Efficient Method for Estimating the Presence of a Viral Infectious Disease and How to Control its Rapid Spread in a Population

SIRENGO JOHN LUCA
Department of Mathematics, Kibabii University

Abstract- ***Lack of sufficient mass testing tools for viral infectious diseases such as Influenza, common cold, covid-19, mumps, Ebola and many others has contributed to the accelerated spread of the epidemics. Containment of this rapid viral spread requires an extensive testing of the affected individuals in any given society. Current tests for infectious diseases are administered on one – at – a – time basis. These tests are expensive and are limited due to the lack of resources and time. This paper provides a very simple and efficient testing strategy which can significantly reduce the total number of tests compared to individual tests of the entire population. It has been observed that multistage group testing strategy is more economical and efficient in detecting the presence of the virus in a mixed sample of specimens. Multistage method starts by choosing a group from the population to be tested, performing a test on the combined sample from the entire group, and progressively splitting the group further into subgroups. In this paper we develop an adaptive multistage group testing design with the view of studying the behavior of the efficiency of the estimator as the number of stages increases. The method of maximum likelihood estimator is discussed and for comparison with other established estimators, properties of the constructed estimator have also been discussed. From the results it is evident that the asymptotic relative efficiency increases with each additional stage. This proposed design is efficient and simple to be deployed in containment of any infectious disease.***

Indexed Terms- ***Group testing, multistage, asymptotic relative efficiency.***

I. INTRODUCTION

The rapid spread of manyviral infectious diseases such as Influenza, chickenpox, covid-19, mumps, Ebola and many othersin many developing countries such as Kenya urges the concerned authorities to take urgent measures in order to contain the disease or at least, to reduce its spread.

Even though a lot of research is currently being carried out towards a cure of this infectious diseases, to date, the most effective reasonable measure against its spread is the tracing and subsequent isolation of positive cases[9].For instance, at present, the standard tests for the detection of the corona virus is: nucleic acid amplification tests (NAAT), such as the quantitative transcription polymerase chain reaction (qt-PCR). These bio-chemical tests are based on samples from lower or upper respiratory tract of tested individuals [10]. The former is too delicate of an operation to be widely applicable and only visible for hospitalized patients. In the routine laboratory diagnosis, however, sampling the upper respiratory tract with nasal swabs is much more preferred. Therefore the demand for this type of viral testing, is drastically increasing in many health care systems, resulting in shortages of necessary materials to conduct the test [11].

As proposed by a large number of scholars, a primary way to make better use of the available capacities is to mix samples of different individuals before testing, and to first perform the test on these mixtures, the so called pools, as if it was only one sample. Group testing is the procedure of performing joint tests on mixture of samples from several individuals pooled and tested together using a single test [3], thereby requiring significantly fewer tests than the number of individuals to be tested. When a pool tests negative,

this is interpreted as a negative result for all pooled specimens.Only when a group result is positive do the samples need to be tested individually.

In this paper, we will demonstrate and systematically develop a simple testing procedure that can lead to a significant reduction in number of tests hence expanding the capacity of the available infrastructure when large numbers of individuals are to be tested.

II. GROUP TESTING

The testing of pooled samples of biological specimens for disease has a long history, beginning with [3] seminal work consisting of pooling into groups (e.g pooling blood, urine, nasal swabs) and test only those pooled samples first to detect or identify traces of virus. In the second stage, only individuals that belonged to positive groups are tested individually.This procedure is described diagrammatically below in Figure 1.

Groups

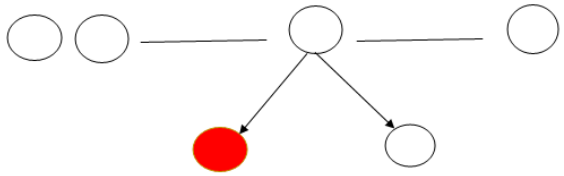

Figure 1: Dorfman (1943) Group testing procedure

Group testing conceals the identity of the subject, since individual units are pooled together hence the identity of the subject is not known in case of a fatal trait [4], thereby preventing stigmatization.

Because of limited information contained in positive response, it is required to test certain units multiple times – either in parallel for all the units or sequentially with additional testing for those units with positive test results. Sequential test designs in which pooling of samples into groups in each stage depend on the results of the former stages, are called adaptive. For non – adaptive methods, in contrast, all the sample poolings are specified in advance, which translates into a one stage design as shown in Figure 1 [3].A special class of adaptive test designs is a hierarchical test, where in the first stage the population is grouped into *'n'* homogeneous groups and each group is subjected to a single test, and in every subsequent stage, groups with positive results are split into smaller groups and retested, units contained in groups with negative results are discarded. This procedure is shown diagrammatically in Figure 2 as proposed by [3] which is an improvement on the original Dorfman test described in Figure 1.

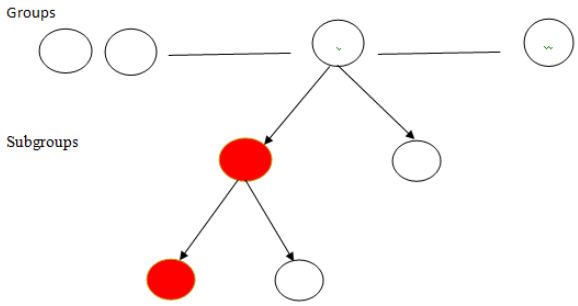

Figure 2: Monzon et. al. (1992) Group testing procedure

The proposed testing scheme is a generalization of this procedure.

III. THE MODEL

The population N understudy is assumed herein as sufficient for the experiment to be considered. Firstly, the population 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 virus. Positive results indicate the presence of at least one positive individual and the negative reading indicates the absence of the virus in all the individuals. 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 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 3 below.

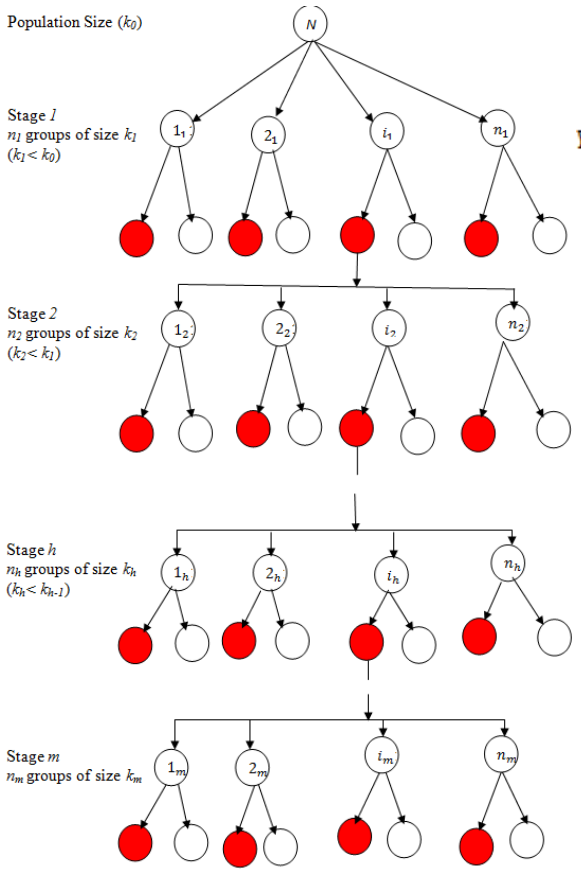

Figure 3:m-stage Group Testing Model.

First we consider the classification an i^{th}group in the h^{th}stage, since h will be allowed to vary from 1 to m as discussed above with the objective of constructing the probability of positive reading at this stage. Notice that $k_m \subseteq k_{m-1} \subseteq \cdots k_2 \subseteq k_1$, this forms a filtration therefore we shall employ the theory of Martingale [1].

The probability of classifying a j^{th}individual from an i^{th}pool in theh^{th} stage is obtained as follows: The j^{th} unit is subjected to testing for the presence of virus, the unit can test positive or negative.

Let

$$Y_{h.j} = \begin{cases} 1 & if\ the\ j^{th} individual\ tests\ positive\ of\ the\ virus \\ 0 & otherwise \end{cases}$$

for simplicity the random variable $Y_{h.j}$ gives binary results with probability of success

$1-(1-p)$(cf Dorfman, 1943). Thus it reduces to

$$P_r(Y_{h.j} = y_{h.j}) = (1-(1-p)^{Y_{h.j}})(1-p)^{1-Y_{h.j}}$$

. (1)

Now computing the probability of classifying thei^{th} group itself. This will be (1) for thei^{th} group scenario.

Let

$$Y_{hi.} = \begin{cases} 1 & if\ the\ i^{th} group\ tests\ positive\ of\ the\ virus \\ 0 & otherwise \end{cases}$$

Also we note that $Y_{hi.}$ is a Bernoulli random variable with probability of success

$1-(1-p)^{k_h}$ (cf Dorfman, 1943). Hence

$$\Pr(Y_{hi.} = y_{hi.}) = (1-(1-p)^{k_h.})^{Y_{hi.}}((1-p)^{k_h})^{1-Y_{hi.}}$$

(2)

The subgroups used at theh^{th}stage come from positive sub groups in stage $h-1$. The probability of interest that is the probability of classifying the i^{th} group as positive given that it comes from a positive sub-group in stage $h-1$ is

$$\Pr(Y_{hi.} = y_{hi.} \backslash Y_{h-1i.} = y_{h-1i.})$$

(3)

Reorganizing this conditional probability we have

$$\Pr(Y_{hi.} = y_{hi.} \backslash Y_{h-1i.} = y_{h-1i.}) = \frac{\Pr(Y_{hi}=y_{hi}, Y_{h-1i}=y_{h-1i.})}{\Pr(Y_{h-1i}=y_{h-1i})}$$

(4)

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

$\Pr(Y_{hi.} = y_{hi.} \backslash Y_{h-1i.} = y_{h-1i.}) = \Pr(Y_{hi.} = y_{hi.}) \backslash Pr(Y_{h-1i.} = y_{h-1i.})$

(5)

From Equation (5) we obtain the probabilities of interest at the h^{th} stage in the model.

We recall that the i^{th}group is positive if at least one of the units in the group is positive, hence

$$\Pr(Y_{hi.} = y_{hi.} | Y_{h-1i.} = y_{h-1i.}) = \frac{(1-(1-p)^{k_h})^{Y_{hi.}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i.}}}$$

(6)

Equations (6) is a truncated probability distribution model which is the probability of classifying an i^{th} group in the h^{th} as positive. This probability isvital the formulation of our model;

$$f(Y,p) = \left(\frac{(1-(1-p)^{k_h})^{Y_{hi.}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i.}}}\right)\left(\frac{(1-(1-p)^{k_h})^{Y_{hi.}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i.}}}\right)$$

(7)

- The Likelihood Function

The likelihood function at this stage is anchored on Equation (7). Thus utilizing the indicator function Y_{hi} as proposed above the likelihood function at the h^{th} stage will be

$$L_h(p) \propto \prod_{i=1}^{n_h} \left(\frac{(1-(1-p)^{k_h})^{Y_{hi}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i}}} \right) \left(\frac{(1-(1-p)^{k_h})^{Y_{hi}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i}}} \right) \quad (8)$$

(8) is a truncated Binomial probability density model. Notice that $h = 1, 2, \ldots, m$, in model (8) thus the M – stage likelihood function is

$$L_m(p) \propto \prod_{h=1}^{m} \cdot \prod_{i=1}^{n_h} \left(\frac{(1-(1-p)^{k_h})^{Y_{hi}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i}}} \right) \left(\frac{(1-(1-p)^{k_h})^{Y_{hi}}}{(1-(1-p)^{k_{h-1}})^{Y_{h-1i}}} \right) \quad (9)$$

Equation (9) holds with $(1-p)^{k_0} = 0$, this is true because at initial stage k_0 is equal to the entire population which is large and $(1-p)^{k_0} \to 0$ as $k_0 \to \infty$ where $k_0 = N$.

Upon setting $m = 1$ in (9) the model reduces to that of [4].

- The Estimator

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

$$\hat{p} = \underset{p}{argmin} \sum_{h=1}^{m} \sum_{i=1}^{n_h} (.)$$

We define

$$f(p) \; as \; \frac{\partial}{\partial p} \log lm(.) \quad (10)$$

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

$$p_{i+1} = p_i - \frac{f(p)}{f'(p)} \quad (11)$$

The iteration ceases if $|p_{i+1} - p_i| < \varepsilon$, for some arbitrary ε. The estimator $\hat{p}$ obtained in (11) is the estimate of p.

- Asymptotic Variance

For a very population say $N \to \infty$, the asymptotic variance of an estimator is obtained by use of the Cramer – Rao lower bound method [8]. Mathematically given as

$$Var(\hat{p}) = -\left[E\left(\frac{\partial^2}{\partial p^2} \log L_m(.) \right) \right]^{-1} \quad (12)$$

Upon utilizing (12) on (9) we get the asymptotic variance of the model as

$$Var(\hat{p}) = \frac{1}{\sum_{h=1}^{m} \sum_{i=1}^{n_h} \left(-\frac{k_h^2 (1-p)^{k_h^{-2}}}{1-(1-p)^{k_h}} + \frac{2K_{h-1}(1-p)^{k_h^{-2}-1}(k_{h-1}-(1-(1-p)^{k_{h-1}}))}{1-(1-p)^{k_{h-1}}} \right)} \quad (13)$$

This equation is vital in the simulation of the asymptotic variance.

- Confidence Interval

The confidence interval gives the limits within which a good estimator lies. We shall consider the closeness of $\hat{p}$, the unbiased estimator of p at the h^{th} stage. Hence without loss of generality we provide the confidence interval of the estimator, $\hat{p}$, as

$$\hat{p} \mp Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{p})} \quad (14)$$

Where $Z \sim Normal\ (0,1)$ and $\hat{p}$ and $Var(\hat{p})$ are by provided by the solutions to Equations (11) and (13) respectively, it follows from Equation (14) that

$$p \in \left[\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{p})} \; , \hat{p}_t - Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{p})} \right]$$ and by the law of central limit theorem we have

$$\sqrt{N} \frac{(\hat{p}-p)}{Var(\hat{p})} \xrightarrow{l} Norma(0,1)$$

IV. RESULTS

- Computation of Asymptotic Variance

In this section we illustrate the computation of the asymptotic variance using a population of size N = 640 individuals initially subdivided into ten groups composed of k_1 = 64 at stage one and, at each successive stage, the positive groups were split into two subgroups by using halving method and tested in parallel. In the computation of the asymptotic variance we utilized Equation (13) and developed R-codes. The codes were used to generate the asymptotic variance for various values of p. These values are shown in Table 1.

Stage	No. of individuals per group (k_h)	p_t					
		0.01	0.02	0.03	0.04	0.05	
1	64	15.534	49.583	130.480	356.27	326.3	
2	32	9.6142	27.550	61.484		127.26	260.12
3	16	7.9441	21.632	44.933		83.981	148.14
4	8	7.2808	19.341	38.850		69.446	115.61
5	4	6.9823	18.323	36.215		63.399	102.84

Table 1: Simulated asymptotic variance for selected values of p_t, where $var(\hat{p}_t)x10^{-6}$

Table 1 gives an illustration of the asymptotic variance of the constructed estimator. It is evident from the table that the asymptotic variance decreases with increase in the number of stages almost three-fold, for instance at stage one the asymptotic variance is $1.5534x10^{-5}$ and at stage five it is $6.9823x10^{-6}$ when $p = 0.01$. It can also be observed that the asymptotic variance increases with increase in the prevalence of the virus. This means that the precision of group testing models decreases with increase in viral prevalence.

Asymptotic Relative Efficiency (ARE)

$$ARE = \frac{1}{\sum_{i=1}^{n}\left(-\frac{k_h^2(1-p_t)^{k_h^{-2}}}{1-(1-p)^{k_h}}+\frac{2k_{h-1}(1-p_t)^{k_{h-1}^{-2}}(k_{h-1}-(1-(1-p)^{k_{h-1}})}{1-(1-p_t)^{k_{h-1}}}\right)} \Big/ \frac{1}{\sum_{h=1}^{m}\sum_{i=1}^{n_h}\left(-\frac{k_h^2(1-p)^{k_h^{-2}}}{1-(1-p_t)^{k_h}}+\frac{2k_{h-1}(1-p)^{k_{h-1}^{-2}}(k_{h-1}-(1-(1-p)^{k_{h-1}})}{1-(1-p_t)^{k_{h-1}}}\right)} \quad (15)$$

Clearly from Equation (15) we have

$$\sum_{h=1}^{m}\sum_{i=1}^{n_h}\left(-\frac{k_h^2(1-p)^{k_h^{-2}}}{1-(1-p_t)^{k_h}}+\frac{2k_{h-1}(1-p_t)^{k_{h-1}^{-2}}(k_{h-1}-(1-(1-p_t)^{k_{h-1}})}{1-(1-p_t)^{k_{h-1}}}\right) \geq \sum_{i=1}^{n}\left(-\frac{k_h^2(1-p_t)^{k_h^{-2}}}{1-(1-p_t)^{k_h}}+\frac{2K_{h-1}(1-p_t)^{k_{h-1}^{-2}}(k_{h-1}-(1-(1-p_t)^{k_{h-1}})}{1-(1-p_t)^{k_{h-1}}}\right)$$

implying ARE ≥ 1. Thus, theoretically our model outperforms [10] strictly for $m > 1$. Furthermore to illustrate the above implication/observation we have computed the asymptotic relative efficiency (ARE) for various values of m as provided in Table 2.

Stage	No. of individuals per group (k_h)	p_t					
		0.01	0.02	0.03	0.04	0.05	
1	64	1.0000	1.0000	1.0000	1.0000	1.0000	
2	32	1.6157	1.7997	2.1222		2.7995	5.0988
3	16	1.9554	2.2921	2.9039	4.2423	8.9530	
4	8	2.1336	2.5636	3.3586		5.1307	11.4722
5	4	2.2248	2.7061	3.6029		5.6195	12.8967

Table 2: AREs of $\hat{p}$, after successive stages.

Table 2 provides generated asymptotic relative efficiency values of the estimator at stages 1, 2, 3, 4 and 5 of an M-Stage pooling study for a range of values of *p*. From this table, we observe that the efficiency of the estimator increases with increase in the number of stages.

CONCLUSION

In this paper, we developed an M-Stage group testing procedure for testing the presence of the virus in a population and constructed a prevalence estimator based on multistage pooling algorithm. The maximum likelihood estimator (MLE) procedure was used in developing the estimator $\hat{p}$. The properties of the maximum likelihood estimator such as the asymptoticvariance and relative efficiency are provided in the discussion.

From our discussion in the previous section, it is clear that the accuracy of the estimator in an M-Stage group testing scheme increases with each additional stage when the prevalence of the virus in the population is rare.

As highlighted in section 4 the asymptotic relative efficiency values are all greater than one for m> 1 showing evidently that the M-stage estimator outperforms the already established estimator by [2, 5]. Results from this section shows that the constructed estimator gains efficiency with each additional stage. This therefore makes the M-Stage adaptive testing scheme more ideal than the two stage testing scheme in estimating presence of the virus in the population.

In summary, the results for the study can be generalized as follows; as the number of stages increases the proposed testing scheme becomes more precise and efficient than the already established scheme with small values of *p*. It also shows how each test can be amplified by applying it to a mixture of samples (blood, urine, nasal swabs etc.) from several individuals thus leading to a significant reduction in the number of tests in estimating the presence of a viral infection in a population.

RECOMMENDATIONS

In view of the foregoing, it would therefore be imperative to recommend the adoption of the M-Stage group testing model in estimating the prevalence of infectious viral diseases in order to significantly reduce the number of tests so that the available facilities are not overstretched and control its rapid spread.

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. Biometrics55, 608 – 612.

[3] Dorfman, R. 1943. The detection of defective members of large population. Annals of Mathematical Statistics, 14, 436-440.

[4] Gastwirth. J. L and Hammick. P.A (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subject by group testing; application estimating the prevalence of Aids antibodies in blood donors. Journal of statistical planning and inference. 22, 15 27.

[5] Hughes-Oliver M J., and Rosenberger F. W., (2000). Efficient estimation of multiple rare traits. Biometrika, 87, 2, 315 - 327

[6] 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.

[7] Nyongesa. K.L (2004). Multistage group testing procedure (batch screening). Communication in statistics-simulation and computation, 33, 621-637.

[8] Tebbs M.J. and Swallow, H.W. 2003. Estimating ordered binomial proportions with the use of group testing. Biometrika, 90, 471-477.

[9] "Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve", The Washington post". https://wapo.st/2wLMbzI. Accessed: 2020-04-28.

[10] C. Fraser, S. Riley, and N. M. Ferguson (2004). "Factors that make an infectious disease outbreak controllable," Proceedings of the National Academy of Sciences, vol. 101, no. 16, pp. 6146–6151.

[11] "Why widespread coronavirus testing isn't coming anytime soon, The New Yorker." https://bit.ly/3dCAHz9. Accessed: 2020-04-28.