International Journal of Statistics and Applications 2017, 7(2): 93-106
DOI: 10.5923/j.statistics.20170702.04

# Consecutively Choosing between Three Experiments for Estimating Prevalence Rate of a Trait with Imperfect Tests

**George M. Matiri[1,*], Kennedy L. 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

**Abstract** The concept of pool testing originated with Dorfman in the context of blood testing as an economical method of testing blood samples of army inductees in order to detect the characteristic of interest. 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, expensive and is bound to errors hence pool testing procedures have been proposed to address these problems. Despite these procedures, when pool testing strategies are used using imperfect kits, there tend to be loss of sensitivity. Lost sensitivity of a test is recovered by retesting pools classified positive in the initial test. This study has developed statistical model which is used to consecutively choosing some combination of the three experiments namely: one–at-a-time, pooled testing and pooled testing with retesting of the positive pools for estimating the prevalence rate of a trait with imperfect tests. 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) is used in obtaining the estimators. The Fisher information for each of the three experiments is compared and the cut-point values where one experiment is better than the other are computed. Properties of the estimators are discussed and compared and the joint model is found to be more efficient.

**Keywords** Proportion, Proportion estimation, Group, Group testing, Cut off Value, Sensitivity, Specificity

## 1. Introduction

In many applications, units can be classified as defective or non defective. Pool testing involves pooling such units into groups or pools, testing the groups, and classifying each group as defective or non-defective. A group is defined as non-defective if non of the unit in the group contains the characteristic of interest otherwise a group is said to be defective. A group testing design has been shown to be a compelling alternative to one-at-a-time testing in many areas where rare traits are of interest. Research has shown that pool studies can be used in plant pathology, genetics and reduction of cost in early stages of drug discovery (Hammick and Gastwirth, 1994; Swallow, 1985; Xie *et al.*, 2001). Pool testing has also been applied in screening the population for the presence of HIV antibody (Kline *et al*., 1989 and Manzon *et al*., 1992). Computational testing that focuses on classifying subjects has been developed Maheswaran *et al*. (2008).

* Corresponding author:
gemuwa@gmail.com (George M. Matiri)
Published online at http://journal.sapub.org/statistics


Recently more research work are focused on estimating the rate of trait. Thomson (1962) considered estimation problem using pool testing which was later considered by Brookmayer (1999) by introducing errors. Sufficiently accurate estimate of the prevalence can be obtained from testing pooled samples as demonstrated by Hammick and Gastwirth (1994) and their procedure provides greater protection of respondent's identity which can be useful in improving the response rate. On the same year, Gastwirth and Johnson (1994) used pool testing to estimate HIV prevalence cost-effectively. Hardwick *et al*., (1998) considered sequentially deciding between two experiments for estimating a common success prevalence rate where he considered the individual Bernoulli trials or the product of $k$ individual independent Bernoulli trials. Nyongesa (2011) 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, 2010; Tamba *et al*., 2012).

Benefits from group testing depend on size of the pools. Swallow (1985) showed that large group sizes can lead to

estimators with enormous bias. In addition, there are biological issues to be considered. For example in HIV testing, enzyme-linked immunosorbent assay tests (ELISA) are commonly used in screening experiments to detect the presence of the virus. However, sensitivity levels for such tests are known to be poor when many blood samples are pooled together. With many of the standard enzyme-linked immunosorbent assay tests (ELISA), group sizes of up to 15 are typically used without experiment dilution effects (Behets et al., 1990; Cahoon-Young et al., 1989; Kline et al., 1989; Tu al., 1995). Recent studies have provided algorithm for the computation of pool sizes (Ding and Xiong, 2015).

This study has focused on estimation of proportions and in particular a better estimation of the prevalence rate. The essence of the study, is to device a method of selecting between three experiments namely:

i) individual testing of items of a population with a view to estimating prevalence rate *p*, with misclassification, this experiment will be denoted by ${}^{1}E$,
ii) pool testing experiment as proposed by Dorfman (1943) but with errors in inspection, this experiment will be denoted by ${}^{2}E$ and
iii) estimating the prevalence rate of the characteristic of interest by retesting the pools declared positive in the first pool test and this experiment will be denoted by ${}^{3}E$.

The rest of the paper is arranged as follows: in Section 2 we shall develop the models and formula for calculating Fisher information, in Section 3 we shall plot the graphs Fisher information against the value of *p*. In Section 4 we shall compute the cut off values. In Section 5 we shall develop the joint model and in section 6 we shall compute the MLE of p of the joint model. In section 7 we shall compare the variances of the models by plotting their graphs. In Section 8 we shall compute the *ARE* values and in section 9 we shall have conclusion of the study.

# 2. The Model

The model have been split into three, that is ${}^{1}E$-, ${}^{2}E$-, and ${}^{3}E$-experiments. *r*, *s* and *t* have been assumed to be the total number of observations from ${}^{1}E$-, ${}^{2}E$-, and ${}^{3}E$-experiments respectively. In typical sequential allocation problems, different experiments give information about different parameters. However, in this study, the three experiments *i.e* ${}^{1}E$-, ${}^{2}E$-, and ${}^{3}E$-experiments give information about the same parameter (*p*), although one experiment have given more information than the other two experiments under consideration depending on the actual value of the parameter, pool size, sensitivity and specificity of the tests. For simplicity, it has been assumed that individual units being pooled are independent and identically distributed Bernoulli random variables.

### 2.1. The ${}^{1}E$-experiment

If ${}^{1}E$-experiment is to be used to estimate the prevalence rate $p$, and if $X_{1i}$ for $i = 1,...,\mathrm{r}$ is a sequence of identically independent distributed random variable, then $X_{1i} \sim \mathrm{B}ernouli(\lambda_1)$ where $\lambda_1$ is the probability of declaring an individual as positive defined by the relation, $\lambda_1 = \eta p + (1-\beta)(1-\mathrm{p})$ given $\eta$ and $\beta$ are sensitivity and specificity of the tests respectively.

For a single experiment, the probability density function is

$$f(x_{1i}, p \mid \eta, \beta) = \lambda_1^{x_{1i}} (1-\lambda_1)^{1-x_{1i}} \tag{1}$$

The Fisher information denoted by $I_{x_1}({}^{1}\mathrm{E})$, on the prevalence rate $p$ contained in a single observation of the ${}^{1}E$-experiment is

$$I_{x_1}({}^{1}\mathrm{E}) = \frac{(\eta+\beta-1)^2}{\lambda_1(1-\lambda_1)} \tag{2}$$

easily obtained by MLE method from (1). If $r$ observations from only the ${}^{1}E$-experiment are used to estimate $p$, then the maximum likelihood estimator of *p*, denoted by ${}^{1}_{r}\hat{p}$, is

$${}^{1}_{r}\hat{p} = \frac{\beta - 1 + \dfrac{\sum_{i=1}^{r} x_{1i}}{r}}{\eta + \beta - 1}. \tag{3}$$

The asymptotic variance of ${}^{1}_{r}\hat{p}$ is obtained from (2) which yields

$$\mathrm{var}({}^{1}_{r}\hat{p}) == \frac{\lambda_1(1-\lambda_1)}{(\eta+\beta-1)^2}. \tag{4}$$

### 2.2. The ${}^{2}E$-experiment

The ${}^{2}E$-experiment involves 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 is denoted by $\lambda_2 = \eta(1-(1-\mathrm{p})^k) + (1-\beta)(1-\mathrm{p})^k$. If $X_{2j}$ denote a sequence of identically independent distributed random variables for $j = 1,...,s$, then $X_{2j} \sim Bernouli(\lambda_2)$. For a single experiment equivalently the probability density function is

$$f(x_{2j}, p \mid \eta, \beta, k) = \lambda_2^{x_{2j}} (1-\lambda_2)^{1-x_{2j}} \tag{5}$$

and the Fisher information denoted by $I_x(^2\mathrm{E})$ contained in a single observation of the $^2E$-experiment is

$$I_x(^2\mathrm{E}) = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\lambda_2(1-\lambda_2)} \quad (6)$$

Suppose there are $s$ pools for the $^2E$-experiment each of size $k$, available for estimating $p$ and suppose $X_{2j}$ pool test positive on the test, then the maximum likelihood estimator of $p$, denoted by ${}^2_s\hat{p}$, is

$${}^2_s\hat{p} = 1-\left(\frac{\eta-\frac{\sum_{j=1}^{s}x_{2j}}{s}}{\eta+\beta-1}\right)^{\frac{1}{k}}. \quad (7)$$

Noted is Thompson (1962) maximum likelihood estimator (*MLE)* of $p$ *i.e*

$\hat{p} = 1-\left(1-\frac{\sum_{j=1}^{n}x_{2j}}{s}\right)^{\frac{1}{k}}$ is a special case of (7).

The asymptotic variance of ${}^2_s\hat{p}$ is

$$\mathrm{var}({}^2_s\hat{p}) = \frac{\lambda_2(1-\lambda_2)}{\mathrm{k}^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}. \quad (8)$$

### 2.3. The $^3E$-experiment

The $^3E$-experiment involves retesting of the pools declared positive in the first pool test inorder to approximate the prevalence rate. Retesting of already tested pools reduce misclassification (Nyongesa, 2011). In this experiment, the probability of declaring a pool of size $k$ positive is denoted by $\lambda_3$ where $\lambda_3 = \eta^2(1-(1-\mathrm{p})^k)+(1-\beta)^2(1-\mathrm{p})^k$. If $X_{3z}$ denote a sequence of identically independent distributed random variables for $z = 1,...,\mathrm{t}$, then $X_{3z} \sim Bernouli(\lambda_3)$. For a single experiment the probability density function is

$$f(x_{3z}, p \mid \eta, \beta, k) = (\lambda_3)^{x_{3z}}(1-\lambda_3)^{1-x_{3z}}. \quad (9)$$

Similarly the Fisher information contained in a single observation of the $^3E$-experiment is

$$I_x(^3\mathrm{E}) = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\lambda_3(1-\lambda_3)}. \quad (10)$$

If $t$ observations from the $^3E$ -experiment are used to estimate $p$ and $X_{3z}$ pool tests positive, then the maximum likelihood estimator of $p$, denoted by ${}^3_t\hat{p}$, is

$${}^3_t\hat{p} = 1-\left(\frac{\eta^2-\frac{\sum_{z=1}^{t}x_{3z}}{t}}{\eta^2-(1-\beta)^2}\right)^{\frac{1}{k}}. \quad (11)$$

Equivalently the asymptotic variance of ${}^3_t\hat{p}$ obtained from (10) is

$$\mathrm{var}({}^3_t\hat{p}) = \frac{\lambda_3(1-\lambda_3)}{\mathrm{k}^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}. \quad (12)$$

## 3. Comparison of $I_x(\cdot)$ of $^1E$-, $^2E$- and $^3E$-experiments

This study compares the Fisher information of $^1E$-, $^2E$- and $^3E$-experiments in this section by plotting the graphs of $I_x(\cdot)$ of $^1E$-, $^2E$- and $^3E$-experiments for values of $k = 2, 5, 10;\ \eta = \beta = 0.9, 0.8$ versus $p$ :

As seen from Figures 1 to 6, the Fisher information for the $^1E$ -experiment is independent of the pool size hence it is not affected by change of the value of $k$ . For the $^2E$ - and $^3E$-experiments, the Fisher information is very high for small values of $p$ and it approaches zero as $p$ increases. As sensitivity and specificity of the test kits increases, the gap between the Fisher information of the $^2E$ - and $^3E$ -experiments shrinks. Holding $k$ constant, increasing sensitivity and specificity of the test kits, the region at which the Fisher information of the $^1E$- and $^3E$-experiments is better shrinks while for $^2E$-experiment increases. Similarly as $k$ increases the region in which the Fisher information of $^1E$ - and $^3E$-experiments is better decreases while that of $^2E$-experiment increases. From Figures 1 to 6 it can be concluded that the $^3E$-experiment is better than $^1E$- and $^2E$-experiments for values of $p$ relatively small, for values of $p$ relatively large, the $^1E$-experiment is better and the $^2E$-experiment is better for some values of $p$ between *0* and *1*. It is also noted that the region in which one experiment is better than the other experiments depends on sensitivity, specificity and the pool size.

## 4. Computation of Cut-Point Values

The cut-point value is defined as the value of $p$ at the point where the Fisher information of one of the experiment surpasses the Fisher information of the other experiment while comparing any two of the $^1E$-, $^2E$- and $^3E$-experiments. If $\xi_{ij}$ is the cut-point value, then $\xi_{ij}$ is a unique root in [0, 1] of the equation $I_x(^i\mathrm{E}) = I_x(^j\mathrm{E})$ for $i, j = 1, 2, 3$, $i \neq j$ and $\xi_{ij} = \xi_{ji}$.

### 4.1. Computation of Cut-Point Values of $I_x(^1\mathrm{E})$ and $I_x(^2\mathrm{E})$

In this section the cut-point values of $^1E$- and $^2E$-experiments are computed by equating the Fisher information of the two experiments. Therefore equating (2) and (6) and simplifying yields

$$\hat{\lambda}_2(1-\hat{\lambda}_2)(1-\mathrm{p})^2 - \mathrm{k}^2(1-\mathrm{p})^{2k}\hat{\lambda}_1(1-\hat{\lambda}_1) = 0. \quad (13)$$

(13) has no solution in closed form therefore the equation is solved iteratively using an R code that we developed which is presented in Appendix A.

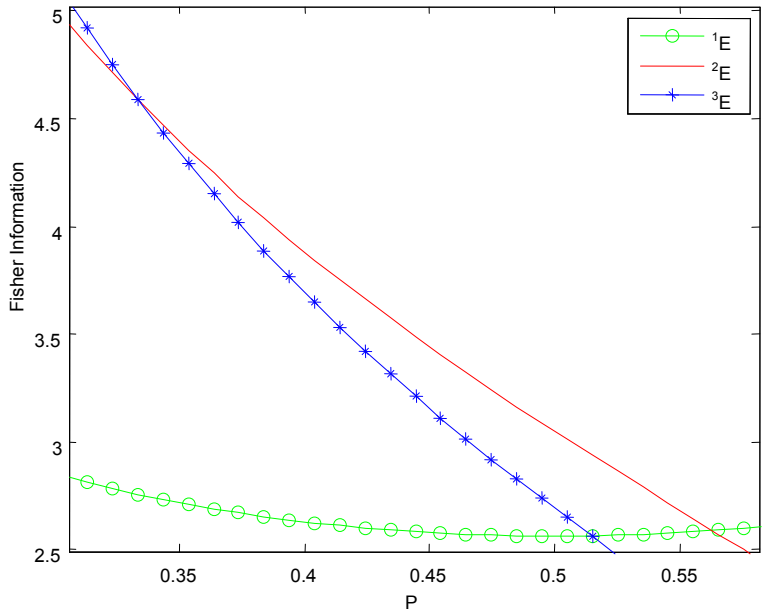


**Figure 1.** A plot of Fisher information against the value of $p$ with $\eta = \beta = 0.90$ and $k = 2$

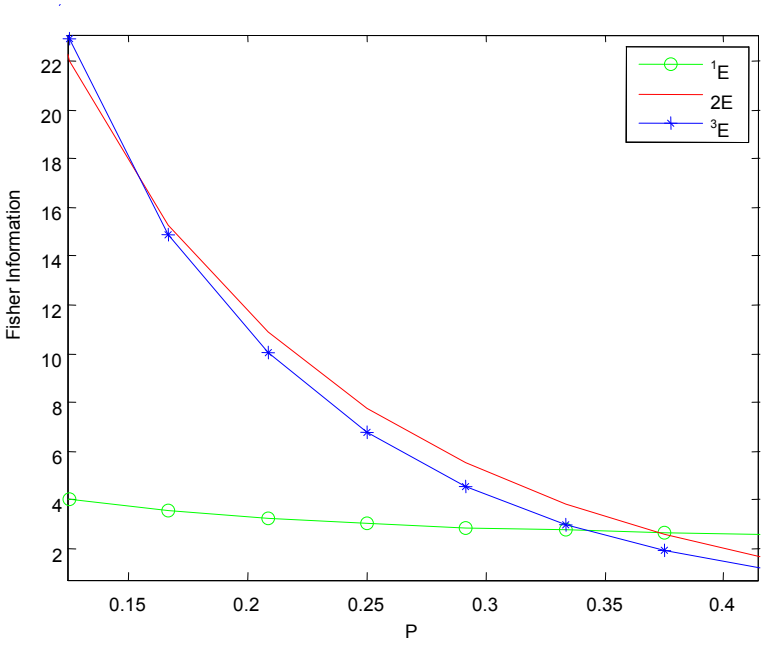


**Figure 2.** A plot of Fisher Information against the value of $p$ with $\eta = \beta = 0.90$ and $k = 5$

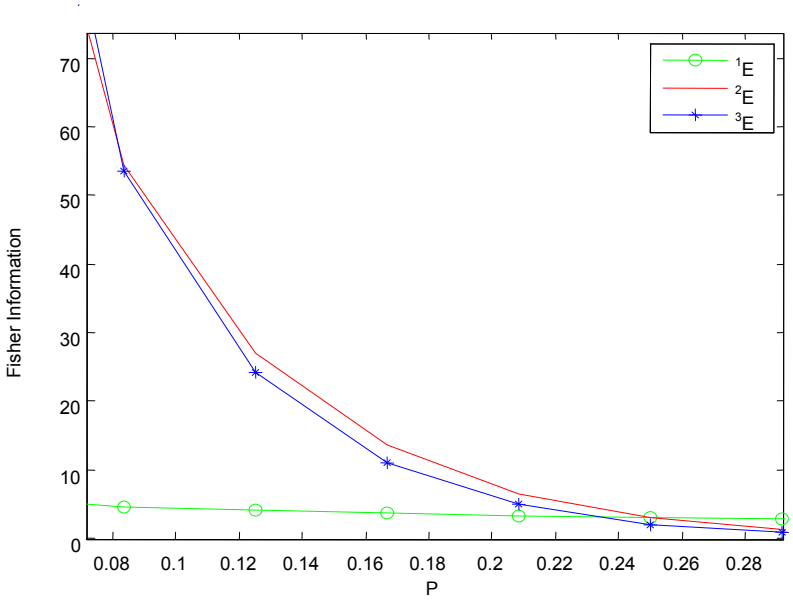


**Figure 3.** A plot of Fisher information against the value of $p$ with $\eta = \beta = 0.90$ and $k = 10$

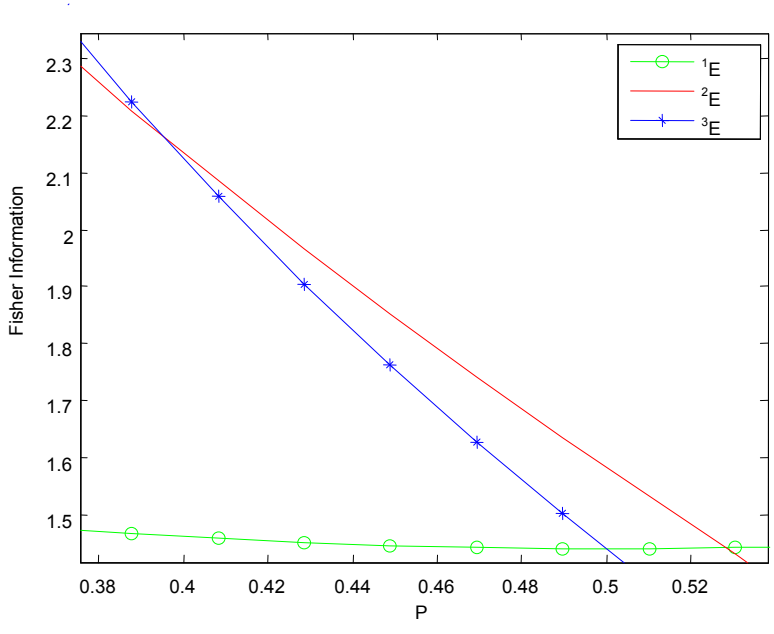


**Figure 4.** A plot of Fisher Information against the value of $p$ with $\eta = \beta = 0.80$ and $k = 2$

### 4.2. Computation of Cut-Point Values of $I_x(^1\mathrm{E})$ and $I_x(^3\mathrm{E})$

The cut-point values of $^1E$- and $^3E$-experiments are computed in this section by equating (2) and (10) which yields

$$\hat{\lambda}_3(1-\hat{\lambda}_3)(1-\mathrm{p})^2 - k^2(1-\mathrm{p})^{2k}(\eta-\beta+1)^2\hat{\lambda}_1(1-\hat{\lambda}_1) = 0 \quad (14)$$

after simplifying. Similarly (14) is solved iteratively using an R code that we developed presented in Appendix B.

### 4.3. Computation of Cut-Point Values of $I_x(^2\mathrm{E})$ and $I_x(^3\mathrm{E})$

Similarly the cut-point values of $^2E$- and $^3E$-experiments are computed by equating (6) and (10) which yields

$$\hat{\lambda}_3(1-\hat{\lambda}_3) - (\eta-\beta+1)^2\hat{\lambda}_2(1-\hat{\lambda}_2) = 0 \quad (15)$$

after simplifying. Equivalently (15) is solved iteratively using an R code developed which is presented in Appendix C.

For various values of $k$, $\eta$ and $\beta$, the values $\xi_{ij}$, the roots of (13), (14) and (15) are given in Table 1.

From Table 1 it can be noted that the cut-point values are sensitive to $k$ (pool size). As specificity and sensitivity increases the cut point value between ${}^{1}E$- and ${}^{3}E$-experiments increases while that between ${}^{2}E$- and ${}^{3}E$-experiments decreases. It is observed from Table 1 that as the pool size ($k$) increases, keeping $\eta$ and $\beta$ the same, the region in which the ${}^{3}E$-experiment is better than the ${}^{2}E$- and ${}^{1}E$- shrinks. Increase in sensitivity and specificity leads to decrease of the area which ${}^{1}E$- and ${}^{3}E$ -experiments is better than ${}^{2}E$-experiment. In general the exact values of $p$ where one experiment is better than the others depends on the pool size ($k$), sensitivity ($\eta$) and specificity ($\beta$) of the tests.

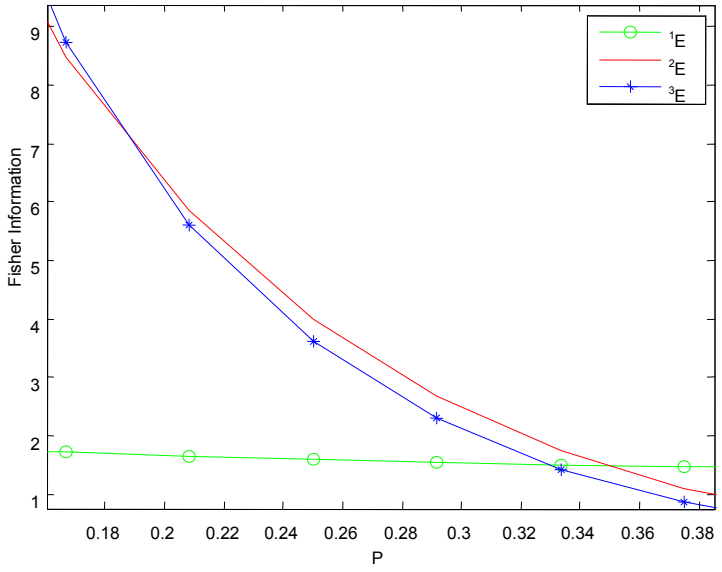


**Figure 5.** A plot of Fisher Information against the value of $p$ with $\eta = \beta = 0.80$ and $k = 5$

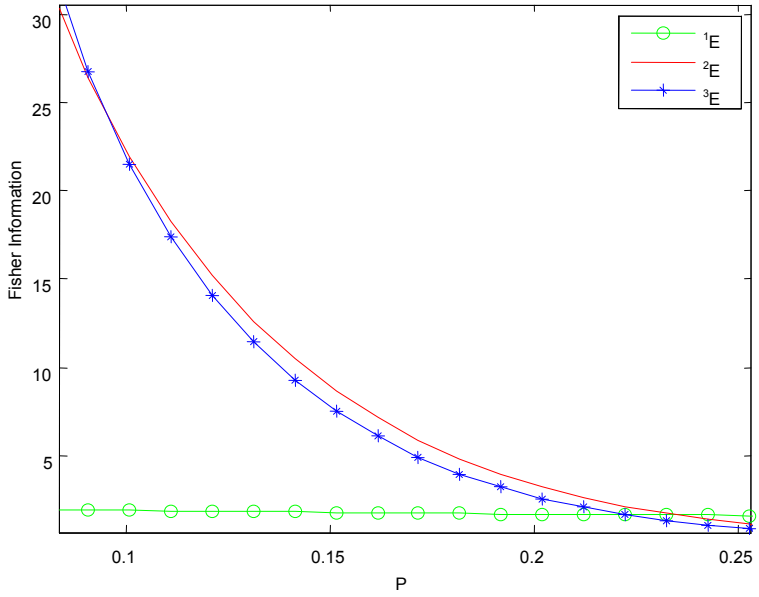


**Figure 6.** A plot of Fisher Information against the value of $p$ with $\eta = \beta = 0.80$ and $k = 10$

**Table 1.** Cut-point point values for various values of $k$, $\eta$ *and* $\beta$

| k | $\eta=\beta=0.80$ | | | $\eta=\beta=0.90$ | | |
|---|---|---|---|---|---|---|
| | $\xi_{12}$ | $\xi_{13}$ | $\xi_{23}$ | $\xi_{12}$ | $\xi_{13}$ | $\xi_{23}$ |
| 2 | 0.528 | 0.500 | 0.394 | 0.563 | 0.515 | 0.334 |
| 3 | 0.446 | 0.422 | 0.284 | 0.477 | 0.438 | 0.237 |
| 5 | 0.348 | 0.329 | 0.182 | 0.371 | 0.342 | 0.150 |
| 10 | 0.234 | 0.222 | 0.095 | 0.248 | 0.229 | 0.078 |
| 15 | 0.181 | 0.172 | 0.065 | 0.190 | 0.177 | 0.053 |
| 20 | 0.150 | 0.142 | 0.049 | 0.156 | 0.145 | 0.040 |
| 50 | 0.078 | 0.075 | 0.020 | 0.080 | 0.075 | 0.016 |

For example if $\eta=\beta=0.80$, $k$ = 3 and $N$ tests are available, then the allocation that maximizes the information about $p$ is:

$$N=\begin{cases} observe\ all\ {}^3E\ if\ \ p<0.283 \\ observe\ all\ {}^2E\ if\ \ 0.283<p<0.446 \\ observe\ all\ {}^1E\ if\ \ p>0.446 \\ arbitrary\ {}^2E\ or\ {}^3E\ if\ \ p=0.283 \\ arbitrary\ {}^1E\ or\ {}^2E\ if\ \ p=0.446 \end{cases}$$

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

$$N=\begin{cases} observe\ all\ {}^3E\ if\ \ p<\xi_{23} \\ observe\ all\ {}^2E\ if\ \ \xi_{23}<p<\xi_{12} \\ observe\ all\ {}^2E\ if\ \ p>\xi_{12} \\ arbitrary\ {}^2E\ or\ {}^3E\ if\ \ p=\xi_{23} \\ arbitrary\ {}^1E\ or\ {}^2E\ if\ \ p=\xi_{12} \end{cases}$$

Note also that the region where one experiment is better than the other depends on the unknown parameter $p$, hence 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-point value.

## 5. The Joint Model

If $r$, $s$ and $t$ are the total number of observations from ${}^1E$-, ${}^2E$- and ${}^3E$-experiment respectively, then the joint probability density function of the random variables $X_{1i}$, $X_{2j}$ and $X_{3z}$ from the ${}^1E$-, ${}^2E$- and ${}^3E$-experiments respectively is a multinomial probability density function. The joint probability density function is given by the product of their respective density functions, since the random variables are assumed to be independent, therefore

$$f(\underline{x},\underline{p}\,|\,k,\eta,\beta)=\lambda_1^{x_{1i}}(1-\lambda_1)^{1-x_{1i}}\times\lambda_2^{x_{2j}}(1-\lambda_2)^{1-x_{2j}}\times\lambda_3^{x_{3z}}(1-\lambda_3)^{1-x_{3z}} \tag{16}$$

The joint likelihood function of (16) is

$$\mathrm{L}(\underline{x},\underline{p}/\mathrm{k},\eta,\beta)\propto\left[\lambda_1\right]^{\sum_{i=1}^{r}x_{1i}}\left[1-\lambda_1\right]^{r-\sum_{i=1}^{r}x_{1i}}\times\left[\lambda_2\right]^{\sum_{j=1}^{s}x_{2j}}\left[1-\lambda_2\right]^{s-\sum_{j=1}^{s}x_{2j}}\times\left[\lambda_3\right]^{\sum_{z=1}^{t}x_{3z}}\left[1-\lambda_3\right]^{t-\sum_{z=1}^{t}x_{3z}} \tag{17}$$

Taking logarithm on both sides of (17) and differentiating with respect to $q$ yields

$$\frac{\mathrm{d}\log L(.)}{dq}=\frac{\sum_{i=1}^{r}x_{1i}-\mathrm{m}\lambda_1}{\lambda_1(1-\lambda_1)}\frac{d\lambda_1}{dq}+\frac{\sum_{j=1}^{s}x_{2j}-n\lambda_2}{\lambda_2(1-\lambda_2)}\frac{d\lambda_2}{dq}+\frac{\sum_{z=1}^{t}x_{3z}-r\lambda_3}{\lambda_3(1-\lambda_3)}\frac{d\lambda_3}{dq} \tag{18}$$

where $\frac{d\lambda_1}{dq}=1-\eta-\beta$, $\frac{d\lambda_2}{dq}=kq^{k-1}\left(1-\eta-\beta\right)$ and $\frac{d\lambda_3}{dq}=kq^{k-1}\left((1-\beta)^2-\eta^2\right)$.

Equating (18) to zero leads to

$$\frac{\sum_{i=1}^{r}x_{1i}-r\lambda_1}{\lambda_1(1-\lambda_1)}\frac{d\lambda_1}{dq}+\frac{\sum_{j=1}^{s}x_{2j}-s\lambda_2}{\lambda_2(1-\lambda_2)}\frac{d\lambda_2}{dq}+\frac{\sum_{z=1}^{t}x_{3z}-t\lambda_3}{\lambda_3(1-\lambda_3)}\frac{d\lambda_3}{dq}=0. \tag{19}$$

The only variable in (19) is $q$. Hence

$$f(\mathrm{q})=\frac{\sum_{i=1}^{r}x_{1i}-r\lambda_1}{\lambda_1(1-\lambda_1)}\frac{d\lambda_1}{dq}+\frac{\sum_{j=1}^{s}x_{2j}-s\lambda_2}{\lambda_2(1-\lambda_2)}\frac{d\lambda_2}{dq}+\frac{\sum_{z=1}^{t}x_{3z}-t\lambda_3}{\lambda_3(1-\lambda_3)}\frac{d\lambda_3}{dq} \tag{20}$$

is a function of $q$ which can be solved iteratively. The value of $q$, computed from (20) is denoted by $\hat{q}_{mle}$, hence the maximum likelihood estimator of $p$, denoted by $\hat{p}_{mle}$, of the joint model is $\hat{p}_{mle}=1-\hat{q}_{mle}$. The asymptotic variance of

$\hat{p}_{mle}$ is obtained by solving $\left\{E\left(\frac{-d^2 \log f(\cdot)}{dp^2}\right)\right\}^{-1}$ where $f(\cdot)$ is the joint probability density function given by (16). Therefore

$$\text{var}(\hat{p}_{mle}) = \frac{\hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_3(1-\hat{\lambda}_1)(1-\hat{\lambda}_2)(1-\hat{\lambda}_3)}{\zeta} \tag{21}$$

where

$$\zeta = r(\eta+\beta-1)^2\hat{\lambda}_2\hat{\lambda}_3(1-\hat{\lambda}_2)(1-\hat{\lambda}_3) + sk^2(1-p)^{2k-2}(\eta+\beta-1)^2\hat{\lambda}_1\hat{\lambda}_3(1-\hat{\lambda}_1)(1-\hat{\lambda}_3)$$
$$+ tk^2(1-p)^{2k-2}(\eta^2-(1-\beta)^2)^2\hat{\lambda}_1\hat{\lambda}_2(1-\hat{\lambda}_1)(1-\hat{\lambda}_2)$$

## 6. Estimator of Prevalence Rate, Its Variance and Confidence Interval

The maximum likelihood estimator $\hat{p}$ of the prevalence rate of the joint model, the variance and 95% Wald-type confidence interval of the MLE for values of $k = 5, 10$ and $\eta = \beta = 80\%, 90\%$ are computed in this section.

From Tables 2 and 3 it is observed that the maximum likelihood estimators of the prevalence rate are very close to the actual value which were 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 $\alpha$ is the level of significance and it is noted from Tables 2 and 3 that the actual value is within the limits.

**Table 2.** Maximum likelihood estimator, variance and Confidence interval for different values of $p$ for $\eta = \beta = 80\%$ and $k = 5, 10$

| | $p$ | $\hat{p}$ | var($\hat{p}$) | 95% *CI* |
|---|---|---|---|---|
| $k = 5$ | 0.01 | 0.01566 | $6.8293 \times 10^{-5}$ | -0.00868, 0.03999 |
| | 0.05 | 0.06397 | $1.7588 \times 10^{-4}$ | 0.01600, 0.11193 |
| | 0.10 | 0.10386 | $2.8551 \times 10^{-4}$ | 0.04407, 0.16366 |
| | 0.15 | 0.17263 | $5.5743 \times 10^{-4}$ | 0.09856, 0.24671 |
| | 0.30 | 0.33119 | $2.1054 \times 10^{-3}$ | 0.23895, 0.42344 |
| $k = 10$ | 0.01 | 0.01745 | $2.8592 \times 10^{-5}$ | -0.00821, 0.04312 |
| | 0.05 | 0.03052 | $4.5378 \times 10^{-5}$ | -0.00319, 0.06428 |
| | 0.10 | 0.08585 | $1.6443 \times 10^{-4}$ | 0.03094, 0.14076 |
| | 0.15 | 0.12212 | $3.2699 \times 10^{-4}$ | 0.05794, 0.18630 |
| | 0.30 | 0.28662 | $4.2189 \times 10^{-3}$ | 0.19800, 0.37525 |

## 7. Comparison of Variances

In this section, the graphs of the variance of $p$ of $^1E$-, $^2E$- and $^3E$-experiments and joint model for various values of $k, \eta$ and $\beta$ versus $p$ are plotted for comparison purposes.

It is observed from Figures 7 to 12 that:

i) var($^3_t\hat{p}$) is smaller than var($^1_r\hat{p}$) and var($^2_s\hat{p}$) for values of $p$ close to *0*,

ii) for values of $p$ close to *1* the var($^1_r\hat{p}$) is smaller than var($^3_t\hat{p}$) and var($^2_s\hat{p}$) while

iii) for some values of $p$ between *0* and *1* the var($^2_s\hat{p}$) is smaller than the variance of the other two models.

It is also noted that holding sensitivity and specificity constant and increasing the value of $k$ from 2 to 10, makes the area in which the var($^3_t\hat{p}$) is smaller than the variance of the other models shrinks. Increasing sensitivity and specificity of the tests shrinks the area between var($^2_s\hat{p}$) and var($^3_t\hat{p}$). It is also observed from Figures 7 to 12 that the variance of the joint model ( var($\hat{p}_{mle}$) ) is smaller than the variance of $p$ of the $^1E$-, $^2E$- and $^3E$-models. Hence the joint model is more reliable compared to the other three models.

**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($\hat{p}$) | 95% *CI* |
|---|---|---|---|---|
| $k = 5$ | 0.01 | 0.02017 | $3.7091 \times 10^{-5}$ | -0.00738, 0.04772 |
| | 0.05 | 0.05229 | $8.2334 \times 10^{-5}$ | 0.00866, 0.09592 |
| | 0.10 | 0.09213 | $1.4452 \times 10^{-4}$ | 0.03544, 0.14881 |
| | 0.15 | 0.16454 | $2.9124 \times 10^{-4}$ | 0.09187, 0.23721 |
| | 0.30 | 0.29088 | $7.6689 \times 10^{-4}$ | 0.20187, 0.37990 |
| $k = 10$ | 0.01 | 0.00971 | $9.3065 \times 10^{-6}$ | -0.00951, 0.02893 |
| | 0.05 | 0.04671 | $4.1100 \times 10^{-5}$ | 0.00535, 0.08807 |
| | 0.10 | 0.10616 | $1.3268 \times 10^{-4}$ | 0.04578, 0.16653 |
| | 0.15 | 0.13960 | $2.2985 \times 10^{-4}$ | 0.07168, 0.20753 |
| | 0.30 | 0.27932 | $1.8171 \times 10^{-3}$ | 0.19139, 0.36726 |

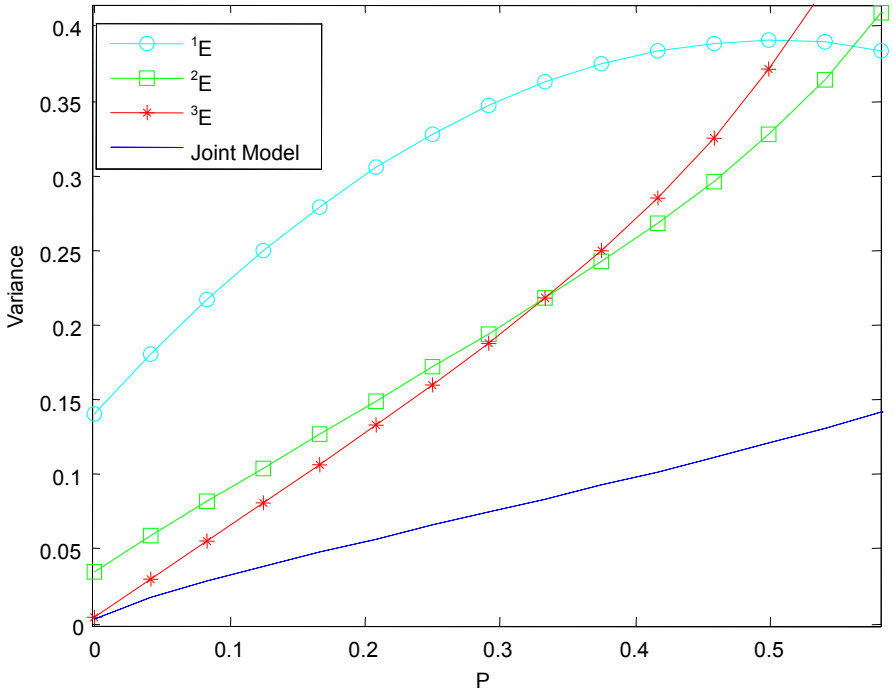


**Figure 7.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.90$ and $k = 2$

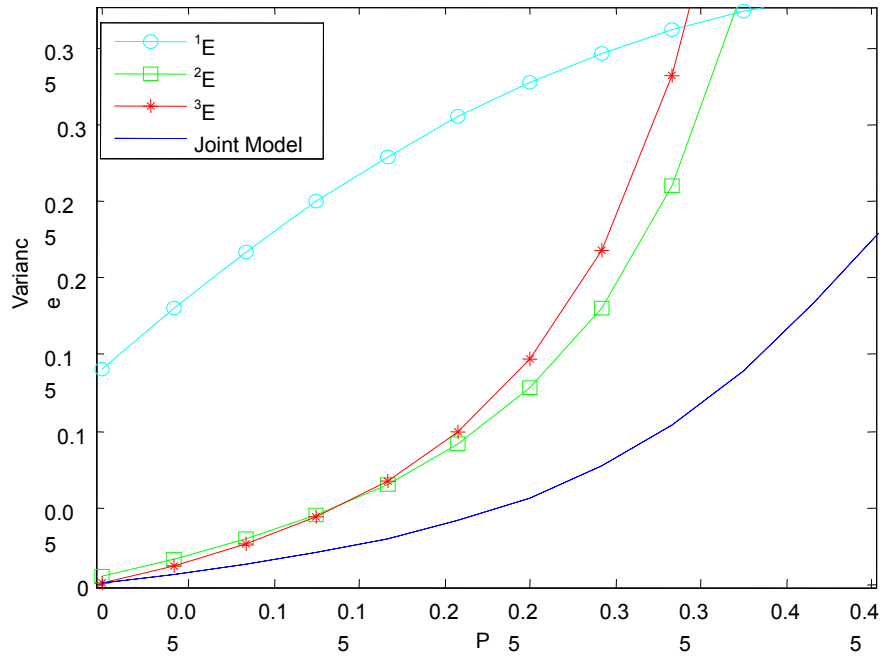


**Figure 8.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.90$ and $k = 5$

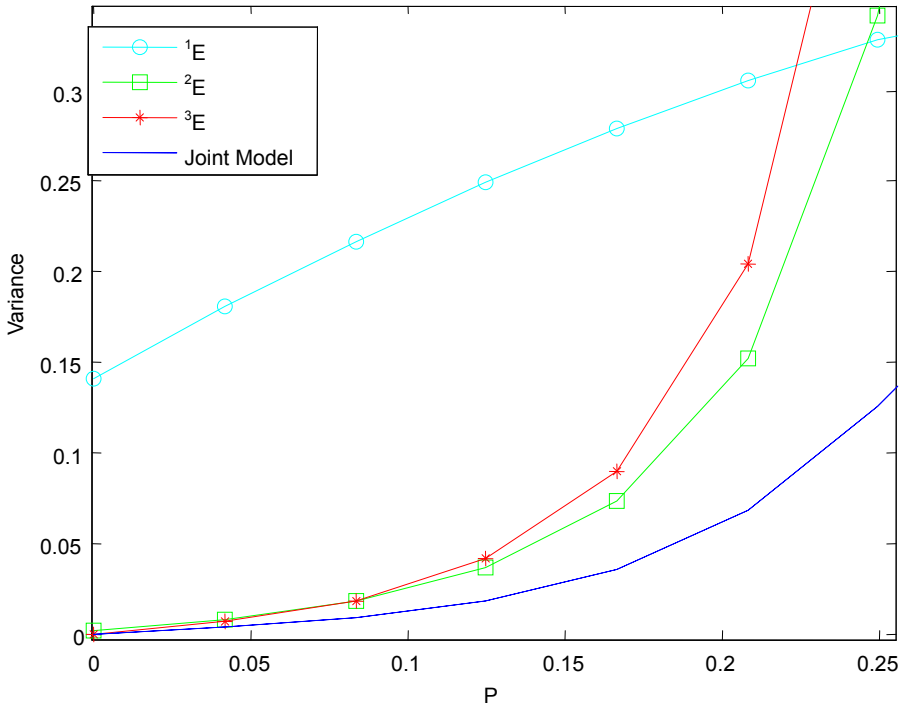


**Figure 9.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.90$ and $k = 10$

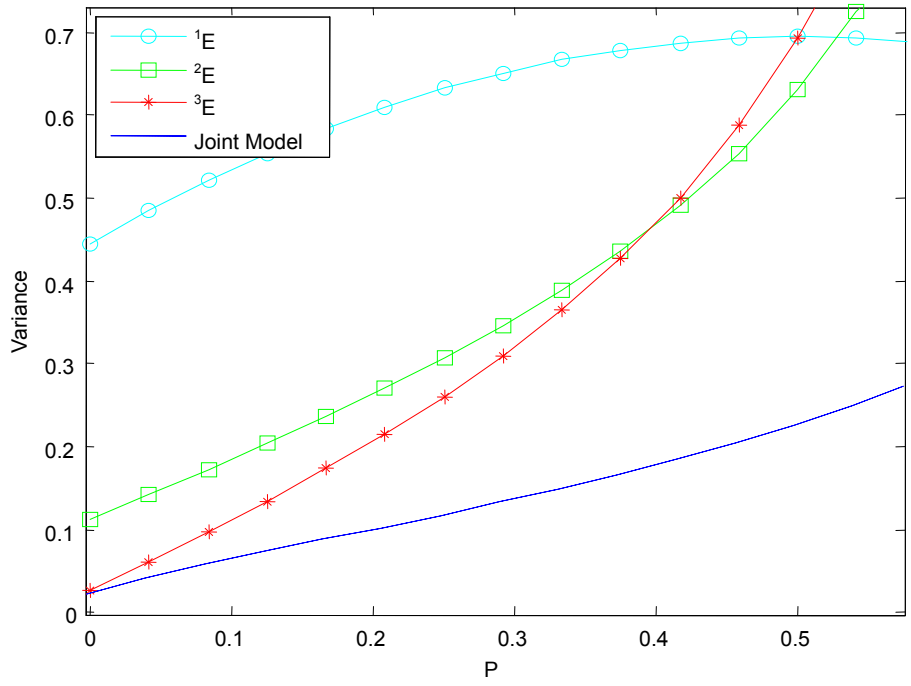


**Figure 10.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.80$ and $k = 2$

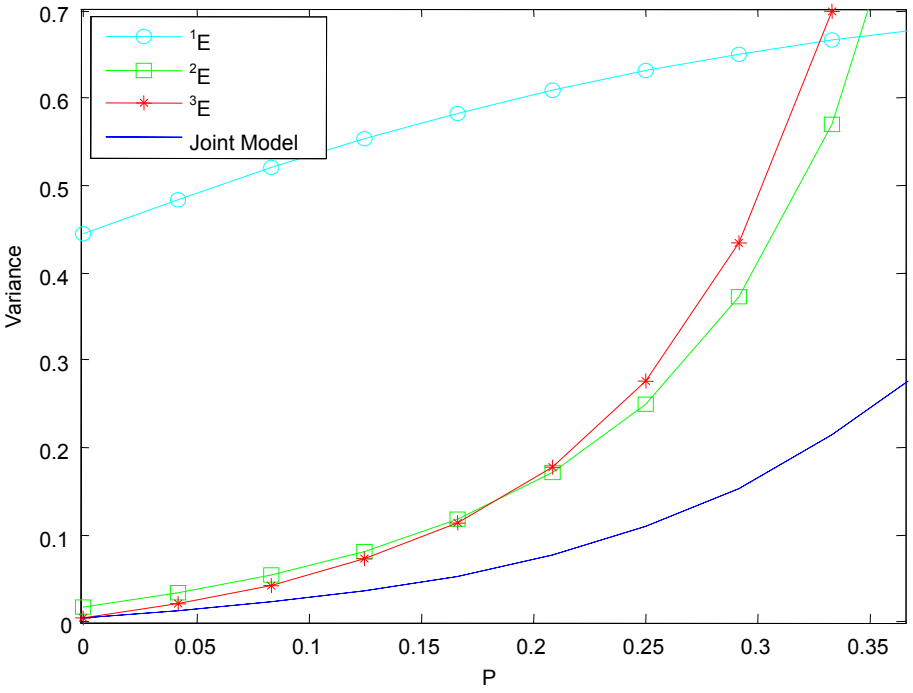


**Figure 11.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.80$ and $k = 5$

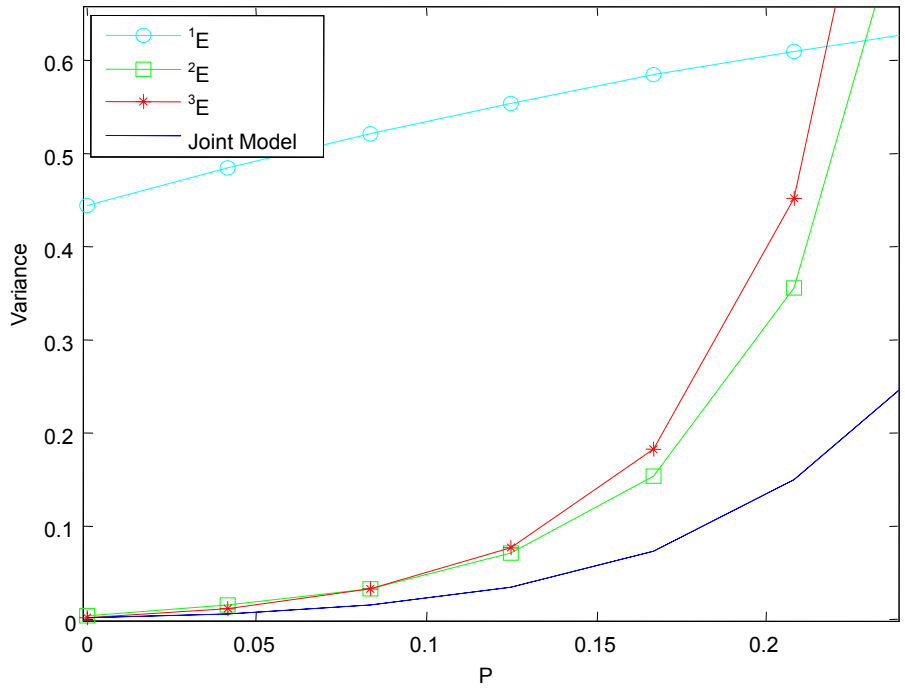


**Figure 12.** A graph of $Var(\hat{p})$ as a function of $p$ with $\eta = \beta = 0.80$ and $k = 10$

# 8. Asymptotic Relative Efficiency (ARE)

The $\text{var}(\hat{p}_{mle})$, $\text{var}({}^{1}_{r}\hat{p})$, $\text{var}({}^{2}_{s}\hat{p})$ and $\text{var}({}^{3}_{t}\hat{p})$ are compared in this section. This is accomplished by computing asymptotic relative efficiency (*ARE*) values for various values of $\eta, \beta, k$ and $p$. Let $ARE^1 = \frac{\text{var}(\hat{p}_{mle})}{\text{var}({}^{1}_{r}\hat{p})}$, $ARE^2 = \frac{\text{var}(\hat{p}_{mle})}{\text{var}({}^{2}_{s}\hat{p})}$ and $ARE^3 = \frac{\text{var}(\hat{p}_{mle})}{\text{var}({}^{3}_{t}\hat{p})}$.

**Table 4.** The *ARE* of the joint model relative to the ${}^{1}E$-, ${}^{2}E$- and ${}^{3}E$-models with $\eta = \beta = 0.80$ and $k = 2, 3, 5$ *and* $10$

| p | | $k=2$ | $k=3$ | $k=5$ | $k=10$ |
|---|---|---|---|---|---|
| 0.01 | $ARE^1$ | 0.056 | 0.028 | 0.012 | 0.004 |
| | $ARE^2$ | 0.215 | 0.235 | 0.264 | 0.314 |
| | $ARE^3$ | 0.729 | 0.736 | 0.723 | 0.682 |
| 0.05 | $ARE^1$ | 0.087 | 0.052 | 0.020 | 0.016 |
| | $ARE^2$ | 0.289 | 0.332 | 0.382 | 0.443 |
| | $ARE^3$ | 0.624 | 0.616 | 0.589 | 0.541 |
| 0.10 | $ARE^1$ | 0.115 | 0.077 | 0.051 | 0.040 |
| | $ARE^2$ | 0.333 | 0.380 | 0.429 | 0.483 |
| | $ARE^3$ | 0.551 | 0.543 | 0.520 | 0.476 |
| 0.15 | $ARE^1$ | 0.139 | 0.102 | 0.079 | 0.05 |
| | $ARE^2$ | 0.356 | 0.401 | 0.446 | 0.484 |
| | $ARE^3$ | 0.505 | 0.497 | 0.475 | 0.421 |
| 0.30 | $ARE^1$ | 0.205 | 0.193 | 0.251 | 0.720 |
| | $ARE^2$ | 0.378 | 0.407 | 0.406 | 0.163 |
| | $ARE^3$ | 0.417 | 0.400 | 0.343 | 0.118 |

From Tables 4 and 5 it is observed that increase in sensitivity and specificity of test leads to increase in the values of *ARE*. For the given values of pool size, sensitivity and specificity the highest value of *ARE* is 0.761. Hence the other models under consideration in the study ${}^{1}E$-, ${}^{2}E$- and ${}^{3}E$-models can only be 76.1% efficiency as the joint model.

# 9. Conclusions

This study focused on construction of the a new model for approximating the prevalence rate of a trait in a population with imperfect tests by consecutively choosing between three experiments namely ${}^{1}E$-, ${}^{2}E$- and ${}^{3}E$ -experiments. The model should select the better experiment and once the better experiment is being used, the estimator should approximate the individual maximum likelihood estimator (MLE) for that experiment. From this study it is clear that the best estimators for small, medium and large values of *p*, respectively, are ${}^{1}_{r}\hat{p}$, ${}^{2}_{s}\hat{p}$ and ${}^{3}_{t}\hat{p}$. From Tables 2 and 3, the computed values of asymptotic relative efficiency (*ARE)* for various values of $\eta, \beta, k$ and $p$ are less than one hence the proposed joint model for sequential choice of the best experiment for optimal estimation of a trait with misclassification is more efficient than the ${}^{1}E$-, ${}^{2}E$- and ${}^{3}E$-models.

**Table 5.** The *ARE* of the joint model relative to the ${}^{1}E$-, ${}^{2}E$- and ${}^{3}E$ -models with $\eta = \beta = 0.90$

| p | | $k=2$ | $k=3$ | $k=5$ | $k=10$ |
|---|---|---|---|---|---|
| 0.01 | $ARE^1$ | 0.051 | 0.029 | 0.015 | 0.006 |
| | $ARE^2$ | 0.188 | 0.224 | 0.277 | 0.350 |
| | $ARE^3$ | 0.761 | 0.746 | 0.708 | 0.644 |
| 0.05 | $ARE^1$ | 0.107 | 0.070 | 0.042 | 0.024 |
| | $ARE^2$ | 0.316 | 0.365 | 0.415 | 0.465 |
| | $ARE^3$ | 0.577 | 0.565 | 0.543 | 0.512 |
| 0.10 | $ARE^1$ | 0.142 | 0.100 | 0.069 | 0.052 |
| | $ARE^2$ | 0.359 | 0.403 | 0.445 | 0.488 |
| | $ARE^3$ | 0.499 | 0.497 | 0.487 | 0.461 |
| 0.15 | $ARE^1$ | 0.165 | 0.124 | 0.096 | 0.101 |
| | $ARE^2$ | 0.375 | 0.414 | 0.452 | 0.488 |
| | $ARE^3$ | 0.460 | 0.461 | 0.452 | 0.411 |
| 0.30 | $ARE^1$ | 0.219 | 0.199 | 0.234 | 0.642 |
| | $ARE^2$ | 0.385 | 0.412 | 0.421 | 0.218 |
| | $ARE^3$ | 0.396 | 0.388 | 0.345 | 0.140 |

It is assumed that the samples being pooled for use in the model are independent and identically distributed Bernoulli random variables. It is also assumed that there is a laboratory test that can determine whether or not a unit or at least a unit in a pool has the characteristic of interest and that the tests are conditionally independent of each other. Sensitivity and specificity of the test kits are also assumed to be the same at each step of testing and for all samples in use in the model.

The findings of the study of a better approximation of the prevalence rate have important health implications for prevention, intervention and treatment of HIV infections in a population. HIV infected population are known to greatly increase the spread of HIV infection. Our improved estimate of the prevalence rate of HIV infection could substantially reduce the potential risks for secondary HIV transmission by

HIV infected population who are unaware of HIV infections. A prompt diagnosis of HIV infection might prevent the infected population from engaging in high-risk behaviours with uninfected population and avoid new HIV infections to occur. But the population with HIV infection should take counselling, regarding risk-reduction strategies such as abstinence and safer sexual behaviours such as 100% use of condoms.

Based on the model developed a pool testing model of retesting of both positive and negative pools can be studied. A model based on cost analysis when sampling from different experiments can also be looked at when using imperfect kits.

## Appendix A: R code to Determine the Cut-Point Value of $I_x(\mathrm{P}^I)$ and $I_x(\mathrm{P}^G)$

```
#Program to compute the cut-point value of I_x(P^I) and I_x(P^G)
# Specify the parameters in use
η = 0.80    #sensitivity of the test
β = 0.80    #specificity of the test
k = 2       # pool size
#Define the function whose root is to be determined
fun1 = τ2*(1-τ2)*(1-p)^2 - k^2*(1-p)^(2*k)*τ1*(1-τ1)
cutoffvalue = uniroot(fun1, c(0,1), tol = 1e-5)
cutoffvalue
```

## Appendix B: R Code to Determine the Cut-Point Value of $I_x(\mathrm{P}^I)$ and $I_x(\mathrm{P}^R)$

```
#Program to compute the cut-point value of I_x(P^I) and I_x(P^R)
# Specify the parameters in use
η = 0.80    #sensitivity of the test
β = 0.80    #specificity of the test
k = 2       # pool size
#Define the function whose root is to be determined
fun2 = τ3*(1-τ3)*(1-p)^2 - k^2*(1-p)^(2*k)*τ1*(1-τ1)(η-β+1)^2
cutoffvalue = uniroot(fun2, c(0,1), tol = 1e-5)
cutoffvalue
```

## Appendix C: R Code to Determine the Cut-Point Value of $I_x(\mathrm{P}^G)$ and $I_x(\mathrm{P}^R)$

```
#Program to compute the cut-point value of I_x(P^G) and I_x(P^R)
# Specify the parameters in use
η = 0.80    #sensitivity of the test
β = 0.80    #specificity of the test
k = 2       # pool size
#Define the function whose root is to be determined
fun3 = τ3*(1-τ3) - (η-β+1)^2*τ2*(1-τ2)
cutoffvalue = uniroot(fun3, c(0,1), tol = 1e-5)
cutoffvalue
```

# REFERENCES


[1] Behets, F., Bertezzi, S., and Kasali, M., Kashamuka, M., Atikala, L., Brrown, C., Ryder, S., and Quinn, C., (1990). Successful use of Pooled Sera to Determine HIV-1 Seroprevalence in Zaire with Development of Cost-effective models. *AIDS* 4, 737-741.

[2] Brookmeyer, R. (1999). Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence. *Biometric* 55, 608 – 612.

[3] Cahoon-Young, B., Chandler, A., Livermore, T., and Benjamir, R. (1989). Sensitivity and Specificity of Pooled versus Individual Sera in a Human Immunodeficiency Virus (HIV) antibody Prevalence Study. *Journal of Clinical Microbiology* 17, 1893-1895.

[4] Ding, J. and Xiong, W. (2015). Robust Groups Testing for Multiple Traits with Misclassifications. *Journal of Applied Statistics.* Vol 42, no 10, 2115-2025.

[5] Dorfman, R. (1943). The Detection of Defective Members of Large Population. *Annals of Mathematical Statistics* 14, 436-440.

[6] 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.

[7] 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.

[8] Hardwick, J., 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.

[9] 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.

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

[13] Swallow, W. H. (1985). Group Testing for Estimating infection Rates and Probability of disease Transmissions. *Phytopathology* 75, 882-889.

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

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

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

[17] Tu, X. M., Litvak, E., and Pagano, M. (1995). On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika* 82, 287-297.

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