TWO-STAGE ADAPTIVE NEGATIVE BINOMIAL GROUP TESTING MODEL FOR ESTIMATING THE PREVALENCE OF A RARE TRAIT

AKOMBOH JACKLINE ONGECHA

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

EGERTON UNIVERSITY

APRIL, 2024

DECLARATION AND RECOMMENDATIONS

Declaration

This Thesis is my original work and has not been presented in this University or any other for the award of a degree.

Signature: …… Date: …02/04/2024……..

Akomboh Jackline Ongecha

SM123/13516/19

Recommendations

This Thesis has been submitted for examination with our approval as University supervisors.

Signature: Date: …02/04/2024……..

Dr. Ronald Waliaula Wanyonyi (Ph.D)

Egerton University

Signature: …… Date: …….02/04/2024……..

Dr. Cox Lwaka Tamba (Ph.D)

Egerton University

COPYRIGHT

DEDICATION

I dedicate this dissertation work to my family; late dad Stanely Akomboh, mom Violet Akomboh, Husband Charles Juma, sisters and brothers especially Christine Akomboh who have given me unwavering support and encouragement during my studies and development of this research work. Having you in my life has been an immense blessing, and I am sincerely thankful.

ACKNOWLEDGEMENTS

I extend my gratitude to my Lord and Savior Jesus Christ, whose abundant grace, good health, wisdom, and favor have been bestowed upon me throughout this endeavor. I would like to express my sincere appreciation to my dedicated supervisors, Dr. Ronald Wanyonyi and Dr. Cox Lwaka Tamba, for their selfless efforts, constructive criticism, and guidance during the preparation of my dissertation. I also want to acknowledge Dr. Justine Obwoge for his valuable resources, consistent advice, and support in programming and scholarly input.

Furthermore, I would like to recognize and thank the Egerton University Council for their belief in me and for granting me the MSc. Statistics scholarship award, which provided me with the opportunity to advance in the field of academia and statistics. To all, I pray that God abundantly blesses you. I am also deeply grateful to my mother, Violet Akomboh, for her continuous encouragement and prayers throughout my academic journey. Lastly, I would like to express my appreciation to my colleagues, Velma Kiprop, Kibet Hillary, Flavian Awere, Francis Kariuki and Collins Otieno, for their unwavering moral support and encouragement throughout the course. Thank you all, and may God bless each and every one of you.

ABSTRACT

Group testing has been found to be an efficient and economical way of classifying observations under study as defective or unsatisfactory depending on the test performed as well as estimating the prevalence rate of a trait in a population. However, groups of appropriate sizes should be used to realize these benefits. Adaptive schemes have been developed to counter the problems brought about by inappropriate choice of group sizes. The available adaptive schemes have been constructed using a binomial sampling model where the number of groups to be tested is fixed, implying that all groups must be tested before recording the number of successes. But in some situations, such as the case of infectious diseases, estimates need to be reported as soon as detection is made, and for that case, the Negative Binomial (NB), sampling model is preferred. Under NB model, the testing procedure stops immediately when the desired number of successes, which is fixed prior, is attained. This study constructed a two-stage adaptive NB group testing model for estimating the prevalence of a rare trait. The adaptation adjusts group sizes from one stage to the next based on the estimate obtained from the previous stage. The group size used in each stage was the optimal one that minimizes the variance of the estimate of the prevalence rate in the previous stages. The maximum likelihood estimation method was used to find the point estimate of the parameter of the developed model and its properties investigated. The study further constructed the Wald confidence intervals, and its performance was investigated using mean interval length. The developed model was compared to the non-adaptive group testing model existing in the literature using relative mean squared error (RMSE) and asymptotic relative efficiency (ARE) to identify the best model. R-programming language version 4.1.2 was used for Monte Carlo simulation and analysis to verify the model. The use of the two-stage adaptive NB model combined with MLE provided lower and precise estimates. The comparative analysis highlighted the superiority of the adaptive model over the non-adaptive model emphasizing the importance of incorporating adaptivity in group testing procedures. The study highly recommends leveraging these findings to enhance the efficiency and reliability of group testing methods across diverse applications, including disease screening and surveillance of viral illnesses such as Covid-19. By incorporating these findings, the effectiveness of this testing strategies can greatly be improved, leading to more accurate and timely identification of infections, ultimately contributing to better public health outcomes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ARE Asymptotic Relative Variance

FMD Foot and Mouth Disease

GT Group Testing

HIV Human Immunodeficiency Virus

MLE Maximum Likelihood Estimator

MSE Mean Squared Error

NB Negative Binomial

P Prevalence rate

PDF Probability Distribution Function

RMSE Relative Mean Squared Error

WNV West Nile Virus

LIST OF SYMBOLS

p	Proportion of positive unit in a population
$\hat{p}_N$	Estimated proportion of positive unit when non-adaptive schemes are used
$\hat{p}_1$	Estimated proportion of positive unit in the first stage
$\hat{p}_A$	Estimated proportion of positive unit in the second stage
E($\hat{p}$)	Expected value of $\hat{p}$
k	Group size used in the non-adaptive group testing schemes
k_1	Group size used in the first stage
k_2	Group size used in the second stage
$\pi(p)$	Probability of a positive group in a population
$\pi_1(p)$	Probability of a positive group in the first stage
$\pi_{2/1}(p)$	Probability of a positive group in the second stage conditioned on the first stage
N	Number of groups to be tested using non-adaptive binomial model
λN	Number of groups to be tested in the first stage of adaptive binomial model
$(1-\lambda)N$	Number of groups to be tested in the second stage of adaptive binomial model
Y	Number of positive groups in the non-adaptive binomial model
Y_1	Number of positive groups in the first stage of adaptive binomial model
Y_2	Number of positive groups in the second stage of adaptive binomial model
T	Number of groups to be tested before the desired positive groups are observed using non-adaptive scheme
T_1	Number of groups to be tested before the desired positive groups are observed in the first stage of adaptive group testing scheme
T_2	Number of groups to be tested before the desired positive groups are observed in the second stage of adaptive group testing scheme
X	Fixed number of positive groups desired using the non-adaptive scheme
X_1	Fixed number of positive groups desired in the first stage of adaptive group testing scheme
X_2	Fixed number of positive groups desired in the second stage of adaptive group testing scheme1

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Group testing also known as pool testing or batch testing refers to the process of pooling the individuals under study into various pools and undertaking the tests simultaneously. The tests may entail checking for infected or defective individuals for the case of an epidemiology study. This method was first proposed by Dorfman (1943) to significantly enhance cost saving techniques in the detection of soldiers with syphilis. In this regard, blood from several solders was mixed to form a sample which when tested at once saved the cost and time of testing.

Since the Dorfman's seminal work, Group testing has been used in various epidemiological studies, quality control procedures as well as in genetics mainly with two objectives (Bilder *et al*., 2010; Hughes-Oliver, 2006). The first objective being the identification or classification of individual specimen based on dichotomous results. In this context, dichotomous results implying that there might be the presence or absence of a trait of interest. The second objective of group testing entails estimation of prevalence of a trait of interest in a population (Kennedy, 2011; Matiri, 2017; Pritchad & Tebbs, 2010; Sobel & Elashoff, 1975; Thompson, 1962). For either objective, group testing has been found to be more effective as compared to individual testing (Berger *et al.,* 2000).

Group testing is categorized into two forms including non-adaptive group testing scheme and adaptive group testing scheme. Non-adaptive group testing schemes of estimation entails testing groups each of size k to obtain dichotomous results that is whether there is the presence or absence of a trait of interest. The results are then used to obtain the non-adaptive model. The interpretation from the model is that if a group test positive then at least one individual in the group has the trait and if the group tests negative then all the individuals in the group are declared free of the trait. On the other hand, adaptive group testing scheme entails testing groups in stages and adjusting or updating the sizes of the groups used from one stage to the next (Hughes-Oliver & Swallow, 1994; Okoth *et al.,* 2017). The results obtained are then used to construct an adaptive model.

Most of the previous research in Group Testing have presumed a binomial model (Hepworth & Watson, 2008; Kennedy, 2004; Okoth *et al.,* 2017; Wanyonyi, 2015). The binomial model is characterized by subjecting a fixed number of groups to testing. Adaptive estimators have

also been developed under the binomial model to improve the efficiency of the estimator by reducing the MSE (Hughes-Oliver & Swallow, 1994; Okoth *et al*., 2017). Other probabilistic models have also been considered in estimation problem of group testing which include the beta-binomial (Turechek & Madden, 2003), geometric which is a special type of negative binomial (George & Elston, 1993) and Hypergeometric model (Bhattacharyya *et al.,* 1979).

However, when there is urgent need of estimating the prevalence rate of a trait without delay for quick response, Negative Binomial sampling model has been considered the best model as compared to the Binomial model (Montesinos-Lopez *et al.*, 2013). This model has been applied in emergency situations for instance natural disasters and disease outbreaks. In such situations, there is need to immediately measure the amount of risk in the bid to cab the spread of infection. Katholi and Unnasch (2006) pioneered the use of Negative Binomial model in group testing under the non-adaptive scheme. Pritchard and Tebbs (2010) estimated the prevalence of a disease using inverse pool testing under the non-adaptive scheme. They found out that Negative Binomial model was a preferable model in group testing when the goal is to estimate small proportions for quick response. In this model, the number of positive groups is fixed before the testing procedure begins. The trait under consideration is considered prevalent if the desired number of positive groups is observed.

Although group testing procedure has been associated with various benefits (in this case cost related benefits in terms of monetary, mean squared error, and timing). It is worth noting that these benefits will only be realized if there is an appropriate group size (Pritchard & Tebbs, 2010; Xiong, 2015). Group testing properties especially when using negative binomial model, tend to be very sensitive to the choice of group sizes as compared to binomial case thus creating deficiencies under this procedure. To counter these deficiencies related to the inappropriate choice of group sizes, various recommendations have been made which includes;

i. Using several group sizes and combining the data to produce a single estimator.
ii. Employing double sampling where the researcher uses information from the first stage to choose the group size to be used throughout the second stage, and using only the second stage in determining the final estimate of p.
iii. Performing a Bayesian analysis.
iv. Adapting or adjusting the group sizes from time to time throughout the testing phase, then using all the accumulated information to obtain a finale estimate of p.

Hughes-Oliver and Swallow (1994) conducted a two-stage adaptive group testing procedure by adjusting the group sizes from time to time although with the use of Binomial model. They found out that adaptive estimators are efficient by restricting the number of the group size to at least 10 units. A multi stage adaptive been established in literature using the binomial model (Okoth *et al.,* 2017). As a preferred model, Negative binomial model has been used in group testing but under non-adaptive scheme. Until now, there is no study that has incorporated adaptive group testing scheme in the estimation of a rare trait using negative binomial model. The adaptation in this case being adjusting the group sizes from time to time throughout the testing phase, then using all the accumulated information to obtain a finale estimate of p. In this regard, this study investigated a two-stage adaptive negative binomial group-testing procedure for estimating the prevalence rate of a rare trait.

1.2 Statement of the Problem

In testing the proportion of individuals that are defective of some trait, group testing can be very effective as compared to testing the individuals one at a time. However, realizing the benefits of group testing critically requires choosing appropriate group sizes since the properties of group testing are sensitive to group size. Incorporating adaptive schemes has been recommended as a way to counter problems brought about by inappropriate choice of group sizes. Adaptation in this case being, the adjustment of group sizes from one stage to the other during the testing phases and using the result to construct an adaptive model. Group testing has been intensively explored using the binomial sampling model, where the number of groups to be tested is fixed before the testing procedure. However, this model has limitations, as the specified number of groups limits the flexibility of the testing procedure and may only be suitable for some situations especially when the prevalence of the trait is low. Adaptive group testing schemes using Binomial sampling model have been constructed in the existing literature and performed better. However, the use of NB sampling model together with group testing has emerged as an excellent option, especially when dealing with situations requiring quick response surveillance of viral diseases outbreaks like Covid-19, Foot and Mouth disease (FMD) and Ebola. Estimation procedures under the NB model in group testing have been developed using the non-adaptive schemes but have suffered problems of inappropriate choice of group sizes. In the existing literature, estimation procedures using NB model under an adaptive scheme have not been constructed and its properties investigated thus, forming the basis of this study.

1.3 Objectives

1.3.1 General Objective

To construct a two-stage adaptive Negative Binomial group testing procedure for estimating the prevalence rate of a rare trait.

1.3.2 Specific Objectives

i. To obtain an estimator for the prevalence of a rare trait by employing the two-stage adaptive Negative Binomial model in group testing using MLE.
ii. To analyze the properties of the derived estimator which are the bias, variance and Mean Squared Error.
iii. To compare the two-stage adaptive negative binomial group testing model to the existing non-adaptive group testing model to identify the most efficient through simulation.

1.4 Justification

Pool testing has proven to have significant benefits over individual testing, such as reduction of cost and time taken to conduct a test. In most cases such as during the onset of Covid-19, the reagents of testing for the disease were expensive and scarce. In such a scenario, Group testing comes in handy economically both cost and time. However, researchers fail to realize the benefits of group testing due to inappropriate group sizes. Unlike the non-adaptive group testing schemes, adaptive group testing comes in handy to counter the problem brought about by the inappropriate choice of group size.

Negative Binomial sampling model with pool testing has been found to be a more appealing strategy that is economical in terms of cost, time and efficiency. In this model, sampling and testing occur sequentially in two stages and the process stops immediately the fixed number of positive groups is observed in each stage. This makes the model more applicable as compared to Binomial model when there is a need to report estimates early in the screening procedures to prompt a quick mitigation response.

Incorporating adaptive schemes in the use of Negative Binomial sampling together with group-testing procedures will, therefore provide a more efficient and cost-effective model for testing and estimating prevalence rate of a trait that can be used for early detection of infectious diseases such as Covid-19 and foot and mouth diseases among other outbreaks. In such disease surveillance situation, early and accurate assessment of the prevalence rate is important in prompting immediate mitigation measures against the outbreak. The results obtained would be

useful in most sectors such as agriculture and health sectors for meaningful planning and provision of immediate and efficient mitigation response in case of an outbreak.

1.5 Assumptions

i. The probability of having the trait of interest is the same for all members of the population.
ii. The test kits are assumed to be perfect.
iii. The population is assumed to be finitely large to allow adaptive sampling using NB.
iv. There is no shielding effect within the groups.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this literature review, the systematically examine the principles, methodologies, and applications of group testing, also known as pool testing, in various domains. Originating from Dorfman's pioneering work during World War II, group testing has evolved into a powerful methodology with broad applications ranging from public health to quality control. A fundamental challenge in group testing lies in selecting optimal group sizes, striking a balance between testing cost and informativeness of results. The study explores recommendations for group size selection and advancements in adaptive schemes, which dynamically adjust group sizes based on previous test outcomes. In addition, it surveyed the diverse applications of group testing, including disease screening, agricultural biotechnology, and public health initiatives such as COVID-19 screening. Through this comprehensive review, the study aimed to provide insights into the evolution, methodologies, and practical implications of group testing across various domains, paving the way for future research and applications in this field.

2.2 Overview Group Testing

Group testing which is also referred to as pool testing has a long history of successful application pioneered by Dorfman (1943) in the context of identifying soldiers infected with syphilis disease. The idea of group testing revolves around the assumption that the sample is collected from a target population in which some individuals have the characteristic of interest. In this regard, the sample is divided into groups and test conducted in each group instead of conducting individual test. The idea is that if a group tests positive, at least one of the individuals in the group has a characteristic of interest or rather it is defective. On the other hand, if a group tests negative, then it is assumed that all individuals in the group are free from the characteristic of interest. Group testing can lead to a significant decrease in the number of tests conducted especially when the prevalence is small. This reduction in in the number of tests performed consequently leads to a decrease in the cost and time incurred when performing the tests. This classification or identification of individuals as either positive or negative of a trait serves as the first objective of Group testing championed by Dorfman (1943).

To enhance cost-effectiveness and reduce the number of tests required, an adaptation of the Dorfman testing scheme has been investigated and extended to incorporate multi-stage testing (Bilder *et al.,* 2010). This modified approach involves testing pooled samples from the

population of interest. A negative result from a pool indicates that all individuals within that pool are free from infection, leading to the discontinuation of further testing. Conversely, if the pool tests positive, random individuals within the pool are selected and retested individually until the first positive individual is identified. The remaining individuals are then combined to form a new pool, and the process is repeated. If the new pool tests positive, the procedure continues until all individuals are classified as either positive or negative.

Researchers have developed the halving algorithm for multi-stage group testing, which utilizes individual covariate information to facilitate information retesting (Black *et al*., 2012). In this algorithm, pools that initially test positive are divided into two halves and subjected to retesting. The process continues until a pool tests negative or individual testing is performed. Subsequent studies have explored the use of individual covariate information to estimate the probability of an individual testing positive. It was found that an informative approach based on the Dorfman's pooling scheme, which considers heterogeneity among individuals, outperforms methods that do not account for such differences (McMahan *et al.,* 2011). Preference is given to the risk probability of an individual, ensuring that a positive pool selected for retesting has the highest probability of containing positive cases.

The test result of a group can be categorized as dichotomous, indicating that it can yield either a positive or negative outcome. Non-hierarchical models are based on utilizing the test results of groups, and they include standard individual testing where each individual is tested individually to determine their status as positive or negative. Another example of a non-hierarchical model is array testing, which considers imperfect tests (Kim *et al.,* 2007). In array testing, specimens are arranged in a grid-like structure, and samples are pooled for testing based on rows and columns. Individual testing is then conducted on specimens that fall at the intersection of a positive row and a positive column. Further advancements in array testing have been explored, such as array testing in multiple directions (Berger *et al.,* 2000) and three-dimensional array procedures with testing errors (Kim & Hudgens, 2009). Apart from classification problem, group testing has also been used extensively in the estimation problem of prevalence rate of a trait. This is the second objective of group testing pioneered by Thompson (1962) and serves as the main focus of this study.

2.3 Estimation of Prevalence Rate

Thompson (1962) introduced another major area of study in group testing which has proved to be important as compared to just identifying defective elements in a sample. The concept of

estimation was used in many applications even before it was introduced in the statistical literature. For instance, estimation was used in finding the rate of disease transmission from an insect to a given plant (Gibbs & Gower, 1960; Thompson, 1962). It should be noted that the idea of estimation of the prevalence rate of a trait under study was motivated by the scarcity of resources. In this regard, information of disease transmission from a larger number would be captured by pooling observations and estimating the prevalence from the pooled samples.

Early studies of the statistical properties of group testing estimators were done by Gibbs and Gower (1960) who established the bias of MLE estimators especially when the population is high. The MLE method of estimation was found to be reliable when the population is small. Statistical properties such as bias and MSE were further addressed by Thompson (1962) and Chiang and Reeves (1962). Subsequent studies on estimation focused on design matters especially basing on selection of group sizes used in group testing procedures. The main aim of putting more emphasis on selection of group sizes was to reduce the chances of obtaining all negative or all positive groups during testing (Kerr, 1971). Another aim was to minimize MSE by incorporating prior information in choosing k (Hughes-Oliver & Swallow, 1994).

Sequential designs in group testing were championed by Kerr (1971). He indicated that if all groups are found to be negative or positive from a fixed sample size, a new sample was to be collected iteratively of predetermined size until a desirable result is obtained. Katholi and Unnasch (2006) later suggested the use of Negative Binomial sampling for estimating efficient p when the prevalence rate is small. Negative binomial sampling arises in a scenario where the number of positive pools is fixed and testing continues until the desired number is observed. Point and interval estimation has been explored under the negative binomial model using equal group sizes and has been found to be efficient in surveillance cases where quick response is desired (Pritchard & Tebbs, 2010). Nevertheless, Pritchard and Tebbs (2010) established debiased estimator by generalizing the Negative Binomial model to the case of unequal group sizes. They found out that the combination of inverse binomial sampling with group testing would be advantageous in situations requiring quick response such as an outbreak of a disease or natural disaster. They considered an example of transmission of WNV by mosquitoes and the impact of an outbreak of foot and mouth disease on the meat industry. Recent works that have considered inverse binomial model include estimation using Bayesian approach (Pritchard & Tebbs, 2011) as well as confidence interval estimation (Hepworth, 2013).

Further, Kariuki *et al.* (2023) presented a significant advancement in sequential group testing methodology using the Negative Binomial sampling model, particularly in the context of dealing with imperfect tests and estimating the prevalence of rare traits. By developing a two-stage negative binomial group testing model and utilizing methods such as Maximum Likelihood Estimation and Cramer-Rao lower bound, the study addressed the need for more efficient estimators in situations where testing processes are not perfect. Through Monte Carlo simulation studies, the research demonstrated the model's effectiveness in providing close approximations of prevalence levels, particularly when group sizes and waiting parameters are large, and at low values of sensitivity and specificity. Moreover, the study highlighted the model's superiority over existing approaches, showing improved efficiency of the estimator even when misclassifications are considered. This study deviated from Kariuki *et al*. (2023) by exploring two-stage adaptive negative binomial group testing model which focused on adjusting group sizes when perfect tests are considered. Moreover, since Thompson (1962), there has not been any adaptive estimation scheme developed under the Negative binomial model.

2.4 Choice of Group Size

The main problem in group testing has always revolved around the appropriate group size to be chosen in the procedure (the number of units or individual observations in a group). The idea of problem regarding the group size is as follows; if the number of k unit in a group is chosen to be too large, the probability π is close to one and the cost of testing is high. On the other hand, when k is chosen to be too small, the probability π is close to zero which may be too uninformative. Due to these unfortunate results, it is desirable to have appropriate group sizes when conducting group testing. Various recommendations have been established for choosing appropriate group sizes based on prior estimates of p. Thompson (1962) recommended k to be;

$$k = \frac{1.6}{p} \tag{2.1}$$

Obtained by minimizing the asymptotic variance of a prior estimator p. However, he recommended that a smaller k than the one he established would be better when the sample size is small to avoid potential bias of the estimate. Elsewhere, Chiang and Reeves (1962) recommended k to be obtained as;

$$k = \log\left(\frac{1}{2}\right)\log(1-p) \tag{2.2}$$

This k could be essential in obtaining equal probability of a negative or a positive result. Swallow (1985) came up with tables of optimal values of k on the basis of MLE of $\hat{p}$ minimization. However, Swallow (1987) later suggested that smaller group sizes may be used if cost per unit of information is to be minimized instead of choosing a group size that minimizes the MSE of a prior estimator.

Montesinos-López *et al*. (2013) conducted a study published in Seed Science Research, which focused on determining the optimal sample size for detecting transgenic plants. Their research employed the innovative approach of inverse binomial group testing, incorporating the dilution effect. This method is particularly pertinent for efficiently screening large plant populations for transgenes. By providing a statistical framework, the authors offered valuable insights into optimizing sampling strategies to reliably detect transgenic plants while minimizing resource expenditure. Their findings contributed significantly to the literature on transgenic plant detection methodologies and have practical implications for agricultural biotechnology research and applications.

Sequential group testing procedures have been established in solving the problem of appropriate group sizes in identification problem but has received little attention when it comes to the estimation problem. Bhattacharyya *et al.* (1979) recommended that more investigations to be done regarding sequential designs. It is worth noting that recent studies have emerged centering on multistage group testing procedure in the bid to clear the deficiencies brought about by inappropriate group size. Hughes-Oliver & Swallow (1994) proposed a two stage adaptive group testing procedure. In their study, the choice of group size depends on the MLE estimator obtained from the previous stages and the number of groups to be tested. From their criterion, MSE of the estimate of p would be minimized if only data from the next stage were used. On the other hand, a multi-stage design was developed by Okoth *et al*. (2017) incorporating retesting in the bid also to counter the problem brought about by inappropriate group sizes. In all these studies, the binomial sampling model was assumed.

The matter of using equal group sizes or unequal group sizes is equally of importance but will not be covered in this research. In this study, the focus is on estimating the prevalence rate from a test where k is adjusted from one stage to the other and negative binomial sampling model is used putting more emphasis on a two-stage procedure.

2.5 Group Testing Schemes

There are two forms of group testing which include non-adaptive and adaptive group testing schemes.

2.5.1 Non-Adaptive Scheme

Non-adaptive schemes have been constructed for both binomial and negative binomial model. Considering a binomial model group testing under this scheme, a large population of size N is grouped into n groups of size k with the aim of testing for the presence or absence of a trait of interest. Suppose that Y pools out of the n pools tested are found to be positive. Then Y follows a binomial distribution with parameters n and $\pi = 1 - (1 - p)^k$

$Y \sim \text{Binomial}(n, \pi)$

Note that π is the probability of a group testing positive. Using this model, the MLE of p is given as

$$\hat{p} = 1 - \left(1 - \frac{y}{n}\right)^{\frac{1}{k}} \tag{2.3}$$

As noted by Thompson (1962). On the other hand, a non-adaptive scheme in the case of Negative Binomial was developed in literature (Katholi & Unnasch, 2006). They established that Negative Binomial model and group testing was suitable in reporting estimates early for quick response especially in low prevalence areas. Suppose that a total of T pools are tested until the X positive pools are observed, T follows a Negative Binomial distribution with parameters X and $\pi = 1 - (1 - p)^k$. Note that π is the probability of a group testing positive just as in the binomial model since they are both formulated from individuals having a Bernoulli distribution. The probability density function (PDF) is given by

$$f(t / p, x, k) = \binom{t-1}{x-1} \left(1 - (1 - p)^k\right)^x (1 - p)^{k(t-x)} \tag{2.4}$$

Similarly, suppose pools of size k are tested for a trait of interest. T is the total pools tested before X positive groups are detected. Then, T follows a Negative Binomial with parameters X and $\pi = 1 - (1 - p)^k$. This can be written as;

$$f\left(t / p\right) = \binom{t-1}{x-1} \left[\pi\left(p\right)\right]^x \left[1 - \pi\left(p\right)\right]^{t-x} \tag{2.5}$$

The Likelihood function (dropping the constants) is then expressed as;

$$L(p/t,x) \propto \left[\pi(p)\right]^{x}\left[1-\pi(p)\right]^{t-x} \tag{2.6}$$

The log Likelihood function to base e is given as;

$$\ln(p/t,x) \propto x\ln\left[\pi(p)\right]+(t-x)\ln\left[1-\pi(p)\right] \tag{2.7}$$

The maximum likelihood estimator of the non-adaptive model is obtained as the solution to

$$\frac{\partial \ln L(.)}{\partial(p)} = 0 \tag{2.8}$$

which is equivalent to;

$$\frac{\partial \ln L}{\partial p} = \frac{x}{\pi}\pi' - \frac{(t-x)}{1-\pi}\pi' \tag{2.9}$$

where $\pi = 1-(1-p)^{k}$ which is the probability of obtaining a positive group and $\pi' = k(1-p)^{k-1}$. Equating Equation 2.9 to zero and solving for $\hat{p}_N$ yields the results obtained by Katholi & Unnasch (2006) as;

$$\hat{p}_N = 1-\left(1-\frac{x}{t}\right)^{\frac{1}{k}} \tag{2.10}$$

The variance of non-adaptive estimator is obtained using the Fisher's information criteria as;

$$\textit{Variance of } \hat{p}_N = \frac{1}{I(\hat{p}_N)} \tag{2.11}$$

where the Fisher's information is given as;

$$\mathrm{I}(\hat{p}_N) = -\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\ln L(.)\right]^{-1} \tag{2.12}$$

The second derivative of the log likelihood is given as;

$$\frac{\partial^2}{\partial p^2}\ln L(.) = \frac{x}{\pi}\pi'' - \frac{x}{\pi^2}(\pi')^2 - \left[\frac{t-x}{1-\pi}\pi'' + \frac{t-x}{(1-\pi)^2}(\pi')^2\right] \tag{2.13}$$

This simplifies to;

$$\frac{\partial^2}{\partial p^2}\ln L(.) = \frac{x}{\pi^2}\left(\pi\pi'' - (\pi')^2\right) - \frac{t-x}{(1-\pi)^2}\left[(1-\pi)\pi'' + (\pi')^2\right] \tag{2.14}$$

It is worth noting that $\pi(p) = 1-(1-p)^{k}$ and $\pi'(p) = k(1-p)^{k-1}$.

But,

$$E(T)=\frac{x}{\pi} \tag{2.15}$$

Taking expectation of Equation 2.14 gives;

$$\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\ln L(.)\right]=\frac{x}{\pi^2}\left(\pi\pi''-(\pi')^2\right)-\frac{x-\pi x}{\pi(1-\pi)^2}\left((1-\pi)\pi''+(\pi')^2\right) \tag{2.16}$$

This simplifies to;

$$\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\ln L(.)\right]=\frac{x}{\pi^2}\left(\pi\pi''-(\pi')^2\right)-\frac{x}{\pi(1-\pi)}\left((1-\pi)\pi''+(\pi')^2\right) \tag{2.17}$$

Putting Equation 36 under one denominator and simplifying further gives;

$$\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\ln L(.)\right]=\frac{x\left[(1-\pi)\pi\pi''-(1-\pi)(\pi')^2-(1-\pi)\pi\pi''-\pi(\pi')^2\right]}{\pi^2(1-\pi)} \tag{2.18}$$

It is equivalent to;

$$\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\ln L(.)\right]=\frac{-x(\pi')^2}{\pi^2(1-\pi)} \tag{2.19}$$

Thus, substituting the values of π, π' and $1-\pi$, multiplying with -1 and taking the inverse yields the variance as;

$$Var(\hat{p}_N)=\frac{\left(1-(1-p)^k\right)^2(1-p)^k}{xk^2(1-p)^{2k-2}} \tag{2.20}$$

This variance was used to construct the Wald confidence interval of the non-adaptive estimator while the MLE of the non-adaptive Negative Binomial group testing model is obtained as;

$$\hat{p}=1-\left(1-\frac{x}{t}\right)^{\frac{1}{k}} \tag{2.21}$$

Both the non-adaptive schemes using Binomial and Negative Binomial models have been found to produce biased estimators (Pritchard & Tebbs, 2010; Thompsons, 1962). Alternative studies have established different estimators to reduce the bias and increase the efficiency of the estimator. However, despite this shortcoming, the non-adaptive schemes have remained applicable especially in studies which involves detection or identification of defective units in a sample and estimation of proportion of defectives items in a population.

2.5.2 Adaptive Schemes

A two-stage adaptive group testing with perfect tests was developed by Oliver-Hughes and Swallow (1994). Their main focus was using ideal test to estimate small proportions in a population. They used MLE method to develop an adaptive estimator. The variance of the estimator was determined using the Cramer-Rao lower bound method. In this study, the total fixed number of pools to be tested was divided into two sets. λn and $(1-\lambda)n$ pools were tested in the first and second stage respectively. λ was used as the parameter of partition of the groups. In the first stage, the MLE was obtained as

$$\hat{p} = 1 - \left(1 - \frac{y}{\lambda n}\right)^{\frac{1}{k_1}} \tag{2.22}$$

where k_1 is the number of units in each group tested in the first stage.

It is worth noting that if $\lambda = 1$ in Equation 2.22, then the equation equals to Equation 2.21 obtained under non-adaptive binomial model with perfect test. The final MLE of prevalence *p* at the second stage was constructed using the joint distribution of y_1 and y_2 and is obtained as a solution to

$$\frac{k_1.y_1}{1-(1-p)^{k_1}} + \frac{k_2(y_1).y_2}{1-(1-p)^{k_2(y_1)}} = n\left(\lambda k_1 + (1-\lambda)k_2(y_1)\right) \tag{2.23}$$

where y_1 and y_2 are the number of positive groups in the first and second stage repectively.

2.6 Application of Group Testing

Due to the benefits associated with group testing since Dorfman (1943) pioneered it in his seminal work, it has been applied in many areas including DNA library blood screening, vector transmission of viruses, plant disease assessment and fisheries among others (Hepworth, 2005; Sobel & Groll, 1966). Group testing was first applied to the identification problem where recruits were tested for syphilis during the world war. The Dorfman (1943) procedure was used by pooling urine specimen and testing in the bid to identify the soldiers who were infected. According to Mundel (1984), group testing can be applied in various industries such as conducting an investigation of gas-filled electrical devices to identify the leaking items. This test is economical in testing a large number of units in a single test. The problem of identification has also been applied in conducting tests on other electronic devices including resistors and condensers.

In addition to the cost-effective benefit, group testing has also been associated with the benefit of confidentiality thus considered as the right method of testing for sensitive diseases to avoid stigmatization. For this reason, pool testing has been applied in screening a group of individuals for the presence or absence of HIV anti-body (Kline *et al.,* 1989; Litvak *et al.,* 1994). In this study, it was shown that group testing offers a feasible way to lower the errors that are associated with labelling units when screening population with low risk of HIV. Likewise, in previous studies, group testing has been employed to screen for chlamydia and gonorrhea, as indicated by Lindan *et al.* (2005). These two bacterial infections have been linked to the development of pelvic inflammatory diseases, ectopic pregnancies, sterility, and infertility. Moreover, research has shown that these infections are also responsible for transmitting other sexually transmitted diseases (STDs) such as HIV and Human papillomavirus (HPV), as highlighted by Lewis *et al.* (2012). Group screening has been utilized for various STDs, including HIV (Pilcher *et al.,* 2005), as well as hepatitis B and hepatitis C (Cardoso *et al.,* 1998).

More advances in agricultural biotechnology have shown that pool testing is an important aspect of risk assessment of products in agriculture. Estimation of proportion of defective units has played an important role in preventing the risk of unintentional mix of genetically modified plants with other plants.

Group testing has been applied in the discovery and the development of a new drug (Xie *et al.,* 2001). In their study, they found out that group testing can be used as a recommended way to reduce the cost incurred when searching for "lead" compound among the very many chemical compounds. It is the lead compound that is essential in the modification of a new drug. In addition, group testing has been applied in screening for West Nile Virus (WNV) (Busch *et al.,* 2005; Rutledge *et al*., 2003) as well as screening for HIN 1 influenza virus (Van *et al.,* 2012).

Quality control processes have also made use of group testing (Fang *et al.,* 2007; Wanyonyi *et al.,* 2021) to identify defective items. In this study, they realized that substantial savings can be realized if group testing is applied.

Recently, group testing has been applied in large-scale Covd-19 screening. This procedure has enabled many countries to deal with the problem of shortage of testing kits which might have hindered the effort of identification and isolation of infected individuals. Due to these numerous applications of group testing procedure, it is with no doubt that more improved and

efficient models should be developed to increase the benefits of conducting group testing and two-stage adaptive Negative Binomial model is constructed in this study for this great purpose.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Introduction

This chapter introduces the foundational framework of this study, centered on the development and evaluation of a two-stage adaptive negative binomial model. This adaptive scheme, assuming perfect tests serves as the basis of the research, aiming to enhance the efficiency and accuracy of prevalence estimation in group testing scenarios. The chapter commences with a detailed exposition of the two-stage adaptive scheme illustrating the sequential testing process and the dynamic adjustment of group sizes based on preceding stage outcomes. A diagrammatic illustration reinforces the essence of the model and explaining the progression from stage one to stage two. Stage one entails the formation of groups with an optimized size, determined to minimize the variance of the non-adaptive estimator, leveraging prior information on the trait's prevalence. Subsequently, stage two introduces adaptive adjustments in group size based on stage one outcomes, fostering refined estimations through an iterative testing approach. The chapter further delves into the estimation of prevalence rates, explaining the likelihood functions and methodologies employed in both stages. Additionally, it expounds on the computation of confidence intervals and the assessment of estimator properties, encompassing bias, variance, and mean squared error. The chapter culminates in a comparative analysis the proposed adaptive model with existing non-adaptive schemes.

3.2 Two-Stage Adaptive Negative Binomial Model

A two stage adaptive scheme assuming perfect test is described as it is the backbone of this study and thereafter perform a comparison analysis with other existing non-adaptive schemes which have also considered perfect tests.

The adaptive scheme involves testing groups in stages and adjusting the group sizes from one stage to the next. The group size used at a stage depends entirely on the outcome of the preceding stages. This implies that $k's$ are determined sequentially as the experiment progresses. The model is presented diagrammatically as illustrated in figure 3.1.

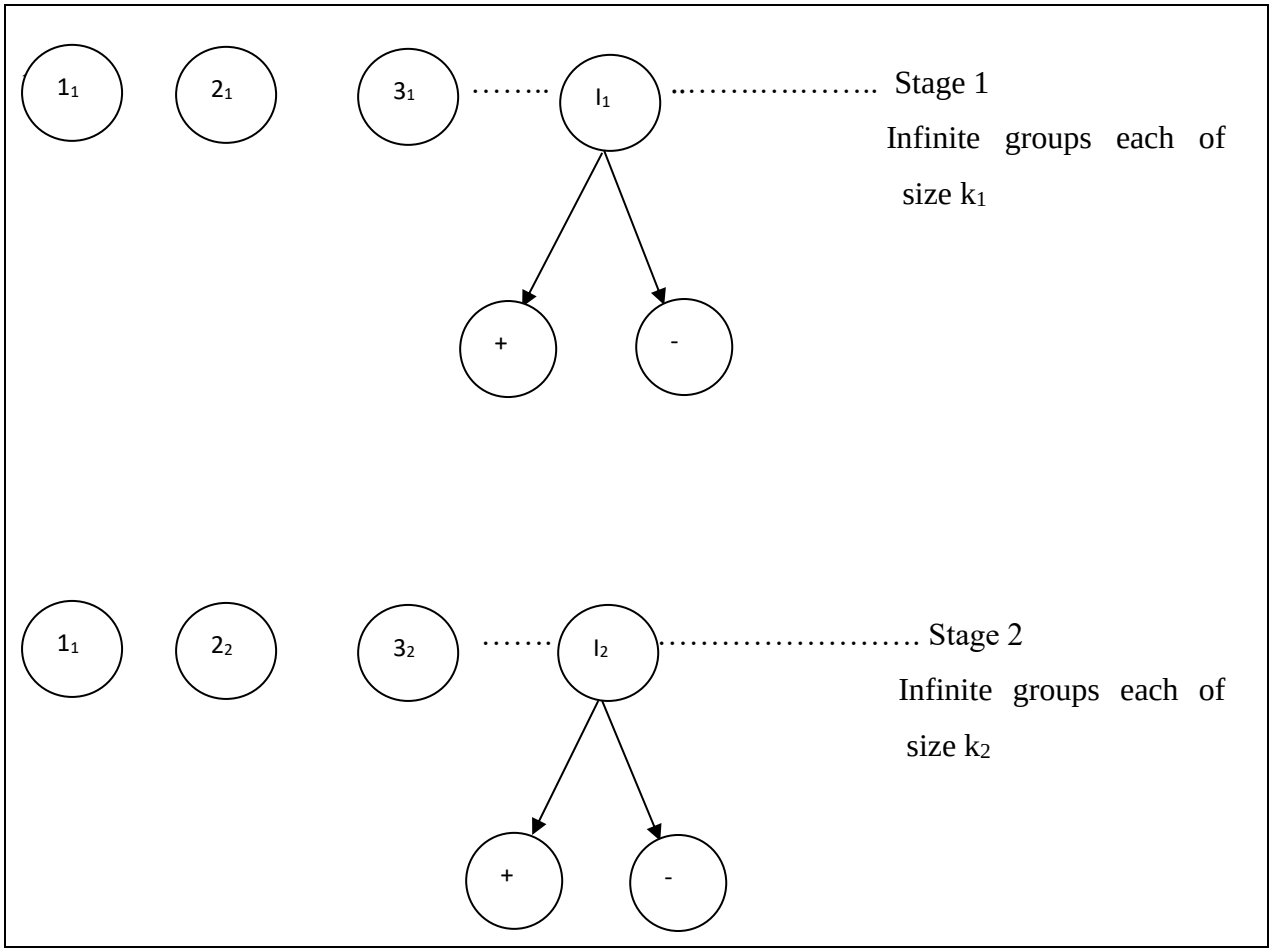

Figure 3. 1: Diagrammatic Illustration of the Model

3.2.1 Stage One

Groups of size k_1 are formed. The group size was obtained by minimizing the variance of the non-adaptive estimator as;

$$k_1 = \text{argmin}_k \left[\text{var}\left(\hat{p}_N\right) \right]_{p=p_0} \quad (3.1)$$

where k_1 was the value which minimizes the variance and is evaluated at p_0. It is worth noting p_0 is obtained from the prior information about the proportion of trait of interest. Suppose that x_1 the number of positive pools to be observed is fixed before the experiment begins and is determined based how severe a characteristic of interest is from the existing information. While T_1 is the number of groups to be tested before x_1 pools are observed. Then T_1 follows a Negative binomial distribution with parameters x_1 and $\pi_1 = 1 - (1-p)^{k_1}$ That is;

$$T_1 \sim Negative\ Binomial\ (x_1, \pi_1) \quad (3.2)$$

The prevalence estimator at stage one was obtained using the probability distribution function formed as the Negative Binomial distribution. Subsequently, the estimator of the second stage was obtained using stage one's estimator as the prior information.

3.2.2 Stage Two

In stage two, k_2 which is the group size to be used in this phase was constructed as

$$k_2 = \text{argmin}_k \left[\text{var}\left(\hat{p}_1\right) \right]_{p=\hat{p}_1} \tag{3.3}$$

where $\hat{p}_1$ is the estimator obtained in stage one.

Suppose x_2 is the fixed number of desirable positive pools in stage two. While T_2 is the number of pools to be tested to get x_2 positive pools. Then, T_2 follows a Negative Binomial distribution with parameters x_2 and $\pi_{2/1} = 1 - (1-p)^{k_2(t_1)}$ and is dependent on the output of the first stage. Then, T_2/T_1 can be written as,

$T_2/T_1 \sim NegativeBinomial\left(x_2, \pi_{2/1}\right)$

Now using the definition of conditional probability, it follows that;

$$f\left(t_2 / t_1\right) = \frac{f\left[t_2, t_1\right]}{f\left[t_1\right]} \tag{3.4}$$

Equation 11 yields the joint distribution of T_1 and T_2 as;

$$f\left(t_2, t_1\right) = f\left(t_2 / t_1\right) \times f\left(t_1\right)$$

$$= \text{Negative Binomial}\left(x_2, \pi_{2/1}\right) \times \text{Negative Binomial}\left(x_1, \pi_1\right) \tag{3.5}$$

Equation 3.5 was used to derive the final two-stage adaptive estimator $\hat{p}_A$.

3.3 Estimation of Prevalence Rate

The point estimator of the prevalence was obtained using the method of maximum likelihood. This is a common method that has been used in inferential statistics especially group testing involving Binomial model. The method involves maximizing the likelihood function of a distribution.

3.3.1 The Likelihood Function of p in Stage One

Dropping the constant and using proportionality sign, the likelihood function in stage one is given as;

$$L(p / t_1) \propto \left(\pi_1\right)^{x_1} (1-\pi_1)^{t_1 - x_1} \tag{3.6}$$

This Equation 3.6 was maximized to obtain stage one estimator of the proportions.

3.3.2 The Likelihood Function of p in Stage Two

The likelihood function for the joint distribution in stage two was given as

$$L(p/t_1,t_2) \propto \left(\pi_1\right)^{x_1}(1\text{-}\pi_1)^{t_1-x_1}\times\left(\pi_{2/1}\right)^{x_2}(1\text{-}\pi_{2/1})^{t_2-x_2} \tag{3.7}$$

The final estimate which is the adaptive estimator is obtained as the value that maximizes the Equation 3.7 where the second elements after the product sign in the distribution are the parts dependent on the first stage since $\pi_{2/1} = 1-(1-p)^{k_2(t_1)}$.

3.3.3 The Wald Confidence Interval

Further, the Wald confidence interval was constructed and its performance investigated by constructing interval length. When X which is the fixed desired positive groups is large, the MLE is approximately normal and thus the CI is given as;

$$CI = \hat{p} \pm Z_{\frac{\alpha}{2}} I\left(\hat{p}\right)^{-\frac{1}{2}} \tag{3.8}$$

where I($\hat{p}$) is the Fisher's information and $Z_{\frac{\alpha}{2}}$ denotes the upper quantile for N (0,1) distribution. This is the approximate 100 (1- α) percent Wald confidence interval for p.

3.4 Properties of the Maximum Likelihood Estimator

Once the model was developed and the MLE computed, the properties of the derived estimator including bias, variance and Mean Squared Error were determined.

3.4.1 Asymptotic Variance

The asymptotic variance of the proposed model was obtained using the Fisher's Information denoted as

$$\mathrm{I}_{(p)} = -E\left[\frac{\partial^2}{\partial p^2}\ln f\left(t_1,t_2/p\right)\right] \tag{3.9}$$

The variance obtained as the inverse of the Fisher's information.

$$\mathrm{Var}(\hat{\mathrm{p}}) = \frac{1}{\mathrm{I}\left(\hat{\mathrm{p}}_A\right)} \tag{3.10}$$

It is this variance that was applied in finding the Wald interval for p.

3.4.2 Bias of the Estimator

The bias of an estimator refers to the difference between the true value of the parameter and the expected value of the parameter being estimated. It is obtained by

$$\mathrm{Bias}\ (\hat{\mathrm{p}}) = \mathrm{E}(\hat{\mathrm{p}}) - \mathrm{p} \tag{3.11}$$

Or equivalently,

$$\text{Bias}(\hat{p})=\sum[\hat{p}-p]\binom{t_1-1}{x_1-1}(\pi_1)^{x_1}(1-\pi_1)^{t_1-x_1}\times\binom{t_2-1}{x_2-1}(\pi_{2/1})^{x_2}(1-\pi_{2/1})^{t_2-x_2} \quad (3.12)$$

In this study bias was obtained using Monte Carlo simulations utilizing Equation 3.12.

3.4.3 Mean Squared Error of the Estimator

MSE amalgamates the information from both the variance and bias of an estimator. The MSE of an estimator is inflated by either poor precision or inaccuracy of the estimator. The MSE will be obtained as;

$$MSE(\hat{p})=\left(bias(\hat{p})\right)^2+\text{var}(\hat{p}) \quad (3.13)$$

Or alternatively as mentioned by Pritchard and Tebbs (2010);

$$\text{MSE}(\hat{p})=\sum[\hat{p}-p]^2\binom{t_1-1}{x_1-1}(\pi_1)^{x_1}(1-\pi_1)^{t_1-x_1}\times\binom{t_2-1}{x_2-1}(\pi_{2/1})^{x_2}(1-\pi_{2/1})^{t_2-x_2} \quad (3.14)$$

Since the equation do not reduce to anything tractable, the Monte Carlo simulations were used in R software to obtain the MSE of the estimator.

3.5 Model Comparison

The proposed model was compared to the non-adaptive group testing model existing in literature (Katholi & Unnasch, 2006). This was done by computing asymptotic relative efficiency ARE and the relative mean squared error RMSE.

3.5.1 Asymptotic Relative Efficiency

The estimator of non-adaptive model is denoted by $\hat{p}_N$ since it is developed under the non-adaptive Negative Binomial model while the estimator of our proposed model is denoted as $\hat{p}_A$ since is developed under the adaptive Negative Binomial model. Then, ARE is given as

$$ARE=\frac{Var(\hat{p}_N)}{Var(\hat{p}_A)} \quad (3.15)$$

If ARE is found to be greater than one, it implies that our proposed model is more efficient than the non-adaptive model.

3.5.2 Relative Mean Square error

This is a convenient way of comparing the MSE of the estimates obtained using different procedures. It is expected that a good model to produce an estimator with a small MSE. For this study, $\hat{p}_A$ is compared to $\hat{p}_N$ by computing the RMSE as

$$RMSE=\frac{MSE(\hat{p}_N)}{MSE(\hat{p}_A)} \quad (3.16)$$

3.6 Simulation and Analysis

The study applied Monte Carlo simulation using R-software version 4.1.2 to determine the bias, variance, and MSE of the estimator as well as constructing intervals for true population parameter p. In this regard, 1000 Monte Carlo data sets were simulated from the Negative Binomial model in both stages using an appropriate group size. The value of k was obtained by minimizing the variance which was a function of k and p. The resulting function were not linear thus did not have a solution in closed form. Therefore, there was need to use uniroot inbuilt functions in R to solve iteratively. The following is a flow chart and an algorithm of a Negative Binomial Simulation.

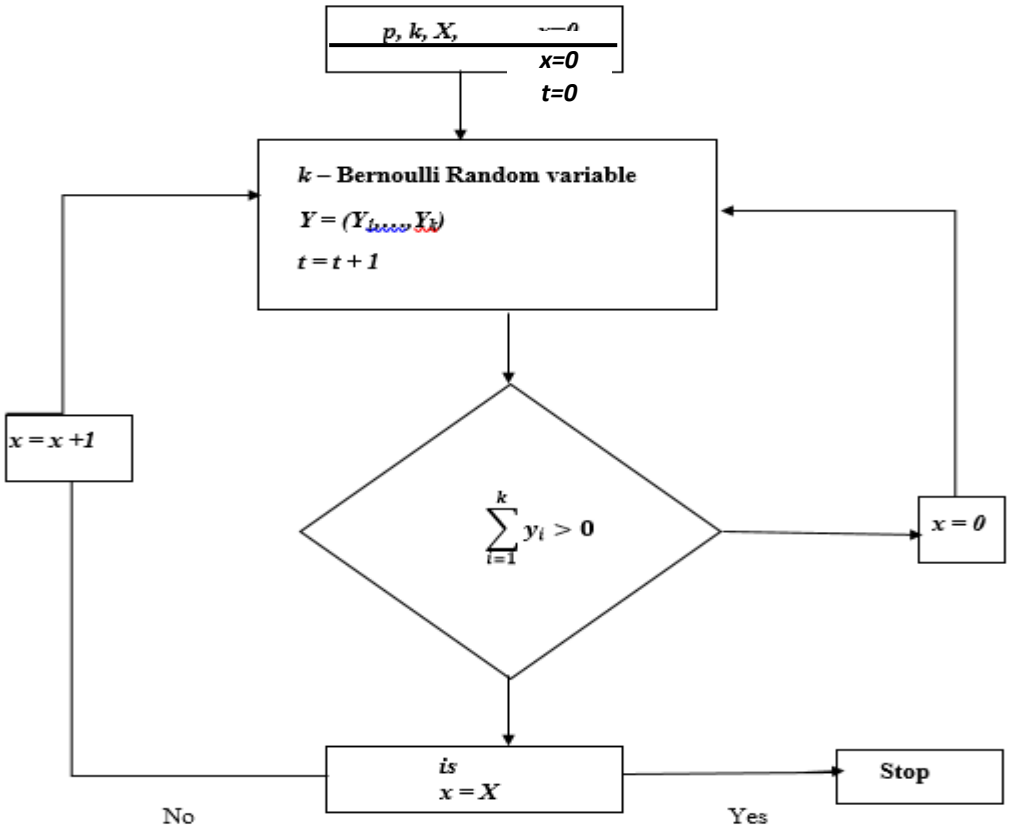

Figure 3. 2: Flow Chart for Negative Binomial Group Testing Simulation

Steps for Simulation

Step 1: Specify p, k and X then set x=0

Step 2: Generate k Bernoulli random variables

Set $Y=(Y_1,\ldots,Y_k)$

Step 3: If the sum of y_{is} is greater than 0. A success is considered. If the sum of y_{is} is less than the group is considered negative.

Step 4: If the success was recorded, repeat the loop if x is not equal to X. If x = X, the procedure stops

Step 5: T is reported and p estimated

3.7 Application to Real Data

Practical application of the model was examined using the West Nile Virus surveillance data that was used in a public health study described by Rutledge *et al.* (2003). This data was collected using the reverse transcription-polymerase chain reaction method where 11948 individuals were pooled into various groups. Out of these pools, 14 groups tested positive for the trait of interest WNV. Although, Negative Binomial model was not used in the collection of this data, this study made use of the data for illustration and verification of the constructed model. The data was assumed to be finitely large to allow for the performance of two stage adaptive stage. It is also worth noting that it was assumed that the 14^{th} positive group was observed in the last trial that was done.

3.8 COVID- 19 Protocols

The coronavirus disease 2019 (COVID-19) is a respiratory communicable disease caused by a new strain of corona virus that affects humans. The World Health Organization (WHO) has described it as a global pandemic that continues to ravage different parts of the world (Guner *et al*., 2021). To mitigate the spread of the infection, some of the COVID-19 protocols as stipulated by the WHO that were observed in this study include:

i. Regularly washing hands with soap and running water, or cleaning them with an alcohol-based hand rub.
ii. Putting on a face mask that covered both the nose, mouth, and chin.
iii. Maintaining at least one-meter social distance between people to reduce the risk of infection when they cough, sneeze or speak.
iv. Ensuring the safe disposal of medical masks in a trash bin.
v. Taking the COVID-19 vaccine and booster shots.
vi. Self-isolation in case of infections, and seeking medical attention for severe symptoms.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents discussion of the results of this study. The derivation of the adaptive negative binomial group testing models is presented. The estimation results of the adaptive estimator are presented which include the constructed estimator and its properties such as variance, bias and mean squared error. The constructed model is also compared to the existing non-adaptive Negative Binomial model using the asymptotic relative efficiency and the relative mean squared error. The chapter also presents application of the constructed model to real data which in this case the study makes use of the West Nile Virus a leading cause of arbovirus encephalitis in the United States.

4.2 Derivation of the Two-stage Adaptive Negative Binomial estimator

This section delves into the derivation of the proposed model focusing on the two-stage adaptive Negative Binomial group testing scheme. By employing the two-stage adaptive Negative Binomial scheme, this approach allows for a more efficient allocation of resources and provides a robust methodology for achieving the desired results. This section therefore, explain the underlying principles and mathematical framework behind the proposed model. It explored how the Negative Binomial distribution is utilized to estimate the probabilities of positive pools. Through the derivation of the proposed model, this study aimed to contribute to the advancement of pool testing strategies offering an effective and practical solution for large-scale screening and identification of positive groups.

The desired number of positive groups desired is divided into two sets namely X_1 and X_2. Pools in an infinitely large population are then tested in two stages where in each stage, testing stops once the desired number of positive pools are attained. In this regard, groups are tested in stage 1 until X_1 positive pools are obtained. Similarly, pools will be tested in stage two until X_1 positive pools are obtained. It is important to note that the specific values of X_1 and X_2 in practice depends on the specific requirements and constraints of the situation, such as the population size, available resources, and the desired level of accuracy in identifying positive groups. These values can be determined based on a thorough understanding of the problem and careful consideration of the trade-offs involved in the testing process. Thus, the study also

considered the impact of varying the desired number of positive groups in the first stage has on the efficiency of the adaptive model.

4.2.1 Stage One Estimator of p

In this stage, groups of size k_1 are obtained from the equation minimizing the variance of the non-adaptive estimator. That is, by substituting the variance of the non-adaptive estimator, the minimizing equation becomes;

$$k_1 = \arg\min_k \left[\frac{\left(1-(1-p)^k\right)^2 (1-p)^k}{xk^2 (1-p)^{2k-2}} \right]_{p=p_0} \tag{4.1}$$

In this case, k_1 is the value that minimizes the variance of the non-adaptive estimator.

Suppose that X_1 is the number of positive pools to be observed in the first stage, while T_1 is the number of groups to be tested before X_1 pools are observed. Then T_1 follows a Negative Binomial distribution with parameters X_1 and π_1. That is;

$$T_1 \sim Negative\ Binomial(X_1, \pi_1) \tag{4.2}$$

Or equivalently;

$$f(t_1, p) = \binom{t_1 - 1}{x_1 - 1} \left[\pi_1(p)\right]^{x_1} \left[1-\pi_1(p)\right]^{t_1 - x_1} \tag{4.3}$$

By employing Equation 37, the prevalence estimator in the first stage is derived as follows;
The Likelihood function of Equation 37 is expressed as;

$$L(p/t_1, x_1) \propto \left[\pi_1(p)\right]^{x_1} \left[1-\pi_1(p)\right]^{t_1 - x_1} \tag{4.4}$$

The log Likelihood function is given as;

$$\log L(p/t_1, x_1) \propto x_1 \log\left[\pi_1(p)\right] + (t_1 - x_1) \log\left[1-\pi_1(p)\right] \tag{4.5}$$

The maximum likelihood estimator of the non-adaptive model is obtained as the solution to;

$$\frac{\partial \log L(.)}{\partial p} = 0 \tag{4.6}$$

This is equivalent to;

$$\frac{x_1}{\pi_1}\pi_1' - \frac{(t_1 - x_1)}{1-\pi_1}\pi_1' = 0 \tag{4.7}$$

Solving Equation 46 yields;

$$\hat{p}_1 = 1 - \left(1 - \frac{x_1}{t_1}\right)^{\frac{1}{k_1}} \tag{4.8}$$

The variance of the estimator obtained in the first stage is obtained as;

$$\text{Variance of } \hat{p}_1 = \frac{1}{\text{Fisher's Information}} \tag{4.9}$$

4.2.2 Stage Two Estimator of p

In this stage, groups are formed of size k_2 which is obtained as;

$$k_2 = \text{argmin}_k \left[\text{var}\left(\hat{p}_1\right)\right]_{p=\hat{p}_1} \tag{4.10}$$

Suppose T_2 is the number of groups to be tested to obtain X_2. Then T_2 conditioned on T_1 follows a Negative Binomial distribution. Specifically,

$$T_2/T_1 \sim NegativeBinomial\left(X_2, \pi_{2/1}\right) \tag{4.11}$$

Using the definition of conditional probability, it follows that;

$$f\left(T_2 / T_1\right) = \frac{f\left[T_2, T_1\right]}{f\left[T_1\right]} \tag{4.12}$$

Equation 46 gives the joint distribution of T_1 and T_2 as;

$$\begin{aligned} f\left(T_2, T_1\right) &= f\left(T_2 / T_1\right) \times f\left(T_1\right) \\ &= \text{Negative Binomial}\left(X_2, \pi_{2/1}\right) \times \text{Negative Binomial}\left(X_1, \pi_1\right) \end{aligned} \tag{4.13}$$

Equation 47 was used to derive the final two-stage adaptive estimator $\hat{p}_A$ as follows.

$$f\left(t_1, t_2\right) = \binom{t_1 - 1}{x_1 - 1}\left[1 - \left(1-p\right)^{k_1}\right]^{x_1}\left(1-p\right)^{k_1(t_1 - x_1)} \times \binom{t_2 - 1}{x_2 - 1}\left[1 - \left(1-p\right)^{k_2}\right]^{x_2}\left(1-p\right)^{k_2(t_2 - x_2)} \tag{4.14}$$

$Since$ $\binom{t_1 - 1}{x_1 - 1}$ and $\binom{t_2 - 1}{x_2 - 1}$ are constants, they are dropped and replaced with proportionality sign giving;

$$f\left(t_1, t_2\right) \propto \left[1 - \left(1-p\right)^{k_1}\right]^{x_1}\left(1-p\right)^{k_1(t_1 - x_1)} \times \left[1 - \left(1-p\right)^{k_2}\right]^{x_2}\left(1-p\right)^{k_2(t_2 - x_2)} \tag{4.15}$$

The log likelihood of the joint function is given as;

$$\ln L \propto x_1 \ln\left[1 - \left(1-p\right)^{k_1}\right] + k_1\left(t_1 - x_1\right)\ln\left(1-p\right) + x_2 \ln\left[1 - \left(1-p\right)^{k_2}\right] + k_2\left(t_2 - x_2\right)\ln\left(1-p\right) \tag{4.16}$$

The MLE of the adaptive estimator is obtained as the solution to;

$$\frac{\partial \ln L(.)}{\partial p} = 0 \tag{4.17}$$

This implies;

$$\frac{\partial \ln L}{\partial p} = \frac{x_1}{\pi_1}\pi_1' - \frac{(t_1 - x_1)}{1-\pi_1}\pi_1' + \frac{x_2}{\pi_2}\pi_2' - \frac{(t_2 - x_2)}{1-\pi_2}\pi_2' \tag{4.18}$$

where $\pi_1 = 1-(1-p)^{k_1}$ and $\pi_2 = 1-(1-p)^{k_2}$. Also $\pi_1' = k_1(1-p)^{k_1-1}$ and $\pi_2' = k_2(1-p)^{k_2-1}$.

Since Equation 4.18 does not have a closed form, it can be solved using iterative methods. This study used uniroot function in R software to obtain the adaptive estimator $\hat{p}_A$ as the solution to Equation 4.18.

4.2.3 Variance of the Adaptive Estimator

To find the variance of the adaptive estimator does not require us to get the value of $\hat{p}_A$. Instead, we find the Fisher's information by getting the second derivative of the Equation 52. Since the first derivative is given by Equation 50, the second derivative was obtained with respect to p as follows;

$$\begin{aligned}\frac{\partial^2}{\partial p^2}\log L(.) = {} & \frac{x_1}{\pi_1}\pi_1'' - \frac{x_1}{\pi_1^2}(\pi_1')^2 - \left[\frac{t_1 - x_1}{1-\pi_1}\pi_1'' + \frac{t_1 - x_1}{(1-\pi_1)^2}(\pi_1')^2\right] + \\ & \frac{x_2}{\pi_2}\pi_2'' - \frac{x_2}{\pi_2^2}(\pi_2')^2 - \left[\frac{t_2 - x_2}{1-\pi_2}\pi_2'' + \frac{t_2 - x_2}{(1-\pi_2)^2}(\pi_2')^2\right]\end{aligned} \tag{4.19}$$

This simplifies to;

$$\begin{aligned}= {} & \frac{x_1}{\pi_1^2}\left(\pi_1\pi_1'' - (\pi_1')^2\right) - \frac{t_1 - x_1}{(1-\pi_1)^2}\left[(1-\pi_1)\pi_1'' + (\pi_1')^2\right] \\ & + \frac{x_2}{\pi_2^2}\left(\pi_2\pi_2'' - (\pi_2')^2\right) - \frac{t_2 - x_2}{(1-\pi_2)^2}\left[(1-\pi_2)\pi_2'' + (\pi_2')^2\right]\end{aligned} \tag{4.20}$$

Finding the variance of $\hat{p}_2$ we use the Fisher's information criteria as;

$$\text{Variance of } \hat{p}_2 = \frac{1}{\text{Fisher's Information}} \tag{4.21}$$

Where the Fisher's information is given as;

$$\text{I=-E}\left[\frac{\partial^2}{\partial p^2}\log L(.)\right] \tag{4.22}$$

The $E(T)=\frac{x}{\pi}$. Thus, the Fishers information is given as;

$$I(\hat{p}_A)=-\left[\begin{array}{l}\frac{x_1}{\pi_1^2}\left(\pi_1\pi_1^{''}-\left(\pi_1^{'}\right)^2\right)-\frac{x_1}{\pi_1(1-\pi_1)^2}\left[(1-\pi_1)\pi_1^{''}+\left(\pi_1^{'}\right)^2\right] \\ +\frac{x_2}{\pi_2^2}\left(\pi_2\pi_2^{''}-\left(\pi_2^{'}\right)^2\right)-\frac{x_2}{\pi_2(1-\pi_2)^2}\left[(1-\pi_2)\pi_2^{''}+\left(\pi_2^{'}\right)^2\right]\end{array}\right] \tag{4.23}$$

Simplifying further gives;

$$I(\hat{p}_A)=-\left[\begin{array}{l}\frac{x_1\left[(1-\pi_1)\pi_1\pi_1^{''}-(1-\pi_1)\left(\pi_1^{'}\right)^2-(1-\pi_1)\pi_1\pi_1^{''}-\pi_1\left(\pi_1^{'}\right)^2\right]}{\pi_1^{\,2}(1-\pi_1)} \\ +\frac{x_2\left[(1-\pi_2)\pi_2\pi_2^{''}-(1-\pi_2)\left(\pi_2^{'}\right)^2-(1-\pi_2)\pi_2\pi_2^{''}-\pi_2\left(\pi_2^{'}\right)^2\right]}{\pi_2^{\,2}(1-\pi_2)}\end{array}\right] \tag{4.24}$$

The Equation 4.24 simplifies to;

$$I(\hat{p}_A)=\frac{x_1\left(\pi_1^{'}\right)^2}{\pi_1^2(1-\pi_1)}+\frac{x_2\left(\pi_2^{'}\right)^2}{\pi_2^2(1-\pi_2)}$$

(4.25)

The variance is then obtained as;

$$Var(\hat{p}_A)=\frac{1}{\sum_{i=1}^{2}\frac{x_i k_i^{\,2}(1-p)^{2k_i-2}}{\pi_i^2(1-\pi_i)}} \tag{4.26}$$

The variance obtained is used to construct the adaptive Wald confidence interval as;

$$\hat{p}_A \pm Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{p}_A)} \tag{4.27}$$

The bias and MSE of the estimators are obtained through Monte Carlo simulations and the simulation's R code are as illustrated in the Appendices. Monte Carlo simulations are used to obtain the bias and mean squared error of estimators because they provide a practical and effective way to estimate these quantities when the sampling distribution of the estimator is not analytically tractable.

4.3 Relationship between T, p, and k

The Negative Binomial sampling method used in group testing experiments involves a random number of trials, while the required positive pools and group size are fixed. The number of tests needed to obtain the required positive groups depends on various variables, so it's important to understand how changing the values of p and k affect the number of testing trials required.

4.3.1 Relationship between T and p when k is fixed

The results provide insights into the relationship between the probability of success p and the number of trials needed to obtain the desired number of positive groups. The desired number of positive groups is set as $X = 30$. The results help in understanding the impact of different parameters on the efficiency and effectiveness of this group testing procedure.

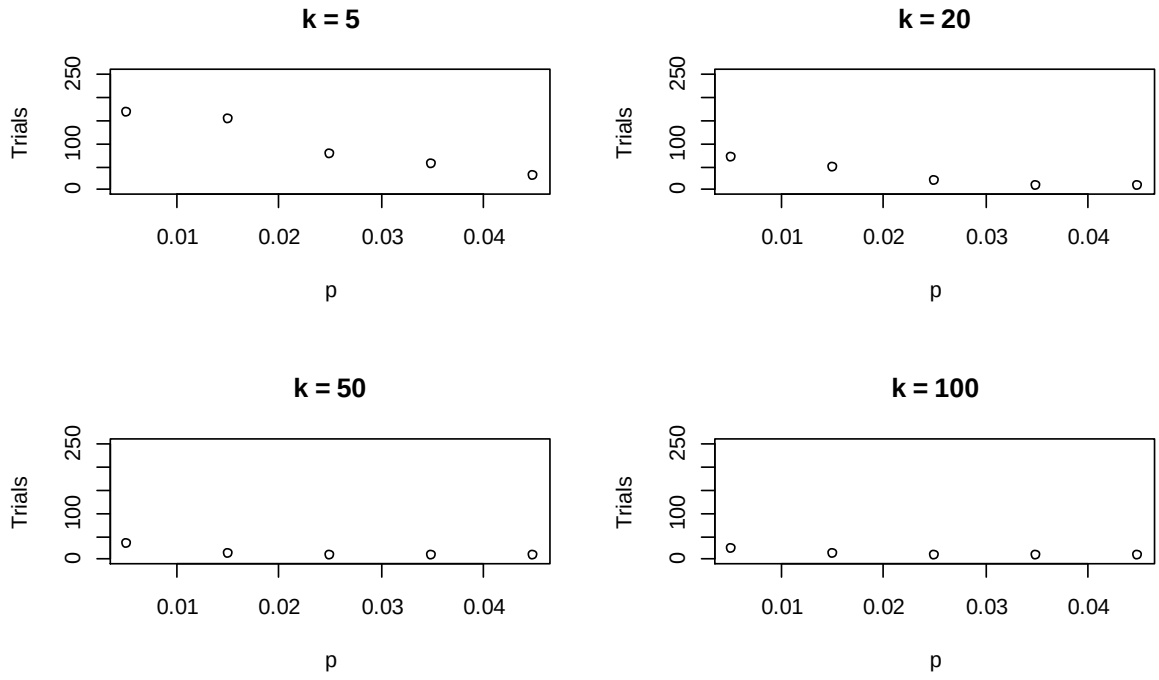

Figure 4. 1: Plots of T versus p for $k = 5,20,50,100$

As the probability of success increases in the negative binomial group testing model, the number of trials required to obtain the desired number of positive groups generally decreases as illustrated in Figure 4.1. Examining the plots, for a fixed value of k, it can be observed that as p increases, the number of trials decreases. This trend holds true across different values of k. This behavior is expected because a higher probability of success implies a greater likelihood

of encountering positive groups during the testing process. Therefore, fewer trials are needed to reach the desired number of positive groups when the probability of success is higher. It is worth noting that increasing the group size from 5 to 100 in the negative binomial group testing model typically leads to a reduction in the number of trials required to achieve the desired number of positive groups as well.

4.3.2 Relationship between T and k when p is fixed

These results obtained in this section provide valuable insights into the relationship between group size and the number of trials needed to achieve the desired number of positive groups in the group testing procedure. By analyzing different values of group size and success probability, the study sheds light on the efficiency and effectiveness of the group testing strategies. Understanding the impact of these parameters is crucial for optimizing the testing process and designing effective group testing procedures.

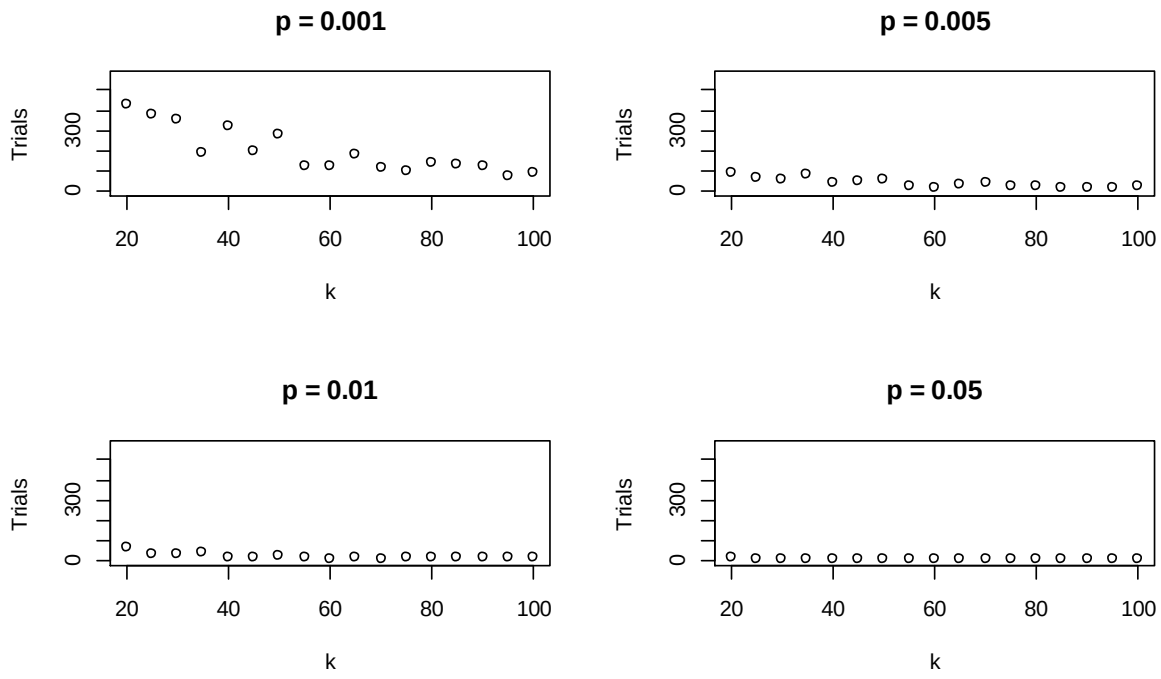

Figure 4. 2: Plots of T versus k for $p = 0{,}001, 0.005, 0.01, 0.05$

The results in figure 4.2 demonstrate that as the group size increases, the number of trials required generally decreases for various values of p. This trend can be observed within each graph were increasing k from 20 to 100 leads to a reduction in the number of trials for all values of $p = 0.001, 0.005, 0.01, 0.05$. Larger group sizes enable more individuals to be tested

simultaneously, resulting in a more efficient testing process and a decreased overall number of trials. The number of trials is also influenced by the success probability, which represents the prevalence of positive cases in the population. Generally, lower values of p like 0.001 require a larger number of trials for a given group size whereas higher values of $p = 0.05$ require fewer trials. A lower success probability implies that positive cases are rare making their detection more challenging and necessitating more trials to identify the desired number of positive groups. Conversely, higher success probabilities indicate a higher prevalence of positive cases, making them easier to detect and resulting in fewer trials needed.

It is important to note that while larger group sizes offer benefits such as increased sensitivity, greater information gain, and reduced probability of false negatives. Very large group sizes may introduce logistical challenges, require larger testing facilities, and incur higher costs. Moreover, if the prevalence of positive cases is low, very large group sizes may result in a higher proportion of negative groups, reducing testing efficiency. Therefore, when determining the optimal group size there is need to strike a balance between the advantages of increased sensitivity and efficiency and the practical considerations as well as constraints of the testing process.

4.4 Adaptive Estimator and Its Properties

This section presents the results of the maximum likelihood estimator and its properties including the variance, bias, and mean squared error for the adaptive group testing model. The results are organized based on different group sizes and true probabilities while the number of predetermined desired positive groups set at $X = 30$ as set by Xiong (2015).

Table 4. 1: Adaptive estimator with its properties for $k = 5,10,20,50,100$ when $X = 30$

P	MLE	Variance	Bias	MSE
		k=5		
0.001	0.000488	7.94E-09	-0.0004	4.0938E-08
0.005	0.002677	2.38E-07	-0.0025	1.02E-07
0.01	0.004862	7.82E-07	-0.004	4.05E-06
0.05	0.035433	3.94E-05	-0.0209	7.98E-05
0.1	0.075796	0.000167	-0.0275	0.000322241
		k=10		
0.001	0.000551	1.01E-08	-0.0005	3.88E-08

0.005	0.002264	1.70E-07	-0.0022	8.89E-07
0.01	0.005933	1.16E-06	-0.0042	4.18E-06
0.05	0.027808	2.31E-05	-0.0177	8.29E-05
0.1	0.070342	0.000142	-0.0258	0.00034779
		k=20		
0.001	0.000634	1.34E-08	-0.0006	3.59E-08
0.005	0.003144	3.28E-07	-0.0019	1.05E-06
0.01	0.005263	9.15E-07	-0.0057	3.63E-06
0.05	0.029922	2.82E-05	-0.0198	9.04E-05
0.1	0.074906	0.000158	-0.0268	0.00036145
		k=50		
0.001	0.000588	1.15E-08	-0.0004	3.84E-08
0.005	0.002204	1.61E-07	-0.0026	8.88E-07
0.01	0.004198	5.84E-07	-0.0042	3.28E-06
0.05	0.026812	2.27E-05	-0.0223	9.28E-05
0.1	0.075437	0.000165	-0.0506	0.001282513
		k=100		
0.001	0.000488	7.94E-09	-0.0005	3.93E-08
0.005	0.002503	2.08E-07	-0.0018	8.49E-07
0.01	0.00432	6.18E-07	-0.0048	3.70E-06
0.05	0.033403	3.49E-05	-0.0164	0.000306669
0.1	0.068592	0.000137	-0.0175	0.004388142

A scrutiny of Table 4.1 shows that the estimated probabilities tend to increase as the probability increases, although the increase is generally small. It is important to note that the MLE values of the adaptive model exhibits monotonic behavior as the model dynamically adjusts the group size based on stage 1 outcomes which leads to more consistent and accurate estimations. The MLE generally increases as p increases for $k = 5$. For example, at $p = 0.001$, the MLE is 0.000488281 and at $p = 0.01$, the MLE is 0.07579572. The variance of the estimated probabilities remains relatively small across different values of p. For instance, at $p = 0.001$, the variance is 7.82E-07, and at $p = 0.01$, it is 0.000166933. The bias of the estimation shows negative values indicating a slight underestimation of the true probability.

However, the bias remains relatively small across all values of p. The MSE combines the variance and bias to provide an overall measure of estimation accuracy. It is generally small implying a better accuracy in the adaptive model. Similar observations can be made for $k = 10,20,50,100$ with similar patterns observed across different group sizes. The bias remains relatively small and consistent across different values of p indicating the effectiveness of the adaptive group testing model in reducing bias compared to the non-adaptive approach. The adaptive model generally provides smaller estimated probabilities of an individual being positive. This indicates that the adaptive model, which adjusts the group size based on the observed testing outcomes, takes into account the adaptive nature of the testing procedure and provides more accurate estimations. The MLE values of the adaptive model vary for different combinations of p with k.

4.4.1 Relationship Between $\hat{p}$ and p

The results presented in this section examines the relationship between the adaptive maximum likelihood estimates and the true probability for different values of group size. The results are represented in four graphs for $k = 5,20,50,100$. These findings in table 4.2 highlight the adaptive nature of the model in adjusting the estimations based on observed outcomes and the varying performance of the adaptive approach across different group sizes.

Table 4. 2: MLE of p for $k = 5,20,50,100$ and $X = 10,20,30$

P	x=10	x=20	x=30
	k=5		
0.001	0.0004	0.00048	0.00049
0.005	0.00259	0.00267	0.00268
0.01	0.00478	0.00486	0.00486
0.05	0.02243	0.03043	0.03543
0.1	0.07072	0.07079	0.07579
	k=10		
0.001	0.00046	0.00054	0.00055
0.005	0.00217	0.00225	0.00226
0.01	0.00584	0.00592	0.00593
0.05	0.0128	0.0228	0.0278
0.1	0.06526	0.06534	0.07034

	k=20		
0.001	0.00055	0.00062	0.00063
0.005	0.00305	0.00313	0.00314
0.01	0.00517	0.00525	0.00526
0.05	0.01692	0.02492	0.02992
0.1	0.06982	0.0699	0.0749
	k=50		
0.001	0.0005	0.00058	0.00058
0.005	0.00211	0.00219	0.0022
0.01	0.00411	0.00419	0.00419
0.05	0.01181	0.02181	0.02681
0.1	0.07035	0.07043	0.07543
	k=100		
0.001	0.0004	0.00048	0.00048
0.005	0.00241	0.00249	0.0025
0.01	0.00423	0.00431	0.00432
0.05	0.0284	0.0284	0.0334
0.1	0.05559	0.06359	0.06859

The results presented in Table 4.2 showcase the flexibility and effectiveness of the adaptive approach in estimating the true probability p by considering the interplay between group size k and the desired number of positive groups X. When X is fixed at 10 and k varies, it was observed that as the group size increased from 5 to 20, the MLE becomes larger, indicating a higher probability of success. Similarly, for a fixed $X = 20$, it was observed that the MLE of p increases as k increases. For example, for $p = 0.01$, the estimated values range from 0.00486 for $k = 5$ to 0.00526 for $k = 20$. This trend suggests that larger group sizes provide more information and lead to more accurate estimates of the true probability. In addition, the impact of X on the estimates was noted. As X increases while keeping k fixed, the MLE of p tends to increase as well. For example, for $p = 0{,}05$ and $k = 20$, the estimated values range from 0.01692 for $X = 10$ to 0.02992 for $X = 30$. These findings demonstrated that aiming for a higher number of positive groups leads to more precise estimates of the true probability. These

numerical values illustrate the interplay between group size, desired number of positive groups, and the estimated probability in the adaptive approach. By considering the appropriate combination of group size and desired positive groups, the adaptive approach enables accurate estimation of the true probability.

The relationship was further investigated by plotting the values of $\hat{p}$ against p while varying the waiting parameter X for different values of group size k.

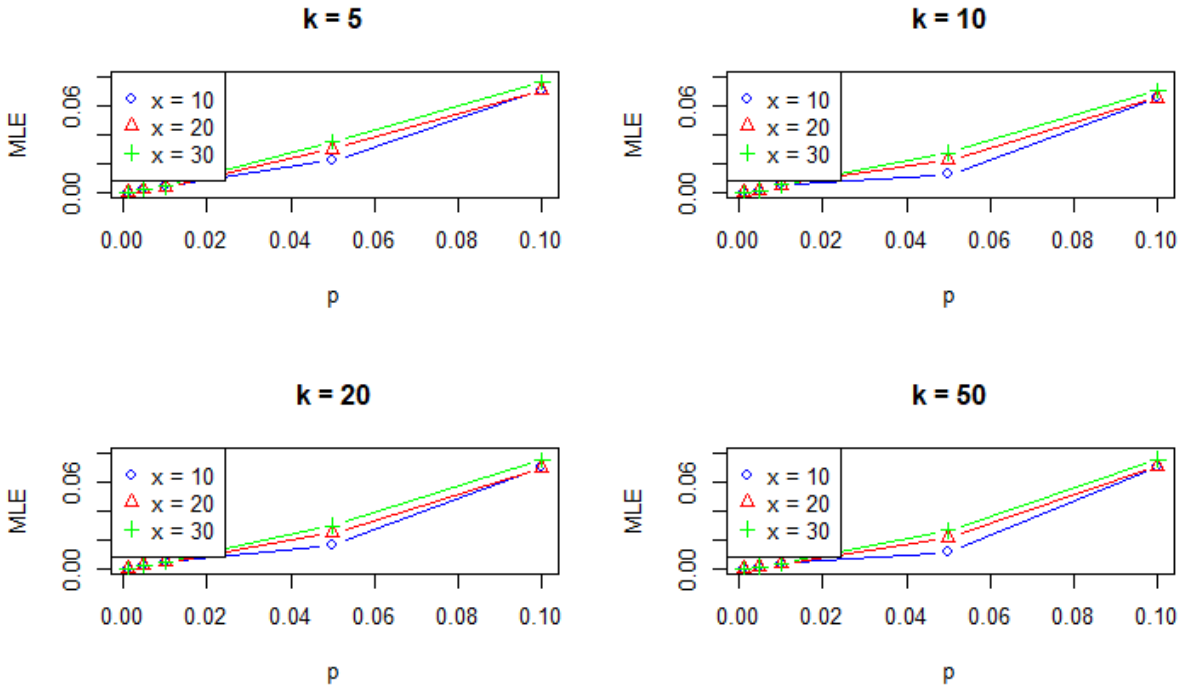

Figure 4. 3: Plots for Adaptive $\hat{p}$ versus p for $k = 5,20,50,100$ and $X = 10,20,30$

Figure 4.3 shows a monotonic increase in MLE as p increases as well as k and X. However, the rate and pattern of this increase vary depending on the specific combination of X and k. For the scenario with X=30, the MLE graph shows a relatively steep curve, indicating a significant change in the MLE values as p varies. This suggests that small changes in p lead to noticeable changes in the estimated probabilities of success. The MLE values obtained using the adaptive approach are generally lower compared to the non-adaptive MLE graph for this scenario, indicating a more conservative estimation. In the case of X=20, the MLE values also increase as p increases. However, the curve is relatively flat, indicating a less pronounced change in the MLE values as p varies. The adaptive MLE in this scenario exhibits a slower rate of increase compared to when X=30, implying a less sensitive response to changes in p. Similar observations can be made for the other combinations of x and k. The adaptive approach tends

to provide more conservative estimates with low MLE values. The MLE values increase with increasing p, but the specific patterns and sensitivities depend on the combination of x and k. The adaptive approach consistently provides more conservative estimates with lower MLE values.

4.4.2 Repeated Sampling

To evaluate the consistency and dependability of the estimates, we also conducted multiple repetitions of the adaptive procedure each utilizing different subsets of the data. Trace plots and histograms of the number of tests required before obtaining the required positive groups and their corresponding MLE of p were constructed. Trace plots are helpful in assessing the variability, stability, and reliability of estimates obtained through an adaptive procedure. They provide a visual representation of the parameter values at each iteration. By examining the patterns and movements of the trace plots, you can gain insights into the convergence, stationarity, and autocorrelation of the estimates. In this regard, the iterations were made for 1000 simulations utilizing the algorithm in figure 2.

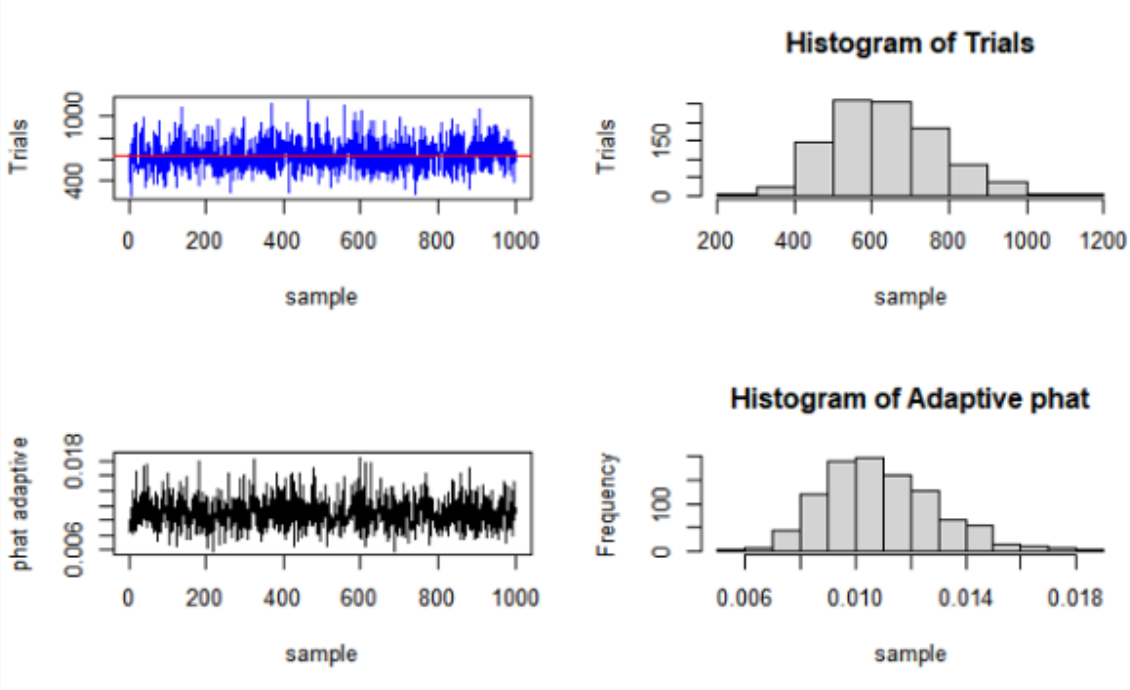

Figure 4. 4: Trace plot and Histogram for adaptive Trials and MLE of p for 1000 Monte Carlo simulations

The trace plot shown in Figure 4.4 indicates that both the trials and maximum likelihood estimates MLE exhibit stable and stationary patterns which suggests that the parameter estimates have converged. The trace plots consistently revolve around a single value of 600 for

the number of groups tested and 0.0003 for the corresponding estimator. This consistency demonstrates that the model has generated reliable estimates of the parameter value. Moreover, the trace plots demonstrate a minimal level of autocorrelation indicating that the model effectively explores the parameter space. The histogram portrays a normal distribution for the number of groups tested and their corresponding MLE indicating that they tend to cluster around a central value with only a few instances significantly deviating from this central value.

4.4.3 Relationship between Variance of $\hat{p}$ and p

In this section, we examine the relationship between the variance of the estimated proportion and the true proportion p in the context of the two-stage adaptive negative binomial model. The values presented in Table 4.3 provide insights into this relationship for different combinations of X and k.

Table 4. 3: Variance of $\hat{p}$ for $k = 5,10,20,50$ and $X = 20,30$

p	x=20	x=30
	k=5	
0.001	3.0598E-08	7.9418E-09
0.005	5.29521E-07	2.37975E-07
0.01	3.96308E-06	7.82165E-07
0.05	0.000112145	3.94449E-05
0.1	0.000325224	0.000166933
	k=10	
0.001	5.01655E-08	1.01251E-08
0.005	1.11251E-07	1.70198E-07
0.01	5.38436E-06	1.16218E-06
0.05	0.000104581	2.30733E-05
0.1	0.000384093	0.000141735
	k=20	
0.001	5.74673E-08	1.33924E-08
0.005	6.18607E-07	3.2789E-07
0.01	3.25448E-06	9.153E-07
0.05	9.59368E-05	2.82459E-05
0.1	0.000212807	0.000157982

	k=50	
0.001	3.76826E-08	1.14977E-08
0.005	5.34763E-07	1.61305E-07
0.01	2.84905E-06	5.83548E-07
0.05	7.05999E-05	2.26898E-05
0.1	0.000100331	0.000100095

Table 4.3 presents the interplay between X, k, and the variance of $\hat{p}$ in the two-stage adaptive negative binomial model. They highlight the impact of the true proportion, the desired number of positive groups X, and the group size k on the variability of the estimates. As the true proportion p increases, the variance of the estimated proportion $\hat{p}$ also tends to increase, although the magnitude of increase varies across different X and k values. This suggests that higher probabilities of success lead to greater variability in the estimates, highlighting the increased uncertainty associated with higher p values. Comparing different X values, we find that as X increases, the variance of $\hat{p}$ generally tends to increase. This implies that aiming for a higher number of positive groups introduces more variability into the estimates. Analyzing the effect of k, we notice that for a fixed X value, as k increases, the variance of $\hat{p}$ tends to decrease. This indicates that larger group sizes result in more precise estimates and lower variability. Larger group sizes provide more information, reducing the sampling error and enhancing the precision of the estimates. The insights obtained from this analysis can inform the selection of appropriate values for X and k to optimize the accuracy and reliability of the model in estimating the proportion of successes in a population.

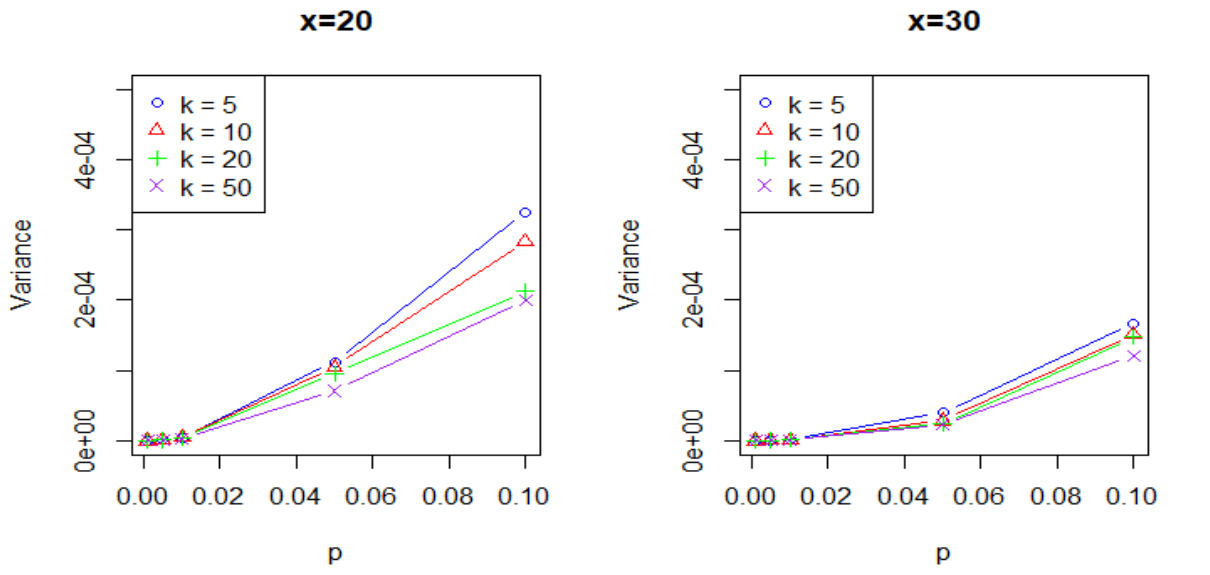

Figure 4. 5: Variance of $\hat{p}$ for $k = 5,10,20,50$ and $X = 20,30$

Figure 4.5 illustrate the variance for different values of k when X is fixed at either 20 or 30. Observing the graphs, several patterns emerge. Across all the tested group sizes, the adaptive estimator consistently demonstrates low variance as the probability increases. This trend is observed for $k = 5,20,50,100$. The adaptive estimator for larger group sizes consistently outperformed the estimator when small group sizes were used in terms of providing more precise estimates. It is worth noting that as the probability of success p increases, the variance tends to increase as well for levels predetermined number of positive groups indicating greater variability in the estimates. This can be attributed to the increased uncertainty associated with higher probabilities. Also, for a fixed X value, increasing k leads to a decrease in variance, suggesting that larger group sizes yield more precise estimates and lower variability. Larger group sizes provide more information, reducing sampling error. Comparing the graphs for $X = 20$ and $X = 30$, we observe that higher X values result in lower variance, indicating improved precision as the desired number of positive groups increases. These observations provide insights into the relationships between group size, desired positive groups, and variance, highlighting the trade-offs involved in selecting appropriate values for k and X in the two-stage adaptive Negative Binomial model.

4.4.4 Fixing k at Stage 1

In this section, we examine the variance of the adaptive estimator when the group size is fixed at stage one, instead of relying on the results from the non-adaptive model. This approach

is considered for practical purposes, and the results are presented in a table comparing the estimator variance in stage one and the adaptive estimator variance for different fixed group sizes $k = 5,15,25,35,45$ and the true population parameter set as $p = 0.001$ suggested by Pritchard and Tebbs (2010).

Table 4. 4: Variance for $\hat{p}$ in stage one and variance of the adaptive estimator for $k = 5,15,25,35,45$

Group size Fixed in stage 1	**Estimator Variance in stage 1**	**Adaptive Estimator Variance**
5	2.65E-08	7.92E-09
15	3.16E-08	7.94E-09
25	4.20E-08	1.17E-08
35	5.29E-08	1.14E-08
45	6.29E-08	7.94E-09

Scrutinizing table 4.4, it was observed that observe the estimator variance in stage one remains relatively consistent across different fixed group sizes. This indicates that when the group size is fixed at stage one, the estimator variance remains stable, irrespective of the actual group size. On the other hand, the adaptive estimator variance shows some variability. As the fixed group size increases from $k = 5$ to $k = 45$, the adaptive estimator variance tends to increase. This suggests that larger fixed group sizes introduce more variability into the estimates obtained from the adaptive model. These results highlight the impact of fixed group sizes in stage one on the estimator variance in adaptive models. The findings suggest that while the estimator variance remains consistent in stage one, the adaptive model offers the potential for improved precision, especially for larger fixed group sizes. These insights can guide the selection of the appropriate approach and group size in practice, depending on the desired level of accuracy and the constraints of the research or sampling process.

The relationship between variance and group size was further invested as presented in Figure 4.6 below.

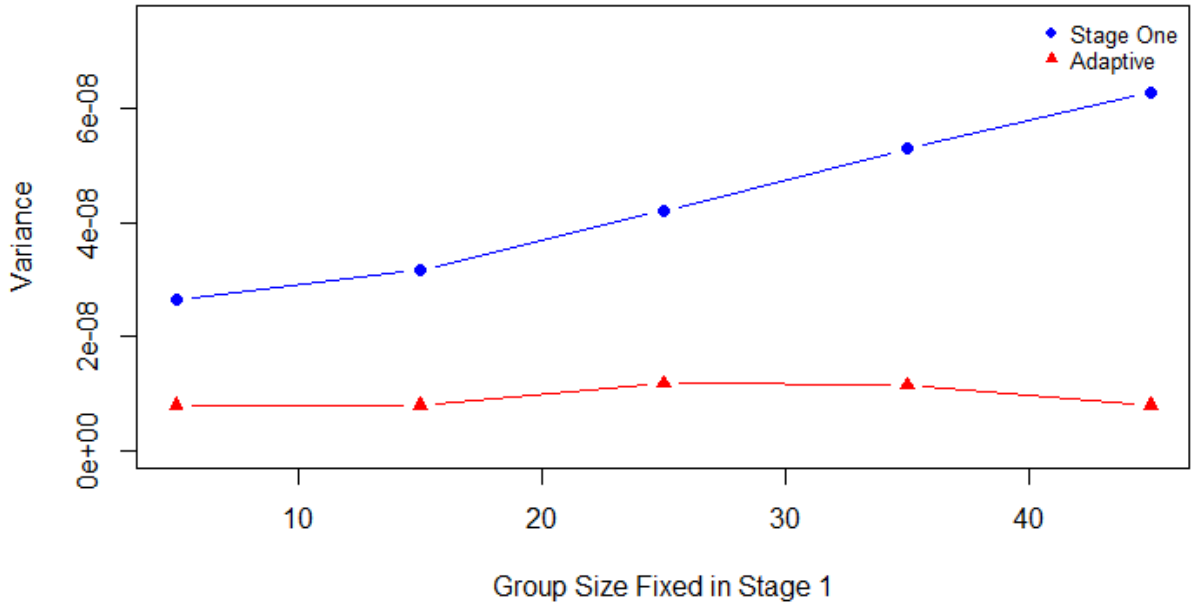

Figure 4. 6: Plots variance against group size for stage 1 and Stage 2

For different fixed group sizes in stage one, the variance of the adaptive estimator consistently outperforms the estimator variance in stage one. This indicates that even with a fixed group size, the adaptive model provides better estimation accuracy compared to relying solely on the results from the non-adaptive model. The difference in variance between the adaptive estimator and the variance in stage one varies depending on the fixed group size. For smaller group sizes such as 5 and 15, the adaptive estimator variance is significantly lower than the estimator variance in stage one indicating a substantial improvement in estimation accuracy. As the fixed group size increases $k = 25,35,45$, the gap between the adaptive estimator variance and the estimator variance in stage one becomes bigger. This indicates that the improvement achieved by the adaptive approach increases as the group size increases. From a practical perspective, fixing the group size in stage one can simplify the implementation of the adaptive model, as it reduces the need for dynamically adjusting the group size based on previous test results from the non-adaptive Negative Binomial group testing model. This trade-off between simplicity and accuracy can be advantageous in certain real-world scenarios where computational resources or time constraints may limit the feasibility of a fully adaptive approach.

4.4.5 Varying Desired Number of Positive Groups in Stage 1 and Stage 2

The adaptive estimator was achieved by obtaining the desired positive groups in stages and adjusting the group size from one stage to the next. For instance, in the first stage, X_1 positive groups each of size k_1 were desired while in the second stage, X_2 positive groups each with size k_2 which was dependent on the results of the first stage were desired, and so forth. This approach is very flexible and allows the number X_i of groups for each stage to be predetermine before the experiment begins, while the group sizes k_i are determined progressively during the experiment. However, the key challenge is to determine how to choose the predetermined number of positive groups required in each stage out of the total number of desired positive groups in a way that leads to using a group size as close to optimal as possible. The results in this section therefore presents the adaptive MLE of p and its variances while varying the number of desired positive groups in the first stage for $p = 0.001$.

Table 4. 5: MLE of p and its variance for varying X_1

X_1	X_2	Mle	Variance
3	27	0.003431105	4.17E-06
6	24	0.000991447	1.65E-07
9	21	0.001055406	1.25E-07
12	18	0.001471974	1.82E-07
15	15	0.001202044	2.68E-07
18	12	0.000765758	3.27E-08
21	9	0.001045769	5.23E-08
24	6	0.001088112	5.95E-08
27	3	0.00126278	6.93E-08

Table 4.5 presents the maximum likelihood estimator MLE of p and its variance while varying the number of desired positive groups in the first stage, represented as X_1. As the number of positive groups in the first stage X_1 increases, the MLE of p tends to decrease. For example, when X_1 is 3, the MLE of p is 0.003431105, whereas for X_1 of 27, the MLE of p decreases to 0.00126278. The variance of the MLE of p also varies with the number of desired positive groups in the first stage. Generally, higher values of X_1 correspond to lower variances,

indicating a greater precision in the estimates obtained. These results highlighted the impact of the desired number of positive groups in stage 1 on the MLE of p and its variance. They demonstrate the trade-off between the number of positive groups and the precision of the estimates obtained, emphasizing the importance of carefully selecting the desired number of positive groups in each stage to optimize the performance of the adaptive estimator.

4.5 Confidence Intervals of p

The analysis of confidence intervals provides valuable insights into the precision and accuracy of parameter estimation in our study. When utilizing Equation 61 for computing the confidence interval, it is assumed that the estimator of p follows an approximately normal distribution. However, when certain combinations of k and p are unsuitable, the group testing estimator lacks symmetry. In such cases, alternative confidence intervals can be derived from the simulated distribution of the estimator of p, assuming that the observed p is indeed the true population parameter. The algorithm for the computation of 95% Confidence intervals is as illustrated in the appendices. The study compared the confidence intervals obtained from the non-adaptive and adaptive estimators across different group sizes for various values of p. The widths of the confidence intervals as well as the lower and upper bounds were examined to assess the performance of the estimators.

Table 4. 6: The 95% CI for different values of p with $k = 5,10,20,50,100$ and $X = 30$

	Non Adaptive			**Adaptive**		
P	**Lower CI**	**Upper CI**	**Width**	**Lower CI**	**Upper CI**	**Width**
			k=5			
0.001	0.000614	0.001301	0.000687	0.000313	0.000663	0.00035
0.005	0.002547	0.005405	0.002858	0.001719	0.003635	0.001916
0.01	0.006923	0.014743	0.00782	0.003125	0.006599	0.003474
0.05	0.035114	0.076711	0.041597	0.023098	0.047768	0.02467
0.1	0.057117	0.127955	0.070837	0.05042	0.101171	0.050751
			k=10			
0.001	0.000786	0.001666	0.00088	0.000354	0.000749	0.000395
0.005	0.003683	0.007826	0.004143	0.001453	0.003074	0.00162
0.01	0.008038	0.017152	0.009115	0.003816	0.008051	0.004235
0.05	0.033346	0.073515	0.04017	0.018374	0.037242	0.018868

0.1	0.058203	0.135185	0.076982	0.04696	0.093724	0.046764
			k=20			
0.001	0.000841	0.001782	0.000942	0.000407	0.000861	0.000455
0.005	0.002747	0.005836	0.003089	0.002019	0.004268	0.002249
0.01	0.006244	0.01333	0.007086	0.003384	0.007142	0.003758
0.05	0.030339	0.068813	0.038474	0.019484	0.04036	0.020876
0.1	0.061052	0.197846	0.136794	0.05022	0.099591	0.049371
			k=50			
0.001	0.00068	0.001443	0.000763	0.000377	0.000798	0.000421
0.005	0.002542	0.005415	0.002872	0.001415	0.002993	0.001578
0.01	0.005753	0.012383	0.00663	0.002698	0.005699	0.003001
0.05	0.021668	0.054673	0.033005	0.017456	0.036167	0.018711
0.1	0.027078	0.105671	0.078592	0.050171	0.100703	0.050532
			k=100			
0.001	0.000557	0.001182	0.000625	0.000313	0.000663	0.00035
0.005	0.003468	0.007484	0.004016	0.001607	0.003399	0.001791
0.01	0.006271	0.013965	0.007694	0.002776	0.005865	0.003088
0.05	0.014115	0.053399	0.039284	0.021796	0.04501	0.023214
0.1	0.014115	0.053399	0.039284	0.045629	0.091554	0.045925

The width of the confidence intervals for the adaptive estimator tends to be smaller than the width for the non-adaptive estimator. This indicates that the adaptive estimator achieves higher precision in estimating the parameter compared to the non-adaptive estimator. In most cases, both the lower and upper bounds of the confidence intervals for the adaptive estimator are closer to the true parameter value compared to the non-adaptive estimator. This suggested that the adaptive approach results in more accurate estimation. The width of the confidence intervals generally increases as the value of p increases. This is expected since a higher value of p introduces more variability and uncertainty into the estimation process. The width of the confidence intervals varies with the group size. In some cases, increasing the group size leads to narrower confidence intervals, indicating improved precision.

4.6 Model Comparison

Model comparison is essential aspects of evaluating the performance of estimation procedures in statistical analysis. In the context of this study of two-stage adaptive negative binomial group testing model for estimating prevalence of a rare trait, it is crucial to assess the accuracy and precision of estimator of the developed model in comparison with the other existing model's estimators. This section focuses on comparing the non-adaptive and adaptive estimators in terms of their ability to estimate the prevalence rate with higher accuracy. The goal is to determine whether the adaptive procedure provides more reliable estimates compared to the non-adaptive procedure. The comparison was performed for different combinations of group size and probability.

4.6.1 Asymptotic Relative Efficiency

Asymptotic Relative Efficiency is a statistical measure used to compare the efficiency of two estimation methods. It quantifies how much more efficient one method is compared to another when the sample size approaches infinity. In this study, the ARE values represent the efficiency of an adaptive model compared to a non-adaptive model for different combinations of p and k. Higher ARE values indicate that the adaptive model is more efficient and provides better estimates or inferences compared to the non-adaptive model.

Table 4. 7: ARE of Two-Stage Adaptive model relative to Katholi and Unnasch (2006) model for $k = 5,10,20,50,100$ and $X = 20,30$

P			**X=20**		
	k=5	**k=10**	**k=20**	**k=50**	**k=100**
0.001	2.979932	3.479932	4.129932	4.657932	4.913432
0.005	2.634808	3.134808	3.784808	4.312808	4.768308
0.01	2.140186	2.640186	3.290186	3.818186	4.018567
0.05	1.012238	1.262238	1.587238	1.851238	2.034456
0.1	0.963791	1.213791	1.538791	1.802791	1.964505
			X=30		
0.001	4.546932	5.046932	5.476932	6.462932	6.437432
0.005	4.201808	4.701808	5.131808	6.117808	6.292308
0.01	3.707186	4.207186	4.637186	5.623186	5.542567
0.05	2.579238	2.829238	2.934238	3.656238	3.558456

0.1	2.530791	2.780791	2.885791	3.607791	3.488505

Table 4.7 showed that as X increased, there was an overall increasing trend in the asymptotic relative efficiency values across all k values, indicating improved performance in detecting the desired outcome. For example, when $X = 20$ and $k = 5$, ARE value was 2.979932, whereas when $X = 30$ and $k = 5$, ARE value increased to 4.546932. This demonstrated that aiming for a higher number of positive groups enhances the model's ability to detect the desired outcome. On the other hand, increasing k for a fixed X value led to a decreasing trend in the ARE values. For instance, when $X = 20$ and $k = 5$, ARE value was 2.979932, but when $X = 20$ and $k = 100$, ARE value decreased to 4.91343. This indicates that larger group sizes result in better performance in detecting the desired outcome, possibly due to the availability of more information and reduced variability in the estimates. The graphs also highlight a trade-off between X and k. When X is low and k is high, the ARE values tend to be higher, suggesting a higher probability of detecting the desired outcome within a smaller number of groups. Conversely, as X increases, achieving better performance requires larger group sizes. These observations underscore the importance of considering both X and k in optimizing the model's performance and emphasize the trade-off between the two factors.

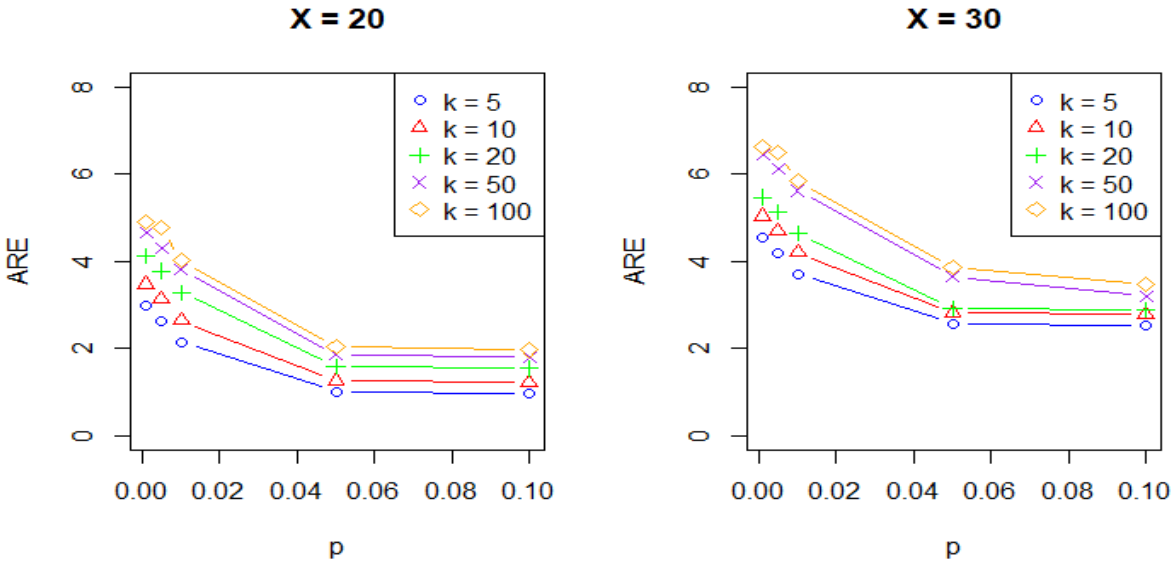

Figure 4. 7: Plots of ARE versus p for $k = 5,1,20,50$ and $X = 20,30$

Figure 4.7 showed that ARE values vary depending on the combination of p with k. Different combinations of p and k result in different efficiencies of the adaptive model compared to the non-adaptive model. The adaptive model tends to have higher ARE values for smaller values of p and larger values of k. This implies that the adaptive model is more efficient in detecting and signaling deviations from the expected process when the probability of an event is low and a larger group size. This confirms the establishment made by Hughes and Swallow (1994) that adaptive group testing procedures improves the efficiency of the estimator.

In addition to analyzing the variances of the non-adaptive and adaptive estimators for different values of p and k, the study also investigated the Asymptotic Relative Efficiency when the group size is fixed at stage one. The purpose of this analysis was to assess the practical efficiency of the adaptive estimator compared to relying solely on the estimator variance in stage one.

Table 4. 8: ARE of the Two-Stage Adaptive model relative to stage 1 estimator for $k = 5,15,25,35,45$, $X = 30$ and $p = 0.001$

Group size Fixed in stage 1	**Asymptotic Relative Efficiency**
5	3.339379
15	3.980165
25	3.588312
35	4.629699
45	7.922492

Table 4.8 showcases the Asymptotic Relative Efficiency ARE values for the adaptive estimator when the group size is fixed at stage one. Analyzing these results reveals several key observations. It was worth noting that the adaptive estimator consistently outperforms the estimator in stage one for all fixed group sizes. This is evident from the lower variance exhibited by the adaptive estimator due to ARE values greater than one, indicating improved estimation accuracy. The dynamic adjustment of the group size in the adaptive approach contributes to this enhanced performance. Higher ARE values signify a greater improvement achieved by the adaptive model, reinforcing its superiority in estimation efficiency. The ARE values range from approximately 3.34 to 7.92, depending on the fixed group size. This indicates a substantial improvement in efficiency when employing the adaptive estimator compared. The

adaptive approach's ability to dynamically adjust the group size contributes to reducing variance and enhancing estimation efficiency. The highest ARE value is attained when the group size is fixed at 45, reaching a value of 7.92. This signifies that the adaptive estimator with a larger fixed group size in stage one exhibits the highest relative efficiency compared to the estimator in stage one. A larger fixed group size facilitates more informative results, leading to improved estimation accuracy.

These results highlight the significant improvement in efficiency achieved by employing the adaptive estimator, even with a fixed group size in stage one. The adaptive model's capability to dynamically adjust the group size based on intermediate results effectively reduces variance and enhances estimation accuracy. These findings underscore the practical benefits of the adaptive approach in group testing scenarios, where efficiency and accuracy are crucial considerations.

4.6.2 Relative Mean Squared Error

The Relative Mean Square Error (RMSE) is a useful metric for comparing the mean squared errors (MSE) of different estimates for the same p obtained through various experimental designs or procedures. The study compared the estimator obtained from the constructed model with that of Katholi and Unnasch (2006). The RMSE is calculated as the ratio of the MSE of the estimator obtained using the method obtained by Katholi and Unnasch (2006) to that obtained using the constructed model, as shown in equation 22. Table 4.9 presents the RMSE values calculated for various values of p , k and X .

Table 4. 9: RMSE of the Two-Stage Adaptive model relative to Katholi's Model for k=20,50 and X=20,30

P	**X=30**		**X=20**	
	k=20	**k=50**	**k=20**	**k=50**
0.001	34.61273	36.66294	30.2668307	32.31705
0.005	16.14118	18.19139	11.7952825	13.8455
0.01	8.599877	10.65009	4.2539797	6.304194
0.05	2.9402	4.990415	1.4834701	2.244517
0.1	0.172221	2.222436	0.11658673	1.625706

Table 4.9 show the values of relative mean squared error RMSE for different combinations of p, group size k and the number of desired positive groups X. Comparing the values for X=30 and X=20, it could be seen that the values of RMSE decrease as X decreases. This suggests that as the number of desired positive groups decreases, the estimated probability of success also decreases. It indicates that achieving a higher number of positive groups leads to higher estimated probabilities. On the other hand, comparing the values for $k = 20$ and $k = 50$, it was observed that the values of RMSE tend to increase as k increases. The results of the mean squared error MSE show that the non-adaptive Negative Binomial model has a much higher MSE compared to the adaptive model in all cases except for p = 0.1 in certain instances. This means that the adaptive model has a better fit to the data and is more accurate in predicting the outcomes.

This result was further illustrated by plotting the values of relative mean squared error against different values of p while varying the number of predetermined desired positive groups and the group sizes as illustrated in Figure 4.8.

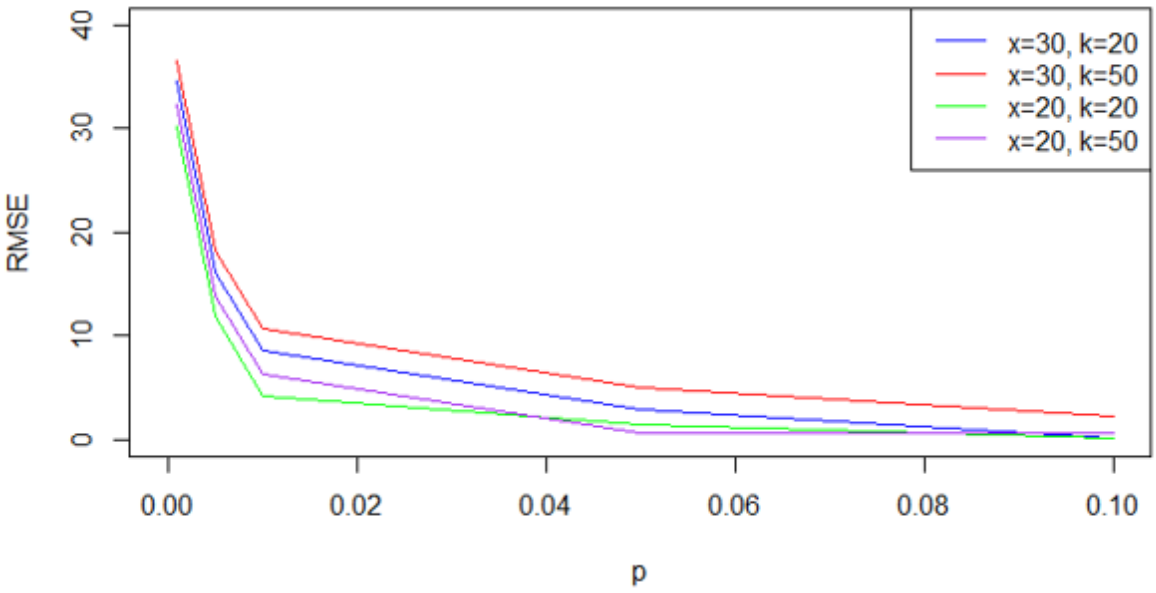

Figure 4. 8: Plots of RMSE versus p for $k = 20{,}50$ and $X = 20{,}30$

The plots in Figure 4.8 provide valuable insights into the relationship between the relative mean squared error RMSE and the corresponding values of p for various combinations of X and k. One notable observation is that as the parameter p increases, the RMSE exhibits a decreasing trend. This implies that as the true probability of success increases, the accuracy of the model in predicting the outcomes improves. The initial high values of RMSE suggest higher

discrepancies between the predicted outcomes and the observed data. However, as p increases, the RMSE gradually decreases, indicating a better fit to the data and improved prediction accuracy.

The smooth and consistent decrease in the RMSE line further supports the negative relationship between p and the relative mean squared error. This suggests that as the true probability of success increases, the model's predictions become more accurate and closer to the actual outcomes. This finding has important implications for practical applications. It indicates that as the likelihood of achieving positive outcomes increases, the model can provide more reliable predictions, enabling better decision-making and more effective resource allocation.

Additionally, the plots reveal a monotonic decrease in the RMSE as the number of predetermined desired positive groups decreases. This indicates that aiming for a lower number of positive groups leads to more accurate predictions and lower discrepancies between the model's estimates and the observed data. This finding highlights the trade-off between the desired number of positive groups and prediction accuracy. By setting a lower number of positive groups as the target, the model can focus its resources on a narrower scope, resulting in more precise estimates and improved accuracy.

Moreover, the plots demonstrate that increasing the group size k also contributes to a decrease in the RMSE. Larger group sizes provide more data and information, leading to more precise estimates and lower prediction errors. This suggests that increasing the group size allows for better sampling and a more comprehensive understanding of the underlying distribution, enhancing the model's accuracy.

4.7 Application of the Model to Real Data

Rutledge *et al*. (2003) conducted a study on West Nile Virus (WNV) surveillance in public health. The study was conducted shortly after a WNV outbreak in Jefferson County, Florida in 2001, with the aim of estimating the prevalence of WNV in the mosquito population. The researchers collected 11,948 individual mosquitoes, formed pools of mosquitoes and then tested them using reverse transcription polymerase chain reaction assays to identify WNV positive pools. They found 14 WNV positive pools, with variable pool sizes being used in the study, but most of the pools were smaller than a specified size. This study was the first field study of WNV vectored by the North American mosquito, Culex nigripalpus Theobald, which is also believed to be a vector for St. Louis and eastern equine encephalitis viruses. WNV had

been detected in Florida in early 2001, and later that year, 12 human cases of WNV meningoencephalitis had been confirmed.

Rutledge *et al.* (2003) did not utilize inverse sampling, but their study is used as a model to demonstrate our methodology. The authors presented only the results from pools that tested positive and reported an estimated prevalence of around 0.005, which may be an upper bound since some negative pools were disregarded. We assume that 0.005 represents the true value of p and generate datasets for waiting parameters of x (no. of positive groups desired) = 5, 10, 15 and 20, using equal pool sizes at stage one which is then adjusted for stage two to obtain the adaptive estimator. The authors used variable pool sizes mostly between 15 and 50 mosquitos, hence we consider $k = 15,25,35,45$ at stage 1 of the model to obtain the adaptive estimator, its variance and Wald confidence intervals.

Table 4. 10: MLE of p for $k = 15,25,35,45$ and $X = 5,10,15,20$

K	MLE	Var	95% Lower CI	95% Upper CI	Width
			X= 5		
5	0.003341214	2.20E-06	0.00035	0.00037	0.00002
15	0.003758223	2.80E-06	0.00038	0.00042	0.00004
25	0.003046124	1.85E-06	0.00031	0.00544	0.00513
35	0.002445127	1.19E-06	0.0053	0.00638	0.00108
45	0.002538891	1.29E-06	0.00581	0.00474	-0.00107
			X = 10		
5	0.002386064	5.61E-07	0.00092	0.00108	0.00016
15	0.002818678	7.87E-07	0.00085	0.00098	0.00013
25	0.002221587	4.91E-07	0.00091	0.00386	0.00295
35	0.002572355	6.58E-07	0.0036	0.00456	0.00096
45	0.002384321	5.66E-07	0.00386	0.00417	0.00031
			X = 15		
5	0.002708607	4.79E-07	0.00135	0.00165	0.0003
15	0.003316924	7.23E-07	0.00124	0.0018	0.00056
25	0.003634372	8.69E-07	0.00134	0.00407	0.00273
35	0.002507602	4.16E-07	0.00499	0.00546	0.00047
45	0.002711335	4.86E-07	0.00377	0.00408	0.00031

		X= 20			
5	0.002274368	2.55E-07	0.00116	0.00128	0.00012
15	0.002069764	2.12E-07	0.00161	0.0017	9E-05
25	0.002861757	4.05E-07	0.00133	0.00327	0.00194
35	0.003015513	4.50E-07	0.00297	0.00411	0.00114
45	0.002359582	2.76E-07	0.00339	0.00433	0.00094

The provided results show the estimated prevalence of West Nile Virus (WNV) in the mosquito population based on the study conducted by Rutledge *et al*. (2003). The results are presented for different waiting parameters x and pool sizes *k* at stage one of the model. The estimated prevalence of WNV is provided, along with the variance of the estimator and the lower and upper confidence intervals CIs as well as the width. The estimated prevalence values vary depending on the waiting parameter and pool size. The estimated prevalence tends to decrease as the waiting parameter increases, indicating that with more positive groups desired, the estimated prevalence decreases. The estimates are also influenced by the pool sizes used in the study. As the pool size k increases from 5 to 45 for all waiting parameters, estimated prevalence tends to decrease indicating that larger pool sizes lead to lower estimated prevalence. Variance also decreases indicating lower variability in the estimates. On the other hand, Confidence intervals become narrower suggesting more precise estimates with reduced uncertainty.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter summarizes the findings drawn from the analyses and offers recommendations for future research endeavors

5.2 Conclusions

In conclusion, this study achieved its objectives and generated valuable insights into the two-stage adaptive negative binomial model in group testing for estimating prevalence of a rare trait using maximum likelihood estimation MLE. The following are the conclusions drawn from the study;

i. The study successfully derived an estimator for the prevalence of a rare trait by employing the two-stage adaptive Negative Binomial model in group testing using Maximum Likelihood Estimation. This achievement provides a robust method for estimating the prevalence of rare traits in populations, enhancing our ability to accurately assess the occurrence of such traits.
ii. The properties of the derived estimator, including bias, variance, and Mean Squared Error, were thoroughly analyzed. The results indicated favorable characteristics of the estimator, with relatively low bias and variance, suggesting its reliability and precision in estimating the prevalence of rare traits.
iii. Through simulation studies, the two-stage adaptive negative binomial group testing model was compared to the existing non-adaptive group testing model. The findings revealed that the adaptive model outperformed the non-adaptive model in terms of efficiency, as it detected more positive cases and yielded smaller estimates. This comparison reveals the superiority of the adaptive approach and highlights its potential for improving the accuracy and effectiveness of group testing methods.

5.3 Recommendations

One limitation of assuming perfect tests in the study is that it does not reflect the real-world scenario where tests may have imperfections, such as false positives and false negatives. By assuming perfect tests, the study overlooks the potential impact of these imperfections on the performance of the adaptive estimator and the accuracy of prevalence estimation in group testing. Besides, in case of dealing with large population, it might be limiting to only consider the two stage approach. In this regard, the study gives the following recommendations;

i. The policy makers should consider adopting the two-stage adaptive negative binomial group testing model over the existing non-adaptive model in situations where efficiency and accuracy are crucial, such as disease surveillance and screening programs. The superior performance of the adaptive model, as demonstrated by the findings suggests that it can significantly enhance the effectiveness of group testing approaches.
ii. Future studies can investigate how the presence of imperfect tests affects the performance of adaptive estimators and the accuracy of prevalence estimation in group testing. Models or algorithms can be developed to account for the sensitivity and specificity of the tests and evaluate their impact on the estimation process. Strategies to mitigate the effects of imperfect tests, such as adjusting decision rules or incorporating additional information, can also be explored to improve the accuracy of the estimators.
iii. The adaptive estimator can be constructed for multiple stages group testing procedures. This study primarily focused on a two-stage Negative Binomial group testing approach. However, real-world scenarios often demand more than two stages to achieve accurate estimates especially when dealing with larger populations. Investigating the performance of adaptive estimators in multi-stage group testing settings presents an opportunity to refine estimation techniques further.

REFERENCES

Berger, T., Mandell, J. W., & Subrahmanya, P. (2000). Maximally efficient two-stage screening. *Biometrics, 56(3),* 833-840. https://doi.org/10.1111/j.0006 341x.2000.00833.x

Bhattacharyya, G. K., Karandinos, M. G., & Defoliart, G. R. (1979). Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies1. *American Journal of Epidemiology*, *109*(2), 124-131. https://doi.org/10.1093/oxfordjournals.aje.a112667

Bilder, C. R., Tebbs, J. M., & Chen, P. (2010). Informative Retesting. *Journal of the American Statistical Association*, *105*(491), 942-955. https://doi.org/10.1198/jasa.2010.ap09231

Black, M. S., Bilder, C. R., & Tebbs, J. M. (2011). Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *61*(2), 277-290. https://doi.org/10.1111/j.1467-9876.2011.01008.x

Busch, M. P., Caglioti, S., Robertson, E. F., McAuley, J. D., Tobler, L. H., Kamel, H., Linnen, J. M., Shyamala, V., Tomasulo, P., & Kleinman, S. H. (2005). Screening the blood supply for west Nile virus RNA by nucleic acid amplification testing. *New England Journal of Medicine*, *353*(5), 460-467. https://doi.org/10.1056/nejmoa044029

Cardoso, M. S., Koerner, K., & Kubanek, B. (1998). Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: Preliminary results. *Transfusion*, *38*(10), 905-907. https://doi.org/10.1046/j.1537-2995.1998.381098440853.x

Chiang, C. L., & Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations1. *American Journal of Epidemiology*, *75*(3), 377-391. https://doi.org/10.1093/oxfordjournals.aje.a120259

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, *14*(4), 436-440. https://doi.org/10.1214/aoms/1177731363

Fang, X., Stroup, W. W., & Zhang, S. (2007). Improved empirical Bayes estimation in group testing procedure for small proportions. *Communications in Statistics - Theory and Methods*, *36*(16), 2937-2944. https://doi.org/10.1080/03610920701386935

George, V. T., & Elston, R. C. (1993). Confidence limits based on the first occurrence of an event. *Statistics in Medicine*, *12*(7), 685-690. https://doi.org/10.1002/sim.4780120707

Gibbs, A. J., & Gower, J. C. (1960). The use of a multiple-transfer method in plant virus transmission studies—Some statistical points arising in the analysis of results. *Annals of Applied Biology, 48*(1), 75-83. https://doi.org/10.1111/j.1744-7348.1960.tb03506.x

Guner, R., Hasanoglu, I., & Aktas, F. (2021). Evaluating the efficiency of public policy measures against COVID-19. *Turkish Journal of Medical Sciences, 51*(11), 3229-3237. https://doi.org/10.3906/sag-2106-301

Hepworth, G. (2005). Confidence intervals for proportions estimated by group testing with groups of unequal size. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*(4), 478-497. https://doi.org/10.1198/108571105x81698

Hepworth, G. (2013). Improved estimation of proportions using inverse binomial group testing. *Journal of Agricultural, Biological, and Environmental Statistics*, *18*(1), 102-119. https://doi.org/10.1007/s13253-012-0126-6

Hepworth, G., & Watson, R. (2008). Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *58*(1), 105-121. https://doi.org/10.1111/j.1467-9876.2008.00639.x

Hughes-Oliver, J. M. (2006). Pooling experiments for blood screening and drug discovery. *Screening, 30*(2),48-68. https://doi.org/10.1007/0-387-28014-6_3

Hughes-Oliver, J. M., & Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association, 89*(427), 982-993. https://doi.org/10.1080/01621459.1994.10476832

Kariuki, F. M., Wanyonyi, R. W., & Islam, A. S. (2023). Analysis of a two-stage negative binomial group testing model for estimating the prevalence of a rare trait. *OALib, 10*(06), 1-20. https://doi.org/10.4236/oalib.1110141

Katholi, C. R., & Unnasch, T. R. (2006). Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches. *The American Journal of Tropical Medicine and Hygiene, 74*(5), 779-785. https://doi.org/10.4269/ajtmh.2006.74.779

Kennedy, N. L. (2004). Testing for the presence of disease by pooling samples. *Australian & New Zealand Journal of Statistics*, *46*(3), 383-390. https://doi.org/10.1111/j.1467-842x.2004.00337.x

Kennedy, N. L. (2011). Dual estimation of prevalence and disease incidence in pool-testing strategy. *Communications in Statistics - Theory and Methods*, *40*(18), 3218-3229.

https://doi.org/10.1080/03610926.2010.493257

Kerr, J. D. (1971). The probability of disease transmission. *Biometrics*, *27*(1), 219 - 234. https://doi.org/10.2307/2528941

Kim, H., Hudgens, M. G., Dreyfuss, J. M., Westreich, D. J., & Pilcher, C. D. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, *63*(4), 1152-1163. https://doi.org/10.1111/j.1541-0420.2007.00817.x

Kim, H., & Hudgens, M. G. (2009). Three-dimensional array-based group testing algorithms. *Biometrics*, *65*(3), 903-910. https://doi.org/10.1111/j.1541-0420.2008.01158.x

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger, S., & Quinn, T. C. (1989). Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, *27*(7), 1449-1452. https://doi.org/10.1128/jcm.27.7.1449-1452.1989

Lewis, J. L., Lockary, V. M., & Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for chlamydia trachomatis and Neisseria gonorrhoeae. *Sexually Transmitted Diseases*, *39*(1), 46-48. https://doi.org/10.1097/olq.0b013e318231cd4a

Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., & Moncada, J. (2005). Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and Neisseria gonorrhoeae in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology*, *43*(4), 1674-1677. https://doi.org/10.1128/jcm.43.4.1674-1677.2005

Litvak, E., Tu, X. M., & Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, *89*(426), 424-434. https://doi.org/10.1080/01621459.1994.10476764

Matiri, G. (2017). Sequentially selecting between two experiments for optimal estimation of a trait with misclassification. *American Journal of Theoretical and Applied Statistics*, *6*(2), 79 – 87. https://doi.org/10.11648/j.ajtas.20170602.12

McMahan, C. S., Tebbs, J. M., & Bilder, C. R. (2011). Informative Dorfman screening. *Biometrics*, *68*(1), 287-296. https://doi.org/10.1111/j.1541-0420.2011.01644.x

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., & Eskridge, K. (2013). Sample size for detecting transgenic plants using inverse binomial group testing with dilution effect. *Seed Science Research*, *23*(4), 279-288. https://doi.org/10.1017/s0960258513000238

Mundel, A. B. (1984). Group testing. *Journal of Quality Technology*, *16*(4), 181-188. https://doi.org/10.1080/00224065.1984.11978916

Okoth, A. W., Nyongesa, L. K., & Kwatch B. O (2017). Multi-stage adaptive pool testing model with test errors; Improved efficiency. *IOSR Journal of Mathematics*, *13*(01), 43-55. https://doi.org/10.9790/5728-1301024355

Pilcher, C. D., Fiscus, S. A., Nguyen, T. Q., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J. O., McPherson, J. T., Stalzer, B., Hightow, L., Miller, W. C., Eron, J. J., Cohen, M. S., & Leone, P. A. (2005). Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine*, *352*(18), 1873-1883. https://doi.org/10.1056/nejmoa042291

Pritchard, N. A., & Tebbs, J. M. (2010). Estimating disease prevalence using inverse binomial pooled testing. *Journal of Agricultural, Biological, and Environmental Statistics*, *16*(1), 70-87. https://doi.org/10.1007/s13253-010-0036-4

Pritchard, N. A., & Tebbs, J. M. (2011). Bayesian inference for disease prevalence using negative binomial group testing. *Biometrical Journal*, *53*(1), 40-56. https://doi.org/10.1002/bimj.201000148

Rutledge, C. R., Day, J. F., Lord, C. C., Stark, L. M., & Tabachnick, W. J. (2003). West Nile virus infection rates in Culex nigripalpus do not reflect transmission rates in Florida. *Journal of Medical Entomology*, *40*(3), 253-258. https://doi.org/10.1603/0022-2585-40.3.253

Sobel, M., & Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika*, *62*(1), 181-193. https://doi.org/10.1093/biomet/62.1.181

Sobel, M., & Groll, P. A. (1966). Binomial group-testing with an unknown proportion of defectives. *Technometrics*, *8*(4), 631. https://doi.org/10.2307/1266636

Swallow, W. H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology*, *75*(8), 882. https://doi.org/10.1094/phyto-75-882

Swallow, W. H. (1987). Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission. *Phytopathology*, *77*(10), 1376. https://doi.org/10.1094/phyto-77-1376

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, *18*(4), 568-578. https://doi.org/10.2307/2527902

Turechek, W. W., & Madden, L. V. (2003). A generalized linear modeling approach for characterizing disease incidence in a spatial hierarchy. *Phytopathology®*, *93*(4), 458-466. https://doi.org/10.1094/phyto.2003.93.4.458

Van, T. T., Miller, J., Warshauer, D. M., Reisdorf, E., Jernigan, D., Humes, R., & Shult, P. A. (2012). Pooling Nasopharyngeal/Throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology*, *50*(3), 891-896. https://doi.org/10.1128/jcm.05631-11

Wanyonyi, R. W. (2015). Estimation of proportion of a trait by batch testing model in a quality control process. *American Journal of Theoretical and Applied Statistics*, *4*(6), 619-629. https://doi.org/10.11648/j.ajtas.20150406.34

Wanyonyi, R. W., Mwangi, O. W., & Mwangi, C. W. (2021). Re-testing in batch testing model based on quality control process for proportion estimation. *Open Journal of Statistics*, *11*(01), 123-136. https://doi.org/10.4236/ojs.2021.111007

Xie, M., Tatsuoka, K., Sacks, J., & Young, S. S. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association*, *96*(453), 92-102. https://doi.org/10.1198/016214501750333009

Xiong, W. (2015). The optimal group size using inverse binomial group testing considering misclassification. *Communications in Statistics - Theory and Methods*, *45*(15), 4600-4610. https://doi.org/10.1080/03610926.2014.923461

APPENDICES

Appendix I: Publication

Open Access Library Journal
2023, Volume 10, e10960
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Analysis of a Two-Stage Adaptive Negative Binomial Group Testing Model for Estimating Prevalence of a Rare Trait

Jackline Akomboh, Wanyonyi Ronald Waliaula, Cox Tamba, Justine Obwoge Okenye

Department of Mathematics, Egerton University, Egerton, Kenya
Email: akombojacky@gmail.com, ronaldwaliaula@gmail.com, cox.tamba@egerton.ac.ke, justinoke69@gmail.com

How to cite this paper: Akomboh, J., Waliaula, W.R., Tamba, C. and Okenye, J.O. (2023) Analysis of a Two-Stage Adaptive Negative Binomial Group Testing Model for Estimating Prevalence of a Rare Trait. *Open Access Library Journal*, 10: e10960.
https://doi.org/10.4236/oalib.1110960

Received: November 3, 2023
Accepted: December 24, 2023
Published: December 27, 2023

Abstract

Group testing is an efficient method for classifying observations and estimating trait prevalence in a population. However, using appropriate group sizes is crucial for maximizing its benefits. Adaptive schemes have been developed to address improper group size selection issues. Existing adaptive schemes are based on a Binomial sampling model, requiring testing of all groups before recording successes. In certain scenarios, like infectious diseases, immediate reporting of estimates upon detection is necessary. A two-stage adaptive Negative Binomial group testing model for such cases was constructed. This adaptive model adjusted group sizes based on estimates from previous stages thus using optimal sizes to minimize the mean squared error and variance of the prevalence rate estimate. The maximum likelihood estimation method was employed to find the model's parameter estimate, and its properties were also investigated. The comparative analysis highlighted the superiority of the adaptive model over the non-adaptive model especially under low prevalence emphasizing the importance of incorporating adaptivity in group testing procedures, particularly in disease screening and surveillance, such as for COVID-19.

Appendix II: Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: 822073

Date of Issue: 31/July/2023

RESEARCH LICENSE

This is to Certify that Ms.. Jackline Ongecha Akomboh of Egerton University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nakuru on the topic: TWO-STAGE ADAPTIVE NEGATIVE BINOMIAL GROUP TESTING MODEL FOR ESTIMATING PREVALENCE OF A RARE TRAIT for the period ending : 31/July/2024.

License No: **NACOSTI/P/23/27893**

822073

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix III: R-codes

R code for Simulation of data from a Negative binomial pooling scheme

```
NEGBIN=function(K,X,p){
# Simulation of the Negative binomial pooling scheme # p the prevalence of the trait
# k the pool size
# X the predetermined number of positive pools
# TT number of pools required to obtain X positive pools tt=1
dd=matrix(rbinom(K,1,p),ncol = K) if (sum(dd)==0){x=0}else{x=1}
while(x<X){ dd1=matrix(rbinom(K,1,p),ncol = K) dd=rbind(dd,dd1) if
    (sum(dd1)==0){x=x}else{x=x+1} tt=tt+1
} return(list(dd=dd,tt=tt))
}
```

R codes for MLE Notations

X1 - Number of predetermined desired positive groups in stage 1

X2 - Number of predetermined desired positive groups in stage 1

K1 – Group size in stage 1

K2 – Group size in stage 2

T1 - Total number of groups tested in stage 1 before obtaining the desired positive groups T2 – Total number of groups tested in stage 2 before obtaining the desired positive groups p – The true population parameter for the prevalence

Var - Variance

Sd – Standard deviation Sqrt - Squareroot

CI – Confidence Interval

62

```
fun<- function(p){(x1 / (1 - (1 - p)^k1) * (k1 * (1 - p)^(k1 - 1)) - ((t1 - x1) / (1 - (1 - p)^k1)) *
    (k1 * (1 - p)^(k1 - 1)) + x2 / (1 - (1 - p)^k2) * (k2 * (1 - p)^(k2 - 1)) - ((t2 - x2) / (1 - (1
    - p)^k2)) * (k2 * (1 - p)^(k2 - 1)))}
phatadaptive<-uniroot(fun,c(0,1)) phatadaptive$root
```

Bias and MSE

```
ptrue
phatadaptive<-rep(1:length(T2))
for (i in 1:length(T2)) { t2<-T2[i]
```

```
fun<- function(p) {(x1 / (1 - (1 - p)^k1) * (k1 * (1 - p)^(k1 - 1)) -((t1 - x1) / (1 - (1 - p)^k1)) *
    (k1 * (1 - p)^(k1 - 1)) + x2 / (1 - (1 - p)^k2) * (k2 * (1 - p)^(k2 - 1)) -((t2 - x2) / (1 - (1
    - p)^k2)) * (k2 * (1 - p)^(k2 - 1)))}
phatadaptive[i]<-uniroot(fun,c(0,1))$root}
phatadaptive
exp<-((sum(phatadaptive))/1000) bias<-(exp-p.true)
bias
MSE<-mean((phatadaptive-p.true)^2) MSE
```

Confidence Interval

```
p<-phatadaptive$root
var <- (1/((x_1 * k_1^2 * (1-p)^(2*k_1-2))/(1-(1-p)^(k_1))^2 + (x_2 * k_2^2 * (1-p)^(2*k_2-
    2))/(1-(1-p)^(k_2))^2))
var
sd<-sqrt(var) sd
Wald CI=c(phatadaptive$root-1.964*sd,phatadaptive$root+1.964*sd) Wald CI
```