 640 individual units that were initially subdivided into ten groups at stage one each of size $k_1 = 64$ and, at each successive stage the groups were further split into half using the halving method and tested in parallel. From the table we observe that the loss in asymptotic variance with each successive stage depend on the prevalence rate of the traits. The computed results from this proposed design gives a generalization of the model proposed by (Hughes-Oliver and Rosenberger, 2000).

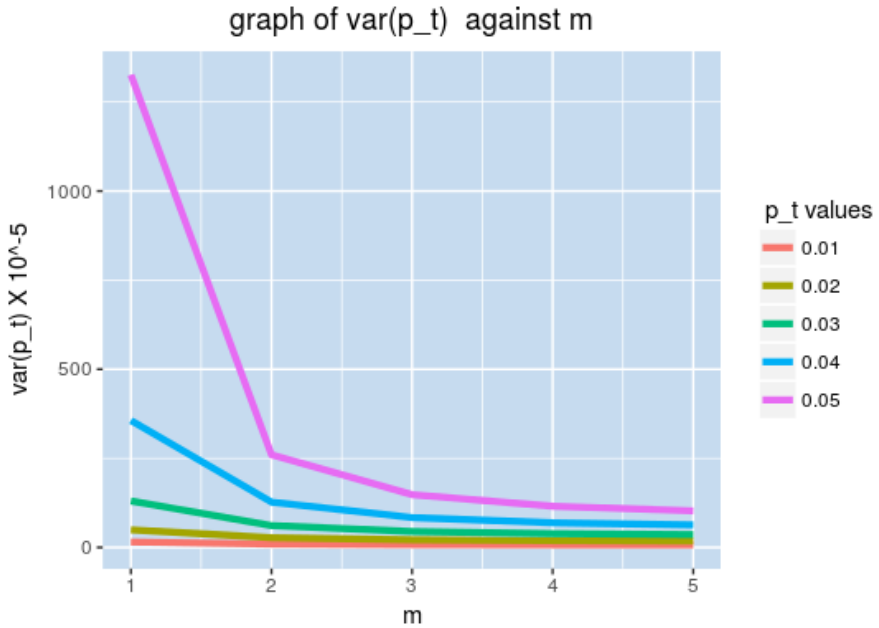

Figure 1. A plot of asymptotic variance for the i^{th} group versus m, the stages.

From the figure we observe that the asymptotic variance for the i^{th} group decreases with each additional stage. The graphs show that as $m \to \infty$ the accuracy of the estimator increases.

VII. Conclusion

In this paper we developed an M – stage hierarchical group testing procedure and constructed the estimator for estimating the occurrence of rare traits in a finite population. To assess the efficiency of our estimator, we computed the asymptotic variance of the i^{th} group at the h^{th} stage. From the computed results we observed that the asymptotic variance decreased with each additional stage. With this in mind, we conclude that the gains in efficiency of the constructed estimator depend on the number of stages involved in testing the presence of rare traits.

References

[1]. Billinggsley P. (1995). Probability and measure Third Edition. John Wiley and Sons, Inc.
[2]. Brookmeyer, R. (1999). Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* **55**, 608 – 612.
[3]. Dorfman, R. 1943. The detection of defective members of large population. *Annals of Mathematical Statistics*, 14, 436-440.
[4]. Hughes-Oliver M J., and Rosenberger F. W., (2000). Efficient estimation of multiple rare traits. *Biometrika*, **87**, 2, 315 - 327
[5]. Hughes-Oliver and Shallow. W.H (1994). A two-stage adaptive group design for group testing of only one trait. *American statistical association*, **89**, 982 – 993.
[6]. Monzon, O. T., Palalin, F. E., Dimaal, E., Balis, A. M., Samson, C., and Mitchel S. (1992). Relevance of antibody content and test formal in HIV testing for pooled sera. *AIDS* 6, 43 – 48.
[7]. Nyongesa. K.L (2004). Multistage group testing procedure (batch screening). Communication in statistics-simulation and computation, 33, 621-637.
[8]. Swallow W.H. 1985. Group testing for estimating infection rates and probability of disease transmission. *Phytopathology*, 75, 882-889. 18.
[9]. Tebbs M.J. and Swallow, H.W. 2003. Estimating ordered binomial proportions with the use of group testing. *Biometrika*, 90, 471-477.

Sirengo L. John, et. al. "An M – Stage Hierachical Group Testing Model for Estimating Multiple Traits in a Population." *IOSR Journal of Mathematics (IOSR-JM)*, 16(3), (2020): pp. 29-34.