COMPUTATIONAL STATISTICAL MODEL FOR GROUP TESTING WITH RETESTING

COX LWAKA TAMBA

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

EGERTON UNIVERSITY

JANUARY 2012

DECLARATION AND RECOMMENDATION

DECLARATION

This thesis is my original work and has not been submitted wholly or in part for any award in any other institution of learning.

Signature …… Date ……

Cox Lwaka Tamba

SM12/2398/09

RECOMMENDATION

We wish to confirm that this thesis has been prepared under our supervision and has our approval to be presented for examination as per the Egerton University regulations.

Signature …… Date ……

Prof. K.L. Nyongesa

Department of Mathematics

Masinde Muliro University of Science and Technology

Signature …… Date ……

Dr. J.W. Mwangi

Department of Mathematics

Egerton University

COPYRIGHT

DEDICATION

To

My dad Simon and mum Filister

ACKNOWLEDGEMENT

I wish to thank God who has granted me the grace, good health, patience, understanding and favour throughout this work. All Glory be unto you Oh God!

I express my heartfelt appreciation to Dr. Nyongesa and Dr. Mwangi, my supervisors, for their constant guidance and advice in coming up with this project. I am also grateful to my lecturers Dr. Orawo, Dr. Ali, and Dr. Kinyanjui for preparing me adequately through the course work.

To my family and in particular dad and mum, for your prayers, inspirations, encouragements and financial assistance, I say thank you. My special appreciation to Ms. Mary Mboga, I thank God for you and I am grateful for the financial support, encouragement you offered me. God bless you. Finally I acknowledge my classmates: Siele, Karisa and Meldrine for their moral support and encouragement throughout the course. God bless you all!

ABSTRACT

Screening of pooled urine specimen was suggested during the Second World War as a method for reducing the cost of detecting syphilis in U.S. soldiers. Recently, pooling has been used in epidemiological studies for screening of human immunodeficiency virus HIV/AIDS antibody to help curb the spread of the virus. Pooling reduces the cost but also – and more importantly – offers a feasible way to lower the misclassifications associated with labeling specimens when imperfect tests are used. Furthermore, misclassifications can be reduced by employing a re-testing design in a pool testing procedure. In this design a large sample from a population of interest is pooled into *n* pools each of size *k* and each pool is subjected to a single test. For pools that test negative further testing are discontinued but those that test positive are given a re-test. Pools that test positive on re-testing, their constituent members are tested individually so as to classify them as either defectives or non-defectives. This study has developed a computational statistical model for classifying a large sample from a population of interest based on the re-testing design described above. This model permits calculation of cost of testing and the number of misclassifications made in this design. Simulated data from a multinomial distribution (specifically a trinomial distribution) has been used to illustrate the computation of cost and the number of misclassifications in the re-testing design. This study has also considered pool testing procedure without re-testing when imperfect tests are used. In this procedure, a sample from the population of interest is pooled into *n* pools of size *k* and each pool subjected to a single test. Pools that test negative further testing are discontinued whereas those that test positive their constituent members are tested individually. Simulation from a binomial distribution has been carried out and statistical moments based on this distribution have been computed to illustrate this testing design. The cost of this testing design and the number of misclassifications made has also been computed. Comparison of the two pool testing designs on the basis of cost and misclassifications has been carried out for the purpose of generalization and improvement. From this study, it has been established that re-testing reduces misclassifications significantly and more so, it is stable at high rates of probability of incidences as compared to Dorfman procedure. However, re-testing comes with a cost i.e. increase in the number of tests. Re-testing considered reduces the sensitivity of the testing scheme but at the same time it improves the specificity; making the model viable in blood donation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.0 Summary

In this chapter we provide a foundation to our study i.e. the background information and thus leading to the formulation of the problem. The entire chapter is arranged as follows: Section 1.1 provides background information to the present study, the statement of the problem is presented in Section 1.2 while the objectives and the justification of the study are presented in Sections 1.3 and 1.4 respectively. Operational definitions and assumptions of this study are presented in Section 1.5 and 1.6 respectively.

1.1 Background Information

The idea of group testing originated with Dorfman (1943) during World War II as an economical method of testing blood specimens of army inductees in order to detect the presence of infection. Group testing has been applied in many areas as outlined by Sobel and Groll (1966). It has also been used to screen the population for the presence of HIV/Aids antibody (Kline, 1989). Dorfman (1943) proposed that, rather than testing each blood specimen individually, portions of each of k specimens can be pooled and the pooled specimen tested first. If the pooled specimen is free of infection, all k inductees are passed with no further tests, otherwise the remaining portions of each of the blood specimens are to be tested individually. If the prevalence of infection is low, the expected number of tests per inductee, and thus the expected cost per inductee, would be reduced. Dorfman (1943) assumed that tests were perfect i.e. a negative reading indicates the group contains no defective item and a positive reading indicates the presence of at least one defective item. Dorfman (1943) did not consider group testing when imperfect tests were used. In real life problem the tests are imperfect (i.e. sensitivity and specificity are less than 100%). This procedure is described diagrammatically in Figure 1.1 as suggested by Dorfman (1943).

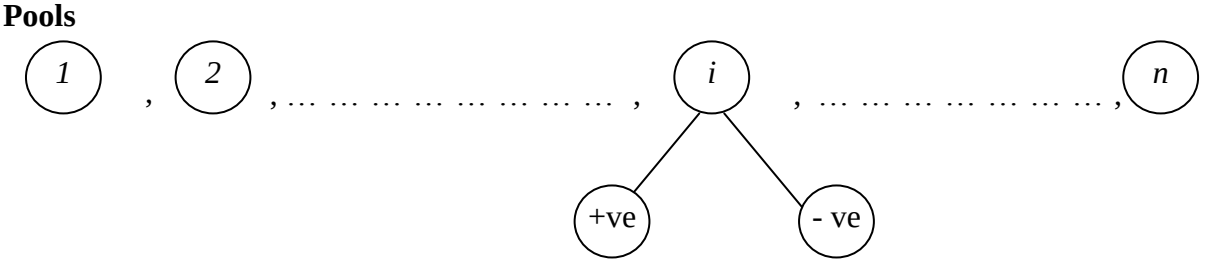

Figure 1.1: Dorfman (1943) pool testing strategy

Based on Dorfman (1943) idea, Monzon *et al.* (1992) proposed another design for screening the population for the presence of HIV/AIDs antibody. In this design the population is pooled into *n* pools and each pool is subjected to a single test. Pools that test negative further testing are discontinued but those that test positive are given a re-test. Pools that test positive on re-testing, their constituent members are tested individually for the presence of HIV/AIDs antibody. This procedure is illustrated diagrammatically in Figure 1.2.

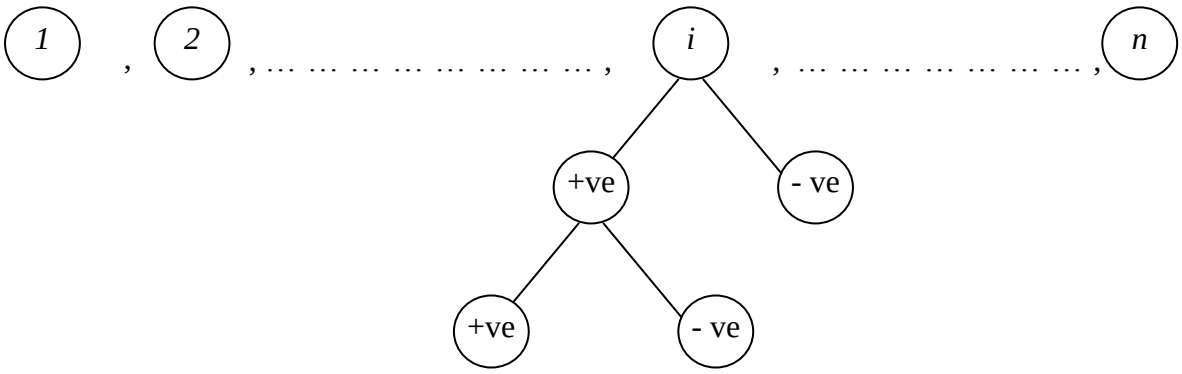

Figure 1.2: Monzon et al., (1992) pool testing strategy

According to the available literature, no statistical model has been developed based on Monzon *et al.* (1992) design of pool testing, in particular the computational aspect unlike the Dorfman (1943) design which has received a lot of attention. This study develops a computational statistical model based on Monzon *et al.* (1992) design of pool testing. The rest of the thesis is organized as follows; literature review of the present study is discussed in Chapter Two and methods used are presented in Chapter Three. Group testing strategies i.e with and without re-testing and their discussions are provided in Chapter Four. Conclusions and recommendations for open problems are presented in Chapter Five.

1.2 Statement of the Problem

Classification of a large sample from a population of interest into defectives and non-defectives can be a tedious and an expensive exercise. This classification can also lead to increased error rates especially when the sample is large and the prevalence rate is high. Pool testing not only reduces cost but also and more importantly offers a feasible way to lower the misclassifications associated with labeling specimens when imperfect tests are used. Furthermore, misclassifications can be reduced by re-testing in the pool testing procedure. The computation of cost of testing and the number of misclassifications at various assumed

values of prevalence rate require a statistical model. A statistical model based on Dorfman procedure of pool testing with the assumption that the test kits are perfect has been developed. In most real cases the test kits are imperfect (i.e. sensitivity and specificity are less than 100%). Based on the Dorfman procedure, a group testing with re-testing design has been proposed. According to the literature available, no computational statistical model has been developed based on this design. Therefore, it is not feasible to determine whether this design reduces the testing cost and/or the number of misclassifications or not. The purpose of this study is to develop a computational statistical model based on the group testing with re-testing design and compare this design with the Dorfman procedure when imperfect test kits are used.

1.3 Objectives

1.3.1 General Objective

To develop a computational statistical model based on Monzon *et al.* (1992) design of pool testing strategy.

1.3.2 Specific Objectives

1 To derive the probability of declaring a pool and a specimen positive/negative before and after re-testing.
2 To carry out simulations from binomial and multinomial distributions based on Dorfman (1943) and Monzon *et al.* (1992) designs of pool testing respectively.
3 To compute the number of misclassifications and the cost of testing.
4 To compare the proposed design with the Dorfman (1943) design.
5 To develop MATLAB codes for computing various statistical measures in the testing designs.

1.4 Justification

Consider a large sample of size N from a population of interest, where N tends to infinity. The idea here is to classify the sample N into two distinct groups, defectives and non-defectives. At the same time, resources are limited to carry out this exercise. Notice that if imperfect tests are applied, one at a time testing can result into fatigue and misclassification as $N \to \infty$ and the resources required will be massive. In this situation, pool testing comes in handy specifically the retesting as this will reduce the misclassification errors.

1.5 Definition of Terms

(i) **Pool/Group**

A set of individuals pooled or grouped together for the purpose of testing.

(ii) **Sensitivity**

This is the probability of correctly classifying a defective group or individual.

(iii) **Specificity**

This is the probability of correctly classifying a non-defective group or individual.

(iv) **Group Test**

This is a test performed on a group of more than one item in which a negative reading indicates the group contains no defective items and a positive reading indicates the presence of at least one defective item.

(v) **Re-testing**

This refers to testing of a group or individual more than once.

1.6 Assumptions

In the entire study the following assumptions have been made;

i) The tests act independent of one another and are not destructive,

ii) Individuals in a group/pool are independent of each other,

iii) Individuals' specimens do not react with each other.

CHAPTER TWO
LITERATURE REVIEW

2.0 Summary

This chapter discusses survey literature of group/pool testing procedure. It further provides the procedure that will be used in the present study such as random number generation which is the backbone of this study. A group/pool test is a test performed on a group of more than one item in which a negative reading indicates the group contains no defective items and a positive reading indicates the presence of at least one defective. The basic idea is to put specimens from individuals for example urine, sera, plasma to form a group and then test the group rather than testing each individual for evidence of a disease. The objectives of group testing are two-fold: classification of the units of a population as either defective or non-defective as presented in Section 2.1 and estimation of the prevalence of a disease in a population presented in Section 2.2. The applications of the testing procedure are presented in Section 2.3. Section 2.4 discusses random number generation.

2.1 Classification of the Units of a Population

A simple procedure was proposed by Dorfman (1943) for classifying members of an population of interest of size N into defective and non-defective items. The idea was to put the population into groups each of size k and perform tests on each group. Dorfman (1943) has given the group size k, depending on the known prevalence rate, which maximizes the expected number of items classified per test. The main benefit of group testing is that it reduces the expense and effort incurred compared to individual testing. Dorfman (1943) showed that if the prevalence rate of a disease is small then group testing can lead to worthwhile saving i.e., reduce the number of tests by about 80%. Dorfman (1943), assumed that if p is the prevalence rate then,

$(1-p)$ = the probability of selecting at random an individual free from infection,

$(1-p)^k$ = the probability of obtaining by random selection a group of size k individuals all of whom are free from infection,

$p' = 1-(1-p)^k$ = the probability of obtaining by random selection a group of k individuals that contains at least one individual infected,

N/k = the number of groups of size k constructed from a population of size N,

$p'\,{}^{N}\!/_{k}$ = the expected number of infected groups of size k in a population of size N with a prevalence rate of p.

The above formulation will be vital in our derivation in subsequent work. Dorfman (1943), showed that the expected number of tests denoted by $E(T)$ obtained by grouping procedure is,

$$E(T) = {}^{N}\!/_{k} + k\left({}^{N}\!/_{k}\right)p' \tag{2.1}$$

that is, the number of groups plus the number of individuals in groups which require individual testing and hence if the prevalence rate is small, then group testing can lead to worthwhile saving since $E(T) \le N$.

Hwang (1975) defined a Dorfman procedure as a partition of units into any number of disjoint groups such that a group test is performed on each of them. Hwang (1975) considered group testing when a population consists of k stochastically independent units where unit i has a probability p_i of being defective which is called a generalized binomial group test (GBGT) problem. When $p_i=p$ for all units then the generalized binomial group test problem reduces to a binary group testing problem considered under Dorfman's procedure. Hwang (1975) was able to give an efficient dynamic programming algorithm for obtaining an optimal Dorfman procedure for the generalized binomial problem with finite k. An optimal group testing procedure in this case implies a procedure which minimizes the expected number of tests. There is an upper bound on the size of a group test which if incorporated into the Dorfman procedure can in effect reduce the amount of computation (Hwang, 1975). Similarly, Hwang (1976) studied group testing model with the presence of a dilution effect i.e. a group containing a few defective items may possibly be misidentified as a group containing no such items, especially when the size of the group is large. When such a misidentification occurs losses are incurred. Hwang (1976) assumed that the dilution effect has a special form, which includes the classical model with no dilution effects as a special case. Hwang (1976) calculated the expected cost of mis-identification for each group of size k under the Dorfman procedure and determined the optimal group size k which minimizes this cost. Hwang (1976) assumed that dilution effect is the only cause of misidentification and by letting $D(k)$ be the probability that a defective group of size k is identified correctly by the test, clearly $D(1)=1$ since there is no dilution and for k being large $D(k)$ should approach p. A class of functions satisfying these two limit conditions is of the form,

$$D(k)=\frac{p}{1-q^{k^d}}\ ,0\le d\le 1 \tag{2.2}$$

where *d* is called a dilution parameter.

When *d*=0 then $D(k)$=1 which is the classical group testing with no dilution and when *d*=1, $D(k)$ can be interpreted as the probability of randomly selecting a unit from a defective group. Further, Hwang (1976) proposed that in most real situation *q* is usually close to one and as such for k not too large $1-q^{k^d}$ is usually approximated by $k^d p$ and $D(k)$ by k^{-d}. The cost of misidentifying a defective group of size *k* was derived as being proportional to the expected number of defective items in the group, i.e. the cost equals $ckp/(1-q^k)$ where *c* is a constant depending on the cost of a test which is taken as a unit cost. By letting *N* be the population size, the expected cost when the group size is *k* was proposed as

$$\begin{aligned} E(k) &= \frac{N}{k}\times(\textit{Expected cost for group of size } k) \\ &= \frac{N}{k}\left\{1+\left(1-q^k\right)\frac{p}{1-q^{k^d}}+\left(1-q^k\right)\left(1-\frac{p}{1-q^{k^d}}\right)\frac{ckp}{1-q^k}\right\} \\ &= N\left\{\frac{1}{k}+\frac{(1-q^k)p}{1-q^{k^d}}+cp-\frac{cp^2}{1-q^{k^d}}\right\} \end{aligned} \tag{2.3}$$

for given (c,p,d), the group size $k(1\le k\le N)$ which minimizes the E(*k*) in Equation (2.3) can be easily found as

$$E(k)\approx N\left(k^{-1}+k^{1-d}p+cp-ck^{-d}p\right). \tag{2.4}$$

By treating $E(K)$ as a continuous function of *k* and taking first derivatives, we have

$$\frac{\partial E(k)}{\partial k}=Nk^{-2}\left(-1+(1-d)k^{2-d}p+cdk^{1-d}p\right). \tag{2.5}$$

Upon defining

$$f(k)=-1+(1-d)k^{2-d}p+cdk^{1-d}p \tag{2.6}$$

then $f(k)$ is a monotone increasing in *k* and hence it has at most one solution at $f(k)=0$ given by

$$-1+(1-d)p+cdp>0. \tag{2.7}$$

In recent years, there has been renewed interest in group testing strategies of biological specimens because of the application in HIV/AIDs epidemiology (Kline *et al.*,

1989). Johnson *et al.* (1992) studied the cost effectiveness of pooling algorithm for the objective of identifying individuals with the trait. In their procedure, each individual group that tested positive was divided into two equal groups, which were tested; groups that tested positive were further subdivided and tested and so on. Litvak *et al.* (1994) extended this work by considering pooling algorithms when there are errors and showed that some of these algorithms can reduce the error rates of the screening procedures (the false positives and false negatives) compared to individual testing. Nyongesa (2004) considered hierarchical pooling studies which involve testing pools and then sequentially subdividing and testing the positive pools. Nyongesa (2005) also considered group testing with re-testing i.e. re-testing of both pools classified as positive and negative. Indeed it was observed that re-testing improves the sensitivity and specificity of the group-testing algorithm. Maheswaran *et al.* (2008) computed statistical measures in their proposed testing strategy. This was the first work of computational statistics in group testing literature. Nyongesa and Syaywa (2011) generalized and extended Maheswaran testing scheme to imperfect testing model and then computed statistical measures. Nyongesa and Syaywa (2010) have developed a computational group-testing strategy with test errors based on Kline *et al.* (1989) design. From their computed results, they showed that when the group size is small, the efficiency of the test kits are high and the prevalence rate is low, then group testing is cost effective. Further, they showed that misclassifications are prominent when the efficiency of the test kits are low and incidence probability high, calling for re-testing. Nyongesa and Syaywa (2010) derived the composite probability of classifying a group as positive denoted by π and gave it as

$$\pi = (1-(1-p)^k)\beta + (1-p)^k(1-\alpha). \tag{2.8}$$

where, β is the sensitivity of the test and α is the specificity of the test. Statistical moments were computed via simulation from binomial distribution based on the above equation. Tamba *et el.* (2011) have considered a computational pool testing strategy when imperfect tests are used based on Dorfman (1943) design. Statistical measures on the number of tests and misclassifications have been computed. From their work, it has been shown that pool testing is only economical when the prevalence rate is low.

2.2 Estimation of the Prevalence Rate using Pool Testing

Sobel and Elashoff (1975) considered group testing with a new goal of estimating the probability *p* of an arbitrary unit being defective. Sobel and Elashoff (1975) observed that a certain class of nested halving procedures is highly efficient and the saving over the one at a

time procedures is even greater for the estimation problem than for the previously treated group testing problems of classifying a given finite set.

Nyongesa (2011) considered estimating the prevalence rate based on a pool-testing scheme with re-testing and showed that when imperfect tests are used in pool testing strategy, there tend to be loss of sensitivity. The loss in sensitivity can be recovered by re-testing the pools classified initially. Apart from improving the sensitivity of the testing scheme, it also improves the efficiency of the estimator. In this procedure, pools classified as negative are retested and the likelihood estimator for the prevalence of the disease derived basing on this scheme. If sensitivity and specificity of the testing scheme are assumed constant, retesting can improve the sensitivity of the testing scheme too. The estimators of prevalence *p* before and after retesting respectively were given as follows;

$$ {}_1\hat{p} = 1 - \left\{ \frac{\beta - \hat{\pi}_1(p)}{\alpha + \beta - 1} \right\}^{\frac{1}{k}} \tag{2.9} $$

where $\hat{\pi}_1(p) = {}^{x_1}\!/_{n}$ and x_1 are the number of pools that test positive before re-testing and α and β are the specificity and sensitivity of the tests respectively. Also a second moment estimator was proposed as

$$ {}_2\hat{p} = 1 - \left\{ \frac{\beta(1-\beta) - \hat{\pi}_2(p)}{\beta(1-\beta) - \alpha(1-\alpha)} \right\}^{\frac{1}{k}} \tag{2.10} $$

where $\hat{\pi}_2(p) = {}^{x_2}\!/_{n}$ and x_2 are the number of pools that test positive after re-testing. Clearly the second estimator of *p* is consistent only if the sensitivity and specificity are not equal.

The maximum likelihood estimator (MLE) of the proportion of infected units in a population using pools is upwardly biased estimator of the population proportion. Hepworth and Watson (2008) investigated this bias of the MLE when testing groups of different sizes using fixed and sequential procedures and observed that the possibility of obtaining all positive groups contributes substantially to the bias and by using analytical method i.e. the simple iterative technique, Hepworth and Watson (2008) were able to correct the bias for fixed procedures satisfactorily but for the sequential procedures with their uneven bias pattern a numerical method which produces an almost unbiased estimator was proposed. The maximum likelihood estimator of the prevalence rate is given as follows. Suppose that, for

$i = 1, 2, ..., m$; n_i groups of size k_i are tested, and x_i of the groups test positive. The likelihood function, denoted as L (p|x) is given by

$$L(p/x) = \prod_{i=1}^{m} \binom{n_i}{x_i} \left\{1-(1-p)^{k_i}\right\}^{x_i} (1-p)^{k_i(n_i-x_i)} \tag{2.11}$$

The $\hat{p}$ that maximizes Log L (p|x) is obtained by solving

$$\sum_{i=1}^{m} k_i n_i = \sum_{i=1}^{m} \frac{k_i x_i}{1-(1-p)^{k_i}}. \tag{2.12}$$

For a constant group of size k, this equation simplifies to $\hat{p} = 1-\left(1-x/n\right)^{1/k}$ and the bias of $\hat{p}$ is given by, $Bias(p) = E(\hat{p}) - p$ which shows that the estimator is biased since, $Bias(p) > 0$ (Hepworth and Watson, 2008).

2.3 Applications of Group Testing

Group testing has been applied in many areas as outlined by Sobel and Groll (1966).The first application of group-testing was to the problem of pooling blood specimens in order to classify each one of a large group of people (e.g. soldiers in an Army unit) as to whether or not they have a particular disease (e.g., syphilis). An interesting feature about the applications of group-testing is the variety of different fields in which they appear. Mundel (1984) has shown that group testing can be applied in many industries for example, in making a "leak test" on a large number of gas-filled (say, with helium) electrical devices, one can test any number of units in a single test and the result of test on x units is that either all x are good or at least 1 of the x is defective.

Group testing has been applied is in testing various electrical devices such as condensers, resistors, etc. The main idea can best be explained with the familiar Christmas tree background. If one assumes that the x bulbs for the tree are all in series so that when he switches on the lights (or plug into the wall socket) he will know by the result that either all the x bulbs are good or at least one of the x 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.

Group 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 group testing in screening HIV antibody to help curb 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 specimens 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 specimen labeled negative in fact has antibodies since each test has a certain sensitivity and specificity.

2.4 Random Number Generators

Our statistical development will be based on random numbers. The theory of generating random numbers is illustrated below (L'Ecuyer, 2004). Random number generators used for simulation are almost always based on deterministic algorithms. A random number generator is a structure (S, μ, f, U, g) where S is a finite set of states (the state space), μ is a probability distribution on S used to select the initial state (or seed) s_0, $f : S \rightarrow S$ is the transition function, U is the output space, and $g : S \rightarrow U$ is the output function. If one assumes that U = (0, 1), the state of the random number evolves according to the recurrence $s_i = f(s_{i-1})$, for $i \geq 1$, and the output at step *i* is $u_i = g(s_i) \in U$. The output values $u_0, u_1, u_2, \ldots$ are the so called random numbers produced by the random number generator. Because the state space S is finite, there are necessarily finite integers $l \geq 0$ and $j > 0$ such that $s_{l+j} = s_l$. Then, for all $i \geq l$, one has $s_{i+j} = s_i$ and $u_{i+j} = u_i$, because both f and g are deterministic. This means that the state and output sequences are eventually periodic. The smallest positive j for which this happens is called the period length of the random number generator, and is denoted by ρ. When $l = 0$, the sequence is said to be purely periodic. Obviously, the period length ρ cannot exceed |S|, the cardinality of the state space. Good random number generators are designed so that their period length ρ is not far from the upper bound. For general recurrences, ρ may depend on the seed s_0, but good random number generators are normally designed so that ρ is the same for all admissible seeds. In practice, it is important that the output be strictly between 0 and 1, because the transformations that generate non-uniform variates sometimes take infinite values when U is 0 or 1.An extremely long period is essential, to make sure that no wrap-around over the cycle can occur. The length of the period must be guaranteed by a mathematical proof. The random number generator must also be efficient (run fast and use little memory), repeatable (able to reproduce exactly the same sequence as many times as one wants), and portable (work the same way in different software/hardware environments). It's also important that the random variables generated are independent and this is verified by the following hypothesis:

H_0: the u_i are realizations of i.i.d. U (0, 1) random variables.

Versus

H_1: the u_i are not realizations of i.i.d. U (0, 1) random variables.

There are several methods of generating random numbers but the most widely used is the multiple recursive generator based on the general linear recurrence,

$$x_i = \left(a_1 x_{i-1} + ... + a_k x_{i-k}\right) \bmod m, \tag{2.13}$$

where m and k are positive integers called the modulus and the order respectively, and the coefficients $a_1, \ldots, a_k$ are in $\mathbb{Z}_m$, interpreted as the set $\{0, \ldots, m-1\}$ on which all operations are performed with reduction modulo m. A multiple recursive generator (MRG) uses the above equation with a large value of m and defines the output as

$$u_i = \frac{x_i}{m} \in (0,1). \tag{2.14}$$

Most methods for generating random variables start with random numbers that are uniformly distributed on the interval (0, 1) (Martinez and Martinez, 2002) .These random variables are denoted by the letter *U*. With the advent of computers, one can easily generate uniform random variable (Hunt et al., 2004) and then through the methods of inverse transform, acceptance-rejection method, random variables from other probability distributions can be generated (Martinez and Martinez, 2002).

CHAPTER THREE
METHODS

3.0 Summary

This chapter presents the methods used in the study. These include statistical packages for random number generation and probability theory. Discussion of statistical packages, specifically MATLAB is presented in section 3.1. The required probability theory for the present study i.e. indicator functions is presented in section 3.2.

3.1 Statistical Packages and Generation of Random Numbers

This is a statistical computational study that encompasses generation of random numbers. Several computer packages can be used to generate random numbers. In this study we have used MATLAB codes to generate these random numbers and statistical measures. Most methods for generating random variables start with random numbers that are uniformly distributed on the interval (0, 1). These random variables are denoted by the letter *U*. With the advent of computers, uniform random variables are easily generated. However the numbers generated by computers are really pseudorandom because they are generated using a deterministic algorithm. The basic MATLAB program has a function **rand** for generating uniform random variables. The function **rand** with no arguments returns a single instance of the random variable *U*. To get an $m \times n$ array of uniform variates, we have used the syntax **rand (m,n)**. The sequence of random numbers that is generated in MATLAB depends on the seed or the state of the generator. The state is reset to the default when it starts up, so that the same sequences of random variables are generated whenever we start MATLAB. From the uniformly distributed random numbers, we have simulated from the binomial and trinomial distribution using the inverse transform procedure and/or using direct methods.

The inverse transform procedure illustrated below has been used to generate binomial random variables. When generating random variables from a binomial distribution with parameters *n* and *p,* this represents the number of successes in *n* independent trials. To obtain a binomial random variable we have generated *n* uniform random numbers and by letting X be the number of those that are less than or equal to *p*. This was easily implemented and the output looped in a row of *N* columns using MATLAB Code as illustrated in the following algorithm:

```
X = zeros (1,N);
U = rand (N,n);
for i = 1:N
ind = find(U(i,:) <= p);
X (i) = length (ind);
end
```

Similarly, we simulated directly from binomial distribution with parameters *n* and *p*, *N* runs directly using MATLAB and this is the easiest procedure as the function is an inbuilt and is given by `X= binornd (n,p,1,N)`.

Simulation of multinomial random variables (e.g. trinomial random variables) with parameters *n* and *p*, *N* times directly has been obtained by defining a vector of probabilities *p* and then generating random variables X (which are vectors) by the inbuilt function `X= mnrnd (n,p,N)`.

3.2 Probability Theory

The probability theory was used to develop the probability distribution based on Monzon *et al.* (1992) design of pool testing. The following indicator functions were cornerstone to our development:

Define

$$T_i = \begin{cases} 1; & \text{if the } i^{th} \text{ group tests positive on the test} \\ 0; & \text{otherwise} \end{cases}$$

$$T'_i = \begin{cases} 1; & \text{if the } i^{th} \text{ group tests positive on the re-test} \\ 0; & \text{otherwise} \end{cases}$$

$$D_i = \begin{cases} 1; & \text{if the } i^{th} \text{ group is positive} \\ 0; & \text{otherwise} \end{cases}$$

$$T_{ij} = \begin{cases} 1; & \text{if the } j^{th} \text{ individual in the } i^{th} \text{ group tests positive on the test} \\ 0; & \text{otherwise} \end{cases}$$

$$D_{ij} = \begin{cases} 1; & \text{if the } j^{th} \text{ individual in the } i^{th} \text{ group is positive} \\ 0; & \text{otherwise} \end{cases}$$

Using the above indicator functions we derived the probability of declaring a group/an individual positive or negative before and after re-testing. These probabilities were used in generating random variables as illustrated above and hence used in developing the statistical

model based on Monzon *et al.* (1992) design. Also, the indicator functions were used in obtaining the cost of the testing scheme.

CHAPTER FOUR
RESULTS AND DISCUSSIONS

4.0 Summary

Consider a large sample of size *N* from a population of interest pooled into *n* pools each of size *k*. The *n* constructed pools are subjected to testing. The testing can be carried with or without re-testing. In this chapter two testing strategies; group testing with and without re-testing have been considered. Section 4.1 presents group testing without retesting when imperfect tests are used whereas group testing with re-testing is presented in Section 4.2. Section 4.3 presents the comparison of the two testing designs.

4.1 Group Testing Without Retesting

In this design, a large sample of size *N* from a population of interest pooled into *n* pools each of size *k.* The *n* pools formed are subjected to testing. If the test result on a given pool is negative, further tests on the pool are discontinued implying that the pool is negative whereas a positive result indicates the presences of at least one defective member then the constituent members of the pool are tested individually. In this design a single testing is sufficient to classify a group or an individual as positive or negative. At the same time we shall assume that the tests applied have sensitivity and specificity values less than 100% thus allowing errors in the design.

4.1.1 Derivation of Probabilities

Suppose we have a set of N individuals to be investigated for some trait and that are to be classified as defectives or non-defectives. Further let ξ be a $\delta-$ field on N and X_1 and X_2 are random variables defined on N, i.e. $X_1 : N \to \mathbb{R}$ and $X_2 : N \to \mathbb{R}$ such that for any Borel set, $\mathrm{B} \in \mathbb{R}$ we have $X_1^{-1}(\mathrm{B}) \in \xi$ and $X_2^{-1}(\mathrm{B}) \in \xi$. Now subdivide N into n partitions in this case representing pools each of size k. We perform a test on each pool as discussed above. For simplicity let X_1 be the number of pools that test positive and X_2 be the number of pools that test negative on testing and hence $X_2 = n - X_1$. Let p be the probability measure on ξ such that an individual is positive (prevalence rate), then $1-p$, the probability that an individual is negative also defined on ξ. Our interest in this study is to consider the testing problem when imperfect tests are used. We derive a new set function $\pi(p,\alpha,\beta,k)$ on ξ, the probability of classifying a pool as positive by the test is

$$\pi(p,\alpha,\beta,k) = Pr(T_i = 1) \tag{4.1}$$

where β and α are the sensitivity and specificity of the test respectively. Upon utilizing the law of total probability we have

$$\begin{aligned}\pi(p,\alpha,\beta,k) &= Pr(T_i=1, D_i=1 \text{ or } D_i=0)\\ &= Pr(T_i=1, D_i=1)+\Pr(T_i=1, D_i=0)\\ &= Pr(T_i=1\mid D_i=1)\Pr(D_i=1)+\Pr(T_i=1\mid D_i=0)\Pr(D_i=0)\\ &= \beta(1-(1-p)^k)+(1-\alpha)(1-p)^k.\end{aligned} \tag{4.2}$$

It is clear that $p\in[0,1]$ and so, $1-\alpha\le\pi(.)\le\beta$ which implies that $\pi(.)$ is a continuous function bounded above by β and below by $1-\alpha$. Also notice that in situation where the pool size is one then Equation (4.2) becomes

$$\pi(p,\alpha,\beta,k)=\beta p+(1-\alpha)(1-p). \tag{4.3}$$

Equation (4.3) can be used in the formulation of individual testing when imperfect tests are employed. Notice that group testing is feasible when the prevalence rate is small. In such situation Equation (4.2) becomes

$$\pi(.)=(\beta+\alpha-1)kp+(1-\alpha)+O(p). \tag{4.4}$$

The probabilities in Equations (4.2), (4.3) and (4.4) will be utilized in the subsequent development.

4.1.2 Expected Number of Tests

In this section we are interested in the construction of an equation for computing the number of test in this design. If we let Z denote the number of tests in this group testing scheme, then clearly:

$$Z=1+n+kX_1 \tag{4.5}$$

where, n is the number of pools, k is the pool size, X_1 is the number of groups that test positive and 1 is the control factor.

Clearly, the expected number of tests is

$$\begin{aligned}E[Z] &= 1+n+kE[X_1]\\ &= 1+n+kn\pi(.).\end{aligned} \tag{4.6}$$

Using Equation (4.4), (4.6) can be written as

$$E[Z]=1+n+k^2np(\beta+\alpha-1)+kn(1-\alpha)+O(p). \tag{4.7}$$

Similarly the variance of the number of tests is obtained as

$$\begin{aligned}Var(Z) &= Var(1+n+kX_1)\\ &= k^2Var(X_1)\\ &= k^2n\pi(.)[1-\pi(.)].\end{aligned} \tag{4.8}$$

Next, we consider the derivation of skewness and kurtosis. Firstly, we consider the derivation of skewness. By definition, skewness denoted by γ_1 is given by

$$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} \tag{4.9}$$

where μ is the respective central moment, therefore

$$\begin{aligned}
\gamma_1 &= \frac{E\left\{Z - E[Z]\right\}^3}{\left\{E\left[Z - E(Z)\right]^2\right\}^{\frac{3}{2}}} \\
&= \frac{k^3 E\left\{X_1 - E[X_1]\right\}^3}{k^3\left\{E\left[X_1 - E(X_1)\right]^2\right\}^{\frac{3}{2}}} \\
&= \frac{E\left\{X_1 - E[X_1]\right\}^3}{\left\{Var(X_1)\right\}^{\frac{3}{2}}}.
\end{aligned} \tag{4.10}$$

But since X_1 is a binomial random variables with parameters n and $\pi(.)$, the third and second moments can easily be obtained and Equation (4.10) becomes

$$\begin{aligned}
\gamma_1 &= \frac{n\pi(.)[1-\pi(.)][1-2\pi(.)]}{\left\{n\pi(.)[1-\pi(.)]\right\}^{\frac{3}{2}}} \\
&= \frac{1-2\pi(.)}{\left\{n\pi(.)[1-\pi(.)]\right\}^{\frac{1}{2}}}.
\end{aligned} \tag{4.11}$$

Secondly, the kurtosis denoted by γ_2 is given by

$$\begin{aligned}
\gamma_2 &= \frac{\mu_4}{\mu_2^{\,2}} \\
\gamma_2 &= \frac{E\left\{Z - E[Z]\right\}^4}{\left\{E\left[Z - E(Z)\right]^2\right\}^2} \\
&= \frac{k^4 E\left\{X_1 - E[X_1]\right\}^4}{k^4\left\{E\left[X_1 - E(X_1)\right]^2\right\}^2} \\
&= \frac{E\left\{X_1 - E[X_1]\right\}^4}{\left\{Var(X_1)\right\}^2}.
\end{aligned} \tag{4.12}$$

Similarly using the fourth and second central moments of a binomial distribution Equation (4.12) becomes

$$\gamma_2 = \frac{n\pi(.)\left[1-\pi(.)\right]\left\{3\left[\pi(.)\right]^2(2-n)+3\pi(.)(n-2)+1\right\}}{\left\{n\pi(.)\left[1-\pi(.)\right]\right\}^2} = \frac{6\left[\pi(.)\right]^2-6\pi(.)+1}{\left\{n\pi(.)\left[1-\pi(.)\right]\right\}}+3. \tag{4.13}$$

These formulations will be useful in the generation of various statistical measures which is the discussion in the next sub-section.

4.1.3 Generation of Moments

If X_1 denotes the number of positive pools then $X_1 \sim binomial(n, \pi(.))$. Hence various statistical measures; mean, standard deviation, kurtosis and skewness are computed by the help of MATLAB Code (1) as outlined in the Appendix. The total number of tests, cost and relative savings are also computed. The above Equations (4.2) and (4.3) helps us to generate moments presented in the Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 where μ is the mean, σ is the standard deviation, γ_1 is the skewness and γ_2 is the kurtosis. Simulations from a sample size of 100 with the group size 10 when the sensitivity and specificity of the test is 99%, the following observations are made based on Table 4.1:

a) The number of defectives increase with increase in the incidence probability *p,*
b) The number of tests increase with increase in incidence probability *p,*
c) Relative savings decrease with increase in incidence probability *p*.

If the sample size is increased to 500 or 1000 from 100 with group size 20 as presented in Table 4.2 and 4.3 respectively similar observations are noted. Furthermore, when the group size is maintained but the sample size is increased more defective are realized but there is no significant difference in relative savings as seen in the case of N =1000, k=20 and $\beta = \alpha = 99\%$ and N =500, k=20 and $\beta = \alpha = 99\%$. Similar observations are made when sensitivity and specificity of the test is varied to 95% as depicted in Table 4.4, 4.5 and 4.6. These observations are true in practice since group testing is only economical when the incidence probability is small (Dorfman, 1943) otherwise individual testing is preferred i.e. relative savings decrease with increase in prevalence rate. Thus the tables provide empirical evidence of group testing scheme.

Table 4.1: Various characteristics for group testing strategy with 1000 runs, N =100, k=10, $\alpha = \beta = 99$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	1.916	1.339	0.609	3.077	5.940	2.318	0.328	2.889	10.787	3.097	0.284	2.905
Number of defective groups	1.041	0.964	0.772	3.377	3.970	1.463	0.140	2.978	6.323	1.563	-0.1200	2.935
Number of group tests	11.000	-	-	-	11.000	-	-	-	11.000	-	-	-
Number of individual tests	10.400	9.640	0.772	3.377	39.700	14.626	0.140	2.978	63.230	15.630	-0.1200	2.935
Total number of tests	21.400	9.640	0.772	3.377	50.700	14.626	0.140	2.978	74.230	15.630	-0.1200	2.935
Total testing cost	21.400	9.640	0.772	3.377	50.700	14.626	0.140	2.978	74.230	15.630	-0.1200	2.935
Percentage savings	78.600	9.640	0.772	3.377	49.300	14.626	0.140	2.978	25.770	15.630	-0.1200	2.935

Table 4.2: Various characteristics for group testing strategy with 1000 runs, N =500, k=20, $\alpha = \beta = 99$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	10.019	3.130	0.354	2.964	29.455	5.361	0.249	3.100	54.003	6.988	0.174	2.841
Number of defective groups	4.584	1.895	0.238	2.903	16.071	2.403	-0.056	2.900	21.791	1.647	-0.442	2.977
Number of group tests	26.000	-	-	-	26.000	-	-	-	26.000	-	-	-
Number of individual tests	91.680	37.91	0.238	2.903	321.42	48.058	-0.056	2.900	435.82	32.930	-0.442	2.977
Total number of tests	117.68	37.91	0.238	2.903	347.42	48.058	-0.056	2.900	461.82	32.930	-0.442	2.977
Total testing cost	23.536	7.582	0.238	2.903	69.484	9.6116	-0.056	2.900	92.364	6.5866	-0.442	2.977
Percentage savings	76.464	7.582	0.238	2.903	30.516	9.6116	-0.056	2.900	7.636	6.5866	-0.442	2.977

Table 4.3: Various characteristics group testing strategy with 1000 runs, *N* =1000, k=20, $\alpha = \beta = 99$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	19.789	4.433	0.141	3.028	59.072	7.481	0.101	2.889	108.10	9.696	0.053	3.233
Number of defective groups	9.462	2.791	0.225	2.817	31.977	3.423	-0.071	2.820	43.592	2.418	-0.306	2.913
Number of group tests	51.000	-	-	-	51.000	-	-	-	51.000	-	-	-
Number of individual tests	189.24	55.82	0.225	2.817	639.54	64.458	-0.071	2.820	871.84	48.354	-0.306	2.913
Total number of tests	240.24	55.82	0.225	2.817	639.54	64.458	-0.071	2.820	922.84	48.354	-0.306	2.913
Total testing cost	24.024	5.582	0.225	3.011	63.954	6.4458	-0.071	2.820	92.284	4.8354	-0.306	2.913
Percentage savings	75.976	5.582	0.225	3.011	30.946	6.4458	-0.071	2.820	7.716	4.8354	-0.306	2.913

Table 4.4: Various characteristics for group testing strategy with 1000 runs, *N* =100, k=10, $\alpha = \beta = 95$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	5.966	2.409	0.455	3.279	9.485	2.903	0.233	2.908	13.866	3.478	0.290	3.014
Number of defective groups	1.402	1.105	0.743	3.531	4.119	1.543	-0.001	3.007	6.378	1.548	-0.310	2.871
Number of group tests	11.000	-	-	-	11.000	-	-	-	11.000	-	-	-
Number of individual tests	14.020	11.05	0.743	3.531	41.190	15.425	-0.001	3.007	63.780	15.484	-0.310	2.871
Total number of tests	25.020	11.05	0.743	3.531	52.190	15.425	-0.001	3.007	74.780	15.214	-0.078	2.832
Total testing cost	25.020	11.05	0.743	3.531	52.190	15.425	-0.001	3.007	74.780	15.214	-0.078	2.832
Percentage savings	74.980	11.05	0.743	3.531	47.810	15.425	-0.001	3.007	25.220	15.214	-0.078	2.832

Table 4.5: Various characteristics for group testing strategy with 1000 runs, *N* =500, k=20, $\alpha = \beta = 95$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	29.693	5.209	0.212	3.023	47.178	6.523	0.109	3.022	69.648	7.909	0.082	2.927
Number of defective groups	5.397	2.122	0.363	3.127	15.640	2.463	-0.104	2.723	20.940	1.871	-0.572	3.367
Number of group tests	26.000	-	-	-	26.000	-	-	-	26.000	-	-	-
Number of individual tests	107.94	42.44	0.363	3.127	312.80	49.360	-0.104	2.723	418.80	37.420	-0.572	3.367
Total number of tests	133.94	42.44	0.363	3.127	338.80	49.360	-0.104	2.723	444.80	37.420	-0.572	3.367
Total testing cost	26.788	8.488	0.363	3.127	67.760	9.872	-0.104	2.723	88.960	7.4840	-0.572	3.367
Percentage savings	73.212	8.488	0.363	3.127	32.240	9.872	-0.104	2.723	11.040	7.4840	-0.572	3.367

Table 4.6: Various characteristics for group testing strategy with 1000 runs, *N* =1000, k=20, $\alpha = \beta = 95$ %

Characteristics	p=0.01				p=0.05				p=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of defectives	58.881	7.655	0.143	2.835	95.216	9.511	0.258	3.092	140.26	10.601	-0.005	3.184
Number of defective groups	10.783	2.900	0.151	2.862	31.348	3.553	-0.091	2.877	42.045	2.404	-0.238	2.919
Number of group tests	51.000	-	-	-	51.000	-	-	-	51.000	-	-	-
Number of individual tests	215.66	58.01	0.151	2.862	626.96	71.054	-0.091	2.877	840.90	48.078	-0.238	2.919
Total number of tests	266.66	58.01	0.151	2.862	677.96	71.054	-0.091	2.877	891.90	4.8078	-0.238	2.919
Total testing cost	26.666	5.801	0.151	2.862	67.796	7.1054	-0.091	2.877	89.190	4.8078	-0.238	2.919
Percentage savings	73.434	5.801	0.151	2.862	32.204	7.1054	-0.091	2.877	10.810	4.8078	-0.238	2.919

4.1.4 Misclassifications in the Testing Scheme

In this sub-section we discuss misclassification that can arise in the experiment, there are two possible cases of misclassifications that may arise in the testing strategy under discussion namely:

a) A defective item is classified as non-defective (this is called false negative)

b) A non-defective item is classified as defective (this is called false positive).

We begin by deriving our probabilities of interest that is, false positive and false negative. First, we derive the probability of correctly classifying a defective individual herein referred to as sensitivity of the testing procedure.

$$
\begin{aligned}
Sensitivity &= \Pr(T_i = 1, T_{ij} = 1 \mid D_{ij} = 1) \\
&= \Pr(T_i = 1, T_{ij} = 1, D_i = 1 or\ D = 0_i \mid D_{ij} = 1) \\
&= \Pr(T_i = 1, T_{ij} = 1, D_i = 1 \mid D_{ij} = 1) + \Pr(T_i = 1, T_{ij} = 1, D_i = 0, \mid D_{ij} = 1) \\
&= \Pr(T_i = 1, T_{ij} = 1, D_i = 1 \mid D_{ij} = 1) + 0 \\
&= \frac{\Pr(T_i = 1, T_{ij} = 1, D_i = 1, D_{ij} = 1)}{\Pr(D_{ij} = 1)} \\
&= \frac{\Pr(T_i = 1 \mid D_i = 1)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 1, D_{ij} = 1)}{\Pr(D_{ij} = 1)} \\
&= \Pr(T_i = 1 \mid D_i = 1)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 1 \mid D_{ij} = 1) \\
&= \beta^2 \left\{(1-(1-p)^k + (1-p)^k\right\} \\
&= \beta^2.
\end{aligned}
\tag{4.14}
$$

The complement of Equation (4.14) gives the probability of false positive

$$f_{se} = 1 - \beta^2. \tag{4.15}$$

The probability of false negative is computed by first calculating the specificity of the testing procedure as

$$Specificity = \Pr(T_i = 1, T_{ij} = 0 \mid D_{ij} = 0) + \Pr(T_i = 0 \mid D_{ij} = 0). \tag{4.16}$$

Let us first consider $\Pr(T_i = 0 \mid D_{ij} = 0)$,

$$\begin{aligned}
\Pr(T_i = 0 \mid D_{ij} = 0) &= \Pr(T_i = 0, D_i = 1 \text{ or } D_i = 0 \mid D_{ij} = 0) \\
&= \Pr(T_i = 0, D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 0, D_i = 0 \mid D_{ij} = 0) \\
&= \frac{\Pr(T_i = 0, D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \frac{\Pr(T_i = 0, D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \frac{\Pr(T_i = 0 \mid D_i = 1)\Pr(D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \frac{\Pr(T_i = 0 \mid D_i = 0)\Pr(D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \Pr(T_i = 0 \mid D_i = 1)\Pr(D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 0 \mid D_i = 0)\Pr(D_i = 0 \mid D_{ij} = 0) \\
&= (1-\beta)(1-(1-p)^{k-1}) + \alpha(1-p)^{k-1}.
\end{aligned} \tag{4.17}$$

Secondly, consider $\Pr(T_i = 1, T_{ij} = 0 \mid D_{ij} = 0)$ that is

$$\begin{aligned}
\Pr(T_i = 1, T_{ij} = 0 \mid D_{ij} = 0) &= \Pr(T_i = 1, T_{ij} = 0, D_i = 1 or\ D_i = 0 \mid D_{ij} = 0) \\
&= \frac{\Pr(T_i = 1, T_{ij} = 0, D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \frac{\Pr(T_i = 1, T_{ij} = 0, D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \frac{\Pr(T_i = 1 \mid D_i = 1)\Pr(T_{ij} = 0 \mid D_{ij} = 0)\Pr(D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \\
&\quad \frac{\Pr(T_i = 1 \mid D_i = 0)\Pr(T_{ij} = 0 \mid D_{ij} = 0)\Pr(D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \Pr(T_i = 1 \mid D_i = 1)\Pr(T_{ij} = 0 \mid D_{ij} = 0)\Pr(D_i = 1, D_{ij} = 0) + \\
&\quad \Pr(T_i = 1 \mid D_i = 0)\Pr(T_{ij} = 0 \mid D_{ij} = 0)\Pr(D_i = 0, D_{ij} = 0) \\
&= (1-\alpha)\alpha(1-p)^{k-1} + \beta\alpha(1-(1-p)^{k-1}.
\end{aligned} \tag{4.18}$$

Combining Equations (4.17) and (4.18) we obtain the specificity as provided by Equation (4.16)

$$\begin{aligned}
Specificity &= \Pr(T_i = 1, T_{ij} = 0 \mid D_{ij} = 0) + \Pr(T_i = 0 \mid D_{ij} = 0) \\
&= (\beta\alpha + (1-\alpha))(1-(1-p)^{k-1}) + \{(1-\alpha)\alpha + \alpha\}(1-p)^{k-1}.
\end{aligned} \tag{4.19}$$

The probability of false negative is then easily obtained from Equation (4.19) as

$$f_{sp} = 1 - Specificity. \tag{4.20}$$

Utilizing Equations (4.15) and (4.20) we compute the misclassifications in this design of pool testing. Computed values of moments of false positives sample sizes 100, 500 and 1000 with group sizes 10 and 20 respectively have been presented in Tables 4.7a and 4.7b when test with specificity and sensitivity of 99% and 95% are employed respectively. It can be seen from these tables that:

a) The number of false positives increases with the increase in incident probability,

b) The false positives realized increase when the sample size is large. In fact when the sample size is doubled false positives almost double,

c) When sensitivity and specificity is increased there is a decrease in false positives; calling for the usage of more accurate test kits.

Simulation of false negatives at sensitivity and specificity of 99% and 95% are presented in Tables 4.8a and 4.8b below. We observe that:

a) The number of false negatives increases with the increase in incidence probability though at a slow rate,

b) Approximately the number of false negatives double when the sample size is doubled,

c) A decrease in sensitivity and specificity leads to an increase in false negatives.

Table 4.7a: Number of false positives in the group testing strategy for different group sizes $\alpha = \beta = 99$ %

Probability,	N =100, k=10				N =500, k=20				N =1000, k=20			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.0389	0.1954	4.9148	23.1331	0.1993	0.4420	2.1725	4.5199	0.3950	0.6222	1.5432	2.2807
0.02	0.0585	0.2394	4.0105	15.4037	0.2942	0.5370	1.7881	3.0620	0.5914	0.7613	1.2612	1.5234
0.03	0.0775	0.2755	3.4851	11.6320	0.3902	0.6184	1.5527	2.3088	0.7881	0.8789	1.0925	1 .1431
0.04	0.0961	0.3069	3.1291	9.3769	0.4900	0.6930	1.3856	1.8386	0.9735	0.9768	0.9830	0.9254
0.05	0.1173	0.3390	2.8323	7.6823	0.5896	0.7601	1.2632	1.5281	1.1737	1.0725	0.8953	0.7676
0.1	0.2168	0.4610	2.0830	4.1553	1.0744	1.0261	0.9357	0.8385	2.1540	1.4530	0.6609	0.4183
0.15	0.3127	0.5536	1.7345	2.8813	1.5578	1.2356	0.7771	0.5783	3.1303	1.7516	0.5482	0.2878

Table 4.7b: Number of false positives in the group testing strategy for different group sizes $\alpha = \beta = 95$ %

Probabilit	N =100, k=10				N =500, k=20				N =1000, k=20			
y, p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.5682	0.7161	1.1241	0.9205	2.8930	1.6158	0.4982	0.1808	5.7572	2.2794	0.3532	0.0908
0.02	0.6524	0.7673	1.0491	0.8017	3.3266	1.7327	0.4646	0.1572	6.6303	2.4462	0.3291	0.0789
0.03	0.7430	0.8188	0.9831	0.7040	3.7336	1.8356	0.4385	0.1401	7.5250	2.6060	0.3089	0.0695
0.04	0.8424	0.8719	0.9232	0.6209	4.1999	1.9469	0.4135	0.1245	8.3590	2.7466	0.2931	0.0626
0.05	0.9277	0.9150	0.8798	0.5638	4.6289	2.0439	0.3939	0.1130	9.2329	2.8866	0.2789	0.0566
0.1	1.3733	1.1133	0.7231	0.3809	6.8563	2.4875	0.3236	0.0763	13.6130	3.5051	0.2297	0.0384
0.15	1.8065	1.2769	0.6305	0.2895	9.0225	2.8536	0.2821	0.0580	18.0418	4.0352	0.1995	0.0290

Table 4.8a: Number of false negatives in the group testing strategy for different group sizes $\alpha = \beta = 99$ %

Probability,	N =100, k=10				N =500, k=20				N =1000, k=20			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.0928	0.3045	3.2780	10.7151	0.8838	0.9392	1.0608	1.1213	1.7678	1.3284	0.7501	0.5606
0.02	0.1679	0.4094	2.4340	5.9036	1.5650	1.2487	0.7957	0.6290	3.1287	1.7660	0.5626	0.3145
0.03	0.2354	0.4846	2.0534	4.1958	2.1158	1.4514	0.6829	0.4622	4.2325	2.0528	0.4829	0.2311
0.04	0.2958	0.5430	1.8301	3.3282	2.5610	1.5960	0.6198	0.3800	5.1221	2.2571	0.4383	0.1900
0.05	0.3504	0.5909	1.6798	2.8005	2.9177	1.7028	0.5800	0.3321	5.8381	2.4087	0.4100	0.1660
0.1	0.5455	0.7363	1.3416	1.7774	3.8231	1.9469	0.5048	0.2504	7.6465	2.7533	0.3570	0.1252
0.15	0.6425	0.7985	1.2333	1.4972	3.9886	1.9877	0.4936	0.2389	7.9714	2.8100	0.3491	0.1195

Table 4.8b: Number of false negatives in the group testing strategy for different group sizes $\alpha = \beta = 95$ %

Probability,	N =100, k=10				N =500, k=20				N =1000, k=20			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.6017	0.7732	1.2767	1.6088	4.8588	2.1929	0.4466	0.1952	9.7125	3.1004	0.3159	0.0977
0.02	0.9309	0.9600	1.0209	1.0208	7.8496	2.7780	0.3478	0.1167	15.7059	3.9295	0.2459	0.0583
0.03	1.2272	1.1004	0.8846	0.7609	10.2818	3.1706	0.3013	0.0865	20.5545	4.4829	0.2131	0.0433
0.04	1.4908	1.2110	0.7988	0.6162	12.2372	3.4510	0.2742	0.0708	24.4703	4.8801	0.1939	0.0354
0.05	1.7305	1.3028	0.7382	0.5228	13.8070	3.6586	0.2566	0.0614	27.6135	5.1740	0.1815	0.0307
0.1	2.5850	1.5834	0.5936	0.3291	17.7867	4.1292	0.2221	0.0447	35.6137	5.8428	0.1570	0.0223
0.15	3.0215	1.7057	0.5428	0.2701	18.5126	4.2037	0.2163	0.0419	37.0504	5.9470	0.1529	0.0209

Indeed group testing is only feasible when the group size is relatively small. This is because when group sizes are large there is a possibility of a dilution effect which makes it impossible to identify defective items in a large group (Hwang, 1976). We have also observed that misclassifications are prominent when the efficiency of the test kits is low and incidence probability high; calling for re-testing and this is the subject of the next section.

4.2 Group Testing With Re-testing

In this section we consider a testing strategy in which the sample is pooled into *n* pools and each pool is subjected to a single test. Pools that test negative further testing is discontinued but those that test positive are given a re-test as presented in Figure 1.2 (cf Monzon *et al.*, 1992). Pools that test positive on re-testing, their constituent members are tested individually for the presence of characteristic of interest. This design herein referred to as group testing with re-testing.

4.2.1 The Probabilities

Let N be a universal set and ξ be a $\delta-$ field on N. Let X_1, X_2, X_{11} and X_{12} be random variables defined on N, i.e. X_1, X_2, X_{11} and X_{12} are functions that take every sample point (outcome) on to the real line. Now subdivide N into n partitions in this case representing pools each of size k. We perform a test on each pool as discussed above. Basically, let X_1 be the number of pools that test positive on the initial test, X_2 be the number of pools that test negative on the initial test, X_{12} be the number of pools that test negative on re-testing and X_{11} be the number of pools that test positive on retesting, hence $X_{11} = n - X_2 - X_{12}$. The discussion is summarized in Figure 1.2. Let p be the probability measure on ξ such that an individual is positive (prevalence rate), then $1-p$, is a set function on ξ, the probability that an individual tests negative. We derive new set functions π_1 and π_2 on ξ, the probabilities of classifying a pool as negative before and after re-testing respectively.

Probability of declaring a pool negative on the initial test

$$\begin{aligned}
\pi_1 &= \Pr(T_i = 0) \\
&= \Pr(T_i = 0, D_i = 0 \text{ or } D_i = 1) \\
&= \Pr(T_i = 0, D_i = 0) + \Pr(T_i = 0, D_i = 1) \\
&= \Pr(T_i = 0 \mid D_i = 0)\Pr(D_i = 0) + \Pr(T_i = 0 \mid D_i = 1)\Pr(D_i = 1) \\
&= (1-p)^k \alpha + (1-(1-p)^k)(1-\beta).
\end{aligned} \tag{4.21}$$

Clearly $p \in [0,1]$ and so $\alpha \le \pi_1 \le 1-\beta$ which implies that π_1 is a continuous function bounded below by α and above by $1-\beta$. Notice also that when the probability *p* is small then Equation (4.21) can be approximated as

$$\pi_1 = \alpha - (\alpha + \beta - 1)kp + O(p). \tag{4.22}$$

Secondly, the probability of declaring a pool as negative on re-testing pools classified positive is

$$\begin{aligned}
\pi_2 &= \Pr(T_i = 1, T'_i = 0) \\
&= \Pr(T_i = 1, T'_i = 0, D_i = 0 \text{ or } D_i = 1) \\
&= \Pr(T_i = 1, T'_i = 0, D_i = 0) + \Pr(T_i = 1, T'_i = 0, D_i = 1) \\
&= \Pr(T_i = 1, T'_i = 0 \mid D_i = 0)\Pr(D_i = 0) + \Pr(D_i = 1)\Pr(T_i = 1, T'_i = 0 \mid D_i = 1) \\
&= \alpha(1-\alpha)(1-p)^k + \beta(1-\beta)(1-(1-p)^k).
\end{aligned} \tag{4.23}$$

Similarly π_2 is bounded below by $\alpha(1-\alpha)$ and above by $\beta(1-\beta)$ i.e. $\alpha(1-\alpha) \le \pi_2 \le \beta(1-\beta)$ and for small values of p Equation (4.23) can be approximated by

$$\pi_2 = \alpha(1-\alpha) + \left[\beta(1-\beta) - \alpha(1-\alpha)\right]kp + O(p) \tag{4.24}$$

and the probability of a pool being classified positive on re-testing of positive pools is

$$\pi_3 = 1 - \pi_2 - \pi_1 \tag{4.25}$$

or this probability of declaring a pool positive on re-testing of initially declared positive pools can be derived directly as $\pi_3 = \Pr(T_i = 1, T'_i = 1)$ and by the law of total probability we have,

$$\begin{aligned}
\pi_3 &= \Pr(T_i = 1, T'_i = 1) \\
&= \Pr(T_i = 1, T'_i = 1, D_i = 0 \text{ or } D_i = 1) \\
&= \Pr(T_i = 1, T'_i = 1, D_i = 0) + \Pr(T_i = 1, T'_i = 1, D_i = 1) \\
&= \Pr(T_i = 1, T'_i = 1 \mid D_i = 0)\Pr(D_i = 0) + \Pr(D_i = 1)\Pr(T_i = 1, T'_i = 1 \mid D_i = 1) \\
&= (1-\alpha)^2(1-p)^k + \beta^2(1-(1-p)^k).
\end{aligned} \tag{4.26}$$

The probabilities $\pi_1, \pi_2, \text{and } \pi_3$ will enable us to compute the joint probability distribution of X_2, X_{11} and X_{12}. Therefore, from the above argument the joint probability density function for X_2, X_{12} and X_{11} is a multinomial probability density

$$f_{X_2 X_{12} X_{11}}(x_2, x_{12}, x_{11}) = \binom{n}{x_2 x_{12} x_{11}} \pi_1^{x_2} \pi_2^{x_{12}} \left(1 - \pi_1 - \pi_2\right)^{n - x_2 - x_{12}}. \tag{4.27}$$

In this retesting strategy, we consider π_2 as the measure which filters out negative pools from the pools that were initially classified as positive in the initial test. The covariance matrix of the random variables X_2, X_{11} and X_{12} is computed as follows

$$C\text{ov}(X_2, X_{12}, X_{11}) = \begin{pmatrix} n\pi_1(1-\pi_1) & -n\pi_1\pi_2 & -n\pi_1\pi_3 \\ -n\pi_1\pi_2 & n\pi_2(1-\pi_2) & -n\pi_2\pi_3 \\ -n\pi_1\pi_3 & -n\pi_2\pi_3 & n\pi_3(1-\pi_3) \end{pmatrix}. \tag{4.28}$$

Equation (4.27) can be used to estimate the prevalence in the re-testing strategy and for further discussion on this subject see Nyongesa (2011).

4.2.2 Expected Number of Tests in Retesting Scheme

In this sub-section we consider the number of tests in the proposed re-testing scheme. If we define Z_r to be the number of tests in this testing strategy then

$$Z_r = 1 + n + X_1 + kX_{11}. \tag{4.29}$$

We obtain the expected number of tests and variance of the number of tests by conditioning X_{11} on X_1 i.e,

$$\begin{aligned} E[Z_r] &= 1 + n + E[X_1] + kE[X_{11}] \\ &= 1 + n + E[X_1] + kE\{E[X_{11} \mid X_1]\} \end{aligned} \tag{4.30}$$

where $X_1 \sim binomial(n, \pi(.))$ and $X_{11} \sim binomial\left(X_1, \frac{\pi_3}{\pi(.)}\right)$.

Using Equation (4.2), the expected number of tests becomes

$$\begin{aligned} E[Z_r] &= 1 + n + E[X_1] + kE\{E[X_{11} \mid X_1]\} \\ &= 1 + n + E[X_1] + k\frac{\pi_3}{\pi(.)}E[X_1] \\ &= 1 + n + n\pi(.) + kn\pi_3. \end{aligned} \tag{4.31}$$

The variance of Z_r is computed as

$$\begin{aligned} Var(Z_r) &= E\{Z_r - E[Z_r]\}^2 \\ &= E\{X_1 - E[X_1] + kX_{11} - kE[X_{11}]\}^2. \\ &= E\{X_1 - E[X_1]\}^2 + 2E\{kX_{11} - kE[X_{11}]\}\{X_1 - E[X_1]\} + E\{kX_{11} - kE[X_{11}]\}^2. \\ &= Var(X_1) + 2kCov(X_1X_{11}) + k^2Var(X_{11}) \end{aligned} \tag{4.32}$$

Using Equations (4.3) and (4.26),

$$Var(Z_r) = Var(X_1) + 2kCov(X_1X_{11}) + k^2Var(X_{11})$$

$$
\begin{aligned}
&= E\left[X_1 - n\pi(.)\right]^2 + 2kE\left[X_1 - n\pi(.)\right]\left[X_{11} - n\pi_3\right] + k^2 E\left[X_{11} - n\pi_3\right]^2 .\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kE\left\{E\left(X_1 - n\pi(.)\right)\left(X_{11} - n\pi_3\right) \mid X_1\right\} + k^2\left\{E\left(Var(X_{11} \mid X_1)\right) + Var\left(E(X_{11} \mid X_1)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kE\left(X_1 - n\pi(.)\right)\left\{E\left(X_{11} - n\pi_3\right) \mid X_1\right\} + k^2\left\{\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1\right) + Var\left(\frac{\pi_3}{\pi(.)}\left(X_1\right)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kE\left(X_1 - n\pi(.)\right)\left\{E\left(X_{11} \mid X_1 - n\pi_3\right)\right\} + k^2\left\{n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right) + \left(\frac{\pi_3}{\pi(.)}\right)^2 Var\left(\left(X_1\right)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kE\left(X_1 - n\pi(.)\right)\left\{E\left(\frac{\pi_3}{\pi(.)}\left(X_1\right) - n\pi_3\right)\right\} + k^2\left\{n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right) + \left(\frac{\pi_3}{\pi(.)}\right)^2 n\pi(.)\left(1-\pi(.)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + \frac{2k\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)\left\{E\left(X_1 - n\pi(.)\right)\right\} + k^2\left\{n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right) + \left(\frac{\pi_3}{\pi(.)}\right)^2 n\pi(.)\left(1-\pi(.)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + \frac{2k\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)^2 + k^2\left\{n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right) + \left(\frac{\pi_3}{\pi(.)}\right)^2 n\pi(.)\left(1-\pi(.)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kn\pi_3\left(1-\pi(.)\right) + k^2\left\{n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right) + \frac{n{\pi_3}^2}{\pi(.)}\left(1-\pi(.)\right)\right\}\\
&= n\pi(.)\left(1-\pi(.)\right) + 2kn\pi_3\left(1-\pi(.)\right) + k^2 n\pi_3(1-\pi_3)
\end{aligned}
$$

(4.33)

Next, we consider the derivation of skewness and kurtosis of Z_r. In general, using the theory of moment generating function of a multinomial distribution the central moments of X_1 and X_{11} can be obtained easily as shown below.

$$
\begin{aligned}
E\left(X_1 - n\pi(.)\right)^2 &= n\pi(.)\left(1-n\pi(.)\right)\\
E\left(X_1 - n\pi(.)\right)^3 &= n\pi(.)\left[1-3\pi(.)+3n\pi(.)+2\left(\pi(.)\right)^2 - 3n\left(\pi(.)\right)^2 + \left(n\pi(.)\right)^2\right]\\
E\left(X_1 - n\pi(.)\right)^4 &= n\pi(.)\left\{1-7\pi(.)+7n\pi(.)+12\left(\pi(.)\right)^2 - 18n\left(\pi(.)\right)^2 + 6\left(n\pi(.)\right)^2 - 6\left(\pi(.)\right)^3\right.\\
&\qquad \left. +11n\left(\pi(.)\right)^3 - 6n^2\left(\pi(.)\right)^3 + \left(n\pi(.)\right)^3\right\}
\end{aligned}
$$

(4.34)

Similarly the central moments for X_{11} are given by

$$
\begin{aligned}
E\left(X_{11} - n\pi_3\right)^2 &= n\pi_3\left(1-n\pi_3\right)\\
E\left(X_{11} - n\pi_3\right)^3 &= n\pi_3\left[1-3\pi_3+3n\pi_3+2{\pi_3}^2 - 3n{\pi_3}^2 + n^2{\pi_3}^2\right]\\
E\left(X_{11} - n\pi_3\right)^4 &= n\pi_3\left\{1-7\pi_3+7n\pi_3+12{\pi_3}^2 - 18n{\pi_3}^2 + 6n^2{\pi_3}^2 - 6{\pi_3}^3 + 11n{\pi_3}^3 - 6n^2{\pi_3}^3 + n^3{\pi_3}^3\right\}
\end{aligned}
$$

(4.35)

Equations (4.34) and (4.35) will aid in the derivation of skewness and kurtosis of Z_r. Firstly, we consider the derivation of skewness. By definition, skewness denoted by γ_{1_r} is given

$$\begin{aligned}\gamma_{1_r} &= \frac{E\left\{Z_r - E\left[Z_r\right]\right\}^3}{\left\{E\left[Z_r - E(Z_r)\right]^2\right\}^{\frac{3}{2}}} \\ &= \frac{E\left\{X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right\}^3}{\left\{E\left[X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right]^2\right\}^{\frac{3}{2}}} \\ &= \frac{E\left\{X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right\}^3}{\left\{n\pi(.)\left(1-\pi(.)\right) + 2kn\pi_3\left(1-\pi(.)\right) + k^2 n\pi_3(1-\pi_3)\right\}^{3/2}}.\end{aligned} \tag{4.36}$$

Evaluating the numerator of Equation (4.36) we have,

$$\begin{aligned}E\left\{X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right\}^3 &= E\left(X_1 - n\pi(.)\right)^3 + 3kE\left(X_1 - n\pi(.)\right)^2\left(X_{11} - n\pi_3\right) \\ &\quad + 3k^2 E\left(X_1 - n\pi(.)\right)\left(X_{11} - n\pi_3\right)^2 + k^3 E\left(X_{11} - n\pi_3\right)^3.\end{aligned} \tag{4.37}$$

Considering the second and third part of Equation (4.37) the following derivations are made;

$$\begin{aligned}E\left(X_1 - n\pi(.)\right)^2\left(X_{11} - n\pi_3\right) &= E\left\{E\left(X_1 - n\pi(.)\right)^2\left(X_{11} - n\pi_3\right) | X_1\right\} \\ &= E\left(X_1 - n\pi(.)\right)^2\left\{E\left(X_{11} - n\pi_3\right) | X_1\right\} \\ &= E\left(X_1 - n\pi(.)\right)^2\left(\frac{\pi_3}{\pi(.)}X_1 - n\pi_3\right) \\ &= \frac{\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)^2\left(X_1 - n\pi(.)\right) \\ &= \frac{\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)^3\end{aligned} \tag{4.38}$$

and,

$$
\begin{aligned}
E\left(X_1-n\pi(.)\right)\left(X_{11}-n\pi_3\right)^2 &= E\left\{E\left(X_1-n\pi(.)\right)\left(X_{11}-n\pi_3\right)^2 \mid X_1\right\} \\
&= E\left(X_1-n\pi(.)\right)\left\{E\left(X_{11}-n\pi_3\right)^2 \mid X_1\right\} \\
&= E\left(X_1-n\pi(.)\right)\left\{E\left({X_{11}}^2-2n\pi_3X_{11}+\left(n\pi_3\right)^2\right) \mid X_1\right\} \\
&= E\left(X_1-n\pi(.)\right)\left\{X_1\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)+{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2-2nX_1\frac{{\pi_3}^2}{\pi(.)}+\left(n\pi_3\right)^2\right\} \\
&= E\left(X_1-n\pi(.)\right)\left\{{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2-2nX_1\frac{{\pi_3}^2}{\pi(.)}+\left(n\pi_3\right)^2\right\}+ \\
&\quad E\left(X_1-n\pi(.)\right)X_1\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right) \\
&= \left(\frac{\pi_3}{\pi(.)}\right)^2 E\left(X_1-n\pi(.)\right)\left\{{X_1}^2-2n\pi(.)X_1+\left(n\pi(.)\right)^2\right\}+ \\
&\quad \frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)X_1 \\
&= \left(\frac{\pi_3}{\pi(.)}\right)^2 E\left(X_1-n\pi(.)\right)^3+\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)^2.
\end{aligned}
\tag{4.39}
$$

Equation (4.37) now becomes

$$
E\left(X_1-n\pi(.)\right)^3\left\{1+3k\frac{\pi_3}{\pi(.)}+3k^2\left(\frac{\pi_3}{\pi(.)}\right)^2\right\}+E\left(X_1-n\pi(.)\right)^2\left\{3k^2\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)\right\}+k^3E\left(X_{11}-n\pi_3\right)^3.
\tag{4.40}
$$

Hence the skewness is given by

$$
\gamma_{1_r}=\frac{E\left(X_1-n\pi(.)\right)^3\left\{1+3k\frac{\pi_3}{\pi(.)}+3k^2\left(\frac{\pi_3}{\pi(.)}\right)^2\right\}+E\left(X_1-n\pi(.)\right)^2\left\{3k^2\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)\right\}+k^3E\left(X_{11}-n\pi_3\right)^3}{\left\{n\pi(.)\left(1-\pi(.)\right)+2kn\pi_3\left(1-\pi(.)\right)+k^2n\pi_3(1-\pi_3)\right\}^{3/2}}.
\tag{4.41}
$$

Now we consider the kurtosis denoted by γ_{2_r} which is given by

$$\gamma_{2_r} = \frac{E\left\{Z_r - E\left[Z_r\right]\right\}^4}{\left\{E\left[Z_r - E(Z_r)\right]^2\right\}^2} = \frac{E\left\{X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right\}^4}{\left\{E\left[X_1 - E(X_1) + k\left(X_{11} - E(X_{11})\right)\right]^2\right\}^2}. \tag{4.42}$$

Expanding the numerator we have,

$$\gamma_{2_r} = \frac{E\left(X_1 - n\pi(.)\right)^4 + 4kE\left(X_1 - n\pi(.)\right)^3\left(X_{11} - n\pi_3\right) + 6k^2E\left(X_1 - n\pi(.)\right)^2\left(X_{11} - n\pi_3\right)^2 + 4k^3E\left(X_1 - n\pi(.)\right)\left(X_{11} - n\pi_3\right)^3 + kE\left(X_{11} - n\pi_3\right)^4}{\left\{E\left[X_1 - n\pi(.) + k\left(X_{11} - n\pi_3\right)\right]^2\right\}^2}$$

$$= \frac{E\left(X_1 - n\pi(.)\right)^4 + 4kE\left(X_1 - n\pi(.)\right)^3\left(X_{11} - n\pi_3\right) + 6k^2E\left(X_1 - n\pi(.)\right)^2\left(X_{11} - n\pi_3\right)^2 + 4k^3E\left(X_1 - n\pi(.)\right)\left(X_{11} - n\pi_3\right)^3 + kE\left(X_{11} - n\pi_3\right)^4}{\left\{n\pi(.)\left(1 - \pi(.)\right) + 2kn\pi_3\left(1 - \pi(.)\right) + k^2n\pi_3(1 - \pi_3)\right\}^2}. \tag{4.43}$$

Now we consider the second, third and fourth parts of Equation (4.43) individually in Equations (4.44), (4.45) and (4.46) respectively:

$$\begin{aligned} E\left(X_1 - n\pi(.)\right)^3\left(X_{11} - n\pi^3\right) &= E\left\{E\left(X_1 - n\pi(.)\right)^3\left(X_{11} - n\pi_3\right) \mid X_1\right\} \\ &= E\left(X_1 - n\pi(.)\right)^3\left\{E\left(X_{11} - n\pi_3\right) \mid X_1\right\} \\ &= E\left(X_1 - n\pi(.)\right)^3\left\{X_1\frac{\pi_3}{\pi(.)} - n\pi_3\right\} \\ &= \frac{\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)^3\left\{X_1 - n\pi(.)\right\} \\ &= \frac{\pi_3}{\pi(.)}E\left(X_1 - n\pi(.)\right)^4 \end{aligned} \tag{4.44}$$

$$E\left(X_1-n\pi(.)\right)^2\left(X_{11}-n\pi_3\right)^2=E\left\{E\left(X_1-n\pi(.)\right)^2\left(X_{11}-n\pi_3\right)^2\mid X_1\right\}$$

$$=E\left(X_1-n\pi(.)\right)^2\left\{E\left({X_{11}}^2-2n\pi_3X_{11}+\left(n\pi_3\right)^2\right)\mid X_1\right\}$$

$$=E\left(X_1-n\pi(.)\right)^2\left\{X_1\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)+{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2-2nX_1\frac{{\pi_3}^2}{\pi(.)}+\left(n\pi_3\right)^2\right\}$$

$$=\left(\frac{\pi_3}{\pi(.)}\right)^2E\left(X_1-n\pi(.)\right)^4+\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)^3+n\pi_3\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)^2.$$

(4.45)

The fourth part of Equation (4.43) becomes

$$E\left(X_1-n\pi(.)\right)\left(X_{11}-n\pi_3\right)^3=E\left\{E\left(X_1-n\pi(.)\right)\left(X_{11}-n\pi_3\right)^3\mid X_1\right\}$$

$$=E\left(X_1-n\pi(.)\right)\left\{E\left({X_{11}}^3-3n\pi_3{X_{11}}^2+3\left(n\pi_3\right)^2X_{11}-\left(n\pi_3\right)^3\right)\mid X_1\right\}$$

$$=E\left(X_1-n\pi(.)\right)\left[X_1\frac{\pi_3}{\pi(.)}-3X_1\left(\frac{\pi_3}{\pi(.)}\right)^2+3{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2+2X_1\left(\frac{\pi_3}{\pi(.)}\right)^3-3{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^3\right.$$

$$\left.+{X_1}^3\left(\frac{\pi_3}{\pi(.)}\right)^3-3n\pi_3\left\{X_1\frac{\pi_3}{\pi(.)}\left(1-\frac{\pi_3}{\pi(.)}\right)+{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2\right\}+3X_1n^2\frac{{\pi_3}^3}{\pi(.)}-n^3{\pi_3}^3\right]$$

$$=E\left(X_1-n\pi(.)\right)\left[X_1\frac{\pi_3}{\pi(.)}-3X_1\left(\frac{\pi_3}{\pi(.)}\right)^2+3{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^2+2X_1\left(\frac{\pi_3}{\pi(.)}\right)^3-3{X_1}^2\left(\frac{\pi_3}{\pi(.)}\right)^3\right.$$

$$\left.+{X_1}^3\left(\frac{\pi_3}{\pi(.)}\right)^3-3X_1n\frac{{\pi_3}^2}{\pi(.)}+3X_1n\frac{{\pi_3}^3}{\pi(.)^2}-3{X_1}^2n\frac{{\pi_3}^3}{\pi(.)^2}+3X_1n^2\frac{{\pi_3}^3}{\pi(.)}-n^3{\pi_3}^3\right]$$

$$=\left(\frac{\pi_3}{\pi(.)}\right)^3E\left(X_1-n\pi(.)\right)^4+3\left(\frac{\pi_3}{\pi(.)}\right)^2\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)^3$$

$$+\frac{\pi_3}{\pi(.)}\left(1-\left(\frac{\pi_3}{\pi(.)}\right)^2\right)E\left(X_1-n\pi(.)\right)^2-3n^2{\pi_3}^2\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)$$

$$=\left(\frac{\pi_3}{\pi(.)}\right)^3E\left(X_1-n\pi(.)\right)^4+3\left(\frac{\pi_3}{\pi(.)}\right)^2\left(1-\frac{\pi_3}{\pi(.)}\right)E\left(X_1-n\pi(.)\right)^3$$

$$+\frac{\pi_3}{\pi(.)}\left(1-\left(\frac{\pi_3}{\pi(.)}\right)^2\right)E\left(X_1-n\pi(.)\right)^2.$$

(4.46)

Hence the kurtosis of Z_r is given by

$$\gamma_{2_r} = \frac{a}{b} \tag{4.47}$$

where,

$$a = E\left(X_1 - n\pi(.)\right)^4 \left\{1 + 4k\frac{\pi_3}{\pi(.)} + 6\left(k\frac{\pi_3}{\pi(.)}\right)^2 + 4\left(k\frac{\pi_3}{\pi(.)}\right)^3\right\} +$$

$$E\left(X_1 - n\pi(.)\right)^3 \left\{6k^2\frac{\pi_3}{\pi(.)}\left(1 - \frac{\pi_3}{\pi(.)}\right) + 12k^3\frac{\pi_3^2}{\pi(.)^2}\left(1 - \frac{\pi_3}{\pi(.)}\right) + 4\left(k\frac{\pi_3}{\pi(.)}\right)^3\right\} +$$

$$E\left(X_1 - n\pi(.)\right)^2 \left\{6k^2 n\pi_3\left(1 - \frac{\pi_3}{\pi(.)}\right) + 4k^3\frac{\pi_3}{\pi(.)}\left(1 - \frac{\pi_3^2}{\pi(.)^2}\right)\right\} + k^4 E\left(X_{11} - n\pi_3\right)^4$$

$$b = \left\{n\pi(.)\left(1 - \pi(.)\right) + 2kn\pi_3\left(1 - \pi(.)\right) + k^2 n\pi_3(1 - \pi_3)\right\}^2$$

These constructions will be useful in the next sub-section.

4.2.3 Generation of Moments in the Group Testing with Retesting Scheme

Suppose in this testing strategy that in the initial test we had X_2 pools that test negative, X_{12} pools that test negative on the re-test, and X_{11} pools test positive on the re-test, then the number of these positive pools on retesting are

$$X_{11} = n - X_2 - X_{12}. \tag{4.48}$$

We are in a position to generate statistical moments based on this design. Using Equations (4.21), (4.23) and (4.25) we generated the moments as presented in Tables 4.9, 4.10 and 4.11 using MATLAB Code (2) as outlined in the Appendix. From the simulation from a sample of size of 100 with group size 10 with $\alpha = \beta = 99\%$ the following observations are made;

a) The number of non-defectives in the initial test decrease with an increase in probability of incidence *p*
b) The number of non-defectives in the re-test is almost invariant with *p*
c) The number of the defectives in the re-test increased with the increase in probability of incidence *p*

Similar observations are made when the sample size is increased to 500 or 1000 with group size of 20 as depicted in Table 4.10 and 4.11. Changing the values of sensitivity and specificity so that $\alpha = \beta = 95\%$, similar observations are made as shown is Tables 4.12, 4.13 and 4.14.

Table 4.9: Various characteristics for group testing with re-testing strategy with *N*=100, k=10, $\alpha = \beta = 99$ %

Probability, p	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	98.0920	1.4050	-.6476	3.0172	0.9500	0.9894	1.1378	3.5644	0.9580	1.0074	0.9558	3.4618
0.02	97.0290	1.7897	-.6307	2.9127	0.9900	1.0182	1.1911	3.2100	1.9810	1.4095	0.5262	3.1625
0.03	96.0800	1.8382	-.4558	3.0429	0.9780	0.9494	1.0860	4.1275	2.9420	1.6311	0.4484	3.0615
0.04	95.0740	2.1485	-.3735	3.3075	0.9970	1.0238	1.0195	3.2108	3.9290	1.9196	0.4097	3.4980
0.05	94.1830	2.3339	-.3541	3.0893	0.9890	1.0280	0.9434	4.9873	4.8280	2.1113	0.4211	3.1513
0.1	89.1630	3.2133	-.3081	3.0934	0.9710	0.9263	0.9727	3.9033	9.8660	3.1019	0.2829	2.9122
0.15	84.2970	3.7271	-.3474	3.0317	0.9890	0.9792	0.9860	3.4233	14.7140	3.6295	0.3911	2.9507

Table 4.10: Various characteristics for group testing with re-testing strategy with *N* =500, k=20, $\alpha = \beta = 99$ %

Probability, p	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	490.140	3.1361	-.3652	3.216	4.9580	2.1692	0.5752	3.1354	4.9020	2.2471	0.6017	3.3866
0.02	485.162	3.8850	-.0997	3.181	5.0930	2.2357	0.4270	2.8934	9.7450	3.1485	0.2345	3.3954
0.03	480.081	4.4258	-.2196	2.612	5.0230	2.2293	0.5108	3.1608	14.8960	3.8797	0.3102	2.9530
0.04	475.440	4.6273	-.0995	3.094	4.9300	2.1919	0.4151	2.8761	19.6300	4.2786	0.2234	3.1743
0.05	470.587	5.0444	-.2584	2.928	4.9320	2.2470	0.4819	3.2585	24.4810	4.5832	0.2202	3.0332
0.1	446.009	6.9683	-.1684	2.920	4.9540	2.2507	0.2901	3.1056	49.0370	6.8149	0.2044	2.7506
0.15	421.582	8.2124	-.0749	2.671	4.9620	2.1031	0.3049	3.1946	73.4560	7.9670	0.0855	2.7040

Table 4.11: Various characteristics for group testing with re-testing strategy with *N* =1000, k=20, $\alpha = \beta = 99$ %

Probability,	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	980.3240	4.4759	-.3669	3.005	9.8430	3.2105	0.4813	2.9751	99.8330	3.1139	0.3662	3.3487
0.02	970.5690	5.3726	-.0938	2.882	9.8750	3.1753	0.2591	3.0031	19.5560	4.2701	0.1761	2.9036
0.03	960.7880	5.8934	-.2732	2.910	9.9020	3.0429	0.4374	3.1523	29.3100	5.1009	0.1921	2.6657
0.04	950.6610	6.7939	-.1119	2.883	10.0150	3.1800	0.2997	3.0618	39.3240	5.9656	0.1615	2.9583
0.05	941.2780	7.4532	-.1036	2.927	10.0860	3.2186	0.3937	2.7049	48.6360	6.8435	0.1086	2.9533
0.1	892.0800	9.5502	0.0167	3.155	9.9200	3.0812	0.1579	3.0927	98.0000	9.1889	-0.0245	3.1685
0.15	842.8170	11.283	0.0979	2.798	9.8890	3.0021	0.3908	3.2034	147.2940	10.9185	-0.0178	2.8292

Table 4.12: Various characteristics for group testing with re-testing strategy with *N* =100, k=10, $\alpha = \beta = 95$ %

Probability,	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	94.2330	2.3077	-.3863	3.221	4.6860	2.0963	0.3360	3.2701	1.0810	1.0135	0.8639	3.7005
0.02	93.2880	2.5032	-.3802	3.454	4.7030	2.0538	0.4316	3.4180	2.0090	1.3538	1.0718	3.7737
0.03	92.3040	2.6365	-.3831	3.122	4.7540	2.1505	0.3693	3.2535	2.9420	1.6480	0.5682	3.1631
0.04	91.4250	2.7476	-.3807	2.833	4.6770	2.0942	0.4644	3.0740	3.8980	1.9067	0.5020	3.0915
0.05	90.5200	3.0546	-.2827	3.137	4.6610	2.2214	0.4218	3.4230	4.8190	2.1677	0.4114	3.3761
0.1	85.9730	3.3674	-.2494	3.134	4.6980	2.0458	0.5056	3.2429	9.3290	2.8658	0.3102	3.1975
0.15	81.6960	3.9026	-.1019	3.093	4.6560	2.2099	0.4902	2.9593	13.6480	3.3720	0.2093	3.1250

Table 4.13: Various characteristics for group testing with re-testing strategy with *N*=500, k=20, $\alpha = \beta = 95$ %

Probability,	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	470.7790	5.2465	-.2005	2.9425	23.4360	4.6969	0.1401	3.0450	5.7850	2.3582	0.4716	3.4082
0.02	465.9130	5.5867	-.1960	2.8804	23.8930	4.8225	0.2235	2.7954	10.1940	3.0358	0.3896	3.2360
0.03	461.0360	5.9309	-.1288	2.6152	24.1130	4.7641	0.1345	2.8208	14.8510	3.7877	0.2435	2.8729
0.04	456.9010	6.1450	-.1321	2.8201	23.8690	4.7349	0.2304	2.8707	19.2300	4.1496	0.2780	2.9681
0.05	452.5610	6.4723	-.1686	3.0232	23.7180	4.7970	0.1268	3.0059	23.7210	4.6795	0.1539	2.8216
0.1	429.7840	7.9616	-.0604	2.8649	23.6830	4.7532	0.1311	3.6184	46.5330	6.7088	0.1274	2.8081
0.15	407.4830	8.6475	0.0104	2.8548	23.6030	4.7146	0.0907	2.8383	68.9140	7.6282	0.0756	3.2026

Table 4.14: Various characteristics for group testing with re-testing strategy with *N* =1000, k=20, $\alpha = \beta = 95$ %

Probability,	Non-defective on the initial test				Non-defective on the retest				Defective on the retest			
p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	940.9280	7.3725	-.217	2.7449	47.4310	6.6779	0.2362	2.7420	11.6410	3.3565	0.2282	3.0047
0.02	931.7290	7.8555	-.120	2.7668	47.9150	6.7480	0.1461	2.8489	20.3560	4.6726	0.1786	2.9884
0.03	922.8680	8.5454	-.123	2.9140	47.7370	6.9302	0.2068	2.7053	29.3950	5.3106	0.1431	2.9788
0.04	913.8860	8.6825	.2473	3.1127	47.7640	6.7042	0.2652	3.0113	38.3500	5.7446	0.2634	3.0295
0.05	904.5560	9.3002	-.135	2.7574	47.9230	6.7043	0.1431	2.8241	47.5210	6.7752	0.1864	2.9703
0.1	860.2880	11.1389	-.121	2.9902	47.1150	6.9920	0.0892	2.9037	92.5970	9.2797	-0.0115	2.8705
0.15	815.2060	12.7498	-.146	2.9033	47.2830	6.8600	0.2130	2.8820	137.511	11.351	0.2134	3.0113

Re-testing the groups that were initially declared positive filters out the negative groups that were misclassified hence reducing the misclassification errors. To compute relative savings in this strategy, we require the total number of tests of which is a function of the defective groups. When we compute our relative savings in our proposed testing strategy when $\alpha = \beta = 99\%$, the following observations are made as shown in Tables 4.15, 4.16, and 4.17:

a) The number of defective groups and number of tests increase with increase in probability of incidence *p*,
b) Relative savings reduce with increase in *p*,
c) Relative savings increase with increase in group size when the probability of incidence is small.

Similar observations are made when sensitivity and specificity values are changed to $\alpha = \beta = 95\%$ as presented in Tables 4.18, 4.19 and 4.20.

Table 4.15: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =100, k=10, $\alpha = \beta = 99$ %

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	8.9470	1.0005	-0.888	3.3359	5.9710	1.4798	-0.205	2.8798	3.5710	1.5288	0.1718	2.7379
Number of non-defective groups on re-test	0.0980	0.3250	0.9265	12.259	0.0900	0.3286	3.0334	12.051	0.1050	0.2986	3.7687	11.722
Number of defective groups on the re- test	0.9550	0.9428	0.9265	3.6347	3.9390	1.4696	0.2439	2.9038	6.3240	1.5490	-0.163	2.6714
Number of group tests	11.026	-	-	-	14.029	-	-	-	16.429	-	-	-
Total number of individual tests	9.550	9.428	0.9265	3.6347	39.390	14.696	0.2439	2.9038	63.240	15.490	-0.163	2.6714
Total number of tests	21.576	9.428	0.9265	3.6347	54.419	14.696	0.2439	2.9038	80.669	15.490	-0.163	2.6714
Total testing cost	21.576	9.428	0.9265	3.6347	54.419	14.696	0.2439	2.9038	80.669	15.490	-0.163	2.6714
Percentage savings	78.424	9.428	0.9265	3.6347	45.581	14.696	0.2439	2.9038	19.331	15.490	-0.163	2.6714

Table 4.16: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =500, k=20, $\alpha = \beta = 99$ %

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	20.275	1.9973	-0.514	3.2247	9.0480	2.2746	0.1715	3.0086	3.1950	1.6644	0.4933	3.0536
Number of non-defective groups on re-test	0.241	0.5119	1.7600	7.3919	0.2340	0.4607	2.0780	7.7788	0.2460	0.4865	1.9591	5.5551
Number of defective groups on the re- test	4.4840	1.9387	0.5252	3.2351	15.718	2.2820	-.1354	2.8377	21.559	1.7219	-.4776	3.0860
Number of group tests	29.725	-	-	-	40.952	-	-	-	46.805	-	-	-
Total number of individual tests	89.680	39.740	0.5252	3.2351	314.36	45.640	-.1354	2.8377	431.18	34.438	-.4776	3.0860
Total number of tests	120.41	39.740	0.5252	3.2351	356.31	45.640	-.1354	2.8377	478.99	34.438	-.4776	3.0860
Total testing cost	24.082	7.9480	0.5252	3.2351	71.262	9.1280	-.1354	2.8377	95.798	6.8876	-.4776	3.0860
Percentage savings	75.918	7.9480	0.5252	3.2351	28.738	9.1280	-.1354	2.8377	4.202	6.8876	-.4776	3.0860

Table 4.17: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =1000, k=20, $\alpha = \beta = 99$ %

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	40.598	2.6509	-.1907	3.1356	17.984	3.3586	0.0764	2.9143	6.3490	2.3364	0.2433	3.0104
Number of non-defective groups on re-test	0.4630	0.6874	1.2219	4.1325	0.4990	0.6753	1.5906	4.1597	0.4590	0.6988	1.4910	6.6991
Number of defective groups on the re- test	8.9390	2.5826	0.2275	3.0620	31.517	3.3787	-.0505	2.9141	43.192	2.4222	-.2026	2.9678
Number of group tests	59.402	-	-	-	82.016	-	-	-	93.651	-	-	-
Total number of individual tests	178.78	51.652	0.2275	3.0620	630.34	67.574	-.0505	2.9141	863.84	48.444	-.2026	2.9678
Total number of tests	239.18	51.652	0.2275	3.0620	713.36	67.574	-.0505	2.9141	958.49	48.444	-.2026	2.9678
Total testing cost	23.918	5.1652	0.2275	3.0620	71.336	6.7574	-.0505	2.9141	95.849	4.8444	-.2026	2.9678
Percentage savings	76.082	5.1652	0.2275	3.0620	28.664	6.7574	-.0505	2.9141	4.151	4.8444	-.2026	2.9678

Table 4.18: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =100, k=10, $\alpha = \beta = 95\%$

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	8.6440	1.1700	-.5423	2.9606	5.8830	1.5252	0.0222	2.8162	3.6320	1.5477	0.1486	2.8169
Number of non-defective groups on re-test	0.4970	0.6982	1.2849	4.5230	0.4490	0.6842	1.1915	4.1657	0.4260	0.6766	1.2065	4.8986
Number of defective groups on the re- test	0.8590	0.9300	0.8361	3.3976	3.6680	1.5228	0.0719	3.0232	5.9420	1.5957	-.1297	2.8060
Number of group tests	11.356	-	-	-	14.117	-	-	-	16.368	-	-	-
Total number of individual tests	8.590	9.300	0.8361	3.3976	36.680	15.228	0.0719	3.0232	59.420	15.957	-.1297	2.8060
Total number of tests	20.946	9.300	0.8361	3.3976	51.797	15.228	0.0719	3.0232	76.788	15.957	-.1297	2.8060
Total testing cost	20.946	9.300	0.8361	3.3976	51.797	15.228	0.0719	3.0232	76.788	15.957	-.1297	2.8060
Percentage savings	79.054	9.300	0.8361	3.3976	48.203	15.228	0.0719	3.0232	23.212	15.957	-.1297	2.8060

Table 4.19: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =500, k=20, $\alpha = \beta = 95$ %

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	19.777	2.0304	-.1293	2.8915	9.3350	2.4371	0.0460	2.8004	4.0630	1.8311	0.3887	3.2041
Number of non-defective groups on re-test	1.1230	1.0704	0.9512	2.8250	1.1620	1.0599	0.8855	3.0411	1.2420	1.0706	0.8703	3.3969
Number of defective groups on the re- test	4.1000	1.8146	0.3253	2.9687	14.503	2.5441	0.0251	2.8278	19.695	1.9869	-.3472	3.1641
Number of group tests	30.223	-	-	-	40.665	-	-	-	45.937	-	-	-
Total number of individual tests	82.000	36.292	0.3253	2.9687	290.06	50.882	0.0251	2.8278	393.90	39.738	-.3472	3.1641
Total number of tests	113.22	36.292	0.3253	2.9687	331.73	50.882	0.0251	2.8278	440.84	39.738	-.3472	3.1641
Total testing cost	22.645	7.2584	0.3253	2.9687	66.345	10.176	0.0251	2.8278	88.167	7.9476	-.3472	3.1641
Percentage savings	77.355	7.2584	0.3253	2.9687	33.655	10.176	0.0251	2.8278	11.833	7.9476	-.3472	3.1641

Table 4.20: Various characteristics along with relative savings for group testing with retesting strategy with 1000 runs for *N* =1000, k=20, $\alpha = \beta = 95$ %

Characteristics	P=0.01				P=0.05				P=0.1			
	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
Number of non- defectives group on the 1st test	39.328	2.8819	-.1709	3.0407	18.604	3.4090	0.1603	3.0714	7.9660	2.6810	0.2306	2.8109
Number of non-defective groups on re-test	2.3860	1.4937	0.6457	3.3194	2.3970	1.5047	0.5207	3.7208	2.3680	1.5230	0.5571	3.7657
Number of defective groups on the re- test	8.2860	2.6445	0.1395	3.0004	28.999	3.5002	-.2052	2.8431	39.666	2.8472	-.0960	2.8837
Number of group tests	60.672	-	-	-	81.396	-	-	-	92.034	-	-	-
Total number of individual tests	165.72	52.890	0.1395	3.0004	579.98	70.004	-.2052	2.8431	793.32	56.944	-.0960	2.8837
Total number of tests	227.39	52.890	0.1395	3.0004	662.38	70.004	-.2052	2.8431	886.35	56.944	-.0960	2.8837
Total testing cost	22.739	5.2890	0.1395	3.0004	66.238	7.0004	-.2052	2.8431	88.635	5.6944	-.0960	2.8837
Percentage savings	77.261	5.2890	0.1395	3.0004	33.762	7.0004	-.2052	2.8431	11.365	5.6944	-.0960	2.8837

4.2.4 Misclassifications in the Retesting Scheme

As observed in the earlier design of Dorfman procedure when imperfect tests are used misclassifications are bound to occur. Similarly misclassifications may occur in this proposed testing design. This is the discussion of the present section. The sensitivity of the retesting procedure;

$$
\begin{aligned}
Sensitivity_r &= \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 1 \mid D_{ij} = 1) \\
&= \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 1, D_i = 1 \; or \; D_i = 0 \mid D_{ij} = 1) \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 1, D_i = 1, D_{ij} = 1)}{\Pr(D_{ij} = 1)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 1, D_i = 0, D_{ij} = 1)}{\Pr(D_{ij} = 1)} \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 1)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 1, D_{ij} = 1)}{\Pr(D_{ij} = 1)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 0)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 0, D_{ij} = 1)}{\Pr(D_{ij} = 1)} \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 1)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 1, D_{ij} = 1)}{\Pr(D_{ij} = 1)} \\
&= \Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 1)\Pr(T_{ij} = 1 \mid D_{ij} = 1)\Pr(D_i = 1 \mid D_{ij} = 1) \\
&= \beta^3 \left\{ (1 - (1-p)^k + (1-p)^k \right\} \\
&= \beta^3.
\end{aligned}
\tag{4.49}
$$

Equation (4.49) provides the sensitivity of the testing procedure; hence the false positive probability is given by

$$
\begin{aligned}
f_{p_r} &= 1 - Sensitivity_r \\
f_{p_r} &= 1 - \beta^3.
\end{aligned}
\tag{4.50}
$$

Note that $\beta^3 < \beta$ since $0 \le \beta \le 1$ thus testing procedure lowers the sensitivity. Similarly $\beta^3 \le \beta^2$, and hence the sensitivity of this re-testing procedure is less than that of the procedure discussed in the preceding section.

Now the specificity of this testing procedure is given by

$$
Specificity_r = \Pr(T_i = 0 \mid D_{ij} = 0) + \Pr(T_i = 1, T_{i}^{'} = 0 \mid D_{ij} = 0) + \Pr(T_i = 1, T_{i}^{'} = 1, T_{ij} = 0 \mid D_{ij} = 0).
\tag{4.51}
$$

We derive the probability $\Pr(T_i = 0 \mid D_{ij} = 0)$ as follows;

$$\begin{aligned}
\Pr(T_i = 0 \mid D_{ij} = 0) &= \Pr(T_i = 0, D_i = 1 \text{ or } D_i = 0 \mid D_{ij} = 0) \\
&= \Pr(T_i = 0, D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 0, D_i = 0 \mid D_{ij} = 0) \\
&= \frac{\Pr(T_i = 0, D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \frac{\Pr(T_i = 0, D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \frac{\Pr(T_i = 0 \mid D_i = 1)\Pr(D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \frac{\Pr(T_i = 0 \mid D_i = 0)\Pr(D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \Pr(T_i = 0 \mid D_i = 1)\Pr(D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 0 \mid D_i = 0)\Pr(D_i = 0 \mid D_{ij} = 0) \\
&= (1-\beta)(1-(1-p)^{k-1}) + \alpha(1-p)^{k-1}.
\end{aligned} \tag{4.52}$$

Next we consider $\Pr(T_i = 1, T_i^{'} = 0 \mid D_{ij} = 0)$,

$$\begin{aligned}
\Pr(T_i = 1, T_i^{'} = 0 \mid D_{ij} = 0) &= \Pr(T_i = 1, T_i^{'} = 0, D_i = 1 \text{ or } D_i = 0 \mid D_{ij} = 0) \\
&= \Pr(T_i = 1, T_i^{'} = 0, D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 1, T_i^{'} = 0, D_i = 0 \mid D_{ij} = 0) \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 0, D_i = 1\ , D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 0, D_i = 0\ , D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 0 \mid D_i = 1)\Pr(D_i = 1\ , D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 0 \mid D_i = 0)\Pr(D_i = 0\ , D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \Pr(T_i = 1, T_i^{'} = 0 \mid D_i = 1)\Pr(D_i = 1 \mid D_{ij} = 0) + \\
&\quad \Pr(T_i = 1, T_i^{'} = 0 \mid D_i = 0)\Pr(D_i = 0 \mid D_{ij} = 0) \\
&= \beta(1-\beta)(1-(1-p)^{k-1}) + (1-\alpha)\alpha(1-p)^{k-1},
\end{aligned} \tag{4.53}$$

and $\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0 \mid D_{ij} = 0)$,

$$\begin{aligned}
\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0 \mid D_{ij} = 0) &= \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0, D_i = 1 or\, D_i = 0 \mid D_{ij} = 0) \\
&= \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0, D_i = 1 \mid D_{ij} = 0) + \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0, D_i = 0 \mid D_{ij} = 0) \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0, D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0, D_i = 0 \mid D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \frac{\Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 1) \Pr(T_{ij} = 0 \mid D_{ij} = 0) \Pr(D_i = 1, D_{ij} = 0)}{\Pr(D_{ij} = 0)} + \\
&\quad \frac{\Pr(T_i = 1, T_i^{'} = 1 / D_i = 0) \Pr(T_{ij} = 0 / D_{ij} = 0) \Pr(D_i = 0, D_{ij} = 0)}{\Pr(D_{ij} = 0)} \\
&= \Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 1) \Pr(T_{ij} = 0 \mid D_{ij} = 0) \Pr(D_i = 1 \mid D_{ij} = 0) + \\
&\quad \Pr(T_i = 1, T_i^{'} = 1 \mid D_i = 0) \Pr(T_{ij} = 0 \mid D_{ij} = 0) \Pr(D_i = 0 \mid D_{ij} = 0) \\
&= \beta^2 \alpha (1 - (1-p)^{k-1}) + (1-\alpha)^2 \alpha (1-p)^{k-1},
\end{aligned} \tag{4.54}$$

thus combining Equations (4.52), (4.53) and (4.54), we obtain the specificity of the procedure as

$$\begin{aligned}
specificity_r &= \Pr(T_i = 0 \mid D_{ij} = 0) + \Pr(T_i = 1, T_i^{'} = 0 \mid D_{ij} = 0) + \Pr(T_i = 1, T_i^{'} = 1, T_{ij} = 0 \mid D_{ij} = 0) \\
&= \left\{3\alpha - 3\alpha^2 + \alpha^3\right\}(1-p)^{k-1} + \left\{1 - \beta^2 + \beta^2\alpha\right\}(1 - (1-p)^{k-1}) \\
&= 1 - \left\{(1-\alpha)^3 (1-p)^{k-1} + \beta^2 (1-\alpha)(1 - (1-p)^{k-1})\right\} \\
&= 1 - (1-\alpha)\left\{(1-\alpha)^2 (1-p)^{k-1} + \beta^2 (1 - (1-p)^{k-1})\right\}.
\end{aligned} \tag{4.55}$$

This design improves the specificity. The false negative probability is given as

$$f_{n_r} = 1 - specificity_r. \tag{4.56}$$

Utilizing the above Equations (4.50) and (4.56) we compute the moments of misclassifications at different group sizes for various probabilities of incidence *p*. The moments of false positives are presented in Tables 4.21a and 4.22b. The tables present the false positives for sensitivity and specificity of 99% and 95% for sample sizes; 100, 500 and 1000 with group sizes 10, 20 and 20 respectively. The following observations are made;

a) The number of false positive increase with increase in the incidence probability *p*,
b) The number of false positive increase when the group size is large, in fact when the group size is doubled false positives increase by at least two fold,
c) Increase in the efficiency of the test kits results in reduction in false positive i.e. the higher the sensitivity and specificity the fewer the false positives realized.

Computations of false negatives are provided in Tables 4.22a and 4.22b. From these tables, we observe that:

a) The number of false negative increase at a slow rate with increase in the incidence probability *p*,
b) The number of false negative approximately doubles when the sample size is doubled and group size is maintained,
c) If the efficiency of the test is increased, less false negatives are realized.

Table 4.21a: Number of false positives in the group testing with retesting strategy for different group sizes for $\alpha = \beta = 99$ %

Probabili	N =100, k=10				N =500,k=20				N =1000,k=20			
ty, p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.0298	0.1699	5.5352	28.6422	0.1427	0.3721	2.5277	5.9728	0.2913	0.5317	1.7691	2.9258
0.02	0.0602	0.2416	3.8927	14.1656	0.2926	0.5328	1.7652	2.9131	0.5893	0.7562	1.2439	1.4465
0.03	0.0870	0.2905	3.2375	9.7984	0.4428	0.6555	1.4350	1.9250	0.8841	0.9262	1.0156	0.9642
0.04	0.1188	0.3396	2.7700	7.1731	0.5861	0.7541	1.2472	1.4542	1.1762	1.0683	0.8804	0.7247
0.05	0.1435	0.3732	2.5206	5.9395	0.7344	0.8442	1.1142	1.1606	1.4654	1.1924	0.7888	0.5817
0.1	0.2930	0.5332	1.7642	2.9095	1.4624	1.1912	0.7896	0.5829	2.9246	1.6846	0.5584	0.2915
0.15	0.4361	0.6505	1.4460	1.9546	2.1784	1.4539	0.6470	0.3913	4.3606	2.0570	0.4573	0.1955

Table 4.21b: Number of false positives in the group testing with retesting strategy for different group sizes for $\alpha = \beta = 95$ %

Probabili	N =100,k=10				N =500,k=20				N =1000,k=20			
ty, p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.1623	0.3730	1.9160	1.9137	0.8085	0.8326	0.8585	0.3841	1.6098	1.1748	0.6084	0.1929
0.02	0.2988	0.5061	1.4121	1.0395	1.4606	1.1191	0.6387	0.2126	2.9355	1.5865	0.4505	0.1058
0.03	0.4168	0.5978	1.1957	0.7453	2.1019	1.3424	0.5324	0.1478	4.2150	1.9010	0.3760	0.0737
0.04	0.5555	0.6901	1.0357	0.5591	2.7227	1.5279	0.4678	0.1141	5.4826	2.1681	0.3297	0.0567
0.05	0.6822	0.7648	0.9346	0.4553	3.3942	1.7059	0.4190	0.0915	6.7955	2.4138	0.2961	0.0457
0.1	1.3217	1.0645	0.6714	0.2350	6.6171	2.3819	0.3001	0.0469	13.2175	3.3664	0.2123	0.0235
0.15	1.9276	1.2856	0.5560	0.1611	9.8323	2.9034	0.2462	0.0316	19.5610	4.0953	0.1745	0.0159

Table 4.22a: Number of false negatives in the group testing with retesting strategy for different group sizes for $\alpha = \beta = 99$ %

Probabil	N =100,k=10				N =500,k=20				N =1000,k=20			
ity, p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.0840	0.2897	3.4459	11.8539	0.8439	0.9179	1.0858	1.1749	1.6878	1.2981	0.7678	0.5874
0.02	0.1599	0.3995	2.4949	6.2043	1.5314	1.2356	0.8043	0.6428	3.0626	1.7473	0.5687	0.3214
0.03	0.2283	0.4772	2.0856	4.3293	2.0905	1.4428	0.6871	0.4681	4.1797	2.0400	0.4860	0.2341
0.04	0.2897	0.5375	1.8494	3.3994	2.5414	1.5900	0.6223	0.3831	5.0812	2.2482	0.4401	0.1916
0.05	0.3444	0.5858	1.6947	2.8510	2.9018	1.6983	0.5816	0.3341	5.8008	2.4011	0.4114	0.1671
0.1	0.5419	0.7339	1.3461	1.7900	3.8251	1.9475	0.5048	0.2504	7.6480	2.7538	0.3570	0.1252
0.15	0.6433	0.7990	1.2327	1.4961	3.9910	1.9884	0.4935	0.2389	7.9783	2.8113	0.3490	0.1195

Table 4.22b: Number of false negatives in the group testing with retesting strategy for different group sizes for $\alpha = \beta = 95$ %

Probabil	N =100, k=10				N =500,k=20				N =1000,k-20			
ity, p	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
0.01	0.3971	0.6289	1.5773	2.4675	3.9282	1.9741	0.4985	0.2445	7.8566	2.7918	0.3525	0.1222
0.02	0.7447	0.8597	1.1455	1.2918	7.0875	2.6429	0.3674	0.1309	14.1722	3.7373	0.2598	0.0655
0.03	1.0594	1.0236	0.9556	0.8925	9.6540	3.0760	0.3122	0.0933	19.3125	4.3507	0.2207	0.0466
0.04	1.3424	1.1505	0.8449	0.6931	11.7331	3.3833	0.2811	0.0749	23.4679	4.7849	0.1988	0.0374
0.05	1.5948	1.2522	0.7718	0.5747	13.4060	3.6095	0.2615	0.0642	26.8125	5.1047	0.1849	0.0321
0.1	2.5157	1.5640	0.6040	0.3428	17.7311	4.1278	0.2233	0.0455	35.4220	5.8343	0.1580	0.0228
0.15	2.9931	1.6998	0.5475	0.2766	18.5761	4.2162	0.2168	0.0423	37.1598	5.9632	0.1533	0.0212

4.3 Comparison of the Two Testing Designs

As observed in the study, variations in the incidence probability impact on the parameters of the two models. Truly, the gist of group-testing is to minimize the number of tests and errors in the experiment. The discussion below is a comparison of the two group-testing models considered in the study.

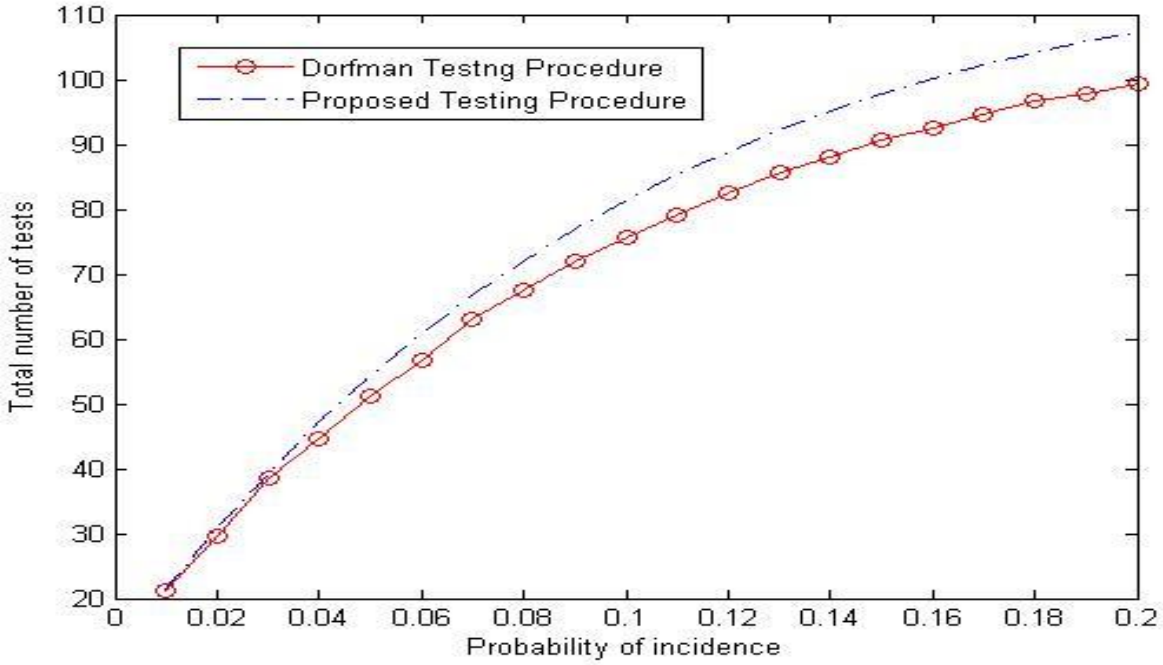

Figure 4.1: Total Number of Tests in the Dorfman Testing Procedure and the Proposed Testing Procedure

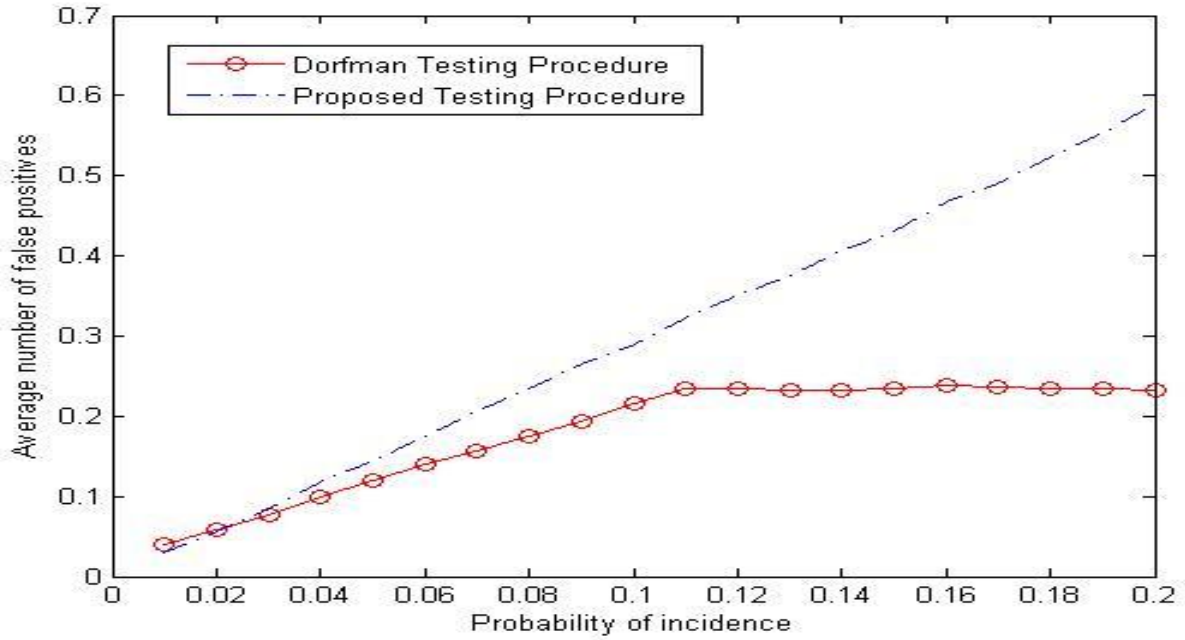

Figure 4.2: Average Number of False Positives in the Dorfman Testing Procedure and the Proposed Testing Procedure

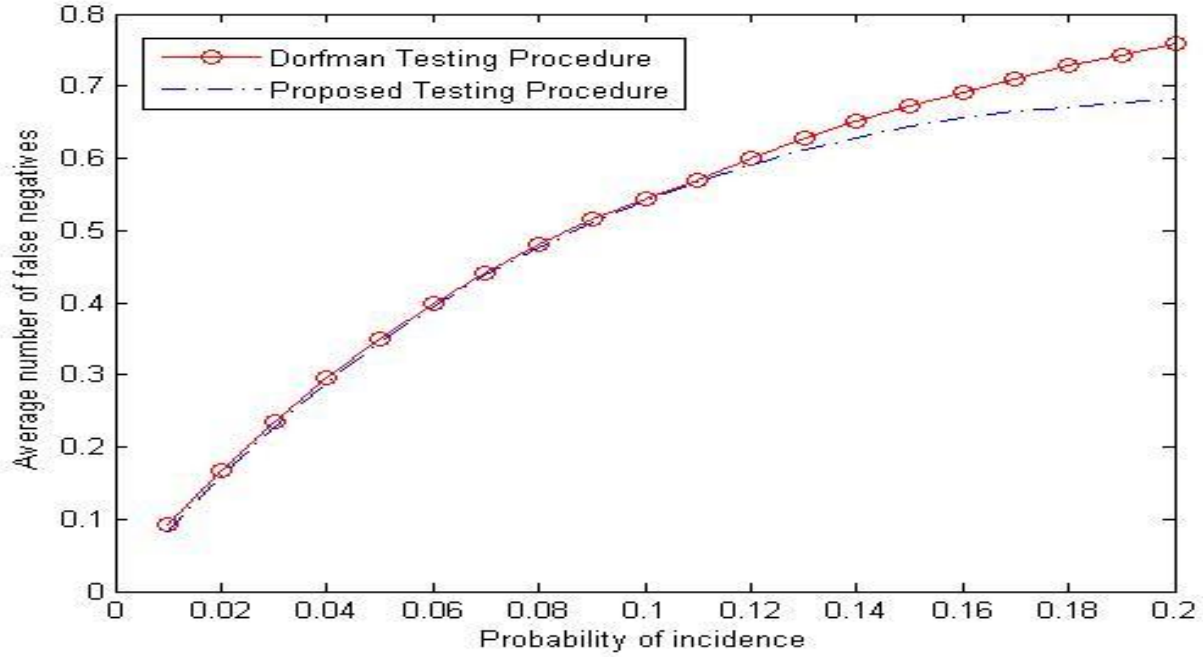

Figure 4.3: Average Number of False Negatives in the Dorfman Testing Procedure and the Proposed Testing Procedure

Figures 4.1, 4.2 and 4.3 show the shape taken when we vary the sample sizes, group sizes and the efficiency of the tests in each testing procedure. We observe that the two testing procedures have the following similarities:

a) As the incidence probability increases, the number of defectives increase which leads to an increase in the cost of testing hence decreasing the relative savings; this is evident in Figure 4.1,
b) The number of defectives realized increase when tests with lower sensitivity and specificity are used; this in turn increases the misclassifications,
c) As the incidence probability increases, the number of misclassifications increases as depicted in Figures 4.2 and 4.3.

On the other hand the two testing strategies have the following differences:

a) The Dorfman testing strategy has fewer numbers of tests than the proposed testing strategy of pool testing with re-testing as shown in Figures 4.1 hence re-testing comes with a cost,
b) There is a significant reduction in the number of false negatives in the proposed testing strategy as compared to the Dorfman (1943) testing strategy as depicted in Figures 4.3,
c) The proposed testing strategy reduces the sensitivity of the testing proposed and hence the number of false positives is high in this strategy as compared to the Dorfman a procedure; this is evident in Figure 4.2,
d) When $p \geq 0.12$, the numbers of false positives in the Dorfman procedure become redundant. This calls for re-testing since the Dorfman procedure fails to depict any misclassifications beyond this value of *p*. The proposed design of group testing with re-testing is viable in such situations.

CHAPTER FIVE
CONCLUSIONS AND RECOMMENDATIONS

5.0 Summary

In this chapter we present a conclusion to our study. This will help in drawing recommendations and open problems for further research. Section 5.1 presents the conclusions to the present study whereas the recommendations and further research are presented in Sections 5.2 and 5.3 respectively.

5.1 Conclusions

From the results obtained in both testing schemes, group testing is an economical testing strategy if the prevalence rate is low. It has been observed that when the value of p increases, the number of tests increase hence a decrease in savings. The higher the prevalence rate not only increases the cost of testing but also increases the misclassifications. It was also observed that when the group size increases, the number of tests and error rates increase. The increased misclassifications can be attributed to a dilution effect especially when the prevalence rate is low. A large group with a few defective individuals may be misclassified as non-defective. From the results of this study, it has been observed that group testing is only viable in situations where the efficiency of the test-kits is high and low prevalence rate. The lower the sensitivity and specificity of the test-kits, the higher the number of tests and misclassifications via Tables (4.7) and (4.8). The group testing scheme without re-testing when imperfect tests are used fails to detect any false positives for values of $p \geq 0.12$ as seen in Figure (4.2). At this point the Dorfman (1943) design becomes redundant hence calling for a better design and thus re-testing comes in handy.

The Monzon et al. (1992) design of group testing with re-testing comes with more cost since it increases the number of tests as compared to the Dorfman (1943) design. However, the design considerably reduces the number of false negatives since it increases the specificity of the testing procedure. Conversely, this procedure lowers the sensitivity of the testing procedure since only the groups that were initially classified as positive are re-tested. This calls for re-testing both the pools initially classified as positive and negative as suggested an open problem for further research.

5.2 Recommendations

Group testing is feasible in low prevalence populations. The proposed re-testing model reduces the misclassifications to some extent; in particular the false negatives, making

the model viable in blood donation. Re-testing only the initially declared positive groups improves the specificity of the testing procedure. Groups that were initially declared as negatives should also be re-tested so as to improve the sensitivity of the testing scheme. Single re-testing cannot eliminate misclassifications completely. Therefore, these calls for repeated testing on the same subject, for minimal errors, although it is not easy to establish the optimal number of re-tests that can be performed.

5.3 Further Research

Single re-testing fails to eliminate misclassifications completely calling for repeated testing although it is not easy to establish the optimal number of re-tests that can be performed. This is an open area for further investigation. Hierarchical computation can also help minimize the number of tests and errors, and according to the literature available no computational hierarchical models exists in group-testing literature. Furthermore, models can be considered when destruction is allowed in modeling, in such situation Monte-Carlo methods can be in handy especially for the computational models.

6.0 REFERENCES

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

Hepworth, G., and Watson R., (2008). Debiased estimation of proportions in group testing. *Royal Statistical Society*, **58**,105–121.

Hunt, R. B., Ronald, L. L., Jonathan, M. R. (2004). *A Guide to MATLAB for Beginners and Experienced Users*. Cambridge University Press, pp.101-119.

Hwang, F.K., (1976). Group testing with a dilution effect. *Biometrika* **63**, 611-613.

Hwang. F.K., (1975). A Generalized Binomial Group Testing Problem. *Journal of the American Statistical Association*, **70**,923- 926.

Johnson N. L., Kotz, S. and Wu, X., (1992). *Inspection errors for attributes in quality control.* London; Chapman and Hall.

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

L'Ecuyer, P., (2004). *Uniform Random Number Generation. Handbook of Computational Statistics*. Springer-Verlag, Berlin, pp. 35–70.

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.

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

Martinez, W.L. and Martinez, A.L., (2002). *Computational Statistics Handbook with MATLAB.* Chapman & Hall/CRC, 96-115.

Monzon, 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.

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

Nyongesa, L. K., (2004). Multistage group testing procedure (Group screening). *Communication in Statistics-Simulation and computation,* **33**, 621-637.

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

Nyongesa, L. K., (2011). Dual Estimation of Prevalence and Disease Incidence in Pool-Testing Strategy. *Communication in Statistics Theory and Method.* **40**, 1-12.

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

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.* (Submitted).

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.

Sobel, M., and Elashoff, R.M., (1975). Group-testing with a new goal, Estimation. *Biometrika,* **62**, 181-193.

Tamba C. L., Nyongesa L. K., Mwangi J. W., (2011). Computational Pool Testing Strategy. *Egerton University Journal of Science and Technology*. (Submitted).

APPENDIX

In this appendix we provide MATLAB codes for simulations purposes for the two models.

MATLAB Code (1): Generation of Moments for the Group Testing without Retesting Procedure

```
%This program generates moments for the Dorfmans' procedure we used se and sp for sensitivity %and specificity respectively, p is the incidence probability, k is the group size, N is sample size, n is number of groups and pro is probability of classifying a group positive.
p=0.2;
se=0.99;
sp=0.99;
k=10;
N=100;
n=N/k;
pro=(1-p)^k*(1-se)+(1-(1-p)^k)*(sp);
pro2=(1-p)*(1-se)+(1-(1-p))*(sp);
%to obtain the number of the defective groups from 1000 simulations
x=binornd(n,pro,1,1000);
m=mean(x);
s=std(x);
k1=kurtosis(x);
s1=skewness(x);
%to obtain the number of defective individuals then we assume that the group size is 1
x1=binornd(N,pro2,1,1000);
m1=mean(x1);
s2=std(x1);
k11=kurtosis(x1);
s12=skewness(x1);
% since the average number of def. might be less than 1 we find the false negatives and false %positives by using the computation formulae for finding mean and variance for a binomial %distribution. Let fp be the number of false positives, fn be number of false negatives fse be %false sensitivity and fsp be false specificity
fse=1-se^2;
fp=m1*fse
```

```
sp1=(1-(1-p)^(k-1))*(se*sp+(1-sp))+(1-p)^(k-1)*((1-sp)*sp+sp);
fsp=1-sp1;
stdfp=sqrt(m1*fse*(1-fse));
skwefp=(1-2*fse)/(sqrt(m1*fse*(1-fse)));
kurfp=(1-6*fse*(1-fse))/(m1*fse*(1-fse));
%to obtain the negatives it's obviously the difference between the population and the positive
%individuals and hence false negatives and false negatives were obtained as;
m2=N-m1;
fn=m2*fsp;
fp=m1*fse;
stdfn=sqrt(m2*fsp*(1-fsp));
skwefn=((1-2*fsp))/(sqrt(m2*fsp*(1-fsp)));
kurfn=(1-6*fsp*(1-fsp))/(m2*fsp*(1-fsp));
```

MATLAB Code (2): Generation of Moments for the Group Testing with Retesting Procedure

```
%This program generates moments for the Proposed procedure we used se and sp for
%sensitivity and specificity respectively, p is the incidence probability, k is the group size, N
%is sample size, n is number of groups and p1 is the probability of declaring a group
%positive on the initial test p2, p3 are probabilities of classifying a group negative before and
%after re-testing respectively, p4 be the probability of classifying a group negative on
%retesting of initially declared positive groups and pro be the vectors of probabilities
p=0.2;
se=0.99;
sp=0.99;
k=10;
N=100;
n=N/k;
p1=(1-p)^k*(1-se)+(1-(1-p)^k)*(sp);
p2=(1-p)^k*se+(1-(1-p^k))*(1-sp);
p3=sp*(1-sp)*(1-p)^k+(1-(1-p)^k)*se*(1-se);
p4=1-p2-p3;
pro=[p2 p3 p4];
m=mean(mnrnd(n,pro,1000));
```

```
s=std((mnrnd(n,pro,1000)));
skw=skewness((mnrnd(n,pro,1000)))
kur=kurtosis((mnrnd(n,pro,1000)))
%to find the number of defective individuals we assume that the group size is 1 and now find
%the various probabilities represented by p
pr2=(1-p)*se+(1-(1-p))*(1-sp);
pr3=sp*(1-sp)*(1-p)+(1-(1-p))*se*(1-se);
pr4=1-pr2-pr3;
pro1=[pr2 pr3 p4];
m1=mean(mnrnd(N,pro1,1000));
s1=std((mnrnd(N,pro1,1000)));
skw1=skewness((mnrnd(N,pro1,1000)));
kur1=kurtosis((mnrnd(N,pro1,1000)));
%note that we can derive directly p4
p4=(1-sp)^2*(1-p)^k+se^2*(1-(1-p)^k);
%to find the sensitivity of the testing procedure let se1, sp1 be the sensitivity and specificity
%of this testing procedure respectively and fse1, fsp1 be the false sensitivity probability and
%false specificity probability of this procedure
se1=(se)^3;
sp1=1-(1-sp)*((se)^2*(1-(1-p)^(k-1))+(1-sp)^2*(1-p)^(k-1));
fse1=1-se1
fsp1=1-sp1
% let mm1 be the number of defectives in the retest so we
calculate its
% moments
mm1=m1(3);
fp1=mm1*fse1;
stdfp1=sqrt(mm1*fse1*(1-fse1));
skwefp1=(1-2*fse1)/(sqrt(mm1*fse1*(1-fse1)));
kurfp1=(1-6*fse1*(1-fse1))/(mm1*fse1*(1-fse1));
% let mm2 be the number of non-defectives in the retest so we
calculate its
% moments
mm2=m1(1)+m1(2);
```

```
fn1=mm2*fsp1
stdfn1=sqrt(mm2*fsp1*(1-fsp1))
skwefn1=((1-2*fsp1))/(sqrt(mm2*fsp1*(1-fsp1)))
kurfn1=(1-6*fsp1*(1-fsp1))/(mm2*fsp1*(1-fsp1))
```