American Journal of Theoretical and Applied Statistics
2017; 6(2): 79-89
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20170602.12
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

SciencePG
Science Publishing Group

Sequentially Selecting Between Two Experiment for Optimal Estimation of a Trait with Misclassification

George Matiri[1], Kennedy Nyongesa[2], Ali Islam[1]

[1]Department of Mathematics, Egerton University, Nakuru, Kenya

[2]Department of Mathematics, Masinde Muliro University of Science and Technology, Kakamega, Kenya

Email address:

gemuwa@gmail.com (G. Matiri), knyongesa@hotmail.com (K. Nyongesa), asislam54@yahoo.com (A. Islam)

To cite this article:

George Matiri, Kennedy Nyongesa, Ali Islam. Sequentially Selecting Between Two Experiment for Optimal Estimation of a Trait with Misclassification. *American Journal of Theoretical and Applied Statistics*. Vol. 6, No. 2, 2017, pp. 79-89. doi: 10.11648/j.ajtas.20170602.12

Received: January 18, 2017; **Accepted**: February 3, 2017; **Published**: February 27, 2017

Abstract: The idea of pool testing originated with Dorfman during the World War II as an economical method of testing blood samples of army inductees in order to detect the presence of infection. Dorfman proposed that rather than testing each blood sample individually, portions of each of the samples can be pooled and the pooled sample tested first. If the pooled sample is free of infection, all inductees in the pooled sample are passed with no further tests otherwise the remaining portions of each of the blood samples are tested individually. Apart from classification problem, pool testing can also be used in estimating the prevalence rate of a trait in a population which was the focus of our study. In approximating the prevalence rate, one-at-a-time testing is time consuming, non-cost effective and is bound to errors hence pool testing procedures have been proposed to address these problems. This study has developed statistical model which is used to sequentially switching between two experiments when the sensitivity and specificity of the test kits is less than 100%. The experiments are selected sequentially, so that at each stage, the information available at that stage is used to determine which experiment to carry out at the next stage. The method of maximum likelihood estimator (MLE) was used in obtaining the estimators. The fisher information of different experiments is compared and the cut off values where one experiment is better than the other are calculated. The variance of the estimators has also been compared. The joint model has been compared to one-at-a-time and pool testing models by computing *ARE*. The joint model is found to be more efficient.

Keywords: Pool, Pool Testing, Cut off Value, Prevalence Rate, Sensitivity, Specificity

1. Introduction

Sequential testing of a population in the form of pools began by Dorfman [2] as an economical method of testing blood samples of army inductees in order to detect the presence of infection. Johnson *et al.* [6] and Nyongesa [14] extended Dorfman [2] work to multistage with the aim of reducing the number of tests. Computational testing with the first objective of classifying subjects has been developed by Maheswaran *et al.* [10]. Recently more research work are focused on the second objective for estimating the rate of trait. Thomson [18] studied the estimation problem using pool testing. This was later considered by Brookmayer [1] by introducing errors.

Sufficiently accurate estimate of the prevalence can be obtained from testing pooled samples as demonstrated by Hammick and Gastwirth [4]. Their procedure provides greater protection of respondent's anonymity which can lead to greater participation in the survey. On the same year, Gastwirth and Johnson [3] used pool testing to estimate HIV prevalence cost-effectively. Of recent Xie *et al.* [19] have demonstrated how pool testing can reduce costs in early stages of drug discovery. Janis *et al.* [5] considered sequentially deciding between two experiments for estimating a common success prevalence rate where he considered the individual Bernoulli *(p)* trials or the product of k individual independent Bernoulli (p^k) trials. Nyongesa [13] proposed pool testing when members that form the population under investigation are pooled together in pools and these pools are given a test. Pools that test negative, further testing are discontinued but if the reading is positive the pool is divided into blocks of equal sizes. The blocks are further tested and those that test positive the constituent members are tested individually for the presence or absence of the trait under investigation. Pools that test negative are given a retest and those that test positive on retest member constituents are tested individually. Nyongesa [13] used moment method to estimate the prevalence and he observed that his proposed testing procedure reduced misclassification, particularly the false positives. Computational statistics has been used in pool testing to compute the statistical measures when perfect and imperfect tests are used (Syaywa and Nyongesa [16]; Tamba *et al.* [17]).

Pool testing can be applied in many areas as outlined by Sobel and Groll [15]. The first application of pool-testing was to the problem of pooling blood samples in order to classify each one of a large group of people as to whether or not they have a particular disease. Mundel [12] showed that group testing can be applied in industries for example, in making a "leak test" on a large number of gas-filled electrical devices, one can test any number of units in a single test and the result of test on *k* units is that either all *k* are good (no leak) or at least 1 of the *k* is defective. Another application is in testing various electrical devices such as condensers, resistors, *etc*. Pool testing has been applied in screening the population for the presence of HIV antibody (Kline *et al.* [8] and Manzon *et al.* [11]). Litvak *et al.* [9], applied pool testing in screening HIV antibody to help curb the further spread of the virus. Litvak *et al.*, [9] showed that pooling offers a feasible way to lower the error rates associated with labelling samples when screening low risk HIV population. For instance, given the limited precision of the available test kits, it has been shown that screening pooled sera can be used to reduce the probability that a sample labelled negative in fact has antibodies since each test has a certain sensitivity and specificity. Juan and Wenju [7] have provided algorithm for the computations of pool sizes.

The essence of this study is to device a method of selecting between two experiments namely:

i) individual testing of items of a population with a view to estimating prevalence rate *p*, in this experiment we shall assume the tests are imperfect that is to say the test have $\eta,\beta < 100\%$, where η and β are sensitivity and specificity respectively, this experiment here in denoted by P^I,

ii) pool testing experiment as proposed by Dorfman [2] but with errors in inspection. This experiment here in denoted by P^G.

The rest of the paper is arranged as follows: in Section 2 we shall develop the models and formula for calculating their Fisher information, in Section 3 we shall plot the graphs of Fisher information against the value of p. In Section 4 we shall compute the cut off values. In Section 5 we shall develop the maximum likelihood estimators of *p* and their asymptotic variance. Section 6 we shall compare the asymptotic variances of the maximum likelihood estimators by plotting their graphs. In Section 7 we shall compute the *ARE* values and in section 8 we shall have discussion and conclusion of the study.

2. The Models

The model have been split into two that is P^I-experiment and P^G-experiment. P^I-experiment means estimating the prevalence rate of the characteristic of interest with testing each individual under study while P^G-experiment means estimating the prevalence rate of the characteristic of interest by putting together items or individuals to form a pool and testing the pool rather than testing each subject. Throughout the study m and n have been assumed to be the number of observations from the P^I-experiment and the P^G-experiment respectively with $N = m + n$, the total number of observations from both experiments.

2.1. The P^I Experiment

In our study the P^I-experiment will involve estimating prevalence rate of the characteristic of interest with testing each individual under study. Suppose the P^I-experiment is to be used to estimate the prevalence rate p of interest and if X_{1i} for $i = 1,\dots,\mathrm{m}$ is a sequence of identically independent distributed random variable, then $X_{1i} \sim \mathrm{B}\mathit{ernouli}(\tau_1)$ where τ_1 is the probability of declaring an individual as positive *i.e* $\tau_1 = \eta p + (1-\beta)(1-\mathrm{p})$.

For a single experiment, the probability density function is

$$f(x_{1i}, p \mid \eta,\beta) \\ = (\eta p + (1-\beta)(1-\mathrm{p}))^{x_{1i}}((1-\eta)\mathrm{p} + \beta(1-p))^{1-x_{1i}} . \quad (1)$$

The Fisher information on the prevalence rate p contained in a single observation denoted by $I_{x_1}(\mathrm{P}^I)$ is

$$I_{x_1}(\mathrm{P}^I) = \frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} . \quad (2)$$

If m observations from only the P^I-experiment are used to estimate p, then the likelihood function of Equation (1) is

$$\mathrm{L}(\mathrm{x}_{1i}, \mathrm{p} \mid \eta,\beta) = \prod_{i=1}^{m} \tau_1^{x_{1i}}(1-\tau_1)^{1-x_{1i}} \\ = \tau_1^{\sum_{i=1}^{m} x_{1i}}(1-\tau_1)^{m-\sum_{i=1}^{m} x_{1i}} .$$

Therefore the estimator of p from P^I-experiment is

$$\hat{p}_m^I = \frac{\beta - 1 + \frac{\sum_{i=1}^{m} x_{1i}}{m}}{\eta+\beta-1} \quad (3)$$

and the asymptotic variance of $\hat{p}_m^1$ is

$$\mathrm{var}(\hat{p}_m^1) = \frac{\tau_1(1-\tau_1)}{(\eta+\beta-1)^2} . \quad (4)$$

2.2. The P^G-experiment

The P^G-experiment involve putting together items to form a pool and testing the pool rather than testing each individual for the evidence of a characteristic of interest. A negative reading indicates that the pool contains no defective item and a positive reading indicates at least one defective item in the pool. Pooling procedures have proved to reduce the cost of testing when the prevalence rate is low. In this experiment, the probability of declaring a pool of size k positive will be denoted by τ_2 and for analysis purposes, we shall assume that the constituent members of a pool act independent of each other with $\tau_2 = \eta(1-(1-\mathrm{p})^k)+(1-\beta)(1-\mathrm{p})^k$. Let X_{2j} denote a sequence of identically independent distributed random variable for $j=1,...,n$, then $X_{2j} \sim Bernouli(\tau_2)$. For a single experiment equivalently the probability density function is

$$f(x_{2j}, p \mid \eta, \beta, k) = (\tau_2)^{x_{2j}}(1-\tau_2)^{1-x_{2j}} \tag{5}$$

from which the fisher information denoted by $I_{x_2}(\mathrm{P}^G)$ is

$$I_{x_2}(\mathrm{P}^G) = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}. \tag{6}$$

Suppose there are n pools from the P^G-experiment each of size k, available for estimating p and suppose X_{2j} pool test positive on the test. Then from Equation (5), the maximum likelihood estimator of p from the P^G-experiment is

$$\hat{p}_n^G = 1 - \left(\frac{\eta - \frac{\sum_{j=1}^{n} x_{2j}}{n}}{\eta+\beta-1} \right)^{\frac{1}{k}} \tag{7}$$

and the asymptotic variance of $\hat{p}_n^G$ is

$$\mathrm{var}(\hat{p}_n^G) = \frac{\tau_2(1-\tau_2)}{\mathrm{k}^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}. \tag{8}$$

2.3. The Joint Model

If m is the number of observations from P^I-experiment and n is the number of observations from P^G-experiment, assuming independence, then the joint probability density function of the random variables X_{1i} and X_{2j} from the P^I-experiment and P^G-experiment respectively is a multinomial probability density function given by the product of their density functions

$$f(\underline{x}, \underline{p} \mid k, \eta, \beta) = \tau_1^{x_{1i}}(1-\tau_1)^{1-x_{1i}} \times \tau_2^{x_{2j}}(1-\tau_2)^{1-x_{2j}}. \tag{9}$$

The joint likelihood function of Equation (9) is

$$\begin{aligned}&\mathrm{L}(\underline{x}, \underline{p} \mid \mathrm{k}, \eta, \beta)\\ &= \left\{ [\tau_1]^{\sum_{i=1}^{m} x_{1i}} [1-\tau_1]^{m-\sum_{i=1}^{m} x_{1i}} \times [\tau_2]^{\sum_{j=1}^{n} x_{2j}} [1-\tau_2]^{n-\sum_{j=1}^{n} x_{2j}} \right\}\end{aligned}$$

where the maximum likelihood estimator (MLE) is obtained by solving

$$\frac{\sum_{i=1}^{m} x_{1i} - \mathrm{m}\tau_1}{\tau_1(1-\tau_1)} \frac{d\tau_1}{dq} + \frac{\sum_{j=1}^{n} x_{2j} - n\tau_2}{\tau_2(1-\tau_2)} \frac{d\tau_2}{dq} = 0. \tag{10}$$

Since k, β *and* η are known constants, then Equation (10) is a continuous function of $q = 1-p$ and a unique value of *q,* that satisfy the equation exists since its plot cuts the *q*-axis at a point as q varies from 0 to 1. The value of *q,* denoted by $\hat{q}_{mle}$, that satisfy Equation (10) can be solved iteratively as follows:

Let

$$f(\mathrm{q}) = \frac{\sum_{i=1}^{m} x_{1i} - \mathrm{m}\tau_1}{\tau_1(1-\tau_1)} \frac{d\tau_1}{dq} + \frac{\sum_{j=1}^{n} x_{2j} - n\tau_2}{\tau_2(1-\tau_2)} \frac{d\tau_2}{dq},$$

then a unique value of q exists such that $f(\mathrm{q}) = 0$. Consider a tangent line of $f(\mathrm{q})$ that passes through the point $(\mathrm{q}_0, \mathrm{f}(\mathrm{q}_0))$ and $(\mathrm{q}_1, 0)$ where q_0 is the initial approximation of the root of $f(\mathrm{q})$, then the gradient of the tangent line at the point $(\mathrm{q}_0, \mathrm{f}(\mathrm{q}_0))$ denoted by $f'(\mathrm{q}_0)$ is given by $f'(\mathrm{q}_0) = \frac{f(\mathrm{q}_0)}{q_0 - q_1}$ and solving for q_1 leads to $q_1 = q_0 - \frac{f(\mathrm{q}_0)}{f'(\mathrm{q}_0)}$. Similarly $q_2 = q_1 - \frac{f(\mathrm{q}_1)}{f'(\mathrm{q}_1)}$, $q_3 = q_2 - \frac{f(\mathrm{q}_2)}{f'(\mathrm{q}_2)}$. In general $q_{i+1} = q_i - \frac{f(\mathrm{q}_i)}{f'(\mathrm{q}_i)}$ where $f'(\cdot)$ is the derivative of the function $f(\mathrm{q})$ which is not equal to zero for any value of q_i for $i = 0, 1, 2, \cdots\cdots$. The iteration will stop if $|q_{i+1} - q_i| < \varepsilon$ for some arbitrary value ε, and since the series converges, q_{i+1} is taken as an approximate value of $\hat{q}_{mle}$ which is the solution of Equation (10). The 'while' matlab loop was used for solving Equation (10).

The asymptotic variance of $\hat{p}_{mle}$ of the joint model where $\hat{p}_{mle} = 1 - \hat{q}_{mle}$ is

$$\mathrm{var}(\hat{p}_{mle}) = \frac{\tau_1\tau_2(1-\tau_1)(1-\tau_2)}{Q} \tag{11}$$

where $Q = (\eta+\beta-1)^2\left\{m\tau_2(1-\tau_2) + nk^2(1-p)^{2k-2}\tau_1(1-\tau_1)\right\}$.

3. Comparison of $I_x(\cdot)$ of P^I and P^G Experiments

In this section we compare the performance of each of the two procedures by plotting the graphs of $I_x(\cdot)$ of P^I and P^G-

experiment for various values of k, η, β versus p.

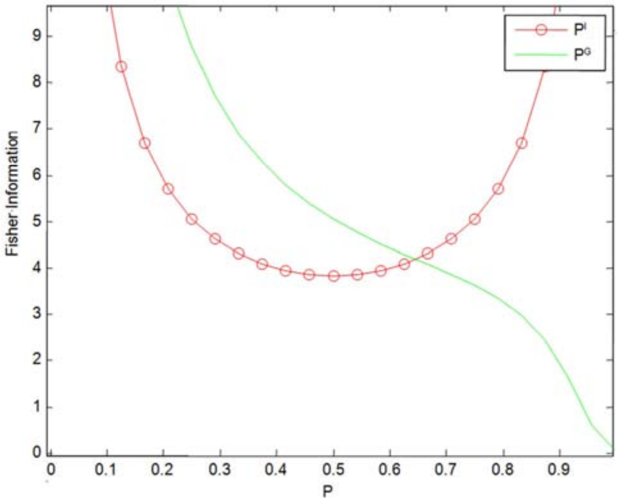

***Figure 1.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.99$ and $k = 2$.*

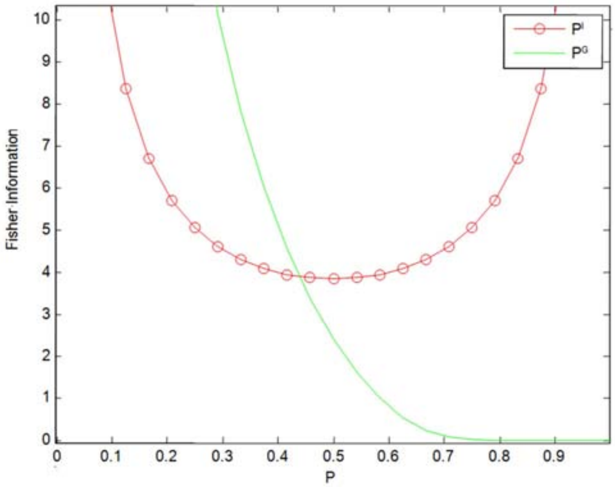

***Figure 2.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.99$ and $k = 5$.*

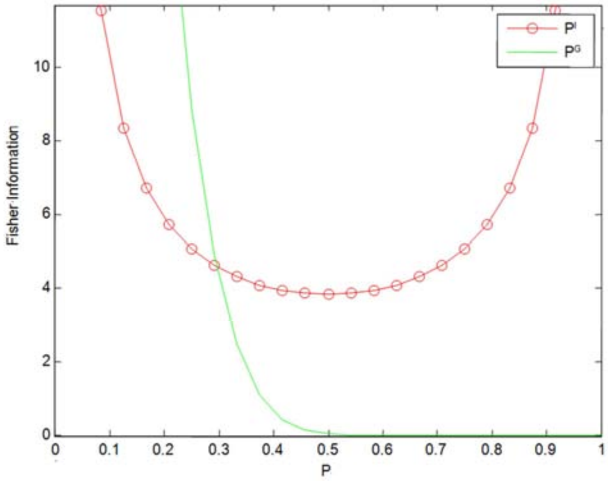

***Figure 3.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.99$ and $k = 10$.*

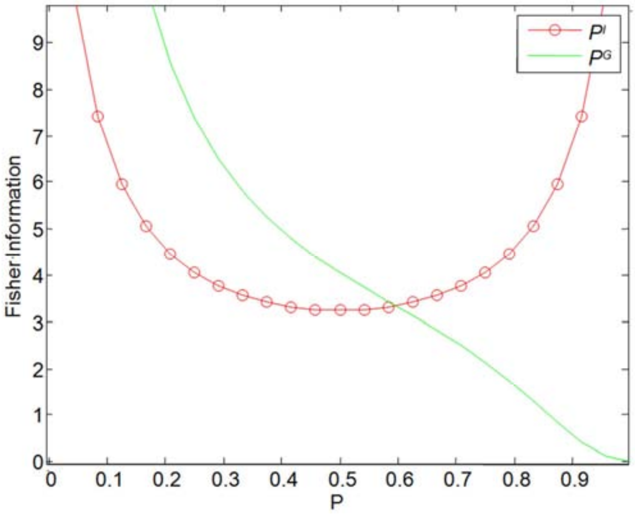

***Figure 4.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.95$ and $k = 2$.*

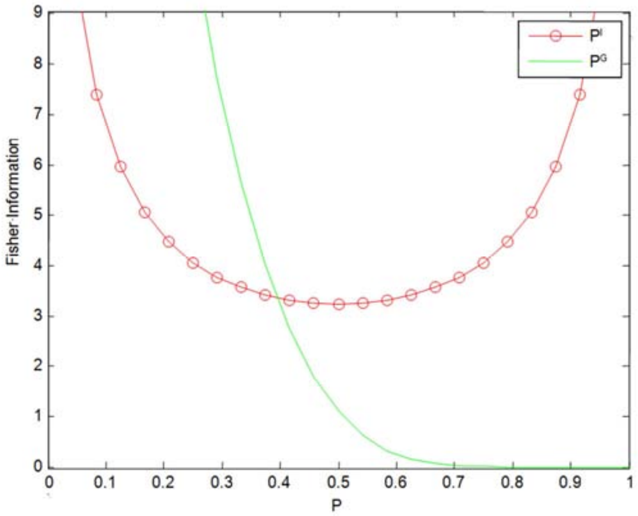

***Figure 5.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.95$ and $k = 5$.*

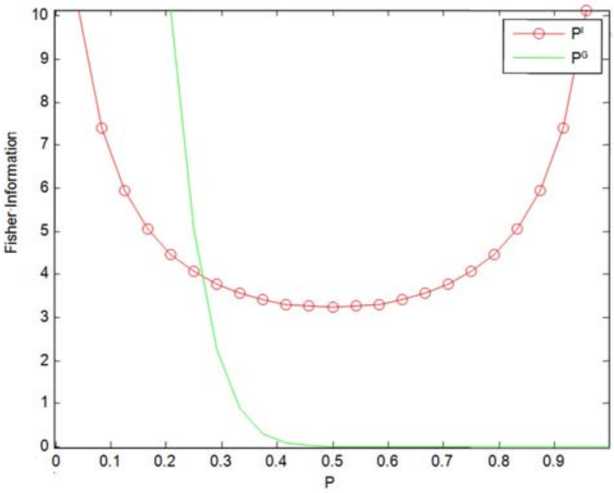

***Figure 6.** A graph of Fisher Information against the value of p with $\eta = \beta = 0.95$ and $k = 10$.*

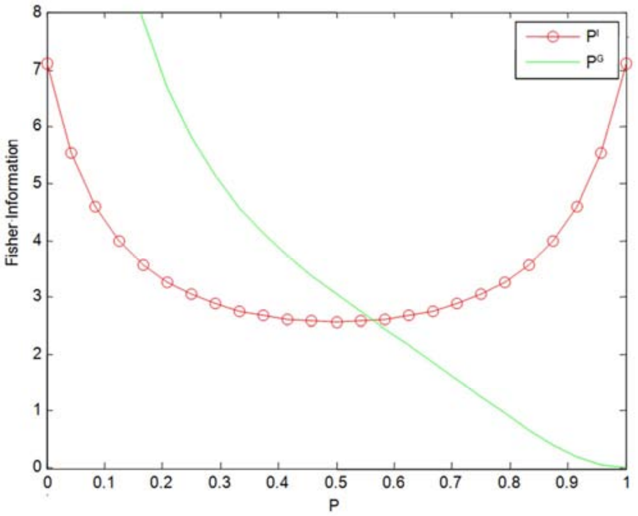

Figure 7. *A graph of Fisher Information against the value of p with* $\eta=\beta=0.90$ *and* $k=2$.

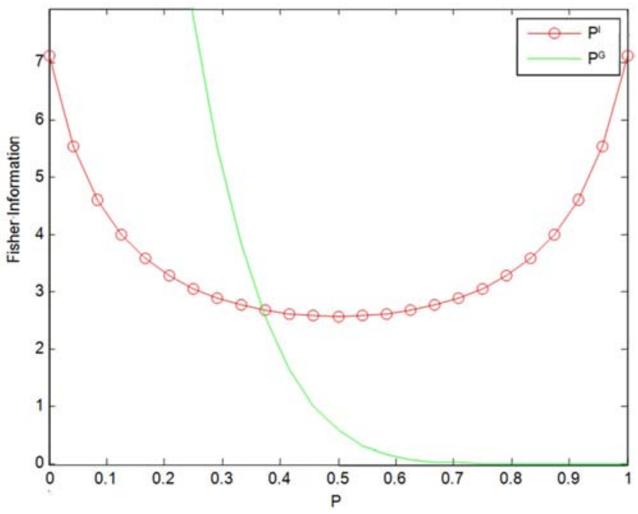

Figure 8. *A graph of Fisher Information against the value of* p *with* $\eta=\beta=0.90$ *and* $k=5$.

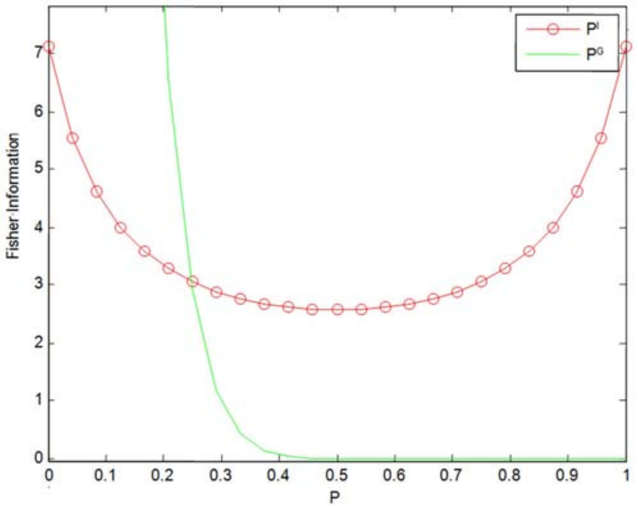

Figure 9. *A graph of Fisher Information against the value of p with* $\eta=\beta=0.90$ *and* $k=10$.

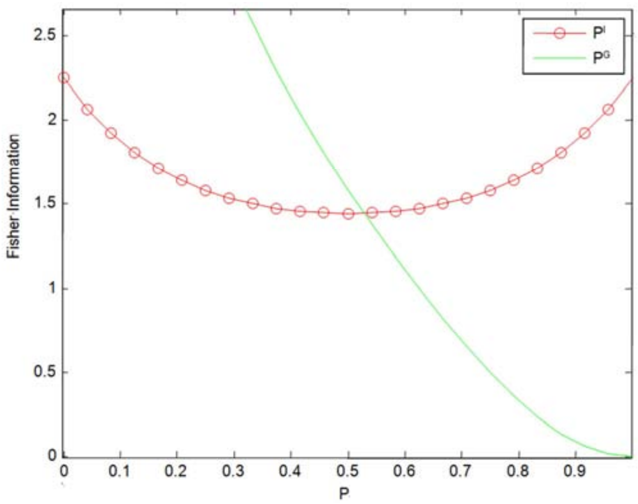

Figure 10. *A graph of Fisher Information against the value of* p *with* $\eta=\beta=0.80$ *and* $k=2$.

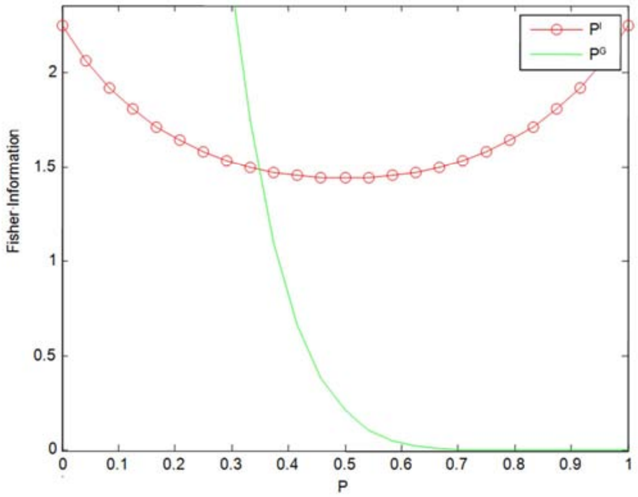

Figure 11. *A graph of Fisher Information against the value of* p *with* $\eta=\beta=0.80$ *and* $k=5$.

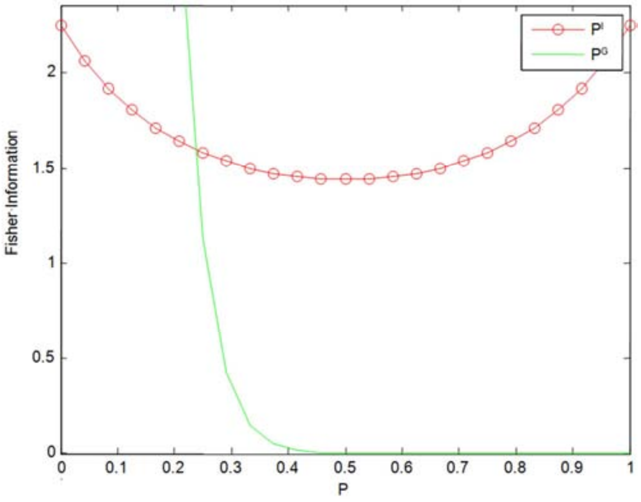

Figure 12. *A graph of Fisher Information against the value of p with* $\eta=\beta=0.80$ *and* $k=10$.

As seen from Figures 1 to 12, the plot of the Fisher information of the P^I-experiment is symmetric and concave upwards *i.e* the Fisher information is very high for values of p close to 0 and for the values of p close to 1. It is minimum for the values of p about 0.5. It can also be noted that the change of the value of *k* does not affect the Fisher information of the P^I-experiment since the P^I-experiment is independent of *k*. As sensitivity and specificity of the tests increases the Fisher information for P^I-experiment also increases. The graph of Fisher information of the P^G-experiment is found to be strictly decreasing as the value of the parameter p increases from 0 to 1. A striking feature also to note is that the relationship between the Fisher information and the parameter p is sensitive to *k* as the slope of the curve changes with varying *k*. The curve become steeper as *k* increases but the slope become less steep and almost levelises as p approaches 1. It is also noted that as *k* increases the curve of the Fisher information of the P^G-experiment shift to the left of the graph meaning that the region for which P^G-experiment is better than the P^I-experiment shrinks. As sensitivity and specificity of the tests increases the region at which the Fisher information of P^G-experiment is higher than for the P^I-experiment increases. It can also be observed that pool testing is only visible and better than individual testing strategy where the prevalence rate is small which concurs with the idea of Dorfman [2] that pool testing is only viable if the prevalence rate is low otherwise the use of P^I-experiment is recommended.

4. Computation of Cut off Values

The cut off value shall be defined as the value of p at which the Fisher information for the P^I -experiment and the P^G -experiment are equal or the value of p at the point of intersection of the graphs of $I_x(\mathrm{P}^I)$ *and* $I_x(\mathrm{P}^G)$.

If we let $'a'$ be the cut off value, then $'a'$ is a unique root in $(0,1)$ of the equation $I_x(\mathrm{P}^I) = I_x(\mathrm{P}^G)$ *i.e*

$$\frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} = \frac{k^2(1-p)^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}$$

$$\frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} - \frac{k^2(1-p)^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)} = 0$$

$$\frac{1}{\tau_1(1-\tau_1)} - \frac{k^2(1-p)^{2k-2}}{\tau_2(1-\tau_2)} = 0$$

$$\tau_2(1-\tau_2) - \mathrm{k}^2(1-\mathrm{p})^{2k-2}\tau_1(1-\tau_1) = 0$$

$$\tau_2(1-\tau_2)(1-\mathrm{p})^2 - \mathrm{k}^2(1-\mathrm{p})^{2k}\tau_1(1-\tau_1) = 0 \qquad (12)$$

since k, β *and* η are known constants, then Equation (12) is a function of p , of which the value of p can be solved iteratively as follows:

Let

$$f(\mathrm{p}) = \tau_2(1-\tau_2)(1-\mathrm{p})^2 - \mathrm{k}^2(1-\mathrm{p})^{2k}\tau_1(1-\tau_1)\,, \qquad (13)$$

then the function $f(\mathrm{p})$ is continuous in the interval $(0,1)$ and from Figures 1 to 12 of the graphs of Fisher information, there exist a value *p*, such that Equation (13) is equal to zero which is the point of intersection of the two curves. Consider a tangent line of $f(\mathrm{p})$ that passes through the point $(\mathrm{p}_0, \mathrm{f}(\mathrm{p}_0))$ and $(\mathrm{p}_1, 0)$ where p_0 is the initial approximation of the root of $f(\mathrm{p})$, then the gradient of the tangent line at the point $(\mathrm{p}_0, \mathrm{f}(\mathrm{p}_0))$ denoted by $f'(\mathrm{p}_0)$ is given by $f'(\mathrm{p}_0) = \dfrac{f(\mathrm{p}_0)}{p_0 - p_1}$ and solving for p_1 yields $p_1 = p_0 - \dfrac{f(\mathrm{p}_0)}{f'(\mathrm{p}_0)}$. Similarly $p_2 = p_1 - \dfrac{f(\mathrm{p}_1)}{f'(\mathrm{p}_1)}$, $p_3 = p_2 - \dfrac{f(\mathrm{p}_2)}{f'(\mathrm{p}_2)}$. In general $p_{i+1} = p_i - \dfrac{f(\mathrm{p}_i)}{f'(\mathrm{p}_i)}$ where $f'(\cdot)$ is the derivative of the function $f(\mathrm{p})$ which is not equal to zero for any value of p_i for $i = 0, 1, 2, \ldots\ldots$. The iteration will stop if $|p_{i+1} - p_i| < \varepsilon$ for some arbitrary value ε which is the error term which should be small. If the series converges, p_{i+1} is taken as an approximate value of $'a'$ which is the solution of Equation (12). The 'while' matlab loop was used for solving Equation (13).

For various values of k, η and β the values of the roots of Equation (13) or the cut off values are given in Table 1:

Table 1. *Cut off values for various values of* k, η *and* β .

	a			
k	$\eta = \beta = 0.99$	$\eta = \beta = 0.95$	$\eta = \beta = 0.90$	$\eta = \beta = 0.80$
2	0.646	0.596	0.563	0.528
3	0.555	0.507	0.477	0.446
5	0.439	0.395	0.371	0.348
10	0.296	0.263	0.248	0.234
15	0.227	0.201	0.190	0.181
20	0.185	0.164	0.156	0.150
50	0.092	0.082	0.080	0.078

From Table 1 it can be observed that as the pool size (*k*) increases, the cut off point value decreases for various values of η and β *i.e* the region in which the P^G-experiment is better shrinks. This concurs with the conclusion that pool testing is only feasible when the pool size are reasonably small. It can also be observed that as sensitivity and specificity of the test kits increases the region in which the P^G-experiment is better also increases.

For example at $\eta = \beta = 0.90$, *k* = *5* and if *N* tests are available, the maximum information about *p* is obtained when

$$N = \begin{cases} \textit{observe all } P^G \textit{ if } p < 0.371 \\ \textit{observe all } P^I \textit{ if } p > 0.371 \\ \textit{arbitrary } P^I \textit{ or } P^G \textit{ if } p = 0.371 \end{cases}$$

In general, if *N* tests are available, then the allocation that maximizes the information about p is

$$N = \begin{cases} observe\ all\ P^G\ \ if\ \ p < a \\ observe\ all\ P^I\ \ if\ \ p > a \\ arbitrary\ P^I\ or\ P^G\ \ if\ \ p = a. \end{cases}$$

Note that the region where one experiment is better than the other depends on the unknown parameter p. Thus the obvious adaptive rule is suggested where p is estimated at each stage and the next observation is allocated depending on the relationship between the estimated p and the cut off point value.

5. Estimator of Prevalence Rate, Its Variance and Confidence Interval

In this section we compute the maximum likelihood estimator $\hat{p}$ of the prevalence rate, the variance and 95% Wald-type confidence interval of the maximum likelihood estimator for various values of sensitivity, specificity and pool size.

Table 2. *Maximum likelihood estimator, variance and Confidence interval for different values of p for* $\eta = \beta = 99\%$ *and* $k = 5, 10$.

	p	$\hat{p}$	**var(p̂) ×10^{-4}**	**95% *CI***
$k = 5$	0.01	0.0160	0.3266	-0.0086, 0.0407
	0.05	0.0465	0.8728	0.0052, 0.0878
	0.10	0.1190	2.291	0.0556, 0.1825
	0.20	0.2027	4.226	0.1239, 0.2815
$k = 10$	0.01	0.0113	0.1224	-0.0094, 0.0319
	0.05	0.0567	0.6592	0.01138, 0.1021
	0.10	0.1119	1.605	0.0501, 0.1736
	0.20	0.2337	6.136	0.1500, 0.3168

Table 3. *Maximum likelihood estimator, variance and Confidence interval for different values of p for* $\eta = \beta = 90\%$ *and* $k = 5, 10$.

	p	$\hat{p}$	**var(p̂) ×10^{-4}**	**95% *CI***
$k = 5$	0.01	0.0034	0.6200	-0.0081, 0.0150
	0.05	0.0561	1.9000	0.0110, 0.1013
	0.10	0.0831	2.6000	0.0290, 0.1373
	0.20	0.1634	5.1800	0.0909, 0.2359
$k = 10$	0.01	0.0073	0.2310	-0.0094, 0.0238
	0.05	0.0597	1.1100	0.0133, 0.1061
	0.10	0.1106	2.6000	0.0491, 0.1720
	0.20	0.2039	9.6000	0.1249, 0.2828

Table 4. *Maximum likelihood estimator, variance and Confidence interval for different values of p for* $\eta = \beta = 80\%$ *and* $k = 5, 10$.

	p	$\hat{p}$	**var(p̂) ×10^{-4}**	**95% *CI***
$k = 5$	0.01	0.0148	2.1900	-0.0089, 0.0385
	0.05	0.0542	3.6400	0.0098, 0.0986
	0.10	0.1164	6.5640	0.0535, 0.1793
	0.20	0.1789	10.748	0.1038, 0.2547
$k = 10$	0.01	0.0172	0.0780	-0.0083, 0.0428
	0.05	0.0306	0.0940	-0.0032, 0.0644
	0.10	0.1000	4.120	0.0412, 0.1588
	0.20	0.2767	4.128	0.1890, 0.3644

From Tables 2 to 4 it can be noted that the maximum likelihood estimators of the prevalence rate are very close to the actual value which was used to simulate the estimators. The population estimators resulting from the experiments are used to evaluate the $(1-\alpha)100\%$ confidence limits of the confidence interval of the simulated estimators where α is the level of significance and it can be noted from Tables 2 to 4 that the actual value is within the upper and the lower limits.

6. Comparison of Variances

In this section we shall plot the graphs of the variance for P^I, P^G-experiments and joint model for various values of k, η and β versus p values.

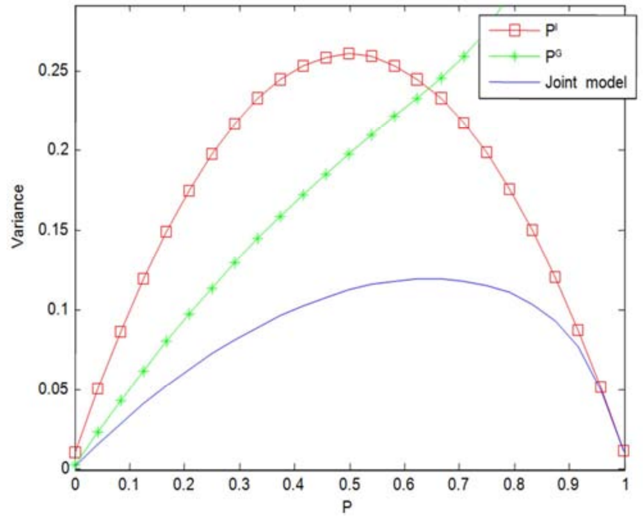

Figure 13. *A graph of* $Var(\hat{p})$ *as a function of* P *with* $\eta = \beta = 0.99$ *and* $k = 2$.

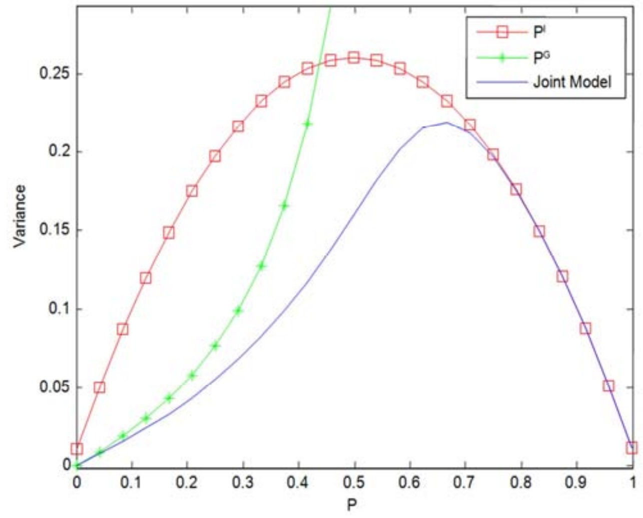

Figure 14. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta = \beta = 0.99$ *and* $k = 5$.

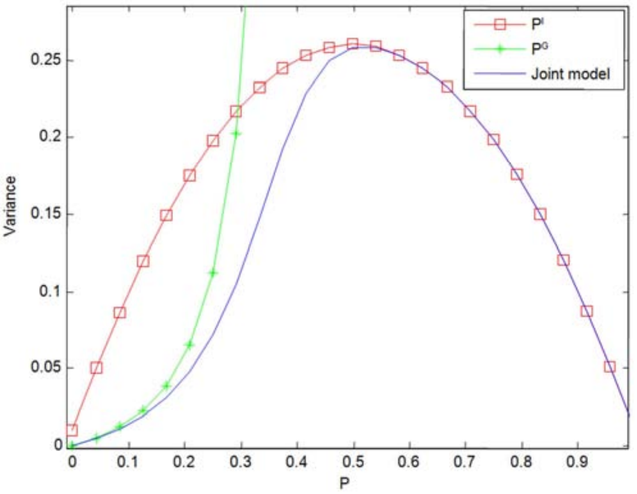

Figure 15. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.99$ *and* $k=10$.

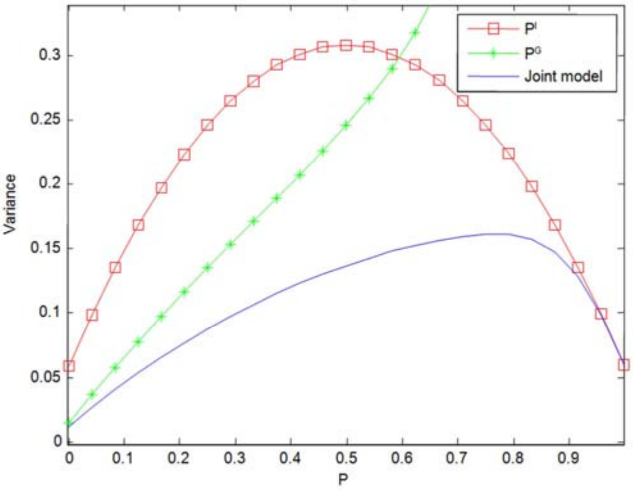

Figure 16. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.95$ *and* $k=2$.

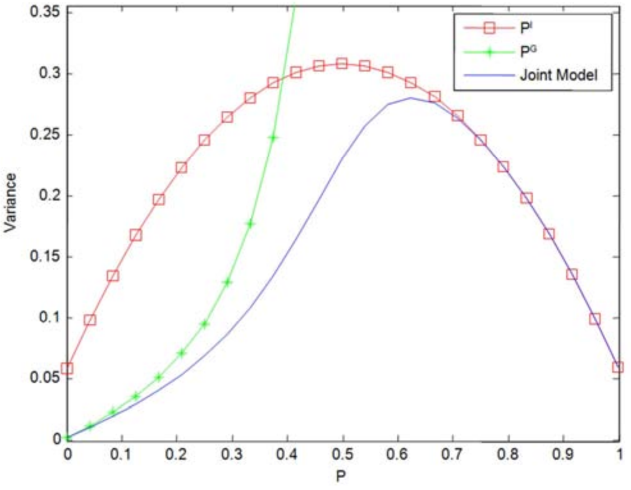

Figure 17. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.95$ *and* $k=5$.

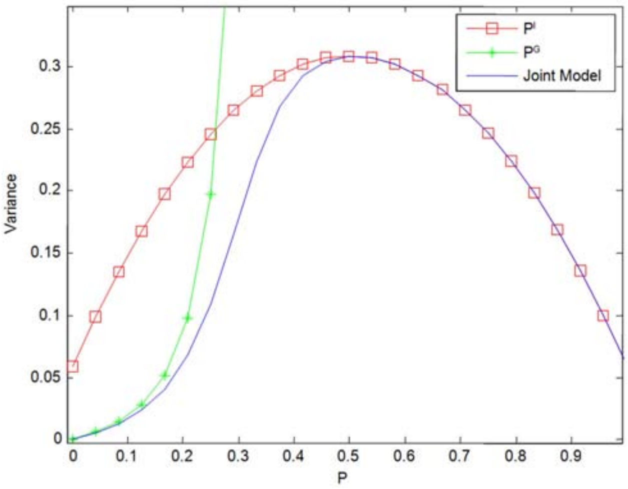

Figure 18. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.95$ *and* $k=10$.

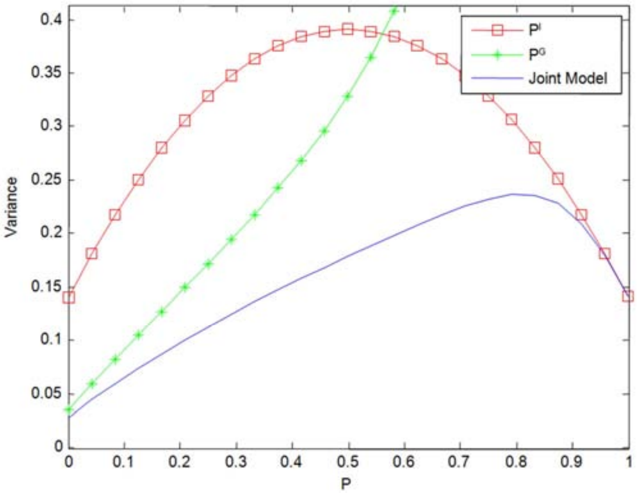

Figure 19. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.90$ *and* $k=2$.

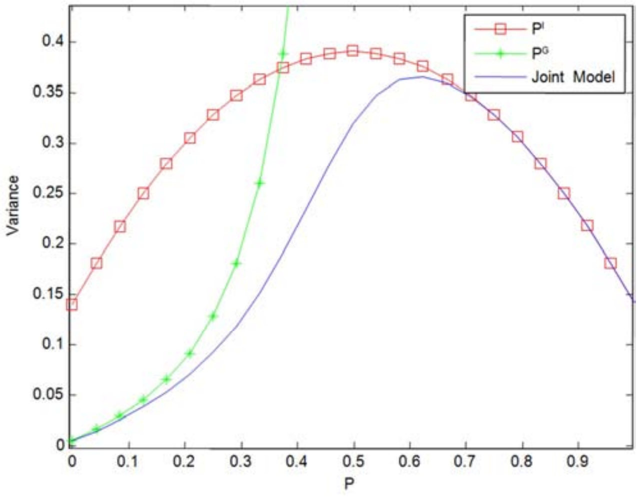

Figure 20. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.90$ *and* $k=5$.

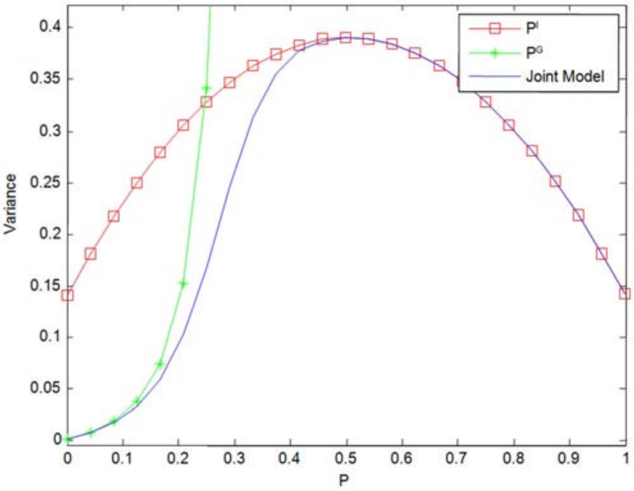

Figure 21. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.90$ *and* $k=10$.

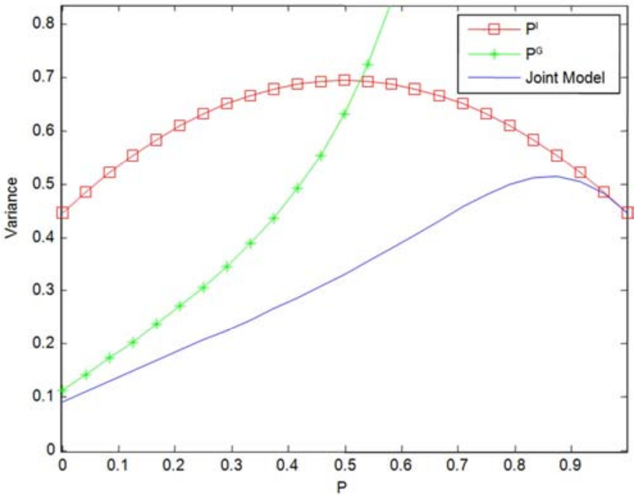

Figure 22. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.80$ *and* $k=2$.

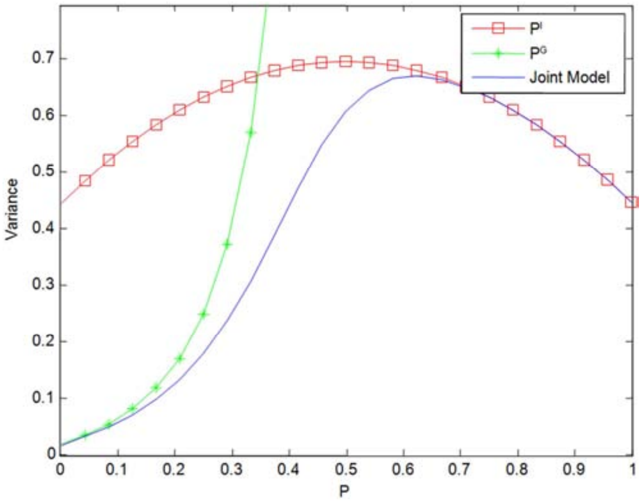

Figure 23. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.80$ *and* $k=5$.

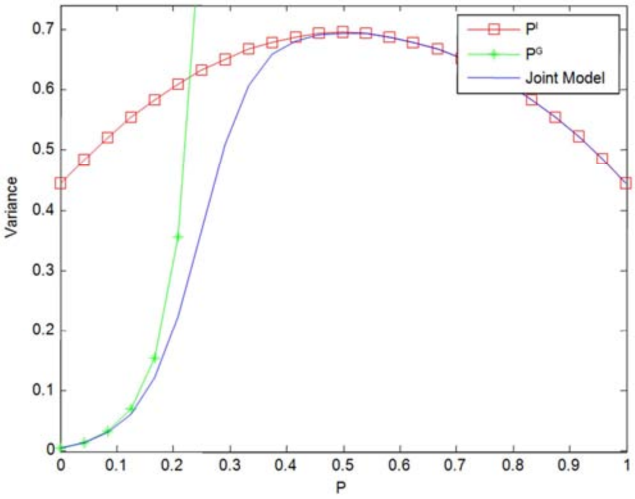

Figure 24. *A graph of* $Var(\hat{p})$ *as a function of* p *with* $\eta=\beta=0.80$ *and* $k=10$.

As seen from Figures 13 to 24 the plot of $\mathrm{var}(\hat{p}_m^I)$ is concave downwards and symmetric, maximum at approximate value of *p* equal 0.5. The $\mathrm{var}(\hat{p}_m^I)$ is unaffected by the change of the value of *k* holding specificity and sensitivity constant since the model is independent of *k*. As specificity and sensitivity of the tests increases the $\mathrm{var}(\hat{p}_m^I)$ decreases. It can also be noted that the $\mathrm{var}(\hat{p}_n^G)$ increases exponentially as the value of the parameter *p* increases from 0 to 1. As *k* increases, the $\mathrm{var}(\hat{p}_n^G)$ decreases keeping sensitivity and specificity constant while holding *k* constant, increasing sensitivity and specificity of the tests decreases the $\mathrm{var}(\hat{p}_n^G)$. The $\mathrm{var}(\hat{p}_{mle})$ increases as the value of the parameter *p* increases but thereafter it starts decreasing as *p* gets closer to 1. The $\mathrm{var}(\hat{p}_{mle})$ increases as the value of *k* increases keeping sensitivity and specificity constant while holding *k* constant, increasing sensitivity and specificity decreases the value of $\mathrm{var}(\hat{p}_{mle})$. As the value of *k* increase the plot of the $\mathrm{var}(\hat{p}_n^G)$ shifts to the left meaning the region in which the $\mathrm{var}(\hat{p}_n^G)$ is higher than the $\mathrm{var}(\hat{p}_m^I)$ decreases. As sensitivity and specificity of the tests increases the area in which $\mathrm{var}(\hat{p}_n^G)$ is higher than the $\mathrm{var}(\hat{p}_m^I)$ increases. For small values of the parameter *p*, the $\mathrm{var}(\hat{p}_{mle})$ is smaller than the $\mathrm{var}(\hat{p}_m^I)$ and $\mathrm{var}(\hat{p}_n^G)$ but is equal to the $\mathrm{var}(\hat{p}_m^I)$ for the values of *p* close to 1. The region in which the $\mathrm{var}(\hat{p}_{mle})$ is higher than the $\mathrm{var}(\hat{p}_n^G)$ increases exponentially as the value of *p* increases from 0 to 1 however the region in which it is better than $\mathrm{var}(\hat{p}_m^I)$ increases then it starts decreasing again and they are equal for the values of *p* close to 1. As the value of *k* increases, the region in which the $\mathrm{var}(\hat{p}_{mle})$ and $\mathrm{var}(\hat{p}_m^I)$ are equal increases. In general we observed that the $\mathrm{var}(\hat{p}_{mle})$ is smaller or equal to the $\mathrm{var}(\hat{p}_m^I)$ or $\mathrm{var}(\hat{p}_n^G)$ for $0 \le p \le 1$.

7. Asymptotic Relative Efficiency (ARE)

In this section, $\text{var}(\hat{\text{p}}_{mle})$, $\text{var}(\hat{\text{p}}_m^I)$ and $\text{var}(\hat{\text{p}}_n^G)$ have been compared. This is accomplished by computing asymptotic relative efficiency (*ARE*) values for various values of η, β, k and p. Let $ARE^1 = \frac{\text{var}(\hat{\text{p}}_{mle})}{\text{var}(\hat{\text{p}}_m^I)}$ and $ARE^2 = \frac{\text{var}(\hat{\text{p}}_{mle})}{\text{var}(\hat{\text{p}}_n^G)}$ then, $ARE < 1.0$ implies that the joint model is more efficient than the other two models namely P^I- and P^G-procedures.

Table 5. *The ARE of the joint model relative to P^I- and P^G-models with $\eta = \beta = 0.99$.*

p-value		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^1	0.273	0.183	0.109	0.054
	ARE^2	0.727	0.817	0.891	0.946
0.05	ARE^1	0.320	0.238	0.162	0.098
	ARE^2	0.680	0.762	0.838	0.902
0.10	ARE^1	0.336	0.260	0.189	0.136
	ARE^2	0.664	0.740	0.810	0.864
0.15	ARE^1	0.346	0.276	0.215	0.186
	ARE^2	0.654	0.724	0.785	0.814
0.20	ARE^1	0.356	0.293	0.244	0.257
	ARE^2	0.644	0.707	0.756	0.743
0.30	ARE^1	0.376	0.331	0.321	0.513
	ARE^2	0.623	0.669	0.679	0.487

Table 6. *The ARE of the joint model relative to P^I- and P^G-models with $\eta = \beta = 0.90$.*

p-value		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^1	0.213	0.115	0.051	0.018
	ARE^2	0.787	0.885	0.94	0.982
0.05	ARE^1	0.252	0.160	0.092	0.048
	ARE^2	0.748	0.840	0.908	0.951
0.10	ARE^1	0.283	0.199	0.134	0.960
	ARE^2	0.717	0.801	0.866	0.904
0.15	ARE^1	0.306	0.231	0.175	0.172
	ARE^2	0.694	0.769	0.825	0.828
0.20	ARE^1	0.325	0.261	0.223	0.304
	ARE^2	0.675	0.739	0.777	0.696
0.30	ARE^1	0.362	0.326	0.357	0.746
	ARE^2	0.638	0.674	0.643	0.254

Table 7. *The ARE of the joint model relative to P^I- and P^G-models with $\eta = \beta = 0.80$.*

p-value		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^1	0.207	0.108	0.045	0.014
	ARE^2	0.493	0.892	0.955	0.986
0.05	ARE^1	0.231	0.136	0.071	0.034
	ARE^2	0.769	0.864	0.929	0.966
0.10	ARE^1	0.257	0.169	0.107	0.077
	ARE^2	0.743	0.831	0.893	0.923
0.15	ARE^1	0.281	0.202	0.151	0.164
	ARE^2	0.719	0.798	0.849	0.836
0.20	ARE^1	0.304	0.238	0.208	0.332
	ARE^2	0.696	0.762	0.792	0.668
0.30	ARE^1	0.351	0.321	0.383	0.816
	ARE^2	0.649	0.679	0.617	0.184

Tables 5 to 7 of the computed values of *ARE* of the proposed model relative to P^I- and P^G-models reveal the same trend whereby if η and β are held constant, it is observed that as the value of k increases from 2 to 10, ARE^1 decreases for small values of p but as p increases where $p \in (0, 3]$, the ARE^1 decreases and then it starts increasing. ARE^2 increases as the value of k increases from 2 to 10 for small values of p but also as p increases where $p \in (0, 3]$ it starts decreasing. It can also be observed that holding k constant and increasing the value of p increases ARE^1 while ARE^2 decreases. As sensitivity and specificity of the tests decreases ARE^1 decreases while ARE^2 increases. It can also be noted that for the given interval of p $\{ p\varepsilon(0, 0.3] \}$ ARE^1 is less than 0.5 implying that P^I-experiment is less than 50% efficient as the proposed model while ARE^2 is more than 0.5 implying that P^G-experiment is more than 50% as efficient as the proposed model. However it is noted that the computed values of ARE are less than 1 hence the proposed joint model is more efficient than the other two existing models for the given range of p.

8. Discussion

From the study, it is found out that the curve of the Fisher information for the P^I-experiment is concave upwards, symmetric and it is not affected by change of the pooled sample size. Fisher information for the P^G-experiment is very high for small values of p and decreases exponentially as the value of p increases from 0 to 1. Increasing the pool size decreases the value of Fisher information and at the same time shifts the plot of the P^G-experiment to the left. If the pool size is assumed constant, increasing sensitivity and the specificity of the tests increases the value of the Fisher information of both P^I- and P^G-experiments.

The plot of the asymptotic variance of maximum likelihood estimator of p of the P^I-experiment ($\hat{p}_m^I$) against p is concave upwards. The $\text{var}(\hat{\text{p}}_m^I)$ is not affected by change of the pool size assuming sensitivity and specificity remains the same but treating pool size constant and increasing sensitivity and specificity of the tests decreases the variance. Similarly the graph of the asymptotic variance of maximum likelihood estimator of p of the P^G-experiment against p increases exponentially as the value of p increases from 0 to 1. The curve for the P^G-experiment shifts to the left and becomes steeper as the value of k increases from 2 to 10 holding sensitivity and specificity constant. Treating pool size constant and increasing sensitivity and specificity of the tests decreases the asymptotic variance of $\hat{p}_n^G$.

The constructed estimator is affected by change of both pool size and also sensitivity and specificity of the test kits. Increase in pool size increases the variance of the estimator holding specificity and sensitivity constant while increasing sensitivity and specificity, pool size remaining constant decreases the variance. The variance of the constructed estimator is smaller compared to the variances of one-at-a-time experiment and pooled experiment for values of

$p\epsilon[0,1]$ hence the constructed estimator is more efficient than the previous estimators especially for small values of p.

9. Conclusion

This study focused on construction of the new model for approximating the prevalence rate of a trait in a population with imperfect tests by selecting between two experiments namely P^I- and P^G-experiments. Ideally the model should select the better experiment and once the better experiment is being used, the estimator should approximate the individual maximum likelihood estimator for that experiment. From this study it can be concluded that the P^G-experiment is better than the P^I-experiment for values of p close to zero but for values of p close to 1.0 the P^I-experiment is recommended. Hence from the results of the Fisher information, asymptotic variance and *ARE,* the proposed joint model for sequentially selecting between two experiments for estimating the prevalence rate of a trait in a population with imperfect tests is more efficient than P^I- and P^G-models across the entire range of parameter values regardless of the total pool size, sensitivity and specificity of the tests.

The developed model have potential in the application of HIV testing because it gives a superior estimator of the disease prevalence without necessarily identifying the subject. The models may also be applied for use by pharmaceutical companies in discovering drugs in early stages.

Based on the constructed model, one can extend the present work to include a model with more than two experiments with misclassification. The present work can also be extended not only to approximate p but also the value of k (pool size) that will optimize group testing scenario based on the new model. A model based on cost analysis when sampling from different experiments can also be looked at when using imperfect kits.

References

[1] Brookmayer, R. (1999). Analysis of multistage pooling studies of Biological specimens for Estimating Disease Incidence and prevalence. *Biometric* 55, 608–612.

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

[3] Gastwirth, J. L., and Johnson, W. O. (1994). Screening with cost-effective quality control: Estimation of prevalence of a rare disease, preserving the anonymity of the subject by Pool-testing; Application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of statistical planning and inferences*, 22, 15–27.

[4] Hammick, P. A. and Gastwirth, J. L. (1994). Extending the applicability of estimation of prevalence of sensitive characteristics by pool testing to moderate prevalence populations. *International Statistical Review* 62, 319-331.

[5] Janis H., Connie P., and Quentin F. S. (1998). Sequentially deciding between two experiments for estimating a common success probability. *Journal of the American statistical association*. December 1998, vol 93 no 444, 1502-1511.

[6] Johnson, N. L., Kotz, S. and Wu, X. (1991). Inspection errors for attributes in quality control. *London; Chapman and Hall.*

[7] Juan, D. and Wenjun, X. (2015). Robust group testing for multiple traits with misclassifications. *Journal of Applied statistics*, vol 42 no. 10, 2115-2015.

[8] Kline, R. L., Bothus, T., Brookmeyer, R., Zeyer, S., and Quinn, T. (1989). Evaluation of human Immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of clinical microbiology,* 27, 1449-1452.

[9] Litvak, E., Tu, X. M. and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the America statistical Association,* 89, 424-434.

[10] Maheswaran, S., Haragopal, V. V., and Pandit, S. N. N, (2008). Pool-testing using block testing strategy. *Journal of statistical planning and inference* (Submitted).

[11] Manzon, O. T., Palalin, F. J. E., Dimaal, E., Balis, A. M., Samson, C., and Mitchel, S. (1992). Relevance of antibody content and test format in HIV testing of pooled sera. *AIDS,* 6, 43-48.

[12] Mundel (1984). Group-testing. *Journal of quality technology,* 16, 181-187.

[13] Nyongesa, L. K. (2011). Dual Estimation of Prevalence and Disease Incidence in Pool-Testing Strategy. *Communication in Statistics Theory and Method*, 40, 3218-3229.

[14] Nyongesa, L. K. (2004). Multistage Pool Testing Procedure (Pool screening). *Communication in Statistics-Simulation and computation*, 33, 621-637.

[15] Sobel, M. and Groll P. A., (1966). Binomial Group-Testing with an Unknown Proportion of Defectives. *American Statistical Association and American Society for Quality,* 8, 631-656.

[16] Syaywa, J. P and Nyongesa, L. K. (2010). Pool Testing with Test Errors Made Easier. *International Journal of Computational Statistics.*

[17] Tamba C. L, Nyongesa K. L., Mwangi J. W., (2012). Computational Pool-Testing Strategy. *Egerton University Journal*, 11: 51-56.

[18] Thomson, K. H. (1962). Estimation of the Population of Vectors in a Natural Population of Insects. *Biometrics*, 18, 568-578.

[19] Xie, M., Tatsuoka, K., Sacks, J and Young, S. (2001*).* Pool Testing with Blockers and Synergism. *Journal of American Statistical Association* 96, 92-102.