IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 15, Issue 3 Ser. III (May – June 2019), PP 29-37
www.iosrjournals.org

Two-Stage Adaptive Pool Testing with Errors in Inspection

Okoth Annette W
Department of Mathematics Masinde Muliro University of Science and Technology
Corresponding Author: Okoth Annette W

Abstract: *In this study we present a two-stage adaptive estimator p^A of prevalence in the presence of test errors. We assume that tests are not 100 % perfect. We obtain the adaptive estimator using Maximum Likelihood Estimate (MLE) method and use Fisher information to determine the variance of the estimator. We use Matlab, for simulation and veri cation of the model. We analyse and discuss the properties of the constructed estimator in comparison with other existing estimators in the literature of pool testing. We also provide the con dence interval of the estimator. When the test kits have low sensitivity and speci city, we establish that the adaptive estimator outperforms other existing estimators. Further more, we demonstrate that the e ciency of the adaptive estimation scheme improves as the number of stages increases. This makes the adaptive testing scheme more ideal in areas where errors are rampant.*

Key words: *Pooling, Prevalence, Test errors, Likelihood.*

Date of Submission: 21-05-2019 Date of acceptance: 06-06-2019

I. Introduction

Pool testing has been recognized as a sampling scheme that can provide substantial benefits (cf. Hughes-Oliver and Swallow, 1994). Early application of pool testing in estimating prevalence of plant virus transmission by insects (Watson, 1936; Thompson, 1962) was one of the pioneering applications of this concept. Dorfman (1943) introduced the statistical and mathematical concept of pool testing when he used it to estimate the proportion of diseased individuals among the US conscripts. He also derived optimum group sizes assuming that the population was large enough for application of a binomial model and consequently realized significant savings amounting up to 80% in the numbers of tests required. He developed a statistical model as follows: Thompson (1962) discussed estimation in statistics using a pool testing procedure. In the subsequent years this concept has had relevant application in various clinical studies including phytopathology, public health and plant quarantine (cf. Chiang and Reeves, 1962; Bhattacharyya et al., 1979; Swallow, 1987; Yamamura and Sugimoto, 1995; Hughes and Gottwald, 1998; Zenios and Wein, 1998; Remund et al.,2001;). Positively pooled samples can be partitioned into relatively smaller subsets there by reducing on cost and effort, which provides the obvious motive for pooling samples together (Sobel and Elashorff, 1975). On the same subject, Nyongesa (2011) developed an estimation model based on pool testing with retesting the pools that test negative and the model has been shown to be applicable for blood donors. Theobald and Davie (2007) argued that pool

testing need not only be applied to populations where retesting is needed, like in identification of diseased individuals in a human population, but also on other populations with no intention of retesting the individuals contributing to positive pooled samples. For instance if a batch of food items is being tested for contamination, there may be no interest in identifying the particular items which are affected. The emphasis may instead be on estimating the proportion of defective items in a population or on deciding that the number of positive pooled samples justifies removing a food product from the market. In another related study, bacteriological testing of egg-laying hens for salmonella in Great Britain was carried out using organ cultures pooled five at a time. Individual samples contributing to positive pooled samples are not tested again. A population comprises birds in a hen house. If infection was confirmed they were destroyed and compensation paid for the number of birds estimated to be uninfected (Richards, 1991). Oliver-Hughes and Swallow (1994) developed a two-stage adaptive pool testing procedure with perfect tests.The idea here was to estimate small proportions in a population with ideal tests. They used the Maximum Likelihood Estimate (MLE) method to estimate the proportion and Cramer-Rao Lower bound method for determining the variance of the estimator. They realised impressive results with the two-stage adaptive estimator $\hat{p}_A$ being more efficient than the non-adaptive estimator $\hat{p}$. In their study however, they assumed that tests are perfect, yet in real life situations, errors in experiments are inevitable. In real life situations manufactured test kits are never 100% perfect as assumed by Oliver-Hughes and Swallow (1994). In this study, we present a two-stage adaptive pool testing model with imperfect tests that is applicable to real life situations, hence generalizing the Oliver-Hughes and Swallow (1994) model.

II. Adaptive Scheme

Here we obtain a two-stage adaptive estimator $\hat{p}A$ of prevalence of a trait in the presence of test errors. That is we compute the maximum likelihood estimator (MLE) of prevalence and investigate its properties.

The adaptive model involves testing groups in stages and updating group sizes from one stage to the next, with group size at a stage depending on the group size of the preceeding stage(s). That is, testing n1 groups, each of size k1 in the rst stage; n2 groups each of size k2 in the second stage; n3 groups each of size k3 in the third stage; for a three- stage model, and so on; where, k3 depends on both k2 and k1 , while for a two-stage model, k2 depends on k1 . For a general adaptive scheme, at stage i , ni groups each of size ki where the ki depends on ki 1; ki 2, ….. k1 are constructed. The constructed groups are then subjected to testing. These ni groups at this stage are all of equal sizes, ki . The ni is determined before the experiment is carried out while ki 's are sequentially determined as the experiment progresses.

First we introduce the Non-adaptive testing scheme with errors as it will be the basis of our subsequent discussions. Suppose we have a population with the purpose of characterising it into two distinct groups, that is, defective and non-defective. For clarity of this procedure consider a population of size N . Divide this population into n homogenous groups as shown in Figure 1

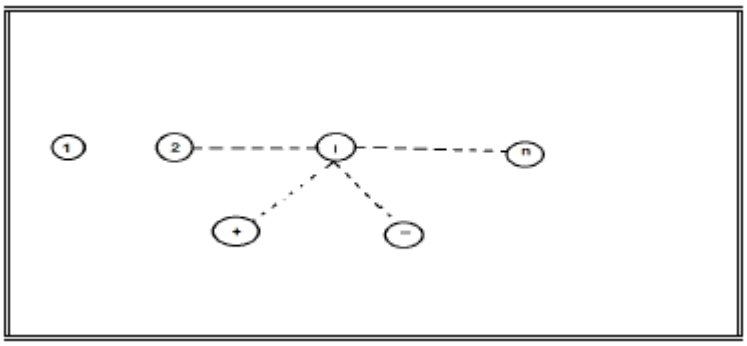

Figure 1: Construction of groups with the purpose of testing.

Each constructed group is subjected to testing as shown in Figure 1. Notice that the test kits in practice have errors (cf. Kline et al., 1989). From Figure 1 if a group is tested, it either yields positive or negative results. We also observe from Figure 1 that there are n groups to be tested, that is, i = 1; 2; 3,……, n . Suppose X out of n groups test positive then X has a binomial distribution simply written as

$$X \sim Binomial(n, \pi(p)). \tag{1}$$

Some authors have used Equation (??)to obtain the estimator p^ of prevalence p, for the non-adaptive scheme with test errors, for instance see Brookmeyer (1999) and Nyongesa (2011) as

$$\hat{p} = 1 - \left[\frac{\eta - \frac{X}{n}}{\eta + \phi - 1}\right]^{\frac{1}{k}}, \tag{2}$$

And the asymptotic variance of Equation (??) can easily be obtained from Equation (??) upon applying Cramer-Rao lower bound method, see for instance (Gupta and Kapoor,1978, p766-770) as

$$Var(\hat{p}) = \frac{(1-p)^2\pi(p)(1-\pi(p))(1-p)^{-2k}}{nk^2(\eta+\phi-1)^2} \tag{3}$$

Equations (??) and (??) will be vital in the development of the succeeding sections.

2.1 Two-stage adaptive scheme

For this Scheme, we set $n_1 = \lambda n$ and $n_2 = (1-\lambda)n$, where n is the total number of groups constructed initially, n_1 the number of groups tested at stage-one and n_2 the number of groups tested at stage two. The group size at stage one, k_1 is determined by

$$k_1 = argmin_l[Var(\hat{p})]|_{p=p_0}. \tag{4}$$

Suppose X_1 groups test positive on the test at stage-one follows a Binomial Distribution written as

$$X_1 \sim Binomial(\lambda n, \pi(p)|_{k=k_1}). \tag{5}$$

The likelihood form of Equation (??) is given by

$$L(p|\lambda n, X_1) \propto \pi^{x_1}(p)[1-\pi(p)]^{\lambda n - x_1}. \tag{6}$$

The MLE of prevalence at stage one is obtained from Equation (??) as

$$\hat{p}_1 = 1 - \left[\frac{\eta - \frac{X_1}{\lambda n}}{\eta + \phi - 1}\right]^{\frac{1}{k_1}}. \tag{7}$$

2.2 Properties of the prevalence estimator in stage-one

In this sub-section, we investigate one key property of the estimator as given by Equation (??). this property is Unbiasedness.

We wish to establish whether or not $\hat{p}_1$ is unbiased. To this end we have

$$E(\hat{p}_1) = 1 - E\left[\frac{\eta - \frac{X_1}{\lambda n}}{\eta + \phi - 1}\right]^{\frac{1}{k_1}}. \tag{8}$$

It is not easy to simplify this equation, therefore we consider special cases. For example if $k_1 = 1$, we have

$$E(\hat{p}_1) = 1 - \left[\frac{\eta - \frac{E(X_1)}{\lambda n}}{\eta + \phi - 1}\right], \tag{9}$$

It is clear from Equation (??) that $E(X_1) = \lambda n \pi(p)$. Upon substituting $\lambda n \pi(p)$ in Equation (??), utilizing Equation (??) and noting that $k_1 = 1$ we have

$$E(\hat{p}_1) = p. \tag{10}$$

Therefore for $k_1 = 1$, $\hat{p}_1$ is an unbiased estimator of p. However, for $k_1 > 1$ it is not easy to solve Equation (??). We shall therefore apply Jensen's inequality (Mood et al., 1974, p72) since Equation (??) is convex on $\Re$ and $E(X_1)$ is bound, that is,

$$E(X_1) = \lambda n \pi(p) < \infty.$$

Applying Jensen's inequality in this case we have,

$$\begin{aligned} E\left[1 - [\frac{\eta - \frac{X_1}{\lambda n}}{\eta + \phi - 1}]^{\frac{1}{k_1}}\right] &\leq 1 - \left[\frac{\eta - \frac{E(X_1)}{\lambda n}}{\eta + \phi - 1}\right]^{\frac{1}{k_1}} \\ \implies E(\hat{p}_1) &\geq 1 - \left[\frac{\eta - \frac{E(X_1)}{\lambda n}}{\eta + \phi - 1}\right]^{\frac{1}{k_1}}. \end{aligned} \tag{11}$$

Noting that $E(X_1) = \lambda n \pi(p)$, Equation (??) becomes

$$E(\hat{p}_1) \geq 1 - \left[\frac{\eta - \pi(p)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}}. \tag{12}$$

Equation (??) can be expanded using Taylor's series about $E(X_1)$ as follows:
Let $\hat{p}_1 = f(X_1)$
Then $\hat{p}_1$ expressed in Taylor's expansion about $E(X_1)$ is

$$\begin{aligned}\hat{p}_1 &= f[E(X_1)] + f'[E(X_1)][X_1 - E(X_1)] + \frac{1}{2}f''[E(X_1)][X_1 - E(X_1)]^2 \\ &+\frac{1}{6}f'''[E(X_1)][X_1 - E(X_1)]^3 + ... + \frac{1}{(\lambda n)!}f^{(\lambda n)}[E(X_1)][X_1 - E(X_1)]^{(\lambda n)},\end{aligned}$$

where $f'[E(X_1)], f''[E(X_1)]$ and $f'''[E(X_1)]$ are first, second and third derivatives respectively of $f[E(X_1)]$ with respect to X_1.

Determining the terms of Taylor's expansion, combining them and taking expectation we get

$$\begin{aligned}E(\hat{p}_1) \geq\ & 1 - (1-p)\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}} + 0 \\ &+ \left(\frac{k_1 - 1}{2k_1^2(\lambda n)^2}\right)\frac{(1-p)}{(1-p)^{2k_1}}\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}-2} Var(X_1) \\ &+ \left(\frac{(k_1 - 1)(2k_1 - 1)}{12k_1^3(\lambda n)^3}\right)\frac{(1-p)}{(1-p)^{3k_1}}\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}-3} \mu_3 + \ldots,\end{aligned}$$

where $\mu_3 = [X_1 - E(X_1)]^3$ is called the third cummulant. Notice from the expansion above that when $k_1 = 1$ and $\eta = \phi = 1$, $E(\hat{p}_1) \geq p$ implying that $\hat{p}_1$ is either an unbiased estimator of p or upwardly biased. However, when $k_1 > 1$,

$$\begin{aligned}E(\hat{p}_1) \geq\ & p\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}} + \left(\frac{k_1 - 1}{2k_1^2(\lambda n)^2}\right)\frac{(1-p)}{(1-p)^{2k_1}}\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}-2} Var(X_1) \\ &+ \left(\frac{(k_1 - 1)(2k_1 - 1)}{12k_1^3(\lambda n)^3}\right)\frac{(1-p)}{(1-p)^{3k_1}}\left[\frac{\eta - (\phi - 1)}{\eta + \phi - 1}\right]^{\frac{1}{k_1}-3} \mu_3 + \ldots,\end{aligned}$$

which implies that $\hat{p}_1$ is upwardly biased.

From the discussion of the properties of our adaptive estimator at stage-one we observe that the properties are similar to those of the non-adaptive estimator with the difference being the parameter (λ). Therefore, for further analysis of the properties of this adaptive estimator just like for the non-adaptive estimator, see Nyongesa (2011) and Brookmeyer (1999). The result in (??) is useful in the construction of k_2 as it will be seen in the following sub-section.Construction of k_2 enables us to estimate the adaptive estimator, $\hat{p}_A$ in stage two.

2.3 Adaptive Estimator at Stage two, $\hat{p}A$

At this stage, as in stage-one, the group size, k_2 is determined by

$$k_2 = argmin_l[Var(\hat{p}_1)]|_{p_1=p}, \tag{13}$$

where k_2 can be obtained by solving the equation

$$\frac{d}{dl}\left(f_{\hat{p}_1(X_1)}(l)\right) = 0,$$

where, k_2 is a function of (n, λ, X_1) or simply written as $k_2 = k_2(n, \lambda, X_1)$.

The $n_2 = (1-\lambda)n$ groups each of size k_2 are tested at stage two. Suppose X_2 groups test positive on the test, then

$$X_2 = 0, 1, 2, ..., (1-\lambda)n$$

Therefore, for fixed X_1 we have

$$X_2|X_1 \sim Binomial((1-\lambda)n, \pi_{2|1}(p)) \tag{14}$$

where

$$\pi_{2|1}(p) = \eta[1-(1-p)^{k_2}] + (1-\phi)(1-p)^{k_2} \tag{15}$$

The joint distribution of X_1 and X_2 is given by

$$f(X_1, X_2) = Binomial(\lambda n, \pi(p)|_{k=k_1}) * Binomial((1-\lambda)n, \pi_{2|1}(p)). \tag{16}$$

The likelihood function of Equation (??) is

$$L(p|n, X_1, X_2) \quad \propto \quad (\pi_1(p))^{X_1}(1-\pi_1(p))^{\lambda n - X_1}$$

$$(\pi_{2|1}(p))^{X_2}(1-\pi_{2|1}(p))^{(1-\lambda)n-X_2}. \tag{17}$$

Notice that X_2 depends on X_1 through the size k_2, that is $k_2 = k_2(X_1)$. Indeed this assumption simplifies Equation (??) as $\pi_{2|1(p)}$ takes the form

$$\begin{aligned} \pi_{2|1}(p) &= \eta[1-(1-p)^{k_2(X_1)}] + (1-\phi)(1-p)^{k_2(X_1)} \\ &= \eta[1-q^{k_2(X_1)}] + (1-\phi)q^{k_2(X_1)} \end{aligned}$$

Utilizing Equation (??) the MLE of p, that is p_A is obtained as the solution to

$$\begin{aligned} &\frac{k_1 X_1 q^{k_1}[(1-\phi)-\eta]}{\eta-(\eta+(1-\phi))q^{k_1}} + \frac{k_2(X_1)X_2 q^{k_2(X_1)}[(1-\phi)-\eta]}{\eta-(\eta+(1-\phi))q^{k_2(X_1)}} \\ = \quad &\frac{k_1 q^{k_1}(\lambda n - X_1)(\eta+(1-\phi))}{1-[\eta-(\eta+(1-\phi))q^{k_1}]} + \frac{k_2(X_1)q^{k_2(X_1)}[(1-\lambda)n-X_2][\eta+(1-\phi)]}{1-[\eta-(\eta+(1-\phi))q^{k_2(X_1)}]}. \end{aligned} \tag{18}$$

III. Asymptotic Variance of $\hat{p}A$

This is done by employing Cramer-Rao Lower bound method. V ar($\hat{p}A$) is obtained as

$$Var(\hat{p}_A) = \frac{\pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))}{(\eta+\phi-1)^2 n[\pi_2(p)(1-\pi_2(p))\lambda k_1^2(1-p)^{2k_1} + \pi_1(p)(1-\pi_1(p))(1-\lambda)k_2^2(X_1)(1-p)^{2k_2(X_1)}} \quad (1$$

IV. Results and discussion

Here we highlight our findings in this study. We estimated prevalence using the adaptive testing scheme. We accomplished this by employing MLE procedure. For us to recommend the suitability of the adaptive estimator, it would be in order to compare with the past constructed estimators in the literature of pool testing and in particular in the presence of test errors. In statistical inference, estimators with small bias are considered to be the best estimators and since in this study we did not discuss the biasness of our two-stage adaptive estimator, we shall not use bias as a measure to compare previous estimators with our adaptive estimator and similarly with mean squared error (MSE). Therefore the only measure of comparison herein is the computation of Asymptotic Relative Efficiency (ARE) values. The ARE in this study is obtained by dividing (??) by (??), that is,

where V ar $\hat{p}A$) and V ar($\hat{p}A$) are as given in (??) and (??) respectively, which upon simpli cation reduces to

$$\text{ARE}_{\hat{p}A} = \frac{\pi(p)(1-\pi(p))\left[\pi_2(p)(1-\pi_2(p))\lambda k_1^2(1-p)^{2k_1} + \pi_1(p)(1-\pi_1(p))(1-\lambda)k_2^2(X_1)(1-p)^{2k_2(X_1)}\right]}{\pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))(1-p)^{2k}}, \quad (20)$$

with k1 and k2 as de_ned by (??) and (??) respectively. Using this Equation andmatlab software Table 1 was generated.

p	$\eta=\phi=0.99$	$\eta=\phi=0.98$	$\eta=\phi=0.97$	$\eta=\phi=0.96$	$\eta=\phi=0.90$
0.1	11.7739	11.8024	11.8304	11.8577	12.0095
0.2	6.1184	6.4223	6.7152	6.9974	8.4904
0.3	3.6109	4.2835	4.8918	5.4448	7.9386
0.4	3.0823	4.3658	5.3725	6.1880	9.0662
0.5	4.4239	6.3134	7.4865	8.3030	10.5715
0.6	7.3875	9.0611	9.9042	10.4289	11.6667
0.7	10.2335	11.1440	11.5420	11.7672	12.2284
0.8	11.9115	12.1858	12.2856	12.3372	12.4329
0.9	12.4422	12.4616	12.4681	12.4713	12.4772

Table 1: ARE values of $\hat{p}_A$ relative to $\hat{p}$ for specified p, η and ϕ

Table 1 provides generated ARE values for given p, η and ϕ. From the tabulated values ARE is high when p is small and decreases as p increases, attaining the minimum at $p = 0.4$ across the board, except for $\eta = \phi = 0.9$ where the minimum is attained at $p = 0.3$. The ARE again improves as p moves away from 0.4 for $\eta, \phi > 0.9$. A similar scenario is observed for the case of $\eta = \phi = 0.9$ where ARE improves as p moves away from 0.3. To depict these observations graphically, see Figure 2.

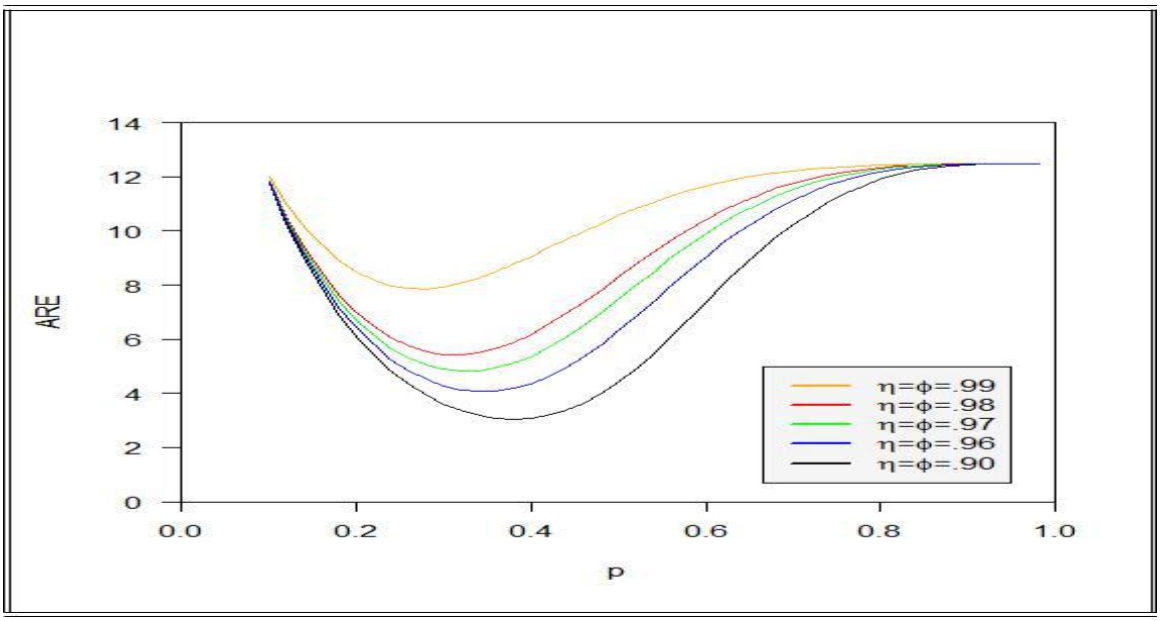

Figure 2: ARE vs probability p .

The figure represents ARE values plotted against prevalence p values. Clearly, as noted in Table 2, the ARE drops significantly as p increases up to the value $p = 0.4$, then it improves as p moves away from 0.4. From Figure 2, the adaptive estimator outperforms the non-adaptive estimator as the sensitivity and specificity of the test kit decreases. Hence in cases where the test kits have low sensitivity and specificity, the adaptive scheme is preferred for more efficient results.

References

[1] Bhattacharyya, G.K., Karandinos, M.G., and DeFoliart, G. R. (1979). Point estimates and con dence intervals for infection rates using pooled organisms in epidemiologic studies. American journal of epidemiology 109,124-131.

[2] Brookmeyer, R. (1999). Analysis of multistage pooling studies of biological spec-imens for estimating disease incidence and prevalence. Biometrics 55,608-612.

[3] Chiang, C. L. and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. American journal on hygiene 75, 377-391

[4] Dorfman R. (1943). The detection of defective members of large populations. Annals of mathematical statistics 14, 436-440.

[5] Hughes-Oliver J.M. and Swallow W.H. (1994), A two-stage adaptive group test-ing procedure for estimating small proportions. American statistical association 89, 982-993.

[6] Hughes, G., and Gottwald, T. R. (1998). Survey methods for assessment of citrus tristeza virus incidence. Phytopathology 88, 715-723.

[7] Kline,R.L.,Brothers,T., Bookmeyer,R.,Teger,S., and Qinn,T.(1989). Evaluation of human immunode ciency virus seroprevalence in population surveys using pooled sera. Journal of clinical microbiology 27, 1449-1452.

[8] Nyongesa, L. K. (2011). Dual estimation of prevalence and disease incidence in pool testing strategy. Communication in statistics theory and methods, 40, 1-12

[9] Remund, K. M., Dixon, D. A., Wright, D. l., and Holden, L. R. (2001). Sta-tistical considerations in seed purity testing for transgenic traits. Seed science Resources. 11, 101-120.

[10] Richards, M. S. (1991). Interpretation of the results of bacteriological testing of egg laying ocks for salmonella enteritidis. Proceedings of the 6th interna-tional symposium on veterinary epidemiology and economics, 124-126, Ottawa, Canada.

[11] Sobel M., Ellasho R.M. (1975).Group testing with a new goal, estimation. Biometrika 62, 181-193.
[12] Swallow W.H. (1987).Relative MSE and cost considerations in choosing group size for group testing to estimate infection rates and probability of disease in-fection. Phytopathology 77, 1376-1381.
[13] Theobald, C. M. and Davie, A.M. (2007). Group testing, the pooled hypergeo-metric distribution and estimating numbers of defectives in small proportions.
[14] Thompson K.H. (1962).Estimation of the proportion of vectors in a natural population of insects. Biometrics 18, 568-578.
[15] Watson, M. A. (1936).Factors a ecting the amount of infection obtained by aphis transmission of the virus Hy. 111. Philos. Trans. R. Soc. Lond., Ser. B 226,457-489.
[16] Yamamura, K., and Sugimoto, T. (1995).Estimation of the pest prevention ability of the important plant quarantine in Japan. Biometrics 51,482-490
[17] Zenios, S. A., and Wein, L. M. (1998).Pooled testing for HIV prevalence esti-mation; exploring the dilution e ect. Statistical Medicine 17, 1447-1467.

Okoth Annette W. " Two-Stage Adaptive Pool Testing with Errors in Inspection." IOSR Journal of Mathematics (IOSR-JM) 15.3 (2019): 29-37.