MODELS FOR SEQUENTIALLY SELECTING BETWEEN TWO AND THREE EXPERIMENTS FOR OPTIMAL ESTIMATION OF PREVALENCE RATE WITH IMPERFECT TESTS

MATIRI GEORGE MUNENE

A Thesis Submitted to the Graduate School in Partial Fulfillment for the Requirements of the Award of Doctor of Philosophy Degree in Statistics of Egerton University

EGERTON UNIVERSITY

SEPTEMBER, 2017

DECLARATION AND RECOMMENDATION

DECLARATION

This thesis is my original work and has not been submitted in part or whole for an award in any institution.

Signature:…… Date:…….

George Munene Matiri
SD12/0407/13

RECOMMENDATION

This thesis has been submitted for examination with our approval as university supervisors.

Signature:…… Date:…….

Prof. Kennedy L. Nyongesa
Department of Mathematics
Masinde Muliro University of Science and Technology

Signature:…… Date:…….

Prof. Ali Islam
Department of Mathematics
Egerton University

DEDICATION

To my beloved mum, the late Gladys W. Matiri, thank you for precious time you sacrificed in molding me to be what I am today. Always in my mind though far off.

ACKNOWLEDGMENT

I would like to thank the Almighty God for having granted me the grace, wisdom, knowledge and good health throughout the entire period of study.

I am grateful to National Commission of Science, Technology and Innovations for giving me a grant that enabled me do most of the work documented herein, thank you for the support. I am also grateful to Egerton University for giving me a PhD tuition waiver.

My heartfelt gratitude goes to my supervisors, Prof K. Nyongesa and Prof Ali Islam whose constant guidance and their precious time dedicated to the work made it possible to complete the thesis on time. I would also like to thank Dr Orawo, Dr Wanyoyi and Dr Kaguchwa of Egerton University for continuous guidance and encouragement throughout the study. Special thanks go to my beloved spouse Alice Munene for putting up with my absence for many hours over the past four years during preparation of this work. Thanks to my beloved and understanding daughters Gladys, Esther, Brenda, Mary, Jane and Gloria for their prayers and support during the course of my study, to all I say Almighty God bless, protect and guide you. To many of my colleagues who contributed to the success of this study directly or indirectly, I say God bless and your support was not in vain.

Thank you and God bless you all.

ABSTRACT

The idea of pool testing originated from 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. 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 models which are used to sequentially select some combination of two or three experiments when the sensitivity and specificity of the test kits is less than 100%. The experiments are selected sequentially at each stage so that the information obtained at a given stage is used to determine the experiment to be carried out in the subsequent stage. To accomplish this, the study has employed the method of maximum likelihood estimation in obtaining the estimators. The Fisher information of different experiments is compared and the cut off values where both experiments have the same Fisher information is calculated. The joint experiment models while choosing between two experiments and joint experiment model while choosing between three experiments obtained in this study are found to be more superior to the existing one–at-a-time, pooled and pooled with retesting experiments. Furthermore, asymptotic relative efficiency (ARE) of the joint two and three experiment models are computed and the joint three experiment model found to perform better.

TABLE OF CONTENT

LIST OF SYMBOLS

$I_x(.)$	Fisher information
k	Pool size
m	Number of observations from the P^1 experiment
n	Number of observations from the P^G experiment
N	Fixed total number of observations
p	Prevalence rate
r	Number of observations from the P^R-experiment
τ_1	Probability of declaring a pool positive in P^I-experiment
τ_2	Probability of declaring a pool positive in P^G-experiment
τ_3	Probability of declaring a pool positive in P^R-experiment
$\hat{p}$	An estimator of prevalence rate p

LIST OF ABBREVIATIONS AND ACRONYMS

ARE Asymptotic Relative Efficiency

MLE Maximum Likelihood Estimator

HIV Human Immunodeficiency Virus

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

1.1 Background information

Under minimum cost, large population say *N*, can be classified as either defective or non-defective. A minimum cost effective method that one can use for classifying such a population was suggested by Dorfman (1943) and herein referred to as pool-testing or group-testing. Pool testing refers to the simultaneous testing of more than one unit by one test and is mostly used when the trait under investigation is rare. In Pool-testing, pools are tested and pools that test negative on the test are dropped from further investigation, while those that test positive, individual constituents are tested. Pool-testing can provide substantial savings as compared to individual testing. More testing strategies have been suggested in pool-testing literature. Recently pool-testing with retesting have been proposed by Nyongesa (2005). Instead of testing the constituents' members of pools that tested positive on the first test, smaller pools can be constructed and tested with the view of reducing the number of tests, however the ultimate aim will be accomplished with minimal cost. This will lower the errors associated with the procedure as it has been widely discussed in pool-testing literature.

Pool-testing has two objectives:

i) To test the pools followed by individual testing in the pools that test positive with the aim of identifying the infected individuals as discussed above and this is known as classification problem. (Dorfman, 1943, Johnson *et al.*, 1991 and Nyongesa, 2004)

ii) The second objective is to estimate the prevalence rate of a trait in a population as advocated by Thomson (1962).

More research work, of recent, are focused on the second objective for estimating the prevalence rate of the characteristic of interest. Thomson (1962) studied estimation problem using pool testing. The model was later considered by Brookmayer (1999) but with sensitivity and specificity less than 100%. By sensitivity we mean the probability of correctly classifying a pool or individual that contains the trait of interest and by specificity we mean the probability of correctly classifying a pool or individual without the trait. Sufficiently accurate estimate of the prevalence can be obtained from testing pooled samples as demonstrated by Hammick and Gastwirth (1994). Xie *et al.*, (2001) demonstrated how pool testing can reduce costs in early stages of drug discovery. Hardwick *et al.*,(1998) considered sequentially deciding between two experiments for estimating a common success prevalence rate. Matiri *et al.*, (2017) introduced the element of errors in combined experiments. Nyongesa (2012) dealt with dual estimation in estimating prevalence rate.

However, in a field experiment, a mixture of experiments can yield worthwhile and more accurate results than the Dorfman (1943) procedure. Hence the main objective of this research work is to present optimal estimation of prevalence rate of a trait and its properties when using a mixture of experiments with imperfect tests.

1.2 Statement of the Problem

If individuals of a large population say $N \to \infty$ are subjected to testing with the aim of estimating the prevalence rate of a trait or characteristic of importance, one-at-a-time testing is time consuming and non-cost effective and is bound to errors. To overcome this problem, pool testing has been suggested in statistical literature. However it has been shown that this procedure is only viable when the prevalence rate is small, in most cases when the prevalence rate is less than 25%. Thus this procedure is not applicable when the prevalence rate is more than 25%. In such case it calls for one at a time testing and this is a drawback to the procedure. Again if the test kits accuracy is not 100%, when applied in the above procedure, there tend to be loss in sensitivity. To recover lost sensitivity some model has been suggested of retesting. Therefore several procedures have been outlined for estimating the prevalence rate of a rare trait. In practice, application of a mixture of experiments gives better results and this is the gap in the literature. Therefore there is need to develop a model of a mixture of experiments namely: one-at-a-time testing, pool testing and pool testing with retesting of the positive pools when the accuracy of the test kits is less than 100%. This is the focus of this study.

1.3 Objectives

1.3.1 General objective

To develop a model of choosing between two and three experiments based on available information in order to estimate the prevalence rate of trait in a population when imperfect tests are used.

1.3.2 Specific objectives

The study will address the following specific objectives:

i) To develop statistical models for estimating prevalence rate.

ii) To derive the Maximum Likelihood Estimator and the Fisher information based on each testing strategy.

iii) To determine the properties of the derived estimators.

iv) To analyze Asymptotic Relative Efficiencies (ARE) of the experiments.

1.4 Justification

To assess the effectiveness of public health measures introduced in Kenya and the world to halt the spread of HIV virus, other rare diseases and the use of drugs in the population, reliable estimates of the prevalence rate of a trait of these characteristics are needed. Due to possible adverse social and economic consequences the degree of voluntary participation by the public in Kenya and the world at large in surveys that screen blood or urine is lower than normal hence pooling comes in since it conceal stigmatization and hence greater voluntary participation of the targeted groups or individuals. Screening individual sample for a particular disease in a large population is not cost effective and is time consuming. To minimize cost and save time taken for the experiments pooling comes in handy and this goes also in drug testing and testing of industrial product. The advantages of pooling go beyond reducing the cost and time, if done properly, it produces more accurate tests.

CHAPTER TWO

LITERATURE REVIEW

2.1 Pool Testing

The idea of pool testing was suggested by Dorfman(1943) during the world war II as an economical method of testing blood samples of army inductees in order to detect the presence of infection. Pool testing involve putting together item or blood samples to form a pool and then testing the pool rather than testing each item or blood sample for evidence of infection or characteristic of interest. A negative reading indicates that the pool contains no defective item or infected blood sample while a positive reading indicates the presence of at least one defective item or infected blood sample. Considering a population of size N pooled into n pools each of size k, Dorfman(1943) considered using pool testing as follows:

If p is the prevalence rate of infection of each unit then:

i) $1-p$ is the probability of selecting at random a unit free of infection,

ii) $(1-p)^k$ is the probability of selecting at random a pool of size k units free from infection,

iii) $1-(1-p)^k$ is the probability of selecting at random a pool of size k units that contain at least one unit infected.

If X pools out of n test positive on the test, then X has a Binomial distribution model with parameters n and π

$$f(x,\pi)=\binom{n}{x}\pi^x(1-\pi)^{(n-x)} \tag{2.1}$$

where $\pi=1-(1-\mathrm{p})^k$. In this model, if T is the number of tests, then $T=n+kX$ and the expected number of tests is $E(T)=n+kn\pi<N$ for small values of p. This implies that the number of pools plus the number of individuals in pools that tested positive which require individual testing is less than the total number of items or individuals under consideration hence pool testing can lead to substantial savings.

Behets *et al.,* (1990), Cahoon-Young *et al.,*(1989) and Kline *et al.,*(1989) presented pooling strategies that creates pools of at most 15 samples. If a pool tests negative, then all the individual samples in that pool are declared as negative. On the other hand, if pool tests positive, then each constituent sample of the pool is subsequently tested individually. For populations with relatively small prevalence rates, such a procedure reduces cost and saves time even with the extra cost incurred due to pooling because it substantially reduces the number of tests that need to be performed. As reported by these authors, in their particular

situations pooling resulted in a cost reduction of up to 80%. The cost reduction is an important issue, especially in light of the recent surge in requests for HIV testing especially in developing countries. Johnson *et al.*,(1991) studied also the cost effectiveness of pooling algorithm whose objective was identifying individuals with characteristic of interest. In their procedure each pool that tested positive was divided into two equal pools, which were tested, pools that tested positive were further subdivided and tested again and so on. Litvak *et al.*,(1994) extended this work by considering pooling algorithms when sensitivity and specificity of the tests was less than 100%. They showed that some of these algorithms can reduce the error rates of the screening procedures compared to individual testing. Nyongesa (2004) generalized a common pooling strategy with retesting and discussed its characteristics. He considered hierarchical pooling strategy that involved testing pools and then sequentially subdividing and retesting the positive pools. Nyongesa (2005) further studied pool testing with retesting in which pools classified as positive and negative are retested. He observed that retesting improves the sensitivity and specificity of the testing procedure. Nyongesa (2012) proposed pool testing where members that form the population under investigation are pooled together in pools and these pools are given a test. Further testing are discontinued for the pools that test negative but if the test is positive the pool is divided into blocks of equal sizes. The blocks are further tested and for those that test positive the constituent members are tested individually for the presence or absence of the trait under investigation.

More research work, of recent, has focused on estimating the prevalence rate of a trait using pool testing without necessarily identifying the individuals or items with characteristic of interest. This is known as estimation. The estimation problem using pool testing was initially investigated by Thomson (1962). The maximum likelihood estimator (MLE) of *p*, using Equation (2.1) is:

$$\hat{p} = 1 - \left(1 - \frac{\sum_{i=1}^{n} x_i}{n}\right)^{\frac{1}{k}}.$$

Thomson (1962) further examined the behavior of the MLE by computing bias and the mean squared error. He found out that it is positively biased estimator of the prevalence rate (*p*). The idea of errors in the Thomson (1962) model was introduced by Brookmeyer (1999) where he modified Equation (2.1) and obtained;

$$f(x,p) = \binom{n}{x}\{\eta(1-(1-p)^k + (1-\beta)(1-p)^k)\}^x\{(1-\eta)(1-(1-p)^k + \beta(1-p)^k)\}^{(n-x)} \qquad (2.2)$$

where η and β are sensitivity and specificity of the test kits respectively.

The likelihood function of Equation (2.2) is

$$L(p/\eta,\beta,\mathrm{n},x)\ \alpha\ \prod_{i=1}^{n}\{\eta(1-(1-p)^k+(1-\beta)(1-p)^k)\}^{x_i}\{(1-\eta)(1-(1-p)^k+\beta(1-p)^k)\}^{(n-x_i)}. \tag{2.3}$$

The maximum likelihood estimator (MLE) based on Equation (2.3) is

$$\hat{p}=1-\left(\frac{\eta-\frac{\sum_{i=1}^{n}x_i}{n}}{\eta+\beta-1}\right)^{\frac{1}{k}}.$$

Sufficiently, accurate estimate of the prevalence can be obtained from testing pooled samples as demonstrated by Hammick and Gastwirth (1994). Their procedure provides greater protection of respondent's anonymity which can lead to greater participation in the survey. They also extended the range of applicability to higher levels of prevalence rate (up to about 30%). They did this by taking two samples of individuals and formed two independent sets of pools which were then tested. The results of the tests were averaged to yield new estimates. Wein and Zenios (1996) developed a hierarchical statistical model that relates the HIV test output to the antibody concentration in the pool, thereby observing the effect of pooling together different samples. Their model was validated using data from a variety of field studies. The simulation results showed that significant cost savings can be achieved without compromising the accuracy of the test in pool testing. However, the efficiency of pool testing depends upon the use of a classification rule that is dependent on pool size, a characteristic that was lacking in the pooled testing procedures implemented at that time. Gastwirth and Johnson (1994) used pool testing to estimate HIV prevalence cost-effectively. Xie *et al.*, (2001) demonstrated how pool testing can reduce costs in early stages of drug discovery. Juan and Wenjun (2015) provided algorithms for the computations of pool sizes.

Nyongesa and Syaywa (2011) used moment method to estimate the prevalence rate. They obtained the following distribution

$$f(p,\underline{x})=\binom{n}{x_1,x_2}p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2}.$$

where x_1 are pools classified as positive on the first test, x_2 are pools classified as negative on retest, *n* is the total number of pools, $\underline{x} = \{x_1, x_2\}$, p_1 is the probability of declaring a pool as positive on the first test and is given by;

$$p_1 = (1-(1-p)^k)\eta^2 + (1-\eta)^2(1-p)^k \tag{2.4}$$

and p_2 is the probability of declaring a pool as positive on the first test and negative on retest and is given by;

$$p_2 = (1-(1-p)^k)\beta(1-\beta) + (1-p)^k\eta(1-\eta)\,. \tag{2.5}$$

The moment estimates of p obtained from Equations (2.4) and (2.5) are

$${}_1\hat{p} = 1-\left(\frac{\eta^2 - p_1}{\eta^2 - (1-\beta)^2}\right)^{\frac{1}{k}}$$

and

$${}_2\hat{p} = 1-\left(\frac{\eta(1-\eta) - p_2}{\eta(1-\eta) - \beta(1-\beta)}\right)^{\frac{1}{k}}$$

respectively where η and β are known constants.

Wanyonyi *et al.*, (2015) developed a model for estimating unknown proportion of a trait in batch testing based on a quality control process. The model developed was found to be superior to the existing models when the proportion of a trait is relatively high. 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).

Hardwick *et al.*, (1998) considered sequentially deciding between two experiments for estimating a common success prevalence rate by considering the individual Bernoulli*(p)* trials or the product of *k* individual independent Bernoulli (p^k) trials. Their goal was to compare the accuracy of their adaptive cut-point estimator for prevalence with those obtained from individual testing and fixed group size testing. The model developed was superior to individual testing and to fixed group size testing. Matiri *et al.*, (2017) introduced errors in Hardwick *et al.*, (1998) model with the same goal of comparing accuracy of the estimator with those obtained from individual testing and pool testing with test errors.

2.2 Applications of Pool Testing

Pool testing can be applied in many areas as outlined by Sobel and Groll (1966). Pool testing is used in pooling blood samples in order to classify each individual of a large population as to whether or not they have a particular disease. 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 *m* units is that either all *m* are good or at least one of the *m* is defective (Mundel, 1984).

Another application is in testing various electrical devices such as condensers, resistors, *etc*. If one assumes that the *m* bulbs for the Christmas tree are all in series so that when he switches on the lights he will know by the result that either all the *m* bulbs are good or at least one of the *m* is defective but he does not know as a result of this test alone how many or which ones are defective. Suppose he had shorter wires (of various sizes) for fewer bulbs he can use these to find out exactly which ones are defective.

Pool testing has been applied in screening the population for the presence of HIV antibody (Kline *et al*., 1989 and Manzon *et al*., 1992). Litvak *et al*., (1994) applied pool testing in screening HIV antibody to help in curbing the further spread of the virus. Litvak *et al*., (1994) showed that pooling offers a feasible way to lower the error rates associated with labeling 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 labeled negative in fact has antibodies since each test has a certain sensitivity and specificity.

Other pool testing scenarios arise in environmental monitoring where sample units of soil or plant matter are combined and tested for toxins. Partly because of its use in different areas of study, pool testing has appeared under other names, such as, batch sampling (Chaubey and Li, 1993), composite sampling (Lovison *et. al*., 1994) and batch testing (Phatarfod and Sudbury, 1994).

2.3 Methods

In this study the following methods have been employed:

a) Maximum Likelihood Estimator (MLE)

For a fixed set of data and underlying statistical model, the method of maximum likelihood selects values of the model parameters that produce a distribution that gives the observed data the greatest probability or parameters that maximize the likelihood function. Maximum-likelihood estimation gives a unified approach to estimation, which is well defined in many distributions. However, in some complex distributions, difficulties do occur and in such problems, maximum-likelihood estimators are unsuitable or do not exist at all. For many models, a maximum likelihood estimator can be found as an explicit function of the observed data or generated data. For some, however, no closed-form solution to the maximization problem is known or available, and a MLE has to be found numerically using optimization methods. In practice it is more often convenient when working with the natural logarithm of the likelihood function, called the log-likelihood:

$$\ln L(\theta; \mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_n) = \sum_{i=1}^{n} \ln f(\mathrm{x}_i \mid \theta).$$

The method of maximum likelihood estimates θ_0 by finding a value of θ that maximizes

$\ln L(\theta; \mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_n)$. If a maximum exists, is MLE estimate and it is the same regardless of whether we maximize the likelihood or the log-likelihood function since log is a monotonically increasing.

b) Fisher Information

In mathematical statistics, Fisher information (FI) sometimes simply called information is a way of measuring the amount of information that an observable random variable *X* carries about an unknown parameter θ of a distribution that models *X*. Formally, it is the variance of the score or the expected value of the observed information. Uses of FI include:

i) Describing the asymptotic behavior of maximum likelihood estimates.

ii) Calculating the variance of an estimator.

iii) Finding priors in Bayesian inference.

The amount of information contained in a random variable *X* is given by

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial \theta^2} \log f(\mathrm{x} \mid \theta)\right)$$

If *X* and *Y* are independent and jointly distributed random variables, then their total FI is

$I\ (\theta) = I_X(\theta) + I_Y(\theta)$.

Consequently, the information in a random sample of *n* independent and identically distributed observations is *n* times the information in a sample of size 1.

c) Statistical Software

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses of MATLAB include:

i) Math and computation,

ii) Algorithm development,

iii) Modeling, simulation and prototyping,

iv) Data analysis, exploration and visualization,

v) Scientific and engineering graphics

vi) Application development, including Graphical User interface building.

In the study solutions to some of the equations were computed using MATLAB R2012b codes developed and presented in the appendices. Graphical representation was also done using MATLAB R2012b codes developed.

d) Simulation

Simulation is a numerical technique for conducting experiments on the computer. It involves random sampling from probability distribution. Usually, when statisticians talk about "simulation" they mean "Monte Carlo Simulation". They are often used in simulating physical and mathematical systems. A typical Monte Carlo simulation involves the following:

i) Generate *S* independent data sets under the conditions of interest.

ii) Compute the numerical value of the estimator/test statistic T (data) for each data set $T_1,\ldots\ldots,T_s$.

iii) If *S* is large enough, summary statistics $T_1,\ldots\ldots,T_s$ should be good approximations to the true sampling properties of the estimator/test statistic under the conditions of interest.

These methods are most suitable to calculation by R version 3.1.2 computer software which was used to simulate data using the formulated models.

CHAPTER THREE

SEQUENTIALLY SELECTING BETWEEN TWO EXPERIMENTS FOR ESTIMATING PREVALENCE RATE WITH IMPERFECT TESTS

3.1 Indicator functions

The indicator function of an event is a random variable that takes values 1 when the event happens and value 0 when the event does not happen. Indicator functions are often used in statistics to simplify notation and to prove theorems.

In construction of the joint experiment model, the following indicator functions are required:

$$\delta_i = \begin{cases} 1, \textit{ if the } i^{th} \textit{ individual is positive} \\ 0, \textit{ otherwise} \end{cases}$$

$$\tau_i = \begin{cases} 1, \textit{ if the } i^{th} \textit{ individual is declared positive by the test} \\ 0, \textit{ otherwise} \end{cases}$$

$$T_j = \begin{cases} 1, \textit{ if the } j^{th} \textit{ pool is declared positive by the test} \\ 0, \textit{ otherwise} \end{cases}$$

$$D_j = \begin{cases} 1, \textit{ if the } j^{th} \textit{ pool is positive} \\ 0, \textit{ otherwise} \end{cases}$$

The parameters, sensitivity and specificity will be assumed remain constant throughout the study where sensitivity, denoted by η, means the probability of correctly classifying a defective pool or defective individual, and is given by

$$\eta = Prob(\tau_i = 1 | \delta_i = 1) = Prob(\mathrm{T}_j = 1 | \mathrm{D}_j = 1)$$

and specificity, denoted by β, means the probability of correctly classifying a non-defective pool or non-defective individual, and is given by

$$\beta = Prob(\tau_i = 0 | \delta_i = 0) = Prob(\mathrm{T}_j = 0 | \mathrm{D}_j = 0).$$

These parameters have been used in subsequent development and are assumed to be based on the manufacturers' specifications.

3.2 The models

The models have been split into two that is P^I-experiment and P^G-experiment. The P^I-experiment means estimating the prevalence rate of the characteristic of interest with testing each individual under study while the 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 item. 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 $N = m+n$ being the total number of observations from both experiments.

3.2.1 Computation of Fisher information from P^I-experiment

If the P^I-experiment is used to estimate the prevalence rate p of interest and X_{1i} for $i = 1, \cdots, m$ is a sequence of identically independent distributed random variable, then $X_{1i} \sim \text{B}ernouli(\tau_1)$ where τ_1 is the probability of declaring an individual as positive *i.e.*

$$\begin{aligned}\tau_1 &= Prob(\text{declaring an individual positive}) \\ &= Prob(\tau_i = 1)\end{aligned}$$

by the law of total probability

$$\tau_1 = Prob(\tau_i = 1,\ \delta_i = 1 \text{ or } \delta_i = 0)$$

thus

$$\begin{aligned}\tau_1 &= Prob(\tau_i = 1,\ \delta_i = 1 \text{ or } \tau_i = 1,\ \delta_i = 0) \\ &= Prob(\tau_i = 1,\ \delta_i = 1) + Prob(\tau_i = 1,\ \delta_i = 0) \\ &= \eta p + (1-\beta)(1-\text{p})\end{aligned}$$

hence

$$\tau_1 = \eta p + (1-\beta)(1-\text{p}).$$

For a single experiment, the probability density function is

$$f(x_{1i}, p/\eta, \beta) = \{\eta p + (1-\beta)(1-\text{p})\}^{x_{1i}} \{(1-\eta)\,\text{p} + \beta(1-p)\}^{1-x_{1i}}. \tag{3.1}$$

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

$$I_{x_1}(\text{P}^I) = \frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} \tag{3.2}$$

which is easily obtained from Equation(3.1).

3.2.2 Computation of Fisher information from P^G-experiment

The P^G-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 that there is 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 where $\tau_2 = Prob(\mathrm{T}_j = 1)$ and for analysis purposes, the constituent members of a pool are assumed to act independent of each other and therefore $Prob(D_j = 0) = (1-p)^k$. Therefore

$$\tau_2 = Prob(\mathrm{T}_j = 1,\ D_j = 1 \text{ or } D_j = 0)$$

by total probability law, whence

$$\begin{aligned}\tau_2 &= Prob(\mathrm{T}_j = 1,\ D_j = 1 \text{ or } \mathrm{T}_j = 1,\ D_j = 0)\\ &= Prob(\mathrm{T}_j = 1,\ D_j = 1) + Prob(\mathrm{T}_j = 1,\ D_j = 0)\\ &= \eta(1-(1-\mathrm{p})^k) + (1-\beta)(1-\mathrm{p})^k\end{aligned}$$

and upon rearranging, we get

$$\tau_2 = \eta(1-(1-\mathrm{p})^k) + (1-\beta)(1-\mathrm{p})^k. \tag{3.3}$$

Let X_{2j} denote a sequence of identically and independent distributed random variable for $j = 1,...,n$, then the outcome which takes positive or negative has a Bernoulli distribution *i.e.* $X_{2j} \sim Bernouli(\tau_2)$. For a single experiment equivalently the probability density function is

$$f(x_{2j}, p \mid \eta, \beta, k) = \{\eta(1-(1-p)^k) + (1-\beta)(1-p)^k\}^{x_{2j}}\{1-\eta(1-(1-p)^k) - (1-\beta)(1-p)^k\}^{1-x_{2j}} \tag{3.4}$$

from which the Fisher information is obtained as follows:

Taking log on both sides of Equation (3.4) yields:

$$\begin{aligned}\log f(\mathrm{x}_{2j}, \mathrm{p} \mid \eta, \beta, k) = \mathrm{x}_{2j}&\log\{\eta(1-(1-\mathrm{p})^k) + (1-\beta)(1-\mathrm{p})^k\} +\\ &(1-\mathrm{x}_{2j})\log\{1-\eta(1-(1-\mathrm{p})^k) - (1-\beta)(1-\mathrm{p})^k\}.\end{aligned} \tag{3.5}$$

Taking the second partial derivative of Equation (3.5) with respect to p leads to

$$\begin{aligned}\frac{\partial^2 \log\mathrm{f}(\cdot)}{\partial \mathrm{p}^2} = &\left(\frac{-x_{2j}}{\tau_2^{\,2}} - \frac{1-x_{2j}}{(1-\tau_2)^2}\right)\left(\eta k(1-\mathrm{p})^k - k(1-\beta)(1-\mathrm{p})^{k-1}\right)^2 +\\ &\left(\frac{-x_{2j}}{\tau_2} - \frac{1-x_{2j}}{(1-\tau_2)}\right)\left(-\eta k(\mathrm{k}-1)(1-\mathrm{p})^{k-2} + k(\mathrm{k}-1)(1-\beta)(1-\mathrm{p})^{k-2}\right).\end{aligned} \tag{3.6}$$

Thus, the expected value of Equation (3.6) is

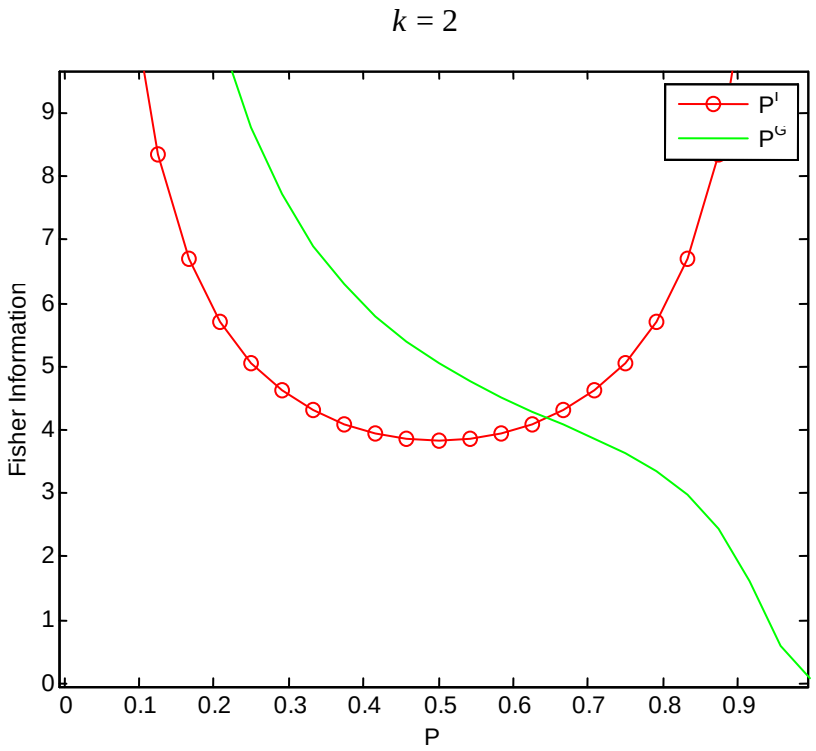
$$E\left(\frac{\partial^2 \log\mathrm{f}(\cdot)}{\partial \mathrm{p}^2}\right) = \left(\frac{-1}{\tau_2} - \frac{1}{(1-\tau_2)}\right)\left(\eta k(1-\mathrm{p})^k - k(1-\beta)(1-\mathrm{p})^{k-1}\right)^2.$$

Hence the Fisher information denoted by $I_{x_2}(\mathrm{P}^G)$ contained in a single observation of the P^G-experiment is

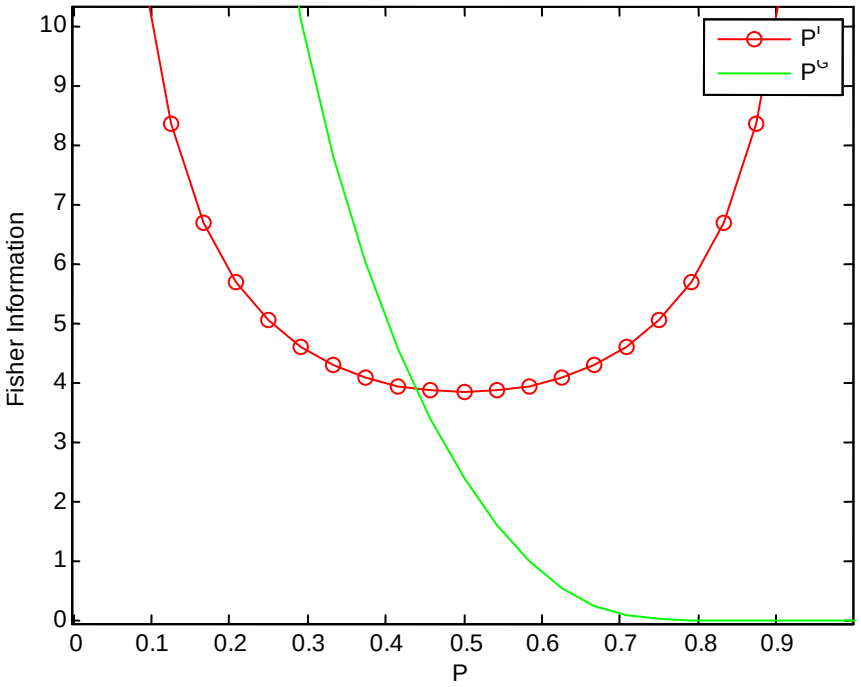
$$I_{x_2}(\mathrm{P}^G) = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}. \tag{3.7}$$

3.3 Comparison of $I_x(\cdot)$ of P^I- and P^G-experiments

In this section the graphs of $I_x(\cdot)$ of P^I- and P^G-experiments for values of $\eta = \beta = 0.99,\ 0.95,\ 0.80$ and $k = 2,\ 5,\ 10$ versus p are plotted.



$k = 5$



$k = 10$

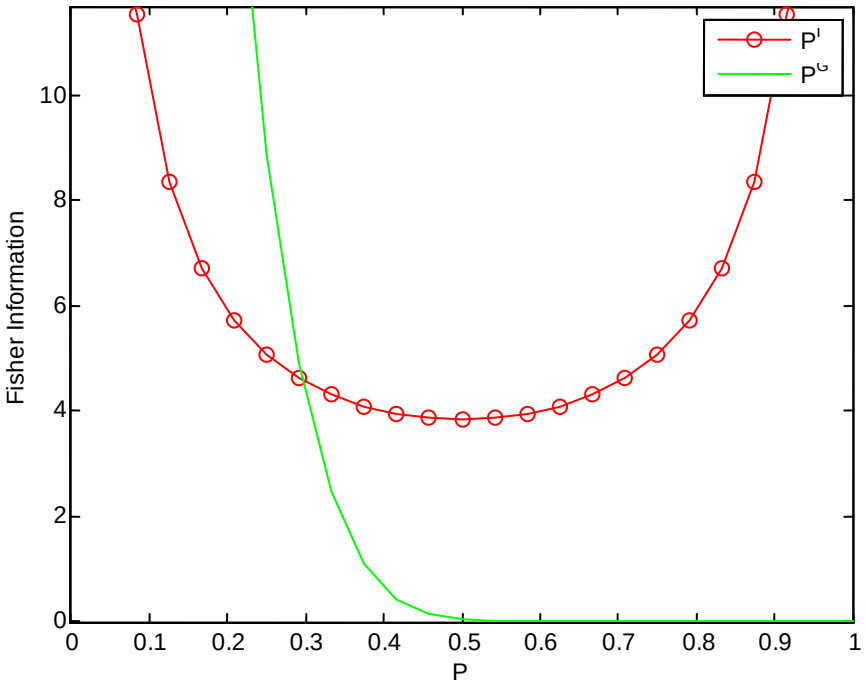
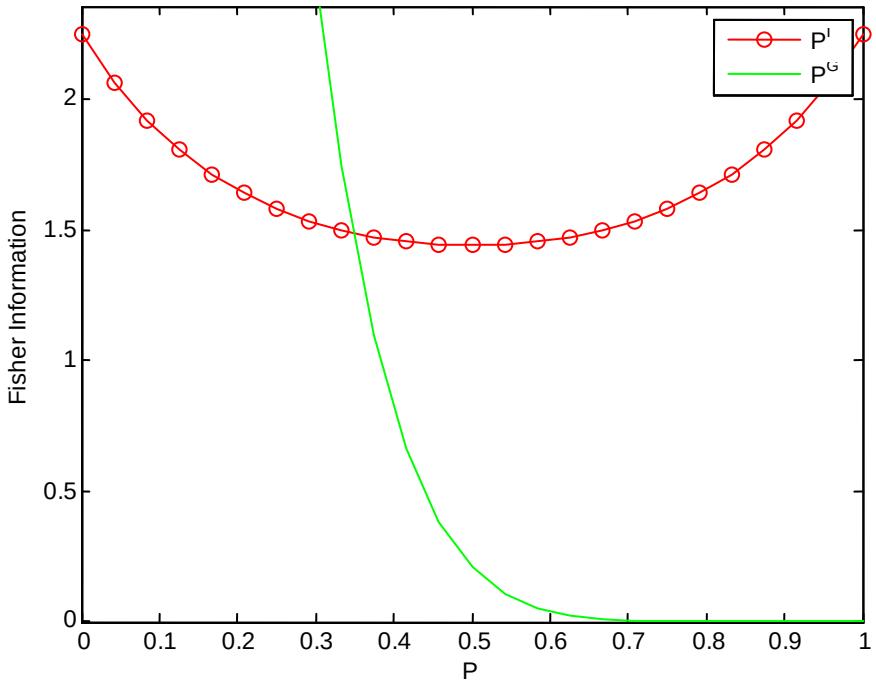


Figure 3.1(a): Plots of $I_x(.)$ versus P for $\eta = \beta = 0.99$ and $k = 2,\ 5,\ 10$

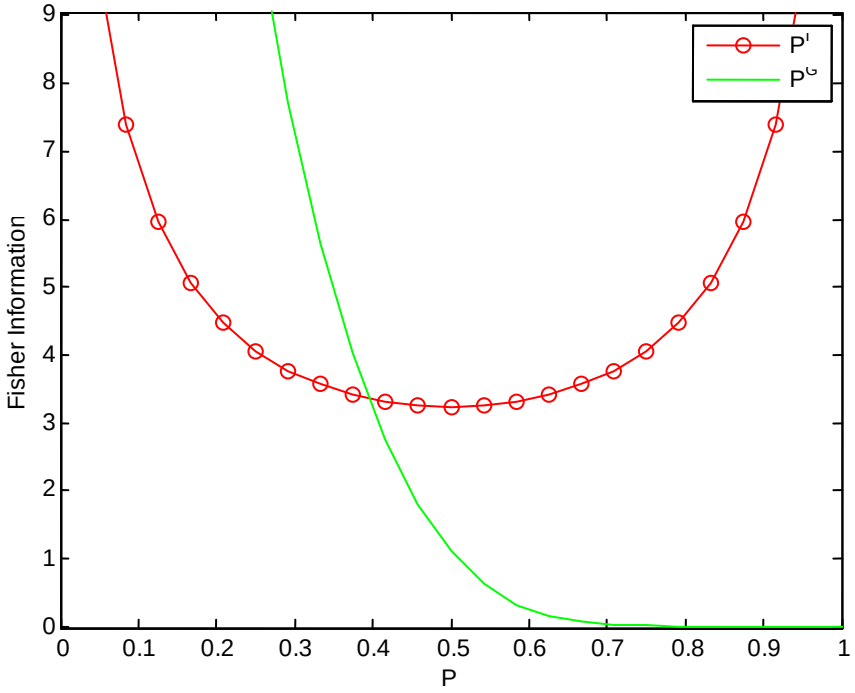
As seen from Figure 3.1(a), that the change in the value of k does not affect the Fisher information of P^I-experiment since P^I-experiment is independent of k. It is observed from the

graph of P^G-experiment 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 one. It is also noted that as k increases the curve of the Fisher information of the P^G-experiment shift to the left meaning that the region for which P^G-experiment is better than the P^I-experiment shrinks.

$$\eta = \beta = 0.80$$



$\eta = \beta = 0.95$



$\eta = \beta = 0.99$

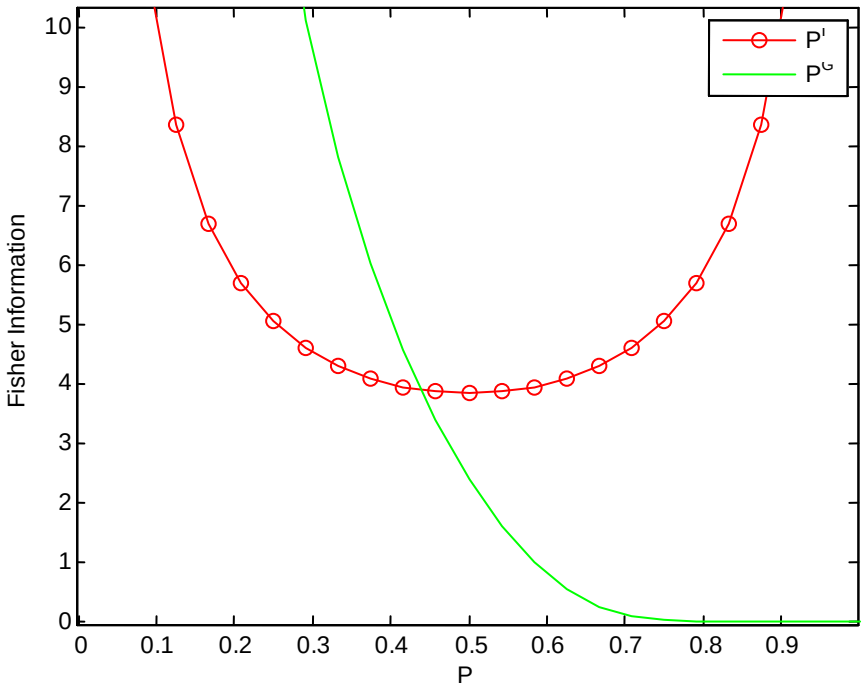


Figure 3.1(b): Plots of $I_x(.)$ versus P for $k = 5$ and $\eta = \beta = 0.80,\ 0.95,\ 0.99$

Figure 3.1(b) shows that as sensitivity and specificity of the tests increases, holding k constant, the Fisher information for both P^I- and P^G-experiments increases. The curve for the P^G-experiment shifts to the right as sensitivity and specificity of the tests increases hence the region in which the Fisher information of P^G-experiment is higher than for the P^I-experiment increases.

In general, as seen from Figure 3.1(a) and 3.1(b), the plot of the Fisher information of the P^I-experiment is concave upwards. The graph of Fisher information of the P^G-experiment is strictly decreasing as the value of p increases. It is also observed that pool testing is only viable and better than individual testing strategy where the prevalence rate is small as observed by (Dorman, 1943), otherwise individual testing is preferred. Thus, the graphs obtained provide empirical evidence of the pool testing scheme for small values of p but for high values, the use of P^I-experiment is recommended.

3.4 Computation of cut off values

The cut off value for the P^I- and P^G-experiment is 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 $I_x(\mathrm{P}^G)$.

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

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

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

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

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

$$\tau_2(1-\tau_2)(1-\mathrm{p})^2 - \mathrm{k}^2(1-\mathrm{p})^{2k}\tau_1(1-\tau_1) = 0 \qquad (3.8)$$

since k, β and η are known constants, then Equation (3.8) is a function of p , of which the value of p can be solved iteratively using Newton-Raphson method as follows:

Let

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

then the function $f(\mathrm{p})$ is continuous in the interval $(0,1)$ and from Figures 3.1(a) and 3.1(b) of the graphs of Fisher information, there exist a value *p*, such that $f(\mathrm{p}) = 0$ 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 $f'(\mathrm{p}_i)$ is the derivative of the function $f(\mathrm{p})$ which is not equal to zero for any value of p_i for $i = 0, 1, 2, \cdots\cdots$. The iteration stops if $|p_{i+1} - p_i| < \varepsilon$ for some arbitrary value ε. If the series converges, p_{i+1} is taken as an approximate value of 'a' which is the solution of Equation (3.8). A **MATLAB** code for the iteration is given in Appendix A.

For various values of k, η and β the values of the roots of Equation (3.8) or the cut off values 'a' are given in Table 3.1:

Table 3.1: Cut off values 'a' of P^I- and P^G-experiments for various values of k, η and β

k	$\eta = \beta = 0.99$	$\eta = \beta = 0.95$	$\eta = \beta = 0.90$	$\eta = \beta = 0.85$	$\eta = \beta = 0.80$
2	0.646	0.596	0.563	0.542	0.528
3	0.555	0.507	0.477	0.458	0.446
5	0.439	0.395	0.371	0.357	0.348
10	0.296	0.263	0.248	0.239	0.234
15	0.227	0.201	0.190	0.185	0.181
20	0.185	0.164	0.156	0.152	0.150
50	0.092	0.082	0.080	0.079	0.078

From Table 3.1 it is 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 is reasonably small (Nyongesa, 2004). 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 increases. Thus pool-testing is most feasible when the test kits are of high accuracy.

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

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

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 value. As k increases to infinity, the cut off values decreases and the region over which the P^G-experiment is better shrinks as shown in Table 3.1 and Figures 3.1(a) and 3.1(b).

3.5 Estimation of *p*

In this section the maximum likelihood estimation method is used to estimate the value of p using the P^I-experiment, P^G-experiment and the joint experiment model separately. The number of observations from P^I- and P^G-experiment are m and n respectively and the fixed total number of observations from both experiments (N) is given by the sum of m and n.

3.5.1 Estimation of *p* from the P^I-experiment

If m observations from the P^I-experiment are used to estimate p, then a random variable $X_{1i} \sim Bernouli(\tau_1)$ *i.e*

$$f(\mathrm{x}_{1i}, \mathrm{p} / \eta, \beta) = \tau_1^{x_{1i}} (1-\tau_1)^{1-x_{1i}} \qquad (3.9)$$

The joint probability density function or likelihood function for the m observations is

$$
\begin{aligned}
\mathrm{L}(\mathrm{p}\,|\,\eta,\beta) &= \prod_{i=1}^{m} f(\mathrm{x}_{1i},\mathrm{p}\,|\,\eta,\beta) \\
&= \prod_{i=1}^{m}\{\tau_1^{\,x_{1i}}(1-\tau_1)^{1-x_{1i}}\} \\
&= \tau_1^{\sum_{i=1}^{m} x_{1i}}(1-\tau_1)^{m-\sum_{i=1}^{m} x_{1i}}
\end{aligned}
\quad .
$$

Upon taking logs on both sides, we get

$$\log L(\cdot) = \sum_{i=1}^{m} x_{1i}\log(\tau_1) + (\mathrm{m} - \sum_{i=1}^{m} x_{1i})\log(1-\tau_1)\,. \tag{3.10}$$

Differentiating Equation (3.10) with respect to q where $q = 1-p$, leads to

$$\frac{d\log L(\cdot)}{d\,q} = \frac{\sum_{i=1}^{m} x_{1i}(1-\eta-\beta)}{\eta(1-\mathrm{q})+(1-\beta)\,\mathrm{q}} + \frac{(\mathrm{m}-\sum_{i=1}^{m} x_{1i})(\eta+\beta-1)}{(1-\eta)(1-\mathrm{q})+\beta\,\mathrm{q}}\,. \tag{3.11}$$

To find the maximum likelihood estimator of q, denoted by $\hat{q}_m^I$, Equation (3.11) is equated to zero and solved for q.

$$\frac{\sum_{i=1}^{m} x_{1i}(1-\eta-\beta)}{\eta(1-\mathrm{q})+(1-\beta)\,\mathrm{q}} + \frac{(\mathrm{m}-\sum_{i=1}^{m} x_{1i})(\eta+\beta-1)}{(1-\eta)(1-\mathrm{q})+\beta\,\mathrm{q}} = 0$$

$$\sum_{i=1}^{m} x_{1i} = m\eta - m\eta q + mq - m\beta q$$

$$\frac{\sum_{i=1}^{m} x_{1i}}{m} = \eta - \eta q + q - \beta q$$

$$\hat{q}_m^I = \frac{\eta - \dfrac{\sum_{i=1}^{m} x_{1i}}{m}}{\beta+\eta-1}\,.$$

Therefore

$$\hat{p}_m^I = \frac{\beta - 1 + \dfrac{\sum_{i=1}^{m} x_{1i}}{m}}{\eta+\beta-1} \tag{3.12}$$

where $\hat{p}_m^I$ is maximum likelihood estimator of p, the same result obtained by Brookmeyer (1999). It should be noted that if $\eta = \beta = 1$ the maximum likelihood estimator reduces to the sample mean.

3.5.2 Estimation of *p* from the P^G-experiment

Suppose that there are n pools for the P^G-experiment, each of size *k*, available for estimating *p* and X_{2j} pool test positive on the test, then:

$$X_{2j} \sim Bernouli(\tau_2) \; i.e$$

$$f(\mathrm{x}_{2j}, \mathrm{p}/\eta, \beta) = \tau_2^{x_{2j}} (1-\tau_2)^{1-x_{2j}} . \tag{3.13}$$

The likelihood function is

$$\begin{aligned} \mathrm{L}(\mathrm{p}|\eta, \beta) &= \prod_{j=1}^{n} (\tau_2^{x_{2j}} (1-\tau_2)^{1-x_{2j}}) \\ &= \tau_2^{\sum_{j=1}^{n} x_{2j}} (1-\tau_2)^{n-\sum_{j=1}^{n} x_{2j}} . \end{aligned}$$

Working similarly as above we have

$$\log L(\cdot) = \sum_{j=1}^{n} x_{2j} \log(\tau_2) + (\mathrm{n} - \sum_{j=1}^{n} x_{2j}) \log(1-\tau_2) . \tag{3.14}$$

Thus upon differentiating with respect to *q* we get

$$\frac{d \log L(\cdot)}{d\, q} = \frac{\sum_{j=1}^{n} x_{2j}}{\tau_2} \frac{d\tau_2}{dq} + \frac{(\mathrm{n} - \sum_{j=1}^{n} x_{2j})}{1-\tau_2} \frac{d(1-\tau_2)}{dq} .$$

From which the maximum likelihood estimator is obtained by

$$\frac{\sum_{j=1}^{n} x_{2j}}{\tau_2} \frac{d\tau_2}{dq} + \frac{(\mathrm{n} - \sum_{j=1}^{n} x_{2j})}{1-\tau_2} \frac{d(1-\tau_2)}{dq} = 0$$

$$\frac{d\tau_2}{dq} \left(\frac{\sum_{j=1}^{n} x_{2j}}{\tau_2} - \frac{(\mathrm{n} - \sum_{j=1}^{n} x_{2j})}{1-\tau_2} \right) = 0 .$$

With similar augment as before

$$\frac{\sum_{j=1}^{n} x_{2j}}{\tau_2} - \frac{(\mathrm{n} - \sum_{j=1}^{n} x_{2j})}{1-\tau_2} = 0$$

$$\eta(1-q^k) + (1-\beta)\mathrm{q}^k = \frac{\sum x_{2j}}{n}$$

and making q the subject

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

whence

$$\hat{p}_n^G = 1 - \left(\frac{\eta - \frac{\sum_{j=1}^{n} x_{2j}}{n}}{\eta + \beta - 1} \right)^{\frac{1}{k}} \tag{3.15}$$

where $\hat{p}_n^G$ is the maximum likelihood estimate of p. This is the result obtained by Nyongesa (2012). When $\eta = \beta = 1$, it leads to the results obtained by Thompson (1962) as

$$\hat{p} = 1 - \left(1 - \frac{\sum_{j=1}^{n} x_{2j}}{n} \right)^{\frac{1}{k}}.$$

3.5.3 Joint experiment model for estimating *p*

If m is the number of observations of the P^I-experiment and n is the number of observations of the P^G-experiment and 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 respective density functions

$$f(\underline{x}, \mathrm{p} / k, \eta, \beta) = \tau_1^{x_{1i}} (1-\tau_1)^{1-x_{1i}} \times \tau_2^{x_{2j}} (1-\tau_2)^{1-x_{2j}}. \tag{3.16}$$

with joint likelihood function given as

$$\mathrm{L}(\underline{\mathrm{x}}, \mathrm{p} / \mathrm{k}, \eta, \beta) = \left\{ (\tau_1)^{\sum_{i=1}^{m} x_{1i}} (1-\tau_1)^{m - \sum_{i=1}^{m} x_{1i}} \times (\tau_2)^{\sum_{j=1}^{n} x_{2j}} (1-\tau_2)^{n - \sum_{j=1}^{n} x_{2j}} \right\}. \tag{3.17}$$

Taking logarithm on both sides of Equation (3.17) and differentiating with respect to q yields

$$\frac{\mathrm{d} \log L(.)}{dq} = \frac{\sum_{i=1}^{m} x_{1i} - \mathrm{m}\tau_1}{\tau_1(1-\tau_1)} \frac{d\tau_1}{dq} + \frac{\sum_{j=1}^{n} x_{2j} - n\tau_2}{\tau_2(1-\tau_2)} \frac{d\tau_2}{dq} \tag{3.18}$$

with $\frac{d\tau_1}{dq} = 1 - \eta - \beta$ and $\frac{d\tau_2}{dq} = kq^{k-1}(1 - \eta - \beta)$.

Equating Equation (3.18) to zero leads to

$$\frac{\sum_{i=1}^{m} x_{1i} - \mathrm{m}\tau_1}{\tau_1(1-\tau_1)} \frac{d\tau_1}{dq} + \frac{\sum_{j=1}^{n} x_{2j} - n\tau_2}{\tau_2(1-\tau_2)} \frac{d\tau_2}{dq} = 0. \tag{3.19}$$

Since k, β *and* η are known constants, then Equation (3.19) is a continuous function of q and a unique root exist. We solve for q 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 iterative solution can be obtained by Newton-Raphson method:

$$q_{i+1} = q_i - \frac{f(\mathrm{q}_i)}{f'(\mathrm{q}_i)}.$$

The **MATLAB** code is provided in Appendix B.

3.6 Variance of the Estimators

In this section, the asymptotic variance of the maximum likelihood estimators (MLE) of p are discussed.

3.6.1 Variance of $\hat{p}_m^1$ of the P^I-experiment

The asymptotic variance of $\hat{p}_m^1$ is obtained from the Fisher information and is given by $\left\{-E\left(\frac{d^2 \log f(\cdot)}{dp^2}\right)\right\}^{-1}$. Finding the logarithm and the second derivative with respect to p of Equation (3.8) yields;

$$\frac{d^2 \log \mathrm{f}(\cdot)}{d\,\mathrm{p}^2} = \left(\frac{-x}{\tau_1^2} - \frac{1-x}{(1-\tau_1)^2}\right)(\eta+\beta-1)^2$$

$$-E\frac{d^2 \log \mathrm{f}(\cdot)}{d\,\mathrm{p}^2} = \left(\frac{1}{\tau_1} + \frac{1}{1-\tau_1}\right)(\eta+\beta-1)^2$$

with

$$\left\{-E\left(\frac{d^2 \log f(\cdot)}{dp^2}\right)\right\}^{-1} = \frac{\tau_1(1-\tau_1)}{(\eta+\beta-1)^2}.$$

The asymptotic variance of $\hat{p}_m^1$ of the P^I-experiment is

$$\mathrm{var}(\hat{p}_m^1) = \frac{\tau_1(1-\tau_1)}{(\eta+\beta-1)^2}. \qquad (3.20)$$

3.6.2 Variance of $\hat{p}_n^G$ of the P^G-experiment

As in Section 3.7.1, finding the logarithm and the second derivative with respect to p of Equation (3.11) and working through the process yields

$$\frac{d^2 \log \mathrm{f}(\cdot)}{d\,\mathrm{p}^2} = \left(\frac{-x}{\tau_2^2} - \frac{1-x}{(1-\tau_2)^2}\right)\left(\eta k(1-\mathrm{p})^{k-1} - k(1-\beta)(1-\mathrm{p})^{k-1}\right)^2 +$$
$$\left(\frac{x}{\tau_2} - \frac{1-x}{(1-\tau_2)}\right)\left(-\eta k(\mathrm{k}-1)(1-\mathrm{p})^{k-2} + k(\mathrm{k}-1)(1-\beta)(1-\mathrm{p})^{k-2}\right).$$

Thus

$$-E\frac{d^2 \log \mathrm{f}(\cdot)}{d\,\mathrm{p}^2} = \left(\frac{1}{\tau_2} + \frac{1}{1-\tau_2}\right)\left(\eta k(1-\mathrm{p})^{k-1} - k(1-\beta)(1-\mathrm{p})^{k-1}\right)^2$$
$$= \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}.$$

Therefore, the variance is

$$\mathrm{var}(\hat{p}_n^G) = \frac{\tau_2(1-\tau_2)}{\mathrm{k}^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}. \tag{3.21}$$

3.6.3 Variance of $\hat{p}_{mle}$ of the joint experiment model

The joint probability density function of the random variables X_{1i} and X_{2j} of P^I- and P^G-experiments respectively is provided by Equation (3.16).

The negative expectation of second derivative with respect to p of the logarithm of the joint probability density function Equation (3.16) is

$$-E\left(\frac{d^2 \log \mathrm{f}(\cdot)}{d\,\mathrm{p}^2}\right) = \frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} + \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}$$
$$= (\eta+\beta-1)^2\left(\frac{1}{\tau_1(1-\tau_1)} + \frac{k^2(1-\mathrm{p})^{2k-2}}{\tau_2(1-\tau_2)}\right).$$

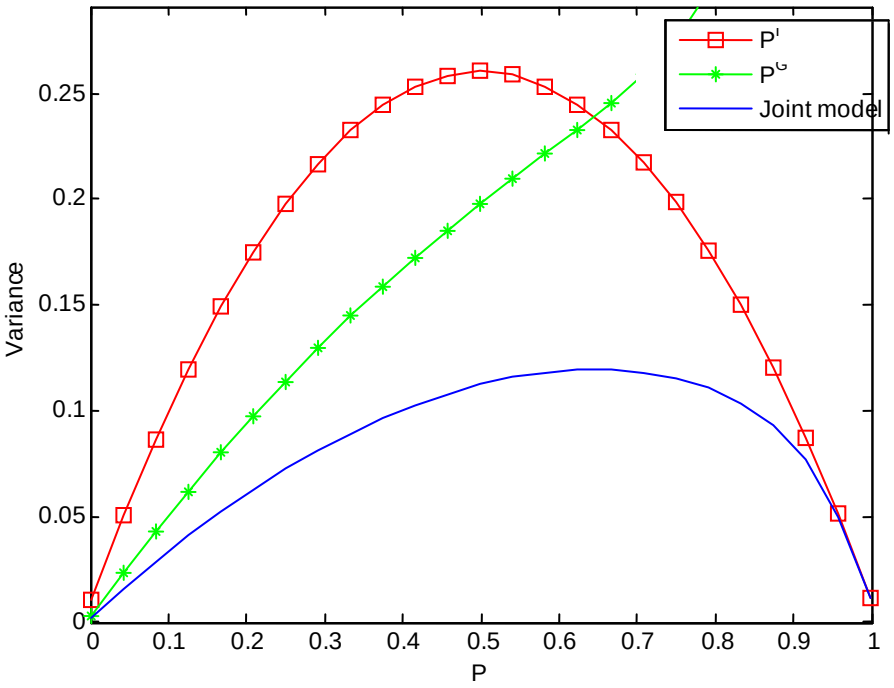
Assuming there are m observations from the P^I-experiment and n observations from the P^G-experiment then asymptotic variance of $\hat{p}_{mle}$ from the joint experiment model is

$$\mathrm{var}(\hat{p}_{mle}) = \frac{\tau_1\tau_2(1-\tau_1)(1-\tau_2)}{(\eta+\beta-1)^2\left\{m\tau_2(1-\tau_2) + nk^2(1-p)^{2k-2}\tau_1(1-\tau_1)\right\}}.$$

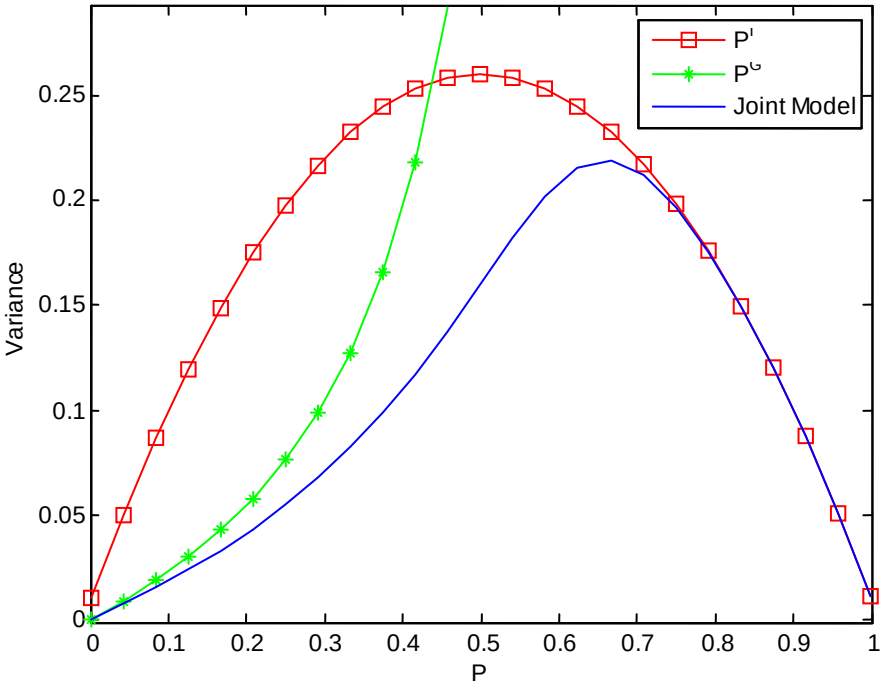
3.7 Comparing the Variances for P^I, P^G-experiments and the joint experiment model

In this section the graphs of the variance for P^I-, P^G-experiments and the joint experiment model for values of $\eta = \beta = 0.99,\ 0.95,\ 0.80$ and $k = 2,\ 5,\ 10$ versus p are plotted.

$k = 2$



$k = 5$.



$k = 10$

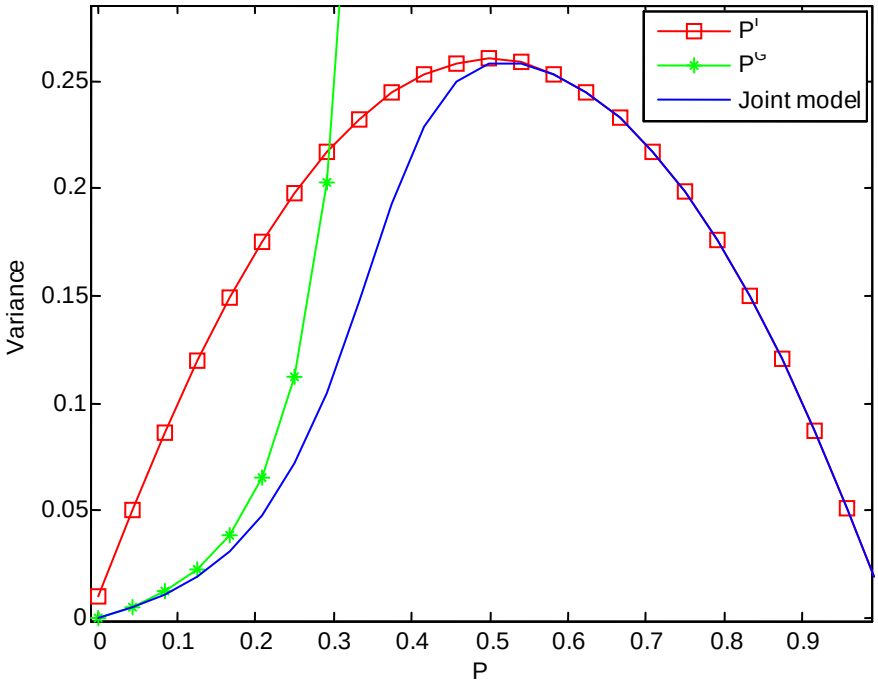


Figure 3.2(a): Plots of $var(\hat{p})$ versus p with $\eta = \beta = 0.99$ and $k = 2,\ 5,\ 10$.

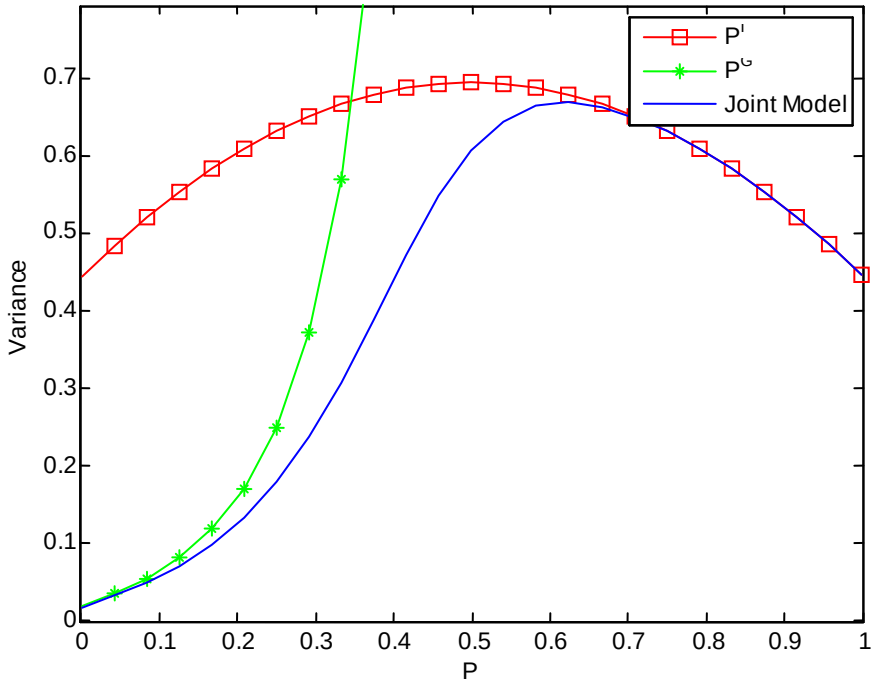
As seen from Figure 3.2(a), $\text{var}(\hat{p}_m^I)$ is unaffected by the changes in k since the model is independent of k. As k increases holding sensitivity and specificity constant:

i) Both $\text{var}(\hat{p}_n^G)$ and $\text{var}(\hat{p}_{mle})$ increases.

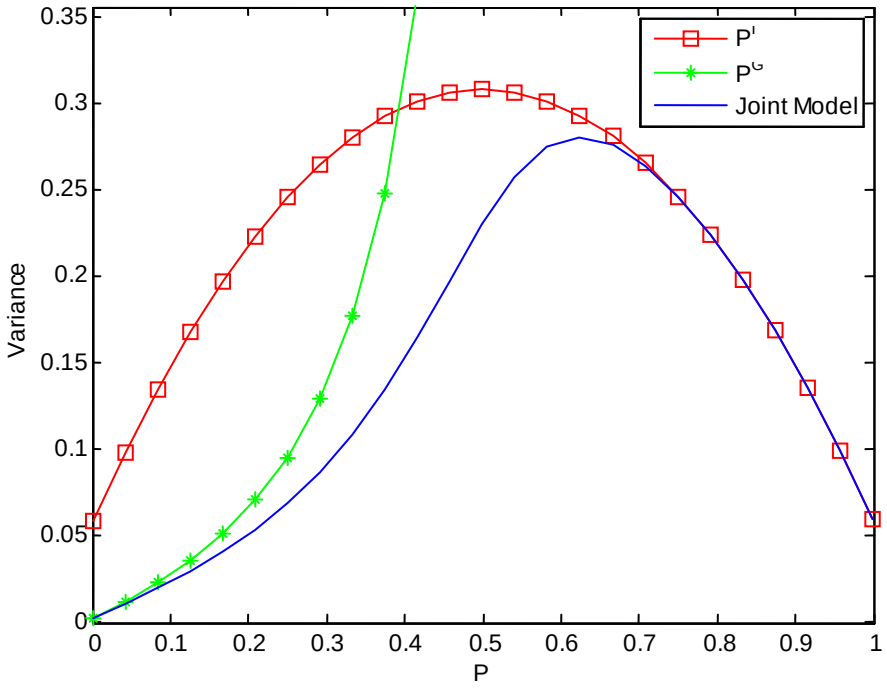
ii) The graph of $\text{var}(\hat{p}_n^G)$ shifts to the left meaning that the region in which $\text{var}(\hat{p}_n^G)$ is higher than $\text{var}(\hat{p}_m^I)$ decreases.

iii) The region in which the plots of $\text{var}(\hat{p}_{mle})$ and $\text{var}(\hat{p}_m^I)$ against the prevalence rate (p) overlaps increases.

$\eta = \beta = 0.80$



$\eta = \beta = 0.95$



$\eta = \beta = 0.99$

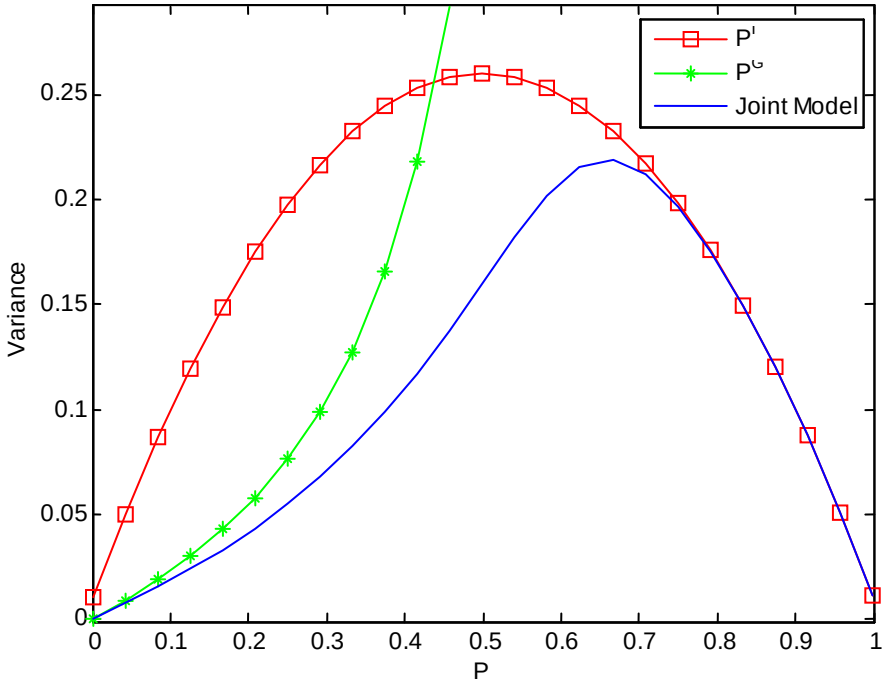


Figure 3.2(b): Plots of $var(\hat{p})$ versus p with $k = 5$ and $\eta = \beta = 0.80,\ 0.95,\ 0.99$.

As seen from Figure 3.2(b), as the specificity and sensitivity of the tests increases:

i) $\text{Var}(\hat{p}_m^I)$, $\text{var}(\hat{p}_n^G)$ and $\text{var}(\hat{p}_{mle})$ decreases.

ii) Area in which $\text{var}(\hat{p}_n^G)$ is higher than $\text{var}(\hat{p}_m^I)$ increases.

In general, as seen from Figures 3.2(a) and 3.2(b), the plot of $\text{var}(\hat{p}_m^I)$ is concave downwards and symmetric with the maximum at a value of p about 0.5. $\text{Var}(\hat{p}_n^G)$ increases exponentially as the value of parameter p increase. The variance of $\hat{p}_{mle}$ of the joint experiment model increases exponentially for small values of p but thereafter it starts decreasing as p gets closer to 1. For small values of the parameter p, $\text{var}(\hat{p}_{mle})$ is smaller than $\text{var}(\hat{p}_m^I)$ and $\text{var}(\hat{p}_n^G)$ but is equal to $\text{var}(\hat{p}_m^I)$ for the values of p close to 1. The region in which $\text{var}(\hat{p}_{mle})$ is higher than $\text{var}(\hat{p}_n^G)$ increases exponentially as the value of p increases. However, the region in which $\text{var}(\hat{p}_{mle})$ is better than $\text{var}(\hat{p}_m^I)$ increases then it starts decreasing and the two variances are equal for the values of p close to 1. It is observed that the variance of the joint experiment is smaller than the variance of the P^I- or P^G-experiments for relatively low values of the prevalence rate.

3.8 Asymptotic Relative Efficiency

In this section, $\text{var}(\hat{p}_{mle})$, $\text{var}(\hat{p}_m^I)$ and $\text{var}(\hat{p}_n^G)$ have been compared. This is accomplished by computing asymptotic relative efficiency (ARE). If $ARE^1 = \frac{\text{var}(\hat{p}_{mle})}{\text{var}(\hat{p}_m^I)}$ and $ARE^2 = \frac{\text{var}(\hat{p}_{mle})}{\text{var}(\hat{p}_n^G)}$ respectively, the results are tabulated in the tables 3.2 – 3.4. The $ARE < 1.0$ in all computed values in tables, implying that the joint experiment model is superior.

Table 3.2: The ARE's of the joint experiment relative to P^I- and P^G-experiments with $\eta = \beta = 0.99$

p		$k=2$	$k=3$	$k=5$	$k=10$
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 3.3: The ARE's of the joint experiment relative to P^I- and P^G-experiments with $\eta = \beta = 0.95$

p		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^1	0.225	0.129	0.062	0.025
	ARE^2	0.775	0.871	0.938	0.938
0.05	ARE^1	0.277	0.189	0.117	0.066

	ARE^2	0.723	0.811	0.883	0.934
0.10	ARE^1	0.306	0.226	0.158	0.114
	ARE^2	0.694	0.774	0.842	0.886
0.15	ARE^1	0.325	0.253	0.194	0.178
	ARE^2	0.675	0.747	0.806	0.822
0.20	ARE^1	0.341	0.277	0.233	0.283
	ARE^2	0.659	0.723	0.767	0.718
0.30	ARE^1	0.369	0.329	0.339	0.663
	ARE^2	0.631	0.671	0.661	0.337

Table 3.4: The ARE's of the joint experiment relative to P^I- and P^G-experiments with $\eta = \beta = 0.80$

p		$k=2$	$k=3$	$k=5$	$k=10$
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 3.2 to 3.4 of the results of AREs of the joint experiment relative to P^I- and P^G-experiments 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 $p \in (0, 0.3]$, the ARE^1 decrease 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

$p \in (0, 0.3]$ 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.

3.9 Estimates of prevalence rate, variance and confidence interval

In this section, the maximum likelihood estimates($\hat{p}$) of the prevalence rate, the variance and 95% Wald-type confidence intervals are computed for various values of sensitivity, specificity and pool size. The parameters used in the simulations are for illustration purposes but for optimal results, generated pool-sizes can be used as suggested by Juan and Wenjun (2015).

Table 3.5: Maximum likelihood estimates, variance and confidence interval for different values of p for $\eta = \beta = 99\%$ and $k = 5, 10$

	p	$\hat{p}$	$\text{var}(\hat{p}) \times 10^{-4}$	95% *CI*
$k = 5$	0.01	0.0160	0.3266	0.0000, 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	0.0000, 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 3.6: Maximum likelihood estimates, variance and confidence interval for different values of p for $\eta = \beta = 90\%$ and $k = 5, 10$

	p	$\hat{p}$	$\text{var}(\hat{p}) \times 10^{-4}$	95% *CI*
$k = 5$	0.01	0.0034	0.6200	0.0000, 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	0.0000, 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 3.7: Maximum likelihood estimates, variance and confidence interval for different values of p for $\eta = \beta = 80\%$ and $k = 5, 10$

	p	$\hat{p}$	$\text{var}(\hat{p}) \times 10^{-4}$	95% *CI*
$k = 5$	0.01	0.0148	2.1900	0.0000, 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	0.0000, 0.0428
	0.05	0.0306	0.0940	0.0000, 0.0644
	0.10	0.1000	4.120	0.0412, 0.1588
	0.20	0.2767	4.128	0.1890, 0.3644

From Tables 3.5 to 3.7 it can be noted that the maximum likelihood estimates of the prevalence rate are very close to the actual values which were used to simulate the estimates. The population estimates resulting from the experiments are used to evaluate the $(1-\alpha)100\%$ confidence limits of the confidence interval of the simulated estimates where α is the level of significance and it can be noted from Tables 3.5 to 3.7 that the actual values are within the upper and the lower limits. Therefore, we can ascertain that our estimates are within the acceptable range with confidence.

CHAPTER FOUR

SEQUENTIALLY SELECTING BETWEEN THREE EXPERIMENTS FOR ESTIMATING PREVALENCE RATE WITH IMPERFECT TESTS

4.1 Sequentially selecting between three experiments

In construction of the model for sequentially selecting between three experiments, apart from the indicator functions defined in Section 3.1, the indicator function

$$T_z^R = \begin{cases} 1, \textit{ if the } z^{th} \textit{ pool is declared positive by retest} \\ 0, \textit{ otherwise} \end{cases}$$

is also required. The model will be split into three, that is P^I-experiment, P^G-experiment and P^R-experiment whereby P^I- and P^G-experiments are as discussed in Chapter 3 while P^R-experiment means estimating the prevalence rate of the characteristic of interest by retesting the pools declared positive in the first test. Throughout the chapter m, n and r have been assumed to be the number of observations from the P^I-, P^G- and P^R-experiments respectively with $N = m+n+r$, the total number of observations from the three experiments.

4.1.1 Computation of Fisher information from the P^R-experiment

In this experiment the probability of declaring a pool positive after retesting the pools declared positive in the first test denoted by τ_3 is $\tau_3 = \text{Prob}(\text{T}_z^R = 1)$.

By the law of total probability

$$\tau_3 = Prob(\text{T}_z^R = 1,\ T_j = 1,\ D_j = 1 \text{ or } D_j = 0),$$

which implies that

$$\begin{aligned} \tau_3 &= Prob(\text{T}_z^R = 1, \text{T}_j = 1,\ D_j = 1 \text{ or } \text{T}_z^R = 1, T_j = 1,\ D_j = 0) \\ &= Prob(\text{T}_z^R = 1, \text{T}_j = 1,\ D_j = 1) + Prob(\text{T}_z^R = 1, \text{T}_j = 1,\ D_j = 0) \\ &= \eta^2(1-(1-\text{p})^k) + (1-\beta)^2(1-\text{p})^k \end{aligned}$$

therefore

$$\tau_3 = \eta^2(1-(1-\text{p})^k) + (1-\beta)^2(1-\text{p})^k. \tag{4.1}$$

Let X_{3z} denote a sequence of identically independent distributed random variables for $z = 1,...,r$, then $X_{3z} \sim Bernouli(\tau_3)$. For a single experiment the probability density function is

$$f(x_{3z}, p \mid \eta, \beta, k) = (\tau_3)^{x_{3z}}(1-\tau_3)^{1-x_{3z}}\ . \tag{4.2}$$

Taking logs on both sides of Equation (4.2) yields

$$\log f(\cdot) = \mathrm{x}_{3z}\log(\tau_3) + (1-\mathrm{x}_{3z})\log(1-\tau_3)\,. \tag{4.3}$$

The second derivative of Equation (4.3) with respect to p is

$$\frac{d^2\,\mathrm{logf}(\cdot)}{d\,\mathrm{p}^2} = \left(\frac{-x_{3z}}{\tau_3{}^2} - \frac{1-x_{3z}}{(1-\tau_3)^2}\right)\left(\eta^2 k(1-\mathrm{p})^{k-1} - k(1-\beta)^2(1-\mathrm{p})^{k-1}\right)^2 +$$

$$\left(\frac{x_{3z}}{\tau_3} - \frac{1-x_{3z}}{(1-\tau_3)}\right)\left(-\eta^2 k(\mathrm{k}-1)(1-\mathrm{p})^{k-2} + k(\mathrm{k}-1)(1-\beta)^2(1-\mathrm{p})^{k-2}\right).$$

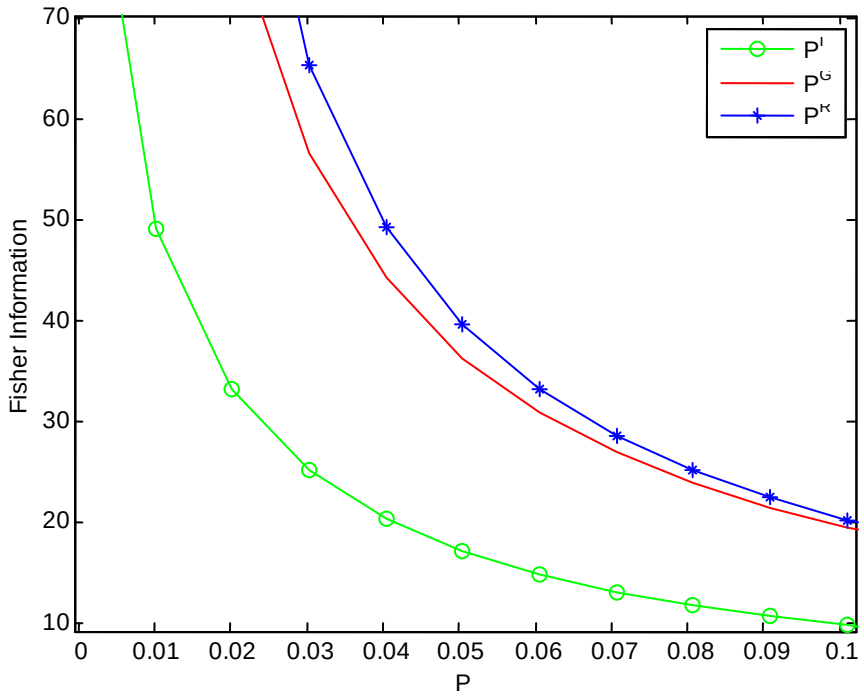
Thus for a single observation in P^R-experiment, the Fisher information is

$$I_x(\mathrm{P}^R) = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\tau_3(1-\tau_3)}\,. \tag{4.4}$$

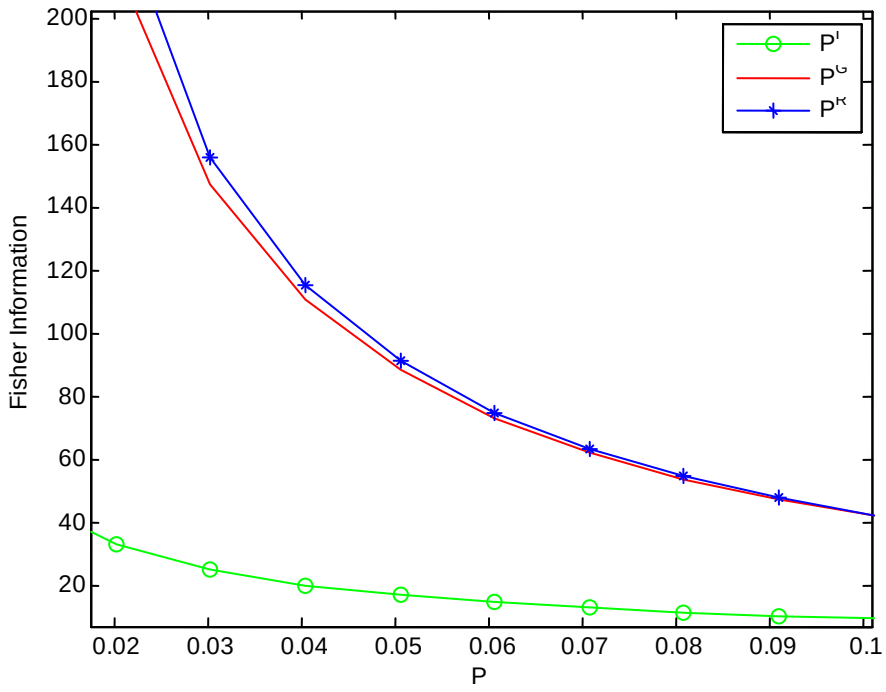
4.2 Comparison of $I_x(\cdot)$ of P^I-, P^G- and P^R-experiments

In this section $I_x(\cdot)$ of P^I-, P^G- and P^R-experiments are compared by plotting the graphs of $I_x(\cdot)$ of P^I-, P^G- and P^R-experiments for various values of η, β *and* k

$$k = 2$$



$k = 5$



$k = 10$

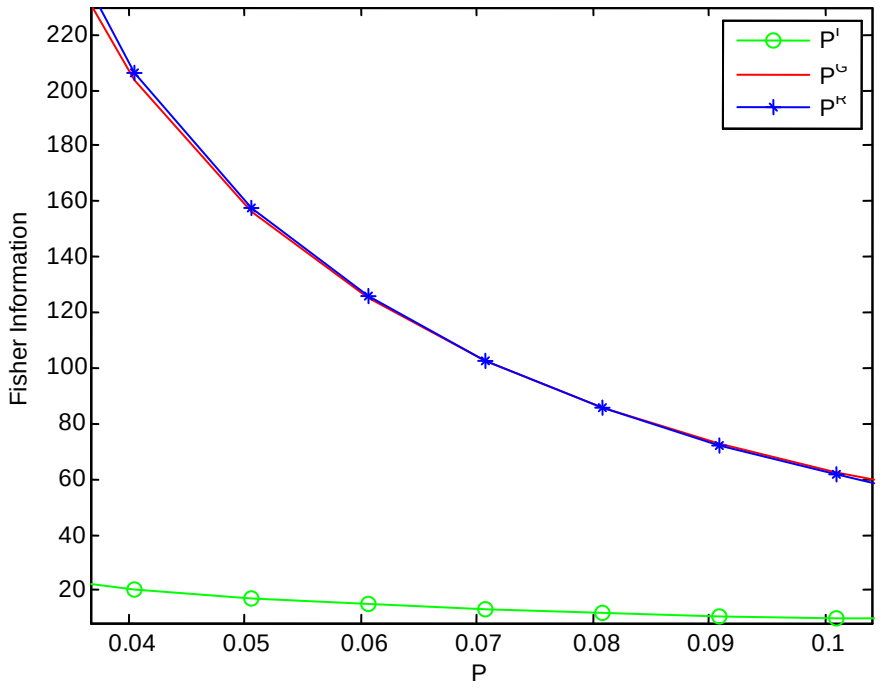
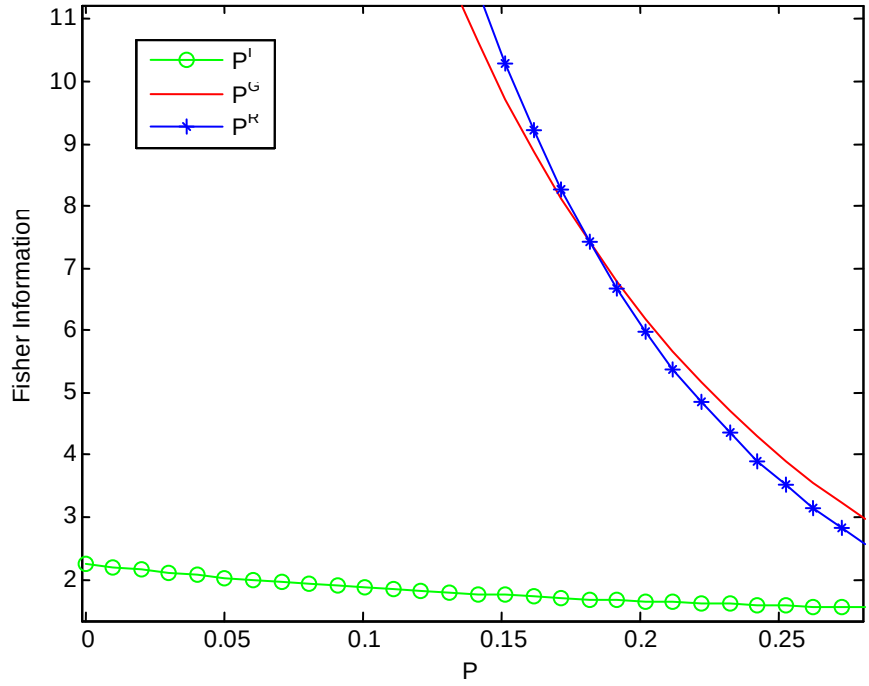


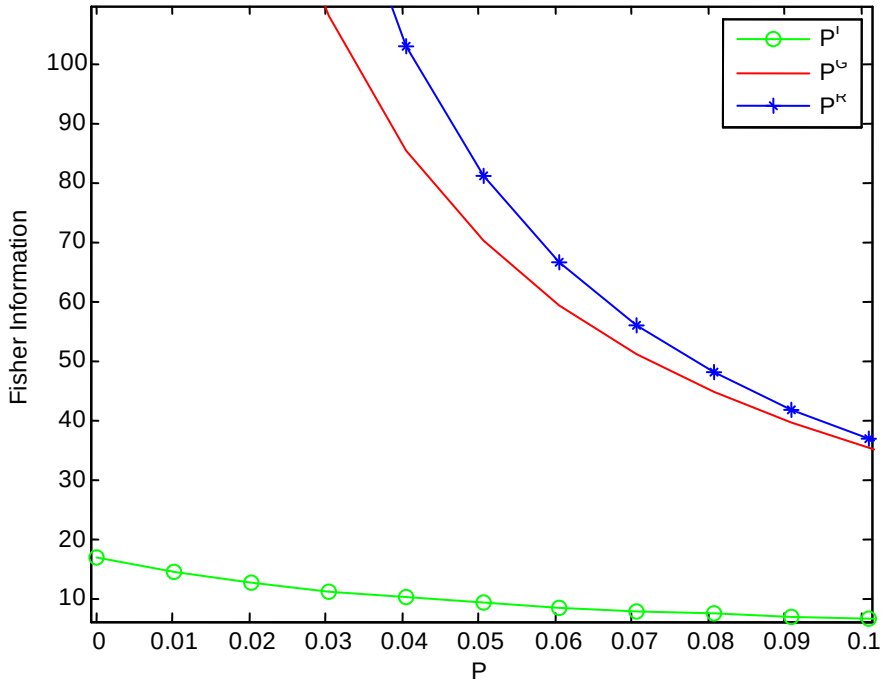
Figure 4.1(a): Plots of $I_x(.)$ versus p with $\eta = \beta = 0.99$ and $k = 2,\ 5,\ 10$

From Figure 4.1(a), increasing the value of k, and upon holding the sensitivity and specificity constant there is a decrease in Fisher information of P^R-experiment and the plot becomes steeper for small values of p but less steep as p increases. It is also observed that the plot of the P^R-experiment shifts to the left meaning the region for which the Fisher information of P^R-experiment is higher than the other two models shrinks. Also noted is that as the value of k increases from 2 to 10 the area in which the Fisher information of P^R-experiment is higher than the P^G-experiment decreases while the area in which the Fisher Information of P^I-experiment is higher than Fisher information of P^R-experiment increases.

$$\eta = \beta = 0.80$$



$\eta = \beta = 0.95$



$\eta = \beta = 0.99$

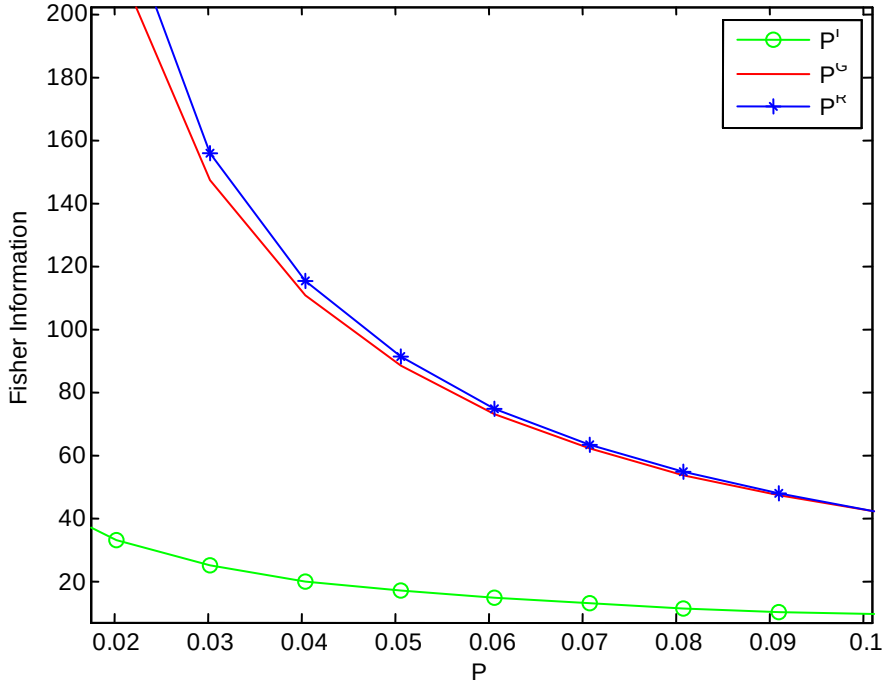


Figure 4.1(b): Plots of $I_x(.)$ versus p with $k = 5$ and $\eta = \beta = 0, 80,\ 0.95,\ 0.99$

From Figure 4.1(b), when k and p are held constant and increasing the values of η and β we observe that:

i) Fisher information of the P^R-experiment increases.

ii) The area between the plots of P^G- and P^R-experiments diminishes meaning that for almost perfect test kits, retesting of the already tested pools is not necessary, which is the case in practice.

iii) The region in which the Fisher information of the P^R-experiment is higher than that of the P^I-experiments increases.

From Figures 4.1(a) and 4.1(b), it can be observed that the plot of the Fisher information of the P^R-experiment is strictly decreasing function as the value of p increases for various values of k, η and β. We can also conclude that the Fisher information of P^R-experiment is higher than the other two models for small values of p. For the values of p close to 1 the Fisher information for P^I-experiment is higher. This means that in situation where p is higher P^I-experiment is viable.

4.3 Computation of cut off values

As defined in Section 3.5, let a_{ij} be the cut off value between $I_x(\mathrm{P}^I)$, $I_x(\mathrm{P}^G)$ and $I_x(\mathrm{P}^R)$ where a_{ij} is a unique root in (0, 1) of the equation $I_x(\mathrm{P}^i) = I_x(\mathrm{P}^j)$ for $i, j = I, G, R$, $i \neq j$ and $a_{ij} = a_{ji}$, we compute the point of intersection for each pair. Notice that we computed the cut off for the pair P^I and P^G in chapter 3, next is the computation of cut off for the pairs (P^I , P^R) and (P^G , P^R).

4.3.1 Computation of cut off values of $I_x(\mathrm{P}^I)$ and $I_x(\mathrm{P}^R)$

In this section, the cut off values of P^I- and P^R-experiment are computed.

If $I_x(\mathrm{P}^I) = I_x(\mathrm{P}^R)$ then

$$\frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)} = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\tau_3(1-\tau_3)}$$

$$\frac{1}{\tau_1(1-\tau_1)} = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta-\beta+1)^2}{\tau_3(1-\tau_3)}$$

and upon re-arranging, we have

$$\tau_3(1-\tau_3)(1-\mathrm{p})^2 - k^2(1-\mathrm{p})^{2k}(\eta-\beta+1)^2\tau_1(1-\tau_1) = 0 \ . \qquad (4.5)$$

To solve for p in Equation (4.5) is not easy, therefore we solve it iteratively as follows:

Define

$$f(\mathrm{p}) = \tau_3(1-\tau_3)(1-\mathrm{p})^2 - k^2(1-\mathrm{p})^{2k}(\eta-\beta+1)^2\tau_1(1-\tau_1), \tag{4.6}$$

then the function $f(\mathrm{p})$ is continuous in the interval $(0,1)$.

From Figure 4.1 of the graphs of Fisher information, a unique value of p for the Equation (4.6) exists such that $f(\mathrm{p}) = 0$ which is the point of intersection of the two curves of P^I- and P^R-experiments and the value of p can be solved iteratively using Newton-Raphson method. The Newton-Raphson **MATLAB** code is provided in Appendix C.

4.3.2 Computation of cut off values of $I_x(\mathrm{P}^G)$ and $I_x(\mathrm{P}^R)$

For $I_x(\mathrm{P}^G) = I_x(\mathrm{P}^R)$, we have

$$\frac{k^2(1-p)^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)} = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\tau_3(1-\tau_3)}$$

$$\frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)} = \frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2(\eta-\beta+1)^2}{\tau_3(1-\tau_3)}$$

$$\frac{(1-\mathrm{p})^{2k-2}}{\tau_2(1-\tau_2)} = \frac{(1-\mathrm{p})^{2k-2}(\eta-\beta+1)^2}{\tau_3(1-\tau_3)}.$$

Upon simplification

$$\tau_3(1-\tau_3)(1-\mathrm{p})^{2k-2} - (1-\mathrm{p})^{2k-2}(\eta-\beta+1)^2\tau_2(1-\tau_2) = 0$$

$$(1-\mathrm{p})^{2k-2}\{\tau_3(1-\tau_3) - (\eta-\beta+1)^2\tau_2(1-\tau_2)\} = 0.$$

Working similarly as in the previous sections in chapter 3, with

$$f(\mathrm{p}) = \tau_3(1-\tau_3) - (\eta-\beta+1)^2\tau_2(1-\tau_2),$$

the iteration is

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

and it stops if $|p_{i+1} - p_i| < \varepsilon$ for some arbitrary ε. The **MATLAB** code is presented in Appendix D.

Next, we compute the roots of *f(p)* in Equation (4.6) and Equation (4.8). The results are presented in Table 4.1.

Table 4.1: Cut off values 'a' of P^I-, P^G- and P^R-experiments for various values of η, β and k

k	$\eta=\beta=0.80$			$\eta=\beta=0.90$			$\eta=\beta=0.95$			$\eta=\beta=0.99$		
	a_{IG}	a_{IR}	a_{GR}	a_{IG}	a_{IR}	a_{GR}	a_{IG}	a_{IR}	a_{GR}	a_{IG}	a_{IR}	a_{GR}
2	0.528	0.500	0.394	0.563	0.515	0.334	0.596	0.549	0.312	0.646	0.623	0.296
3	0.446	0.422	0.284	0.477	0.438	0.237	0.507	0.469	0.221	0.555	0.536	0.209
5	0.348	0.329	0.182	0.371	0.342	0.150	0.395	0.366	0.139	0.439	0.422	0.131
10	0.234	0.222	0.095	0.248	0.229	0.078	0.263	0.244	0.072	0.296	0.282	0.068
15	0.181	0.172	0.065	0.190	0.177	0.053	0.201	0.186	0.049	0.227	0.215	0.046
20	0.150	0.142	0.049	0.156	0.145	0.040	0.164	0.164	0.037	0.185	0.175	0.035
50	0.078	0.075	0.020	0.080	0.075	0.016	0.083	0.077	0.015	0.092	0.087	0.014

From Table 4.1, it can be noted that the cut off value between P^I and P^R and also between P^G and P^R decreases as the value of k increases holding specificity and sensitivity constant. As the specificity and sensitivity increases the cut off value between the P^I- and P^R-experiments increases while that of P^G and P^R decreases. In general from Table 4.1 we can conclude that as k increases, the cut off value decreases for various values of η and β *i.e.* the region in which the P^R-experiment is better than the P^G and P^I shrinks. It can also be noted that the region in which the P^I-experiment is better than the P^G- and P^R-experiments enlarges as the pool size increases. As the sensitivity and specificity of the tests increases, the region in which the P^R-experiment is better decreases and the region in which P^I-experiment is better than the P^G- and P^R-experiments increases. For example if $\eta=\beta=0.80$, k = 3 and N tests are available, then the allocation that maximizes the information about p is

$$N=\begin{cases} observe\ all\ p^R,\ if\ \ p<0.283 \\ observe\ all\ p^G,\ if\ \ 0.283<p<0.446 \\ observe\ all\ p^I,\ if\ \ p>0.446 \\ arbitrary\ p^G\ or\ p^R,\ if\ \ p=0.283 \\ arbitrary\ p^I\ or\ p^G,\ if\ \ p=0.446 \end{cases}$$

For generality, if N tests are available, then the allocation that maximizes the information about p is

$$N=\begin{cases} observe\ all\ \ p^R,\ if\ \ p<a_{GR} \\ observe\ all\ \ p^G,\ if\ \ a_{GR}<p<a_{IG} \\ observe\ all\ \ p^I,\ if\ \ p>a_{IG} \\ arbitrary\ \ p^G\ or\ \ p^R,\ if\ \ p=a_{GR} \\ arbitrary\ \ p^I\ or\ \ p^G,\ if\ \ p=a_{IG} \end{cases}$$

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 off.

4.4 Estimation of *p*

The maximum likelihood estimation method is used in this section to estimate the value of p using P^R-experiment only and the joint experiment model. The number of observations from P^R-experiment is assumed to be r and $N=m+n+r$, total number of observations from the three experiments.

4.4.1 Estimation of *p* from the P^R-experiment

If r observations from the P^R-experiment are used to estimate p and X_{3z} pool test positive. Then $X_{3z} \sim Bernouli(\tau_3)$, and for a single observation

$$f(\mathrm{x}_{3z},\mathrm{p}/\eta,\beta)=\tau_3^{x_{3z}}(1-\tau_3)^{1-x_{3z}} \quad . \tag{4.9}$$

The likelihood function of Equation (4.9) is

$$\begin{aligned} \mathrm{L}(\mathrm{x}_{3z},\mathrm{p}/\eta,\beta) &= \prod_{z=1}^{r}(\tau_3^{x_{3z}}(1-\tau_3)^{1-x_{3z}}) \\ &= \tau_3^{\sum_{z=1}^{r}x_{3z}}(1-\tau_3)^{r-\sum_{z=1}^{r}x_{3z}} \end{aligned}$$

and taking log on both sides we have

$$\log\mathrm{L}(\cdot)=\sum_{z=1}^{r}x_{3z}\log(\tau_3)+(\mathrm{r}-\sum_{z=1}^{r}x_{3z})\log(1-\tau_3)\,. \tag{4.10}$$

The first derivative of equation (4.10) is

$$\frac{d\log\mathrm{L}(\cdot)}{d\,q}=\frac{\sum_{z=1}^{r}x_{3z}}{\tau_3}\frac{d\tau_3}{dq}+\frac{(\mathrm{r}-\sum_{z=1}^{r}x_{3z})}{1-\tau_3}\frac{d(1-\tau_3)}{dq}\,.$$

Equating to zero and upon simplifying we have

$$\frac{\sum_{z=1}^{r} x_{3z}}{\tau_3}\frac{d\tau_3}{dq}+\frac{(\mathrm{r}-\sum_{z=1}^{r} x_{3z})}{1-\tau_3}\frac{d(1-\tau_3)}{dq}=0$$

$$\left(\frac{d\tau_3}{dq}\right)\left(\frac{\sum_{z=1}^{r} x_{3z}}{\tau_3}-\frac{(\mathrm{r}-\sum_{z=1}^{r} x_{3z})}{1-\tau_3}\right)=0.$$

The function $\frac{d\tau_3}{dq}\neq 0$ because it is a function q. Therefore we are left with

$$\frac{\sum_{z=1}^{r} x_{3z}}{\tau_3}-\frac{(\mathrm{r}-\sum_{z=1}^{r} x_{3z})}{1-\tau_3}=0$$

$$\tau_3=\frac{\sum_{z=1}^{r} x_{3z}}{r}$$

$$\eta^2(1-q^k)+(1-\beta)^2\,\mathrm{q}^k=\frac{\sum_{z=1}^{r} x_{3z}}{r}$$

hence

$$\hat{q}_r^R=\left(\frac{\eta^2-\frac{\sum_{z=1}^{r} x_{3z}}{r}}{\eta^2-(1-\beta)^2}\right)^{\frac{1}{k}}.$$

Therefore the maximum likelihood estimator of p is

$$\hat{p}_r^R=1-\left(\frac{\eta^2-\frac{\sum_{z=1}^{r} x_{3z}}{r}}{\eta^2-(1-\beta)^2}\right)^{\frac{1}{k}}. \qquad (4.11)$$

4.4.2 Estimation of p from the joint experiment model

If m is the number of observations from P^I-experiment, n is the number of observations from the P^G-experiment and r is the number of observations from the P^R-experiment, then the joint probability density function of the random variables X_{1i}, X_{2j} and X_{3z} from the P^I-, P^R- and P^R-experiments respectively is a multinomial probability density function given by the product of their respective density functions, since the random variables are assumed to be independent, then

$$f(\underline{x},\underline{p}\,|\,k,\eta,\beta)=\tau_1^{x_{1i}}(1-\tau_1)^{1-x_{1i}}\times\tau_2^{x_{2j}}(1-\tau_2)^{1-x_{2j}}\times\tau_3^{x_{3z}}(1-\tau_3)^{1-x_{3z}} \qquad (4.12)$$

whose joint likelihood function is

$$L(\underline{x},\underline{p}\,|\,k,\eta,\beta)=\left\{[\tau_1]^{\sum_{i=1}^{m}x_{1i}}[1-\tau_1]^{m-\sum_{i=1}^{m}x_{1i}}\times[\tau_2]^{\sum_{j=1}^{n}x_{2j}}[1-\tau_2]^{n-\sum_{j=1}^{n}x_{2j}}\times[\tau_3]^{\sum_{z=1}^{r}x_{3z}}[1-\tau_3]^{r-\sum_{z=1}^{r}x_{3z}}\right\}$$

Proceeding as in the previous sections, we have

$$\frac{d\log L(.)}{dq}=\frac{\sum_{i=1}^{m}x_{1i}-m\tau_1}{\tau_1(1-\tau_1)}\frac{d\tau_1}{dq}+\frac{\sum_{j=1}^{n}x_{2j}-n\tau_2}{\tau_2(1-\tau_2)}\frac{d\tau_2}{dq}+\frac{\sum_{z=1}^{r}x_{3z}-r\tau_3}{\tau_3(1-\tau_3)}\frac{d\tau_3}{dq}$$

where $\frac{d\tau_1}{dq}=1-\eta-\beta$, $\frac{d\tau_2}{dq}=kq^{k-1}(1-\eta-\beta)$ and $\frac{d\tau_3}{dq}=kq^{k-1}\left((1-\beta)^2-\eta^2\right)$.

Thus

$$\frac{\sum_{i=1}^{m}x_{1i}-m\tau_1}{\tau_1(1-\tau_1)}\frac{d\tau_1}{dq}+\frac{\sum_{j=1}^{n}x_{2j}-n\tau_2}{\tau_2(1-\tau_2)}\frac{d\tau_2}{dq}+\frac{\sum_{z=1}^{r}x_{3z}-r\tau_3}{\tau_3(1-\tau_3)}\frac{d\tau_3}{dq}=0.$$

Letting

$$f(q)=\frac{\sum_{i=1}^{m}x_{1i}-m\tau_1}{\tau_1(1-\tau_1)}\frac{d\tau_1}{dq}+\frac{\sum_{j=1}^{n}x_{2j}-n\tau_2}{\tau_2(1-\tau_2)}\frac{d\tau_2}{dq}+\frac{\sum_{z=1}^{r}x_{3z}-r\tau_3}{\tau_3(1-\tau_3)}\frac{d\tau_3}{dq}, \qquad (4.13)$$

then $f(q)$ is a function of q. Since k,β and η are known constants, a unique value of q that satisfy the equation exists since the graph of the Equation (4.13) cuts the q-axis at a point. The unique root of Equation (4.13) is the MLE of q. Since equation (4.13) cannot be put in simpler form to obtain q, we apply Newton-Raphson procedure to obtain the estimates of q from which the MLE of p is obtained. A **MATLAB** for obtaining $\hat{q}_{mle}$ is presented in Appendix E.

4.5 Properties of the Estimators

As in Section 3.7, the properties of the maximum likelihood estimates of the prevalence rate are discussed and their asymptotic variance derived.

4.5.1 Variance of $\hat{p}_r^R$ of the P^R-experiment

As in Section 3.7.1, finding the logarithm and the second derivative with respect to p of Equation (4.12) yields

$$\frac{d^2\,\mathrm{logf}(\cdot)}{d\,\mathrm{p}^2}=\left(\frac{-x_3}{\tau_3{}^2}-\frac{1-x_3}{(1-\tau_3)^2}\right)\left(\eta^2k(1-\mathrm{p})^{k-1}-k(1-\beta)^2(1-\mathrm{p})^{k-1}\right)^2+$$
$$\left(\frac{x_3}{\tau_3}-\frac{1-x_3}{(1-\tau_3)}\right)\left(-\eta^2k(\mathrm{k}-1)(1-\mathrm{p})^{k-2}+k(\mathrm{k}-1)(1-\beta)^2(1-\mathrm{p})^{k-2}\right).$$

Therefore

$$-\mathrm{E}\left(\frac{\mathrm{d}^2\,\mathrm{logf}(\cdot)}{dp^2}\right)=\frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\tau_3(1-\tau_3)}$$

hence the asymptotic variance of $\hat{p}_r^R$ of the P^R-experiment is

$$\mathrm{var}(\hat{p}_r^R)=\frac{\tau_3(1-\tau_3)}{\mathrm{k}^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}\,. \tag{4.14}$$

4.5.2 Variance of $\hat{p}_{mle}$ of the joint experiment model

The asymptotic variance of $\hat{p}_{mle}$ of the joint experiment is obtained by solving $\left\{-E\left(\frac{d^2\log f(\cdot)}{dp^2}\right)\right\}^{-1}$ where $f(\cdot)$ is provided in equation (4.12), the asymptotic variance is

$$-E\left(\frac{d^2\log\mathrm{f}(\cdot)}{d(\mathrm{p}^2)}\right)=\frac{(\eta+\beta-1)^2}{\tau_1(1-\tau_1)}+\frac{k^2(1-\mathrm{p})^{2k-2}(\eta+\beta-1)^2}{\tau_2(1-\tau_2)}+\frac{k^2(1-\mathrm{p})^{2k-2}(\eta^2-(1-\beta)^2)^2}{\tau_3(1-\tau_3)}$$
$$=(\eta+\beta-1)^2\left(\frac{1}{\tau_1(1-\tau_1)}+\frac{k^2(1-\mathrm{p})^{2k-2}}{\tau_2(1-\tau_2)}+\frac{k^2(1-\mathrm{p})^{2k-2}(\eta-\beta+1)^2}{\tau_3(1-\tau_3)}\right)$$

hence assuming m, n and r are the total number of observations of P^I-, P^G- and P^R-experiments respectively then

$$\mathrm{var}(\hat{p}_{mle})=\frac{\tau_1\tau_2\tau_3(1-\tau_1)(1-\tau_2)(1-\tau_3)}{Q} \tag{4.15}$$

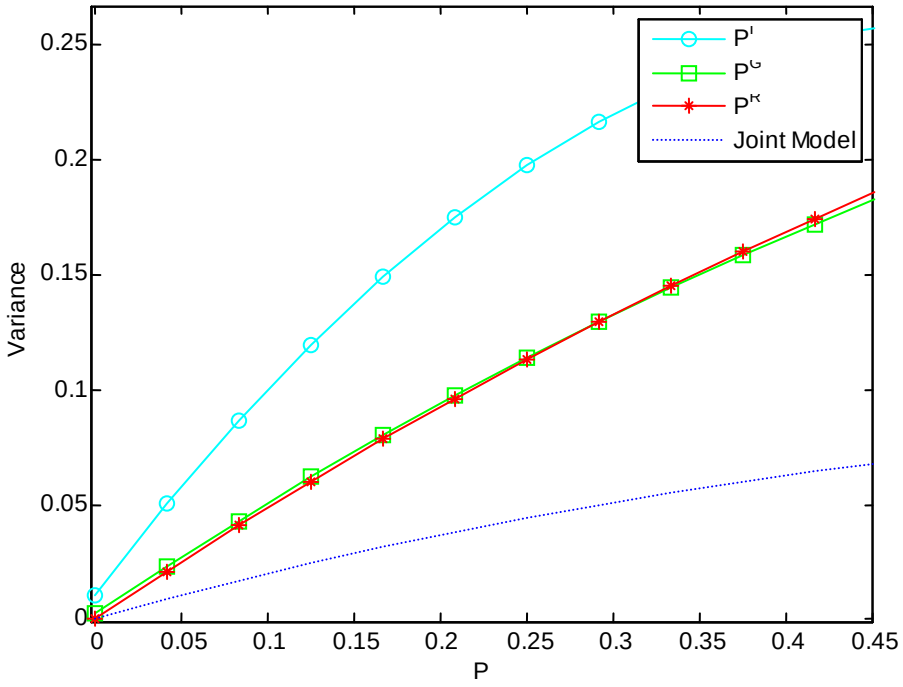
where

$$Q=m(\eta+\beta-1)^2\tau_2\tau_3(1-\tau_2)(1-\tau_3)+nk^2(1-p)^{2k-2}(\eta+\beta-1)^2\tau_1\tau_3(1-\tau_1)(1-\tau_3)$$
$$+rk^2(1-p)^{2k-2}(\eta^2-(1-\beta)^2)^2\tau_1\tau_2(1-\tau_1)(1-\tau_2).$$

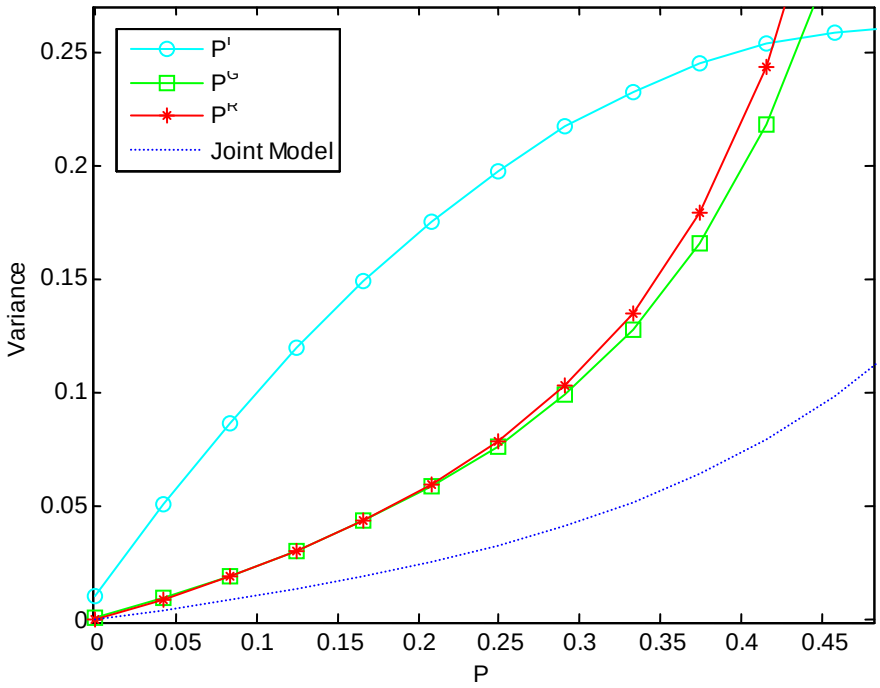
4.6 Comparing the variances for P^I-, P^G-, P^R-experiments and the joint experiment

In this section we shall plot the graphs of the variance of P^I-, P^G- and P^R-experiments and joint experiment model versus p for comparison purposes.

$k = 2$


P^I
P^G
P^K
Joint Model
Variance
P
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45

$k = 5$.


P^I
P^G
P^K
Joint Model
Variance
P
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45

$k = 10$

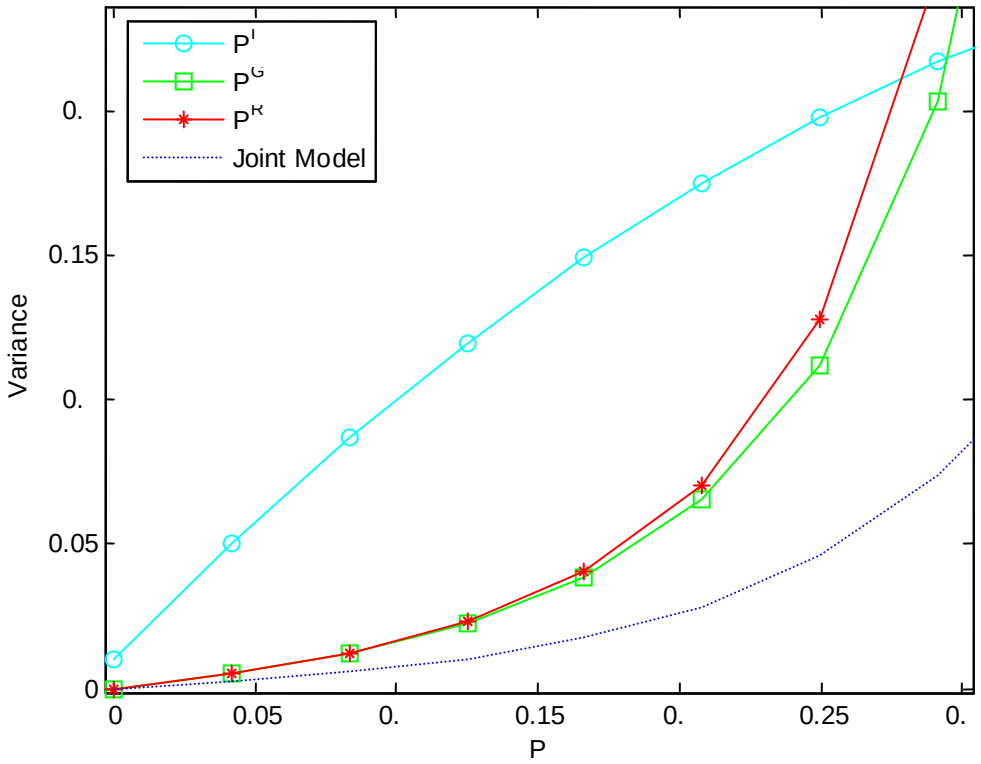


Figure 4.2(a): plots of $Var(\hat{p})$ versus p with $\eta = \beta = 0.99$ and $k = 2,\ 5,\ 10$.

Observed from Figures 4.2(a), is that as the value of k increases from 2 to 10 holding sensitivity and specificity of the tests constant:

i) The graph of $\text{var}(\hat{p}_r^R)$ shifts to the left.

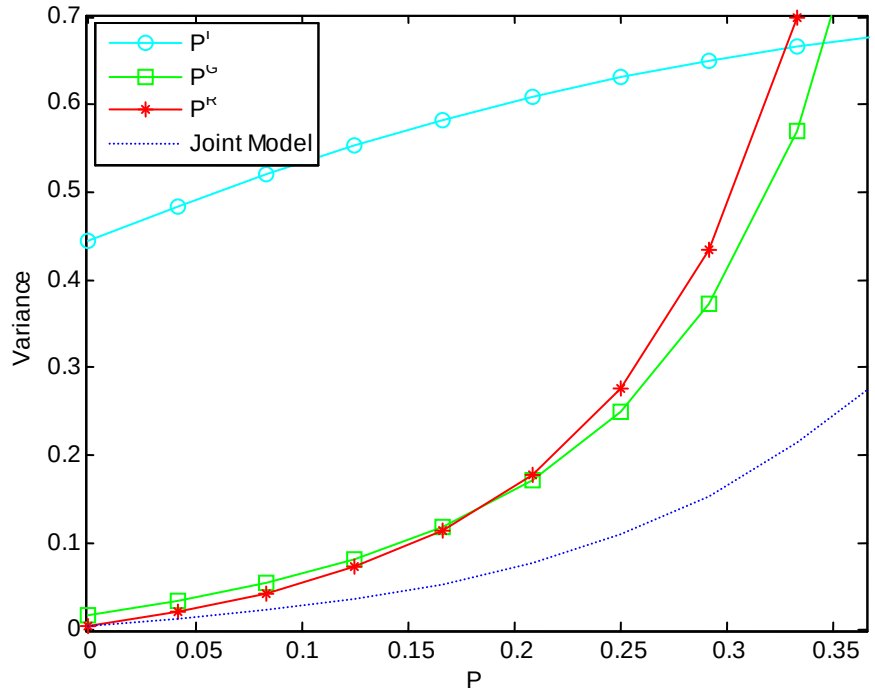
ii) The region in which $\text{var}(\hat{p}_m^I)$ is higher than $\text{var}(\hat{p}_r^R)$ shrinks.

iii) The region in which $\text{var}(\hat{p}_n^G)$ is higher than $\text{var}(\hat{p}_r^R)$ increases.

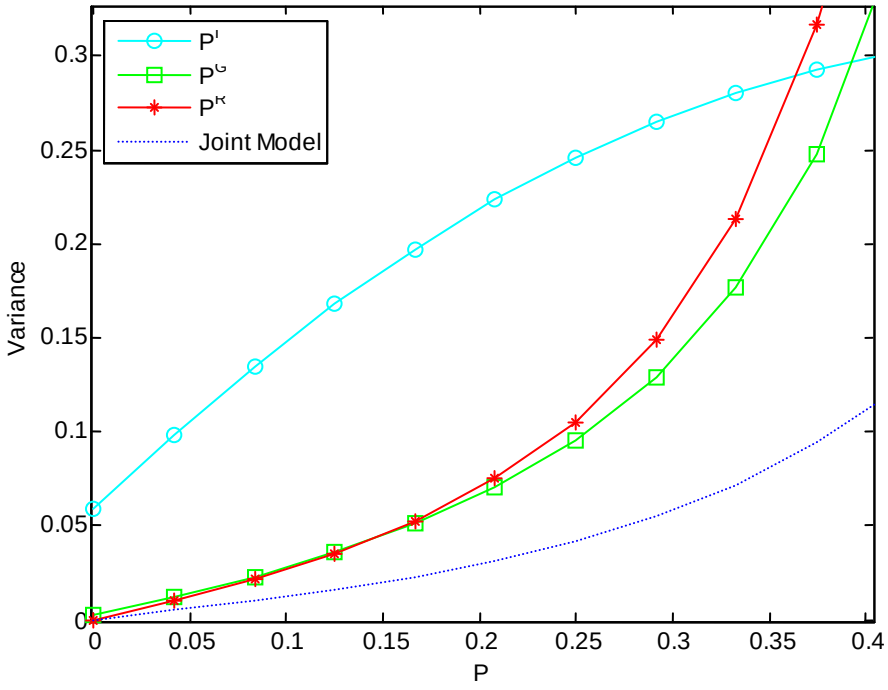
iv) The area in which $\text{var}(\hat{p}_r^R)$ is smaller than variance of the other models decreases.

We also plot $\text{var}(\cdot)$ for fixed k but varying η *and* β. We have the following graphs:

$\eta = \beta = 0.80$



$\eta = \beta = 0.95$



$\eta = \beta = 0.99$

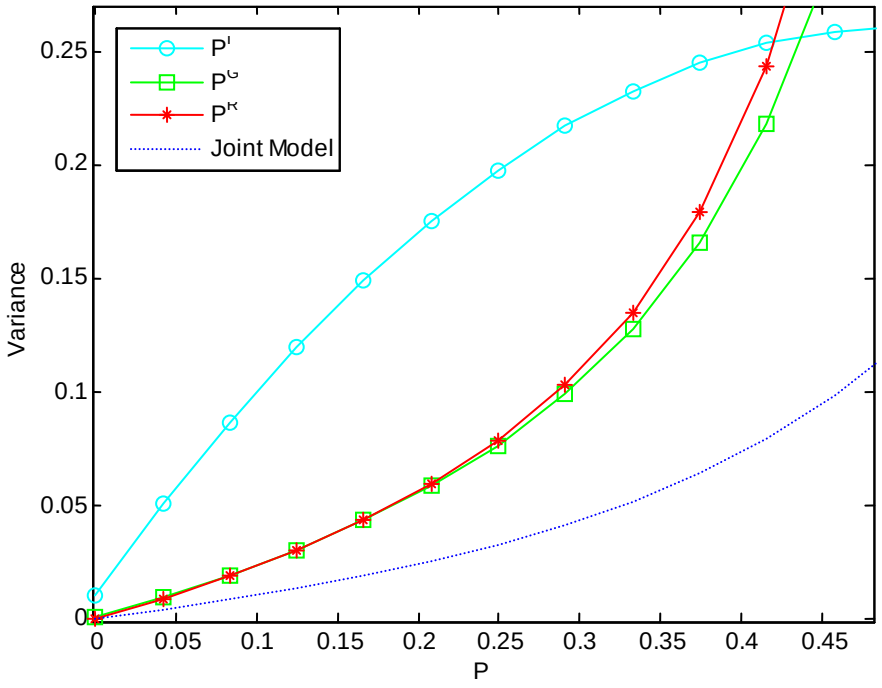


Figure 4.2(b): Plots of $Var(\hat{p})$ versus p with $k = 5$ and $\eta = \beta = 0.80,\ 0.95,\ 0.99$

As observed from Figures 4.2(b), keeping k constant, increasing sensitivity and specificity of the test kits:

i) The graph of $\text{var}(\hat{p}_r^R)$ shifts to the left.

ii) The region in which $\text{var}(\hat{p}_m^I)$ is higher than $\text{var}(\hat{p}_r^R)$ increases.

iii) The region in which $\text{var}(\hat{p}_n^G)$ is higher than $\text{var}(\hat{p}_r^R)$ decreases.

We note that $\text{var}(\hat{p}_r^R)$ increase exponentially as the value of p increases. It is noted also that $\text{var}(\hat{p}_r^R)$ is smaller than $\text{var}(\hat{p}_m^I)$ and $\text{var}(\hat{p}_n^G)$ for values of p close to 0, for values of p close to 1 $\text{var}(\hat{p}_m^I)$ is smaller than $\text{var}(\hat{p}_r^R)$ and $\text{var}(\hat{p}_n^G)$ while for some values of p between 0 and 1, $\text{var}(\hat{p}_n^G)$ is smaller than the variance of the other two models. Noted also is that variance of the combined experiment model of $\hat{p}_{mle}$ is smaller than variance of other three experiments under consideration in this chapter for the given values of η, β and k. Therefore a combination of experiments yields superior estimators.

4.7 Asymptotic Relative Efficiency for the three experiments

In this section, $\text{var}(\hat{p}_{mle})$, $\text{var}(\hat{p}_m^I)$, $\text{var}(\hat{p}_n^G)$ and $\text{var}(\hat{p}_r^R)$ are compared. This is accomplished by computing asymptotic relative efficiency (ARE) values for $\eta = \beta = 0.99,\ 0.80$; $k = 2,\ 3,\ 5,\ 10$ and $p = 0.01,\ 0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.30$. Let $ARE^3 = \dfrac{\text{var}(\hat{p}_{mle})}{\text{var}(\hat{p}_m^I)}$, $ARE^4 = \dfrac{\text{var}(\hat{p}_{mle})}{\text{var}(\hat{p}_n^G)}$ and $ARE^5 = \dfrac{\text{var}(\hat{p}_{mle})}{\text{var}(\hat{p}_n^G)}$. The computed ARE's are tabulated in the following tables.

Table 4.2: The ARE's of the joint experiment relative to the P^I-, P^G- and P^R-experiments with $\eta = \beta = 0.80$

p		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^3	0.056	0.028	0.012	0.004
	ARE^4	0.215	0.235	0.264	0.314
	ARE^5	0.729	0.736	0.723	0.682
0.05	ARE^3	0.087	0.052	0.020	0.016
	ARE^4	0.289	0.332	0.382	0.443
	ARE^5	0.624	0.616	0.589	0.541
0.10	ARE^3	0.115	0.077	0.051	0.040
	ARE^4	0.333	0.380	0.429	0.483
	ARE^5	0.551	0.543	0.520	0.476
0.15	ARE^3	0.139	0.102	0.079	0.05
	ARE^4	0.356	0.401	0.446	0.484
	ARE^5	0.505	0.497	0.475	0.421
0.20	ARE^3	0.161	0.128	0.117	0.216
	ARE^4	0.369	0.410	0.448	0.436
	ARE^5	0.470	0.462	0.434	0.347
0.30	ARE^3	0.205	0.193	0.251	0.720
	ARE^4	0.378	0.407	0.406	0.163
	ARE^5	0.417	0.400	0.343	0.118

Table 4.3: The ARE's of the joint experiment relative to the P^I-, P^G- and P^R-experiments with $\eta = \beta = 0.99$

p		$k=2$	$k=3$	$k=5$	$k=10$
0.01	ARE^3	0.131	0.089	0.053	0.027
	ARE^4	0.349	0.392	0.432	0.465
	ARE^5	0.520	0.520	0.515	0.508
0.05	ARE^3	0.184	0.132	0.087	0.052
	ARE^4	0.390	0.422	0.450	0.472
	ARE^5	0.426	0.446	0.464	0.476
0.10	ARE^3	0.198	0.148	0.104	0.074
	ARE^4	0.393	0.421	0.446	0.466
	ARE^5	0.409	0.431	0.450	0.460
0.15	ARE^3	0.207	0.160	0.121	0.104
	ARE^4	0.392	0.418	0.441	0.456
	ARE^5	0.401	0.422	0.439	0.440
0.20	ARE^3	0.215	0.172	0.140	0.152
	ARE^4	0.390	0.414	0.433	0.439
	ARE^5	0.395	0.415	0.427	0.409
0.30	ARE^3	0.232	0.199	0.195	0.369
	ARE^4	0.384	0.403	0.411	0.351
	ARE^5	0.384	0.398	0.394	0.281

From Tables 4.2 and 4.3 it is noted that the computed values of $ARE's$ are less than 1 for the given values of η, β, k and p hence the joint experiment model is superior to P^I-, P^G- and P^R-experiments.

4.8 Estimates of prevalence rate, variance and confidence interval

The maximum likelihood estimates($\hat{p}$) of the prevalence rate of the joint experiment model, the variance and 95% Wald-type confidence interval for $k = 5, 10$ and $\eta = \beta = 80\%, 90\%$ are computed.

Table 4.4: Maximum likelihood estimates, 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	6.8293×10^{-5}	0.00000, 0.03999
	0.05	0.06397	1.7588×10^{-4}	0.01600, 0.11193
	0.10	0.10386	2.8551×10^{-4}	0.04407, 0.16366
	0.15	0.17263	5.5743×10^{-4}	0.09856, 0.24671
	0.30	0.33119	2.1054×10^{-3}	0.23895, 0.42344
$k = 10$	0.01	0.01745	2.8592×10^{-5}	0.00000, 0.04312
	0.05	0.03052	4.5378×10^{-5}	0.00000, 0.06428
	0.10	0.08585	1.6443×10^{-4}	0.03094, 0.14076
	0.15	0.12212	3.2699×10^{-4}	0.05794, 0.18630
	0.30	0.28662	4.2189×10^{-3}	0.19800, 0.37525

Table 4.5: Maximum likelihood estimates, 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	3.7091×10^{-5}	0.00000, 0.04772
	0.05	0.05229	8.2334×10^{-5}	0.00866, 0.09592
	0.10	0.09213	1.4452×10^{-4}	0.03544, 0.14881
	0.15	0.16454	2.9124×10^{-4}	0.09187, 0.23721
	0.30	0.29088	7.6689×10^{-4}	0.20187, 0.37990
$k = 10$	0.01	0.00971	9.3065×10^{-6}	0.00000, 0.02893
	0.05	0.04671	4.1100×10^{-5}	0.00535, 0.08807
	0.10	0.10616	1.3268×10^{-4}	0.04578, 0.16653
	0.15	0.13960	2.2985×10^{-4}	0.07168, 0.20753
	0.30	0.27932	1.8171×10^{-3}	0.19139, 0.36726

From Tables 4.4 and 4.5 it is observed that the maximum likelihood estimates of the prevalence rate are very close to the actual values which were used to simulate the estimates. The population estimates resulting from the experiments are used to compute the $(1-\alpha)100\%$ confidence limits of the confidence interval of the simulated estimates where α

is the level of significance and it is noted from Tables 4.4 and 4.5 that the actual value is within the limits.

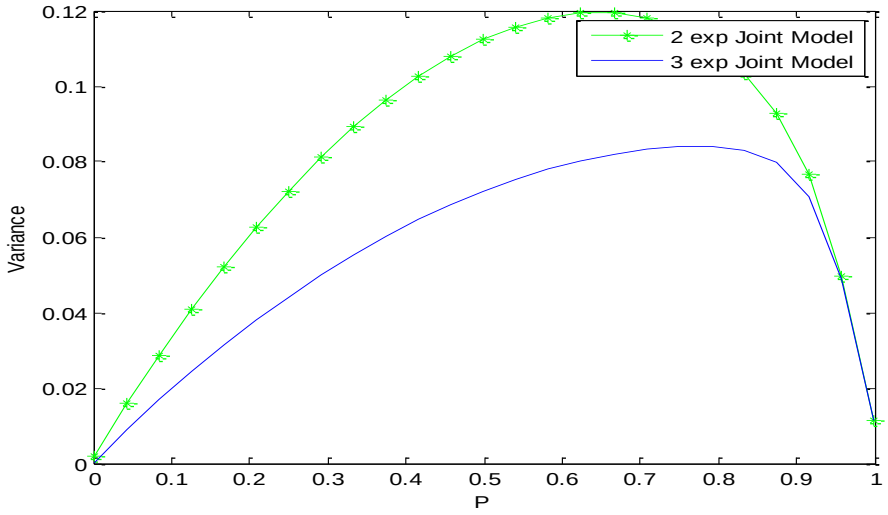
4.9 Comparing the joint experiments

The joint experiment models in Chapter 3 and Chapter 4 are compared in this section by plotting the graphs of their variances. Asymptotic relative efficiency (ARE) values of two and three joint experiment models are computed.

4.9.1 Comparing variances of the joint experiment models.

The variances of two and three experiment joint models are compared in this section by plotting the graphs of their respective variances against p for values of $\eta = \beta = 0.99, 0.80$ and $k = 2, 10$.

$$k = 2$$



$k = 10$

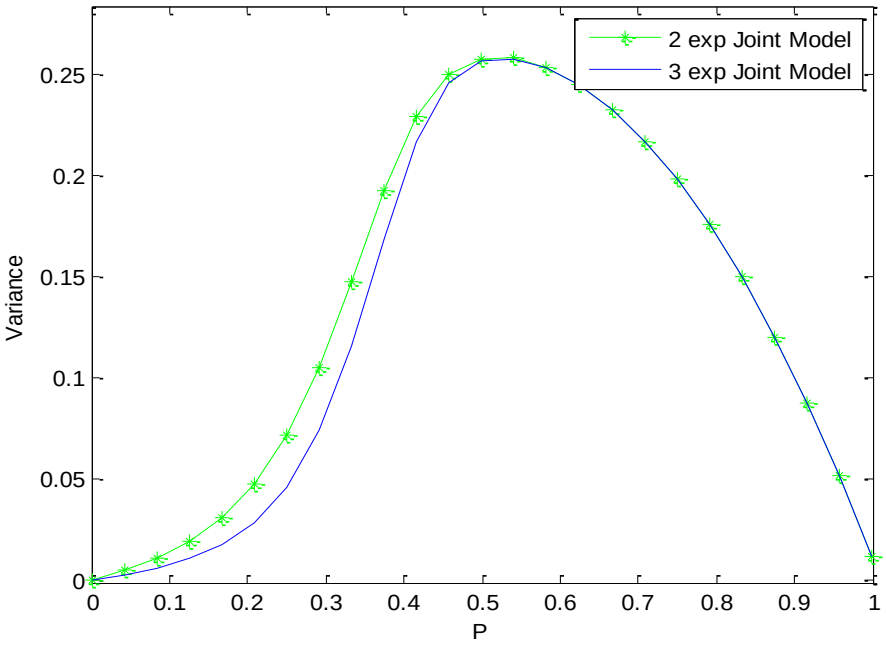
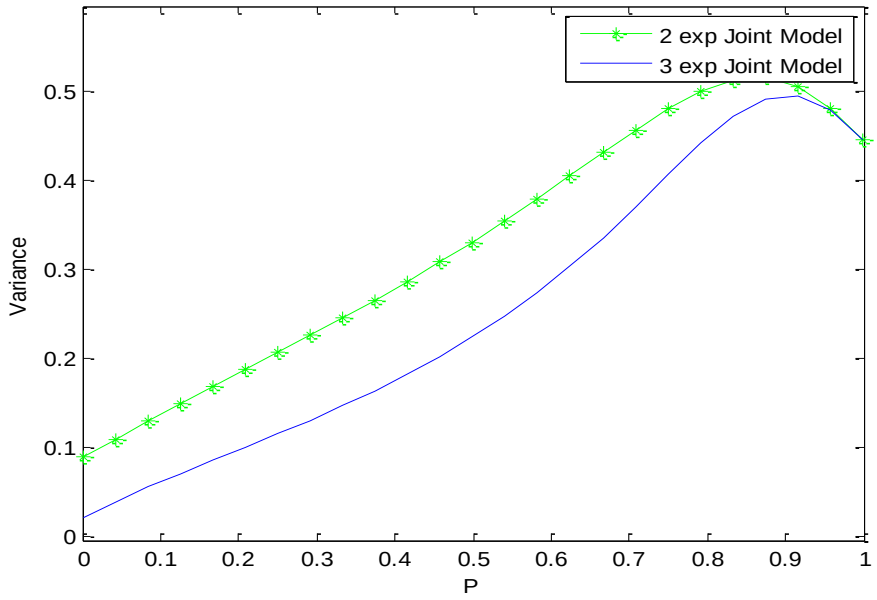


Figure 4.3(a): Plots of $var(\hat{p})$ versus p for joint experiment models with $\eta = \beta = 0.99$ and $k = 2,\ 10$.

As seen from Figures 4.3(a), the area between the two curves decreases as k increases keeping sensitivity and specificity constant

$$\eta = \beta = 0.80$$



$$\eta = \beta = 0.99$$

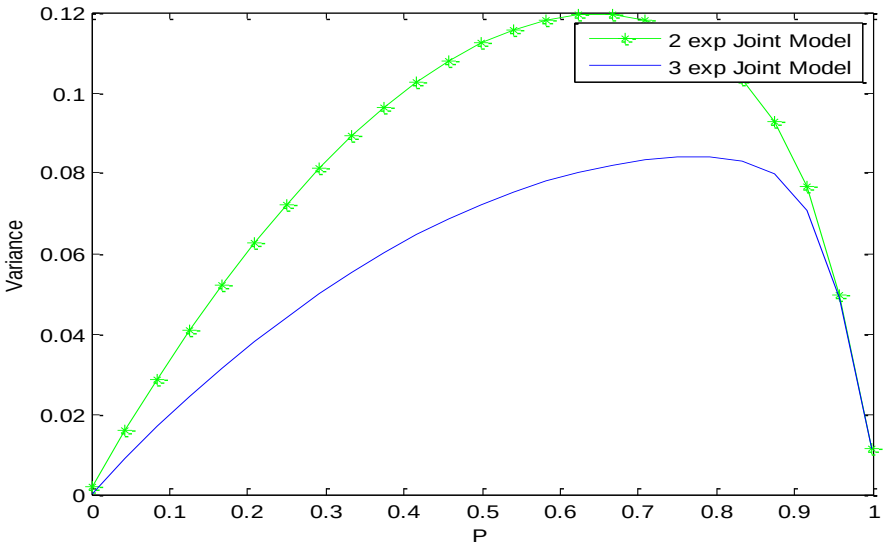


Figure 4.3(b): Plots of $var(\hat{p})$ versus p for joint experiment models with $k = 2$ and $\eta = \beta = 0,80,\ 0.99$.

As seen from Figure 4.3(b), keeping *k* constant, increasing sensitivity and specificity of the test kits increases the area between the two curves. The three experiment joint model has a smaller variance than two experiment model. Thus more information about the prevalence rate will be obtained when the joint three experiment model is applied.

From Figure 4.3(a) and 4.3(b), the general behavior of the two graphs is the same *i.e* increases exponentially for small values of *p* and decreases thereafter as the value of *p* gets closer to 1. It is also observed that as the value of the parameter *p* increases especially for small sample size, the gap between the two curves increases and decreases thereafter as *p* gets closer to 1. For the given values of η, β and k it is observed that the variance of the joint experiment model for three experiments is smaller or equal to variance of the joint experiment model for two experiments for the entire range of *p* hence we can conclude that the three experiment joint model is better than the two experiment joint model.

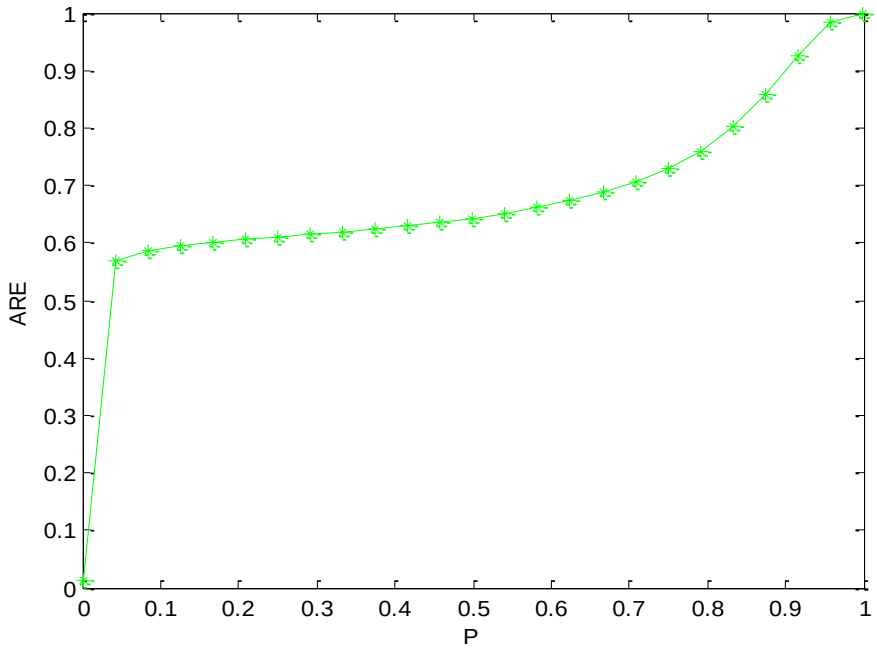
4.9.2 Asymptotic Relative Efficiency of the joint experiment models

In this section the asymptotic relative efficiency (ARE) of two and three experiment joint models is computed. The computed ARE values are plotted against p for values of

$\eta = \beta = 0.99,\ 0.80$ and $k = 2,\ 5,\ 10$ as shown in Figure 4.4(a) and 4.4(b). If the estimator of the two experiment joint model is denoted by $\hat{p}^1_{mle}$ and the estimator of the three experiment joint model is denoted by $\hat{p}^2_{mle}$, then we compute ARE as:

$$ARE = \frac{\text{var}(\hat{\text{p}}^2_{mle})}{\text{var}(\hat{\text{p}}^1_{mle})}.$$

$k = 2$



$k = 10$

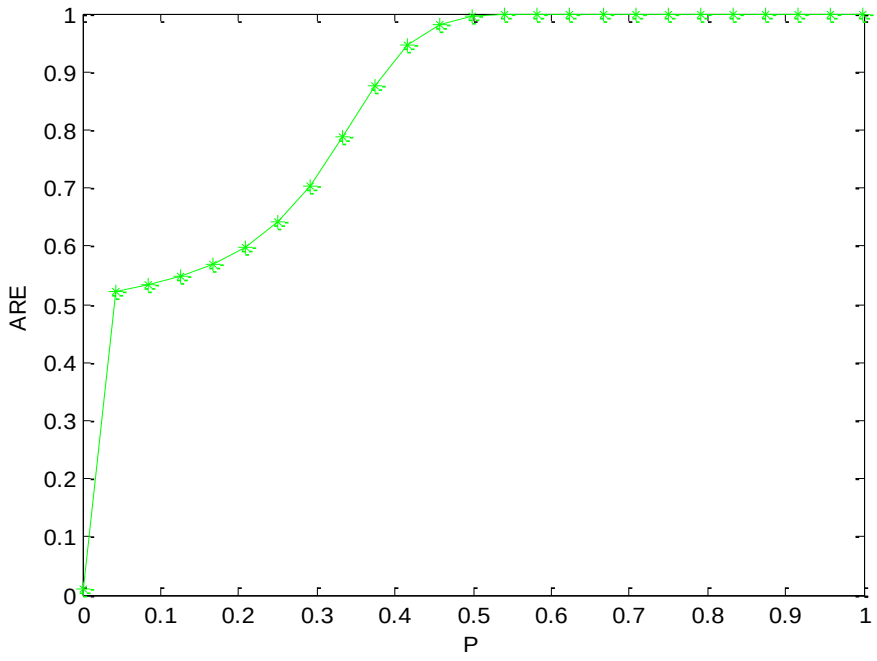
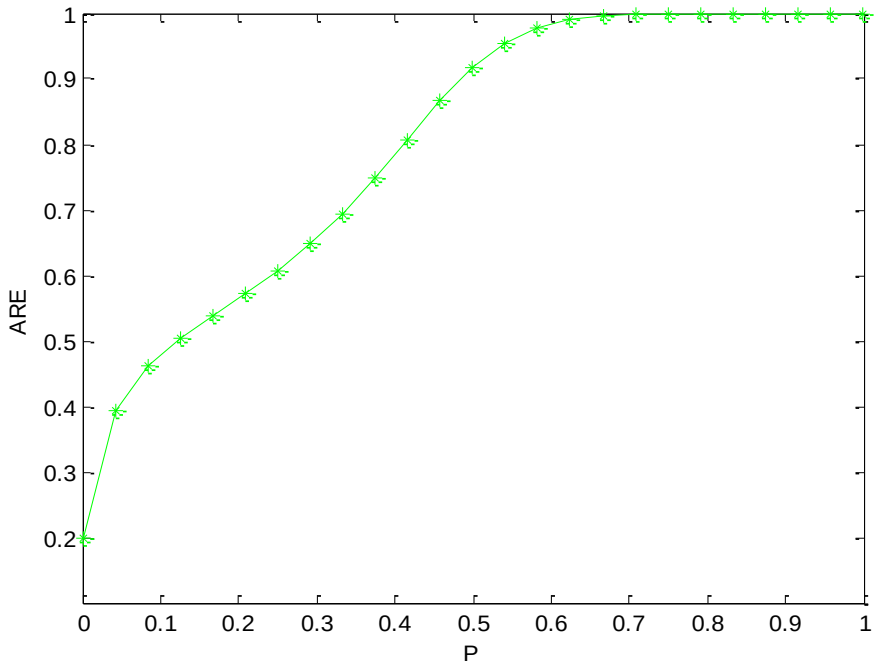


Figure 4.4 (a): ARE plotted against p for $\eta = \beta = 0.99$

As noted from Figure 4.4(a) the region in which the three experiment joint model is better than the two experiment joint model increases with increase in k. This means that as p increases, the two and three joint experiment models provides the same information about p. Now, suppose we change η *and* β, we have this results in Figure 4.4(b).

$\eta = \beta = 80$



$\eta = \beta = 99$

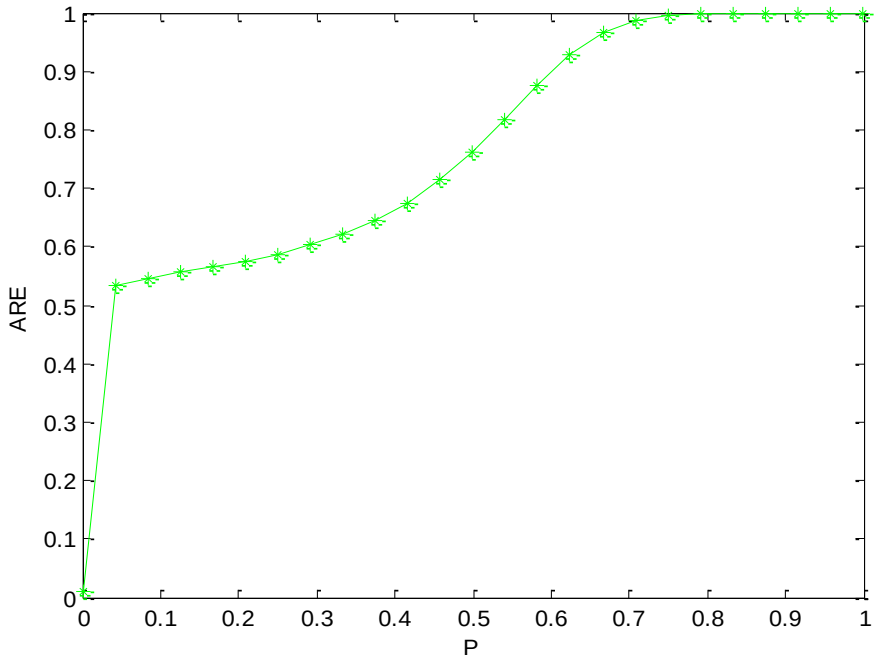


Figure 4.4 (b): ARE plotted against p for $k = 10$

From Figure 4.4(b), increasing sensitivity and specificity of the test kits while keeping k constant the region in which three experiment joint model is superior to two experiment joint model decreases. In general as noted from Figure 4.4(a) and 4.4(b), the computed values of ARE are less than or equal to 1 for the values of the parameter p between 0 and 1 hence the three experiment joint model is more efficient than the developed two experiment joint model as evidenced earlier in the other conclusions.

CHAPTER FIVE

RESULTS AND DISCUSSION

The essence of this study was to develop an adaptive procedure that selects one experiment at a time based on available information for optimal estimation of a common prevalence rate. Thus, a mixture of experiments have been applied mainly to two experiments namely P^I- and P^G-experiments and to three experiments constructed by appending experiment P^R on two earlier ones. A mixture of experiments is widely believed to yield better results in agricultural field. This study has proposed a mixture of experiments to estimate a prevalence rate of a common trait as proposed by Thomson (1962) who only applied P^G-experiment with absence of errors.

To help this discussion, ARE's values have been computed for various pool sizes 2 – 10 and tabulated in Table 5.1.

Table 5.1: The ARE's of P^G-, P^R-, two joint and three joint experiment models relative to the P^I-experiment with $p = 0.05,\ k = 2, 3, 5, 10$ and $\eta = \beta = 99\%$

k	$eff(\mathrm{P}^G, \mathrm{P}^I)$	$eff(\mathrm{P}^R, \mathrm{P}^I)$	$eff(\text{two joint exp}, \mathrm{P}^I)$	$eff(\text{three joint exp}, \mathrm{P}^I)$
2	0.471	0.431	0.320	0.184
3	0.313	0.296	.0238	0.132
5	0.193	0.187	0.162	0.087
10	0.109	0.108	0.098	0.052

The $eff(\mathrm{P}^G, \mathrm{P}^I)$ column of the table gives the ARE's of P^G-experiment to P^I-experiment which is widely discussed in pool testing literature. For instance see Brookmeyer (1999). The $eff(\mathrm{P}^R, \mathrm{P}^I)$ column gives the ARE's of P^R-experiment to P^I-experiment. This column gives the results of Nyongesa (2017). The $eff(\text{two joint exp}, \mathrm{P}^I)$ column gives ARE's of two joint experiment model of P^I- and P^G-experiment to P^I-experiment. A version of no test errors of this ARE's have been discussed (Hardwick *et. al.,* 1998). Hence contribution in this case is introduction of the test error element in Hardwick *et. al.,*(1998) model. Finally the $eff(\text{three joint exp}, \mathrm{P}^I)$ column gives ARE's of the three combined (P^I-, P^G- and P^R-) experiment model to P^I-experiment.

From Table 5.1 all ARE's < 1, this implies that all considered testing procedures in this study are superior to one-at-a-time testing. Hence the procedures are viable for estimating prevalence rate with relative to pool sizes. There is a decline in ARE's across the table with the minimum ARE's being achieved with three joined experiment model. For example for $k = 10,\ p = 0.05, \eta = \beta = 99\%,\ \ eff(\text{three joint exp}, \mathrm{P}^{I}) = 0.052$ which is almost two fold smaller than either procedure. Therefore the proposed three joint experiment model is superior to that of Hardwick *et. al.*,(1998), Nyongesa (2017) and Brookmeyer (1999). This study shows that a mixture of experiments yield superior results for estimating prevalence rate. In fact the more the mixture the experiment, the more superior the model is for estimating a common prevalence rate. This is true in practice as a mixture of experiment can yield worthwhile results particularly in agricultural field.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

This study focused on construction of new models for estimating the prevalence rate of a trait in a population with imperfect tests by selecting between two experiments namely P^I- and P^G-experiments and also by selecting between three experiments namely P^I-, P^G- and P^R-experiments. Ideally the model should select the better experiment and once the better experiment is being used, it should approximate the individual maximum likelihood estimator of the prevalence rate.

From Chapter three of 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. From the graphs of asymptotic variance and tables of ARE's, 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 each of the experiment P^I and P^G across the entire range of parameter estimation regardless of the pool size, sensitivity and specificity of the tests.

From Chapter four, the best estimator for small, medium and high values of the parameter p are $\hat{p}_r^R$, $\hat{p}_m^G$ and $\hat{p}_n^I$ from P^R-, P^G- and P^I-experiment respectively. Also noted is for almost perfect tests *i.e.* sensitivity and specificity of the test kits about 100% retesting of already tested pools is not necessary. From Table 4.2 and 4.3, the computed values of asymptotic relative efficiency (ARE) for various values of η, β, k and p are less than one hence the proposed joint model for sequentially choosing of the best experiment for optimal estimation of a trait with misclassification is more efficient than the P^I-, P^G- and P^R-models separately.

Comparison of the proposed two experiment joint model and three experiment joint model was done by plotting the curves of asymptotic variance of $\hat{p}_{mle}$ against p and by plotting the curves of ARE for various values of η, β, k and p. It was noted that the three experiment joint model does very well across the entire range of the parameter values than the two experiment joint model.

Through simulation using R-code program developed, the computed estimates were very close to the actual values used to simulate the estimates and they were within the upper and the lower $(1-\alpha)100\%$ Wald-type confidence interval. Noted also is that the developed models are suitable for small values of the prevalence rate since the variance of estimates increases as the values of the estimates ($\hat{p}$) increases.

6.2 Recommendation

The models developed in this study of estimating the prevalence rate are recommended for use when incidence probability is relatively small, sensitivity and specificity of the tests high. Better estimation of prevalence rate has important health implications for prevention, intervention and treatment of rare diseases like HIV infections in a population hence the developed models are highly recommended for use. Our improved estimate of the prevalence rate could substantially reduce the potential risks for secondary transmission by infected population who are unaware of infections. Reliable estimates of the prevalence rate are also required in environmental monitoring where samples of units of soil are combined and tested for toxins in order to know the right chemicals or fertilizers to apply. Accurate prevalence rate estimates are required also in prevention and treatment of rare diseases in plant (Graham, 1996) and in early stages of drug discovery (Xie *et al.*, 2001).

6.3 Areas of further research

Mixture of experiments yields better estimators than individual estimators. It could be worthwhile to consider more experiments.

REFERENCES

Behets, F., Bortozzi, S., Kasali, M., Kashamuka, M., Atikala, L., Brown, C., Ryder, R. and Quinn, C. (1990). Successful use of Pooled Sera to Determine HIV-1 Seroprevalence in Zaire with development of Cost Efficiency Models. *AIDS*, **4**, 737 – 741.

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

Cahoon-Young, B., Chandler, A., Livermore, T., Gaudino, J. and Benjamin, R. (1989). Sensitivity and specificity of Pooled versus Individual Sera in a HIV-antibody Prevalence study. *J. Clin. Microbiol*., **27**, 1893 – 1895.

Chaubey, Y.P. and Li, W. (1993). Estimation of Fraction Defectives through Batch Sampling. *American statistical association: proceedings of the quality and productivity section*, 198-206. Alexandria, VA: ASA.

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

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.

Graham Hepworth (1996). Exact Confidence Intervals for Proportions Estimated by Group Testing. *Biometrics*, **52**, 1134-1146.

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.

Hardwick Janis, Connie Page, and Quentin F. Stout (1998). Sequentially Deciding Between Two Experiments for Estimating a Common Success Probability. *Journal of the American statistical association*. December 1998, **93**, 1502-1511.

Hepworth, G. and Watson, R. (2009). Debiased Estimation of Proportions in Group Testing. *Applied Statistics*, **58**, 105 – 121.

Johnson, N. I., Kotz, S. and Wu, X. (1991). Inspection Errors for Attributes in Quality Control. *London; Chapman and Hall*.

Juan Ding and Wenjun Xiong (2015). Robust Groups Testing for Multiple Traits with Misclassifications. *Journal of Applied Statistics,* **42**, 2115-2025.

Kline, R. L., Brothers, T., Brookmeyer, R., Zeger, S., and Quinn, T. (1989). Evaluation of Human Immunodeficiency Virus Seroprevalence in Population Surveys using Pooled Sera. *Journal of clinical microbiology,* **27**, 1449-1452.

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.

Lovison, G., Gore, S. D. and Patil, G. P. (1994). Design and Analysis of Composite Sampling Procedures: *A Review*. *In handbook of statistics, eds G.P. Patil & C.R. Rao,* **12**, 103-166.Amsterdam: Elsevier.

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.

Matiri, G., Nyongesa, K., and Islam, A. (2017). Sequentially Selecting Between Two Experiments for Optimal Estimation of a Trait with Misclassification. *American Journal of Theoretical and Applied statistics*, **6**, 79-89.

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

Nyongesa, L. K. and Syaywa, J. P. (2011). "Block Testing Strategy with Imperfect Tests and its Improved Efficient Testing Model for Donor Blood." *Communication in Statistics-Computational Statistics,* **40**, 3218-3229.

Nyongesa, L. K. (2017). Multiple Test for Estimating Prevalence Rate with Imperfect Test. *Communication in Statistics Theory and Method*. (Submitted).

Nyongesa, L. K. (2012). Dual Estimation of Prevalence and Disease Incidence in Pool-Testing Strategy. *Communication in Statistics Theory and Method*. **40**, 1 – 12.

Nyongesa, L. K. (2005). Hierarchical Screening with Retesting in a Low Prevalence population. *The Indian Journal of Statistics*, **66**, 779 – 790.

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

Nyongesa, L. K. (2004). Testing for the Presence of Disease by Pooling Samples. *Australian and New Zealand Journal of Statistics*, **46,** 383-390.

Phatarfod, R. M. and Sudbury, A. (1994). The Use of a Square Array Scheme in Blood Testing. *Statistics in medicine*, **13**, 2337-2343.

R Core Team, (2014). R: A Language and Environmental for Statistical computing. *R Foundation for Statistical computing*, Vienna, Austria.

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.

Syaywa, J. P. and Nyongesa, L. K. (2010). Pool Testing with Test Errors Made Easier. *International Journal of Computational Statistics*, **1**, 1-9.

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

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

Wanyonyi, R. W., Nyongesa, K. L. and Wasike, A. (2015). Estimation of Proportion of a Trait by Batch Testing Model in a Quality Control Process. *American Journal of Theoretical and Applied statistics*, **4**, 619 – 629..

Wein, L. M. and Zenios, A. S. (1996). Pool Testing for HIV Screening: Capturing the Dilution Effect. *Operations research*, **44**, 543 – 569.

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

APPENDICES

Appendix A: Matlab code for solving Equation (3.8)

```
syms p c k
f(p) = τ2(1−τ2)(1−p)^2 − k^2(1−p)^(2k) τ1(1−τ1)
while i < c
    p = p0 − f(p0)/f'(p0)        %Newton – Raphson method
                                 % f'(·) is the derivative f(p)
                                 % p0 is the initial approximation of the root of f(p)
 if |p − p0| < ε                 % stopping criterion
    fprint('cut off value is %f\n', double(p))
    return
 end
    i = i+1;
    p0 = p
end
```

Appendix B: Matlab code for solving Equation (3.18)

```
syms q c k m n η β
f(q) = (Σ_{i=1}^{m} x1i − m τ1)/(τ1(1−τ1)) dτ1/dq + (Σ_{j=1}^{n} x2j − nτ2)/(τ2(1−τ2)) dτ2/dq
while i < c
    q = q0 − f(q0)/f'(q0)        %Newton – Raphson method
                                 % f'(·) is the derivative f(·)
                                 % q0 is the initial approximation of the root of f(q)
 if |q − q0| < ε                 % stopping criterion
    fprint('mle of q is %f\n', double(q))
    return
 end
    i = i+1;
    q0 = q
end
```

Appendix C: Matlab code for solving Equation (4.6)

```
syms p c k
f(p) = τ3(1−τ3)(1−p)^2 − k^2(1−p)^(2k)(η−β+1)^2 τ1(1−τ1)
while i < c
    p = p0 − f(p0)/f'(p0)        %Newton – Raphson method
                                 % f'(·) is the derivative f(p)
                                 % p0 is the initial approximation of the root of f(p)
  if |p − p0| < ε                % stopping criterion
    fprint('cut off value is %f\n', double(p))
    return
  end
    i = i+1;
    p0 = p
end
```

Appendix D: Matlab code for solving Equation (4.8)

```
syms p c k
f(p) = τ3(1−τ3) − (η−β+1)^2 τ2(1−τ2)
while i < c
    p = p0 − f(p0)/f'(p0)        %Newton – Raphson method
                                 % f'(·) is the derivative f(p)
                                 % p0 is the initial approximation of the root of f(p)
  if |p − p0| < ε                % stopping criterion
    fprint('cut off value is %f\n', double(p))
    return
  end
    i = i+1;
    p0 = p
end
```

Appendix E: Matlab code for solving Equation (4.15)

syms q c k m n r η β

$$f(\mathrm{q}) = \frac{\sum_{i=1}^{m} x_{1i} - \mathrm{m}\tau_1}{\tau_1(1-\tau_1)}\frac{d\tau_1}{dq} + \frac{\sum_{j=1}^{n} x_{2j} - n\tau_2}{\tau_2(1-\tau_2)}\frac{d\tau_2}{dq} + \frac{\sum_{z=1}^{r} x_{3z} - r\tau_3}{\tau_3(1-\tau_3)}\frac{d\tau_3}{dq}$$

while $i < c$

$q = q_0 - \dfrac{f(\mathrm{q}_0)}{f^{'}(\mathrm{q}_0)}$ %*Newton – Raphson method*

% $f^{'}(\cdot)$ is the derivative f(·)

% q_0 *is the initial approximation of the root of* $f(q)$

if $|q - q_0| < \varepsilon$ %*stopping criterion*

fpr int('mle of q is % f\ n', double(q))

return

end

i = i+1;

$q_0 = q$

end