IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 15, Issue 5 Ser. IV (sep – Oct 2019), PP 19-27
www.iosrjournals.org

Three-Stage Adaptive Batch Testing Model for Estimating Prevalence of a Trait Without Truncation

Oduori Victor[1], Okoth Annette W. [2], Wanyonyi Ronald W.[3]
[1;2](Department of Mathematics, Masinde Muliro University of Science and Technology ,Kakamega Kenya)
[3](Department of Mathematics, Egerton University, Nakuru Kenya)

Abstract: *Batch testing is a fundamental testing scheme that results into substantial saving in terms of cost and time. It's mainly applicable in cases with large population sizes and low prevalence rates. Studies on Batch testing have shown that Adaptive Batch testing is more efficient than Non-Adaptive Batch testing particularly as the number of stages increases. Most recent studies on Batch testing have shown that even with Truncation in inspection, Adaptive Batch testing remains more efficient. This study presents a Three-Stage Adaptive Batch testing Model with errors without Truncation with the view to establishing whether or not it's more efficient than the truncated estimator.*

The study made use of Maximum Likelihood Estimate (MLE) method to obtain the Estimator and Crammer-Rao Lower Bound method to determine the variance of the estimator .The efficiency of the Estimator relative to the Adaptive estimator with truncation determined with the view to performing a comparative analysis between the two. R-statistical software was used for Model verification. This study is significant in the sense that it brings forth a new model in the literature of estimation in batch testing, a Model that would find application in various fields including HIV/AIDS , Blood donation, quality control e.t.c.

Keywords: *Batch testing, Truncation, Sensitivity, Specificity, Efficiency*

Date of Submission: 07-10-2019 Date of Acceptance: 22-10-2019

I. Introduction

Batch testing is a fundamental tool used in identifying defective items from a large population with low prevalence rates. It's designed to reduce the number of tests required to identify defective items, thus less time consuming and less expensive. In this scheme items from a given population are pooled and tested as a single entity [1]. Batch testing has been known to reduce the variance of the Estimator thereby making it more efficient, [2], [3].

There are two forms of Batch testing namely: Non-Adaptive and Adaptive. In Non-Adaptive Batch testing scheme, a large population is divided into 'n' batches which are subjected to testing, [1]. The results obtained are then used to construct the Non-Adaptive Estimator. Adaptive Batch testing scheme on the other hand involves partitioning the population into 'n' Batches depending on the number of stages using predetermined partitioning parameters. The Batches are then tested in stages and the results obtained are used to construct the Adaptive model, [4].

II. Literature Review

Batch testing has been widely used as a sampling scheme that is very instrumental in reducing the time and cost of testing especially when the items of interest are rare, [1], [2] and [5]. The scheme has had a rich history, dating back to Dorfman [1] and his work during the world war (II) when they estimated the proportion of diseased individuals among the US soldier.

2.1 Non-Adaptive Batch testing

Dorfman, [1] introduced the statistical and mathematical concepts of batch testing. He used it to estimate the proportion of diseased individuals among the US soldiers by dividing a large population of size say, N into n Batches for the purpose of testing a rare trait and realized substantive savings.

Thompson advanced Dorfman's work and realized impressive results.

Recent studies have seen a large number of scholars engage in Batch testing in different fields such as estimation of HIV/AID's prevalence without necessarily identifying the subject [6], quality control process [5] and phytopathology, [7].

Nyongesa, [9] introduced the idea of error terms thereby altering the Thompson model [3]. He obtained the MLE $\hat{p}$, of prevalence p as

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

And established that the model was more efficient in situations where test kits had low sensitivity and specificity.

2.2 Adaptive Batch testing Scheme

Oliver-Hughes and Swallow, [2] proposed a Two-Stage Adaptive Batch testing model for estimating small proportions. They used Maximum Likelihood Estimation (MLE) method to estimate the proportion and Crammer-Rao Lower Bound method to determine the variance of the estimator. They divided the population in two Batches, that is to sayλn Batches tested at stage one and(1- λ)n Batches tested at stage two; where$\lambda$is a partitioning parameter used to partition the Batches.

The MLE's at stages one$\hat{p}_1$ and two$\hat{p}_2$were established and the results obtained were more impressive compared to the Non-Adaptive estimator $\hat{p}$ because the Adaptive estimator$\hat{p}_2$ at stage two was found to be more efficient than the Non-Adaptive estimator. They further established that, efficiency increases as the number of stages increases from one to two.

Okoth, [11] generalized the Oliver-Hughes and Swallow model by introducing error terms. With introduction of the error terms the probabilities that a batch tests positive at stage one and two were established as,

$$\pi_1(p) = \eta(1 - (1-p)^{k_1} + (1-\phi)(1-p)^{k_1} \tag{2}$$

and

$$\pi_2(p) = \eta(1 - (1-p)^{k_2} + (1-\phi)(1-p)^{k_2} \tag{3}$$

Respectively. Utilizing these probabilities the estimator at stage two was established as the solution to

$$\begin{aligned} & \frac{x_1 k_1 q^{k_1}[(1-\phi) - \eta]}{\eta - (\eta + (1-\phi))q^{k_1}} \\ + & \frac{x_2 k_1(x_1) q^{k_2(x_1)}[(1-\phi) - \eta]}{\eta - (\eta + (1-\phi))q^{k_2(x_1)}} \\ = & \frac{(n\lambda_1 - x_1)k_1 q^{k_1}[(1-\phi) + \eta]}{1 - [\eta - (\eta + (1-\phi))q^{k_1}]} \\ + & \frac{(n\lambda_2 - x_2)k_2 q^{k_1(x_1)}[(1-\phi) + \eta]}{1 - [\eta - (\eta + (1-\phi))q^{k_2(x_1)}]} \end{aligned} \tag{4}$$

and it's variance as , (Var($\hat{p}_A$)),

$$Var(\hat{p_A}) = \frac{\pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))}{R} \tag{5}$$

where,

$$\begin{aligned} R &= (\eta + \phi - 1)^2[\pi_2(p)(1-\pi_2(p))\lambda N k_1^2(1-p)^{2k_1-2} \\ &+ \pi_1(p)(1-\pi_1(p))\lambda n k_2^2(x_1)(1-p)^{2k_1-2}] \end{aligned} \tag{6}$$

2.3 Truncated Models

Okoth et.al [4] did Adaptive Pool Testing at a Multi-stage level and introduced Truncation of models. They established that the adaptive testing scheme is more efficient than the non-adaptive testing scheme with truncation incorporated.

One notable model in their work was the Estimator at stage-three where MLE was found as the solution to

$$\begin{aligned} & \frac{k_1 X_1(1-pk_1+p)}{\pi_1} \\ + & \frac{k_2(X_1)x_2(1-pk_2(X_1)+p)}{\pi_2} \\ + & \frac{k_3(X_2)X_3(1-pk_2(X_2)+p)}{\pi_3} \\ = & \frac{(\lambda_1 n - X_1)k_1(1-pk_1+p)}{(1-\pi_1)} \\ + & \frac{(\lambda_2 n - X_2)k_2(X_1)(1-pk_2(X_1)+p)}{(1-\pi_2)} \\ + & \frac{((1-\lambda_1-\lambda_2)n - X_3)k_3(X_2)(1-pk_3(X_2)+p)}{(1-\pi_3)} \end{aligned} \tag{7}$$

and whose variance was found as

$$Var(\hat{p}_3) = \frac{\pi_1(p)\pi_2(p)\pi_3(p)(1-\pi_2(p))(1-\pi_2(p))(1-\pi_3(p))}{B} \tag{8}$$

Where

$$\begin{aligned} B &= (\eta+\phi-1)^2[\lambda_1 n\pi_2(p)\pi_3(p)(1-\pi_2(p))1-\pi_3(p)(k_1-pk_1^2+pk_1)^2 \\ &+ \lambda_2 n\pi_1(p)\pi_3(p)(1-\pi_1(p))(1-\pi_3(p))(k_2(x_1)-pk_2^2(x_2)+pk_2(x_1))^2 \\ &+ (1-\lambda_1-\lambda_2)n\pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))(k_3(x_1)-pk_3^2(x_1)+pk_3(x_1))^2] \end{aligned} \tag{9}$$

This study presents a Three-Stage Adaptive Estimator of prevalence of a trait without truncation and studies the effect of truncation by comparing with Okoth et.al [4] model at stage three where truncation was done.

III. Model construction

3.1 Three stage Adaptive Batch testing model

In this scheme batches are partitioned into three and tested at three stages.

Suppose we have a population of size N which is pooled into n homogeneous batches each of which has batch size k. In this case λ_1n batches each of size k_1 be tested at stage one, λ_2n each of size k_2 at stage two and λ_3n each of size k_3 at stage three. The diagrammatic representation of this model is as shown in figure 3.1

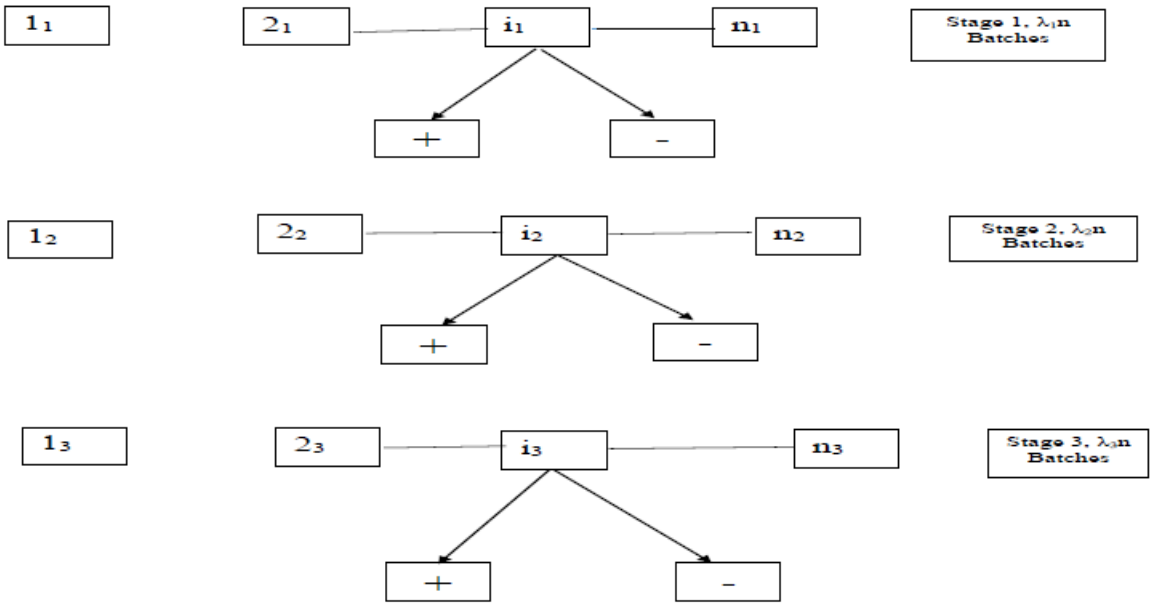

Fig 1. Schematic representation of Three stage Adaptive Batch Testing Model

From Fig 3.1, when a batch is tested, it either yields positive or negative results; with the probability that a batch tests positive as 1-(1-p)^(k) and (1-p)^(k) as the probability that a batch tests negative [11].
The batch size at this stage isk_3 and is determined as

$$k_3 = argmin(Var\hat{p}_2) \tag{10}$$

If X_3 is the number of defective batches at stage 3, conditioned on X_1 and X_2
Then X_3 follows a Binomial distribution with parameters $\lambda_3$nand π_3
Simply written as

$$X_3/X_1X_2 \sim Bi(\lambda_3 n, \pi_3) \tag{11}$$

whereπ_3 is the probability that a batch tests positive at stage three and is given by,

$$\pi_3 = \eta[1-(1-p)^{k_3(x_2)}] + (1-\phi)(1-p)^{k_3(x_2)} \tag{12}$$

The final Three stage adaptive estimator of p,$\hat{p}_B$is the MLE based on the joint distribution of$ X_1, X_2 and X_3.

This joint distribution is given by,

$$\begin{aligned} f(X_3, X_2, X_1) &= Bin(\lambda_1 n, \eta[1-(1-p)^{k_1}] + (1-\phi)(1-p)^{k_1}) \\ &* Bin(\lambda_2 n, \eta[1-(1-p)^{k_1(x_1)}] + (1-\phi)(1-p)^{k_2(x_1)}) \\ &* Bin(\lambda_3 n, \eta[1-(1-p)^{k_3(x_2)}] + (1-\phi)(1-p)^{k_3(x_2)}) \end{aligned} \tag{13}$$

To find MLE we find the likelihood function of equation (13) which is given by

$$
\begin{aligned}
f(X_3, X_2, X_1) &= \begin{pmatrix} n\lambda_1 \\ x_1 \end{pmatrix} [\eta(1-(1-p)^{k_1}) + 1 - \phi)(1-p)^{k_1}]^{x_1} \\
&* \ [1 - (\eta - \eta(1-p)^{k_1} + (1-\phi)(1-p)^{k_1})]^{n\lambda_1 - x_1} \\
&* \begin{pmatrix} n\lambda_2 \\ x_2 \end{pmatrix} [\eta(1-(1-p)^{k_2(x_1)}) + 1 - \phi)(1-p)^{k_2(x_1)}]^{x_2} \\
&* \ [1 - (\eta - \eta(1-p)^{k_2(x_1)} + (1-\phi)(1-p)^{k_2(x_1)})]^{n\lambda_2 - x_2} \\
&* \begin{pmatrix} n\lambda_3 \\ x_3 \end{pmatrix} [\eta(1-(1-p)^{k_3(x_2)}) + 1 - \phi)(1-p)^{k_3(x_2)}]^{x_3} \\
&* \ [1 - (\eta - \eta(1-p)^{k_3(x_2)} + (1-\phi)(1-p)^{k_3(x_2)})]^{n\lambda_3 - x_3} \qquad (14)
\end{aligned}
$$

Applying Maximum Likelihood Estimate method, The Three-stage Adaptive estimator is determined as the solution to

$$
\begin{aligned}
&\frac{x_1 k_1 q^{k_1}[(1-\phi) - \eta]}{\eta - \eta q_1^k + (1-\phi)q_1^k} + \frac{x_2 k_1(x_1) q^{k_2(x_1)}[(1-\phi) - \eta]}{\eta - \eta q^{k_2(x_1)} + (1-\phi)q^{k_2(x_1)}} \\
+ \ &\frac{x_3 k_3(x_1) q^{k_3(x_1)}[(1-\phi) - \eta]}{\eta - \eta q^{k_3(x_1)} + (1-\phi)q^{k_3(x_1)}} \\
= \ &\frac{(n\lambda_1 - x_1) k_1 q^{k_1}[(1-\phi) + \eta]}{1 - [\eta - \eta q^{k_1} + (1-\phi)q^{k_1}]} + \frac{(n\alpha_2 - x_2) k_2 q^{k_1(x_1)}[(1-\phi) + \eta]}{1 - [\eta - \eta q^{k_2(x_1)} + (1-\phi)q^{k_2(x_1)}]} \\
+ \ &\frac{(n\alpha_3 - x_3) k_3 q^{k_3(x_2)}[(1-\phi) + \eta]}{1 - [\eta - \eta q^{k_3(x_2)} + (1-\phi)q^{k_3(x_2)}]} \qquad (15)
\end{aligned}
$$

The derivation of Equation (15) is shown in appendix A

Next we establish the Asymptotic variance of $\hat{p}_B$, Var($\hat{p}_B$) which does not require that we find$\hat{p}_B$

3.2 Asymptotic variance of$\hat{p}_B$

Asymptotic Variance is useful in determining Asymptotic Relative Efficiency ARE. We recall Equation (13) and apply CRLB method to it to obtain
Var($\hat{p}_B$)as

$$
Var(\hat{p}_B) = \frac{\pi_1(p)\pi_2(p)\pi_3(p)(1-\pi_1(p))(1-\pi_2(p))(1-\pi_3(p))}{A} \qquad (16)
$$

where,

$$
\begin{aligned}
A &= (\eta + \phi - 1))^2[\pi_2(p)\pi_3(p)(1-\pi_2(p))(1-\pi_3(p))n\lambda_1 k_1^2(1-p)^{2k_1-2} \qquad (17) \\
&+ \ \pi_1(p)\pi_3(p)(1-\pi_1(p))(1-\pi_3(p))n\lambda_2 k_2^2(x_1)(1-p)^{2k_2(x_1)-2} \\
&+ \ \pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))n\lambda_3 k_3^2(x_3)(1-p)^{2k_3(x_2)-2}
\end{aligned}
$$

3.2 Asymptotic Relative Efficiency

In this section we present results on ARE of the Estimator in the current study relative to the three stage Adaptive batch testing Estimator with truncation.
The computation of ARE was accomplished by dividing Equations (8) by (16) to obtain

$$ARE = \frac{A}{B} \tag{18}$$

where ,

$$\begin{aligned} A &= (\eta+\phi-1))^2[\pi_2(p)\pi_3(p)(1-\pi_2(p))(1-\pi_3(p))n\lambda_1 k_1^2(1-p)^{2k_1-2} \\ &+ \pi_1(p)\pi_3(p)(1-\pi_1(p))(1-\pi_3(p))n\lambda_2 k_2^2(x_1)(1-p)^{2k_2(x_1)-2} \\ &+ \pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))n\lambda_3 k_3^2(x_3)(1-p)^{2k_3(x_2)-2} \end{aligned} \tag{19}$$

and

$$\begin{aligned} B &= (\eta+\phi-1)^2[\lambda_1 n\pi_2(p)\pi_3(p)(1-\pi_2(p))1-\pi_3(p)(k_1-pk_1^2+pk_1)^2 \\ &+ \lambda_2 n\pi_1(p)\pi_3(p)(1-\pi_1(p))(1-\pi_3(p))(k_2(x_1)-pk_2^2(x_2)+pk_2(x_1))^2 \\ &+ (1-\lambda_1-\lambda_2)n\pi_1(p)\pi_2(p)(1-\pi_1(p))(1-\pi_2(p))(k_3(x_1)-pk_3^2(x_1)+pk_3(x_1))^2] \end{aligned} \tag{20}$$

Utilizing equation (18) and R-software, tables 4.1, 4.2 and 4.3 were obtained.

IV. Analysis and Discussion

p	$\eta=\phi=0.99$	$\eta=\phi=0.95$	$\eta=\phi=0.90$	$\eta=\phi=0.85$	$\eta=\phi=0.80$
0.1	2.5427	2.5527	2.5639	2.5727	2.5478
0.2	3.8717	3.9095	3.9522	3.9899	3.8899
0.3	6.9505	7.0844	7.2362	7.3707	7.01027
0.4	11.7912	12.1849	12.6332	13.0321	11.9626
0.5	7.6669	8.0019	3.93498e-4	8.7137	7.7959

Table 1 :ARE values of$\hat{p}_B$ relative to $\hat{p}_3$when$\eta=\phi$ at specified values of p**.**

Table 1 provide a generated ARE values for p, η andϕ at different values. It's evident that ARE values are greater than one at different probabilities when$\eta=\phi$ across the table. This means that estimator in the current study is more efficient than the Truncated Estimator. This scenario is depicted in Table 2 and 3 as illustrated below.

p	$\eta=0.99,$ $\phi=0.80$	$\eta=0.99,$ $\phi=0.85$	$\eta=0.99,$ $\phi=0.90$	$\eta=0.99,$ $\phi=0.95$	$\eta=0.99,$ $\phi=0.98$
0.1	2.4076	2.4583	2.4987	2.5276	2.5396
0.2	3.5657	3.6719	3.7598	3.8295	3.8623
0.3	6.1445	6.4127	6.6.6416	6.8291	6.9229
0.4	9.9034	10.4905	11.0166	11.4757	11.7168
0.5	6.7352	7.0026	7.2530	7.4882	7.6210

Table 2:ARE values of $\hat{p}_B$ relative to$\hat{p}_3$when $\eta$is constant at 99% and ϕ varies.

Table 2 provide a generated ARE values for p, $\boldsymbol{\eta}$ andϕat different values. It's evident that ARE values are greater than one at different probabilities when $\eta$is held constant and ϕ varies across the table. However efficiency is high at higher value ofϕ= 0.98 and lower value of η .

p	$\eta = 0.80,$ $\phi = 0.99$	$\eta = 0.85,$ $\phi = 0.99$	$\eta = 0.90$ $\phi = 0.99$	$\eta = 0.95$ $\phi = 0.99$	$\eta = 0.98,$ $\phi = 0.99$
0.1	2.5250	2.5539	2.5539	2.5592	2.5478
0.2	3.9828	4.0027	4.0027	3.9365	3.8899
0.3	7.4697	7.4614	7.4615	7.1686	7.0102
0.4	13.6263	13.4451	13.4451	12.4291	11.9623
0.5	9.4425	9.11089	9.11089	8.1585	7.7957

Table 3:ARE values of$\hat{p}_B$relative to$\hat{p}_3$ whenϕ is constant at 99% and ηvaries.

Table 3 provide a generated ARE values for p, ηandϕ at different values. It's evident that ARE values are greater than one at different probabilities whenϕ is held constant andη varies across the table. However efficiency is high at lower value ofη = 0.80 and higher value of ϕ=0.99.These observations are graphically illustrated below.

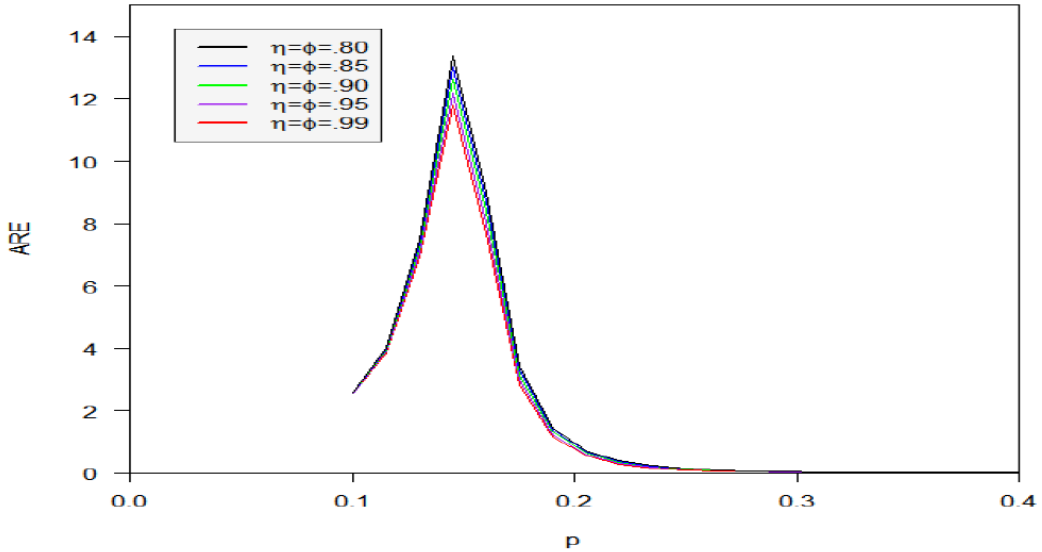

Figure 2: Plot of ARE values vs p whenη= ϕ at specified values of p.

Fig 2 represents a plot of ARE values against p at specified values ofη and ϕ . The figure shows that ARE values are all above one for p≤ 0.2. This means that the estimator in the current study is more efficient than the truncated estimator at those values of p. It's also evident that the efficiency increases with a decrease in sensitivity and specificity; i.e. the highest efficiency is recorded at$\eta = \phi = 0.80$ and lowest at$= \phi = 0.99$. It's also clear that the efficiency reaches the peak at p=0.15 and starts decreasing hitting zero at p=0.3 this shows that the current estimator performs better than the Truncated estimator at low prevalence rates (p≤ 0.2) and low specificity and sensitivity ($\eta = \phi = 0.80$). This same scenario is depicted in Fig 3 and 4 below.

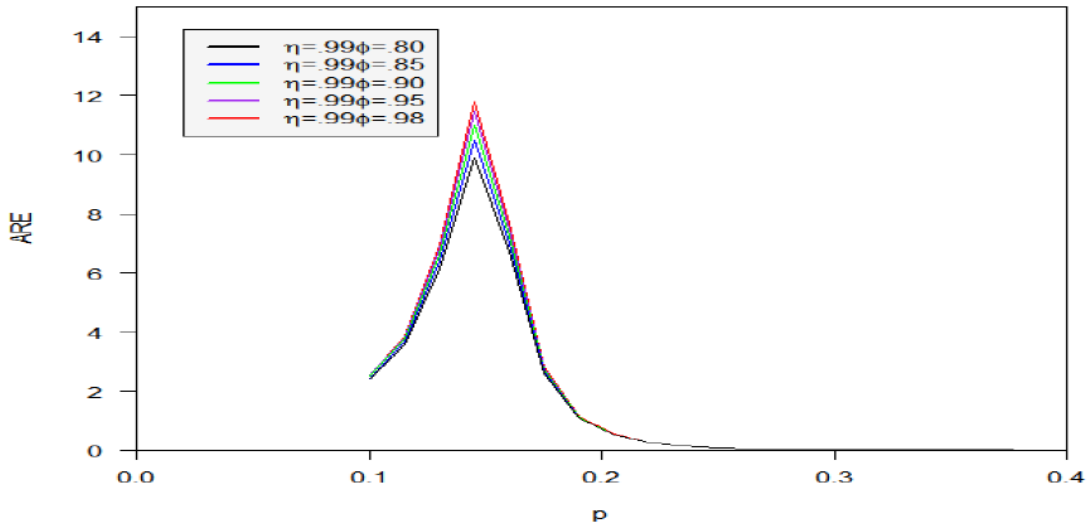

Figure 3: Plot of ARE values vs p whenη is constant at 99% andϕvaries.

In Figure 3, the starting point for ARE is 2.52 and the peak value is 11.7168. This means that when $\eta$is held constant, efficiency increases with increase in ϕ (i.e. highest at ϕ =0.98 and lowest at$\phi = .80$).

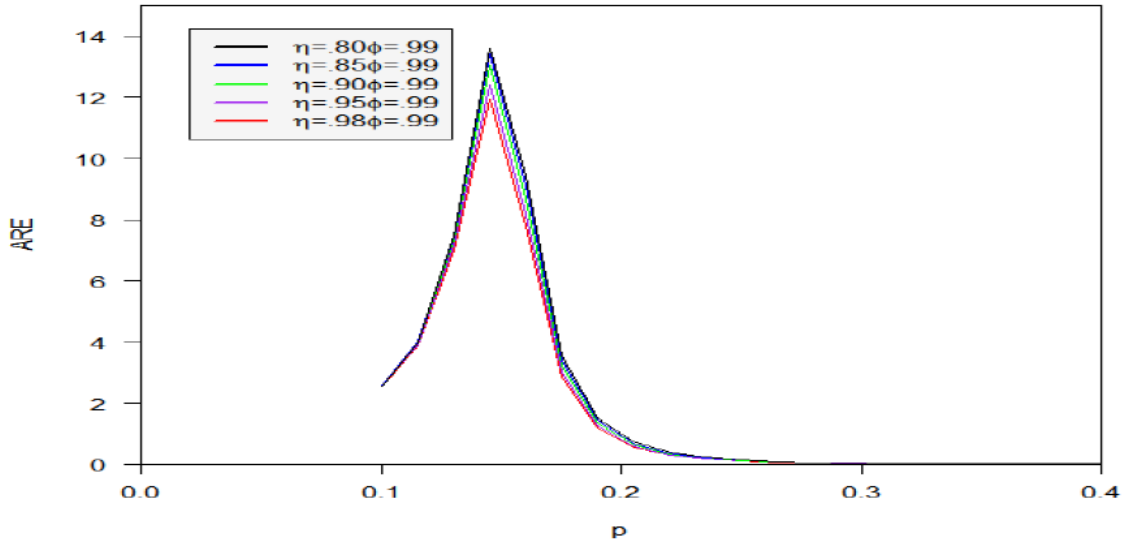

Figure 4: Plot of ARE values vs p whenϕ is constant at 99% andη varies.

In Figure 4, the starting point for ARE is 2.52 and the peak value is 13.6263. This means that when ϕ is held constant efficiency increases with a decrease inηi.e. it's highest at$\eta = 0.80$ and lowest at
$\eta = 0.98$.

V. Conclusion and Recommendation

Evidently from the above discussions the three stage Adaptive Batch testing Estimator without truncation is more efficient than the truncated estimator in presence of errors. The Estimator performs better at lower values of sensitivity and specificity ($\eta = \phi = 0.80$) than at higher values ($\eta = \phi = 0.99$). It also performs better at low prevalence rates (p≤ 0.2). Since Batch testing targets low prevalence rates, the model in this study befits this kind of scenario given that it performs better than the Truncated Model.

After comparison, we realize that Non-truncated models outperforms truncated models. We therefore recommend a generalized use of Non-truncated models of up to 'n' stages in statistical fields that require batch testing as it yields better results.

References

[1]. Dorfman R (1943). The detection of defective members of a large population, annals of mathematical statistics 14, 436-440.
[2]. Oliver-Hughes, JM and Swallow W.H (1994). A two stage adaptive group testing procedure for estimating small proportion. American statistical association. 89,982-993.

[3]. Thompson K.M (1962). Estimating of proportion of vector in a natural population of insect's .Biometrics 18, 568-578.
[4]. Okoth.A.W.(2017). Multi-stage Pool Testing Model With Test Errors: Improved Efficiency .Group testing anew goal Estimation. biometrika, 52-181-194.
[5]. Wanyonyi W.R (2015). Estimation of proportion of a trait by batch testing with Errors in Inspection in a quality control process.
[6]. Biggerstaff, B.(2008). Confidence intervals for the Difference of two proportions Estimated from pooled samples .Journal of Agricultural, Biological and Environmental statistics, 13,478-496.
[7]. Hughes,G., and Gottwald, T.R (1998). Survey methods for assessment of citrus tristeza virus incidence .phytopathology 88. 715-723.
[8]. Brookmeyer, R,(1999). Analysis of multistage Pooling studies of biological specimens for estimating disease incidence and prevalence. Biometrics 55, 608-612.
[9]. Nyongesa L.K (2011). Dual estimation of prevalence and disease incidence in pool testing strategy. Communication in statistics theory and methods 40,1-12.
[10]. Oliver-Hughes J.M (2006). Pooling for Experiments Blood screening and Drug Discovery screening; methods for experimentation in industry,Drug Discovery and Genectics New York, springer.
[11]. Okoth .A.W .{2012}. Two stage Adaptive Pool Testing for estimating prevalence of a trait in the presence of test errors.
[12]. Johnson, N.L., Kotz, S., Balakrishnan, N. (1994) Continuous Univariate Distributions, Volume 1, Wiley. ISBN 0-471-58495-9

Oduori Victor. " Three-Stage Adaptive Batch Testing Model for Estimating Prevalence of a Trait Without Truncation." IOSR Journal of Mathematics (IOSR-JM) 15.5 (2019): 19-27.