Estimating Proportions by Group Retesting with Unequal Group Sizes at Each Stage

ESTIMATING PROPORTIONS BY GROUP RETESTING WITH UNEQUAL GROUP SIZES AT EACH STAGE

BY

YUSANG HU, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS AND

THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2020) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE: Estimating Proportions by Group Retesting with Unequal Group Sizes at Each Stage

AUTHOR: Yusang Hu

B.Sc. (University of Toronto)

SUPERVISOR: Dr. Stephen Walter

NUMBER OF PAGES: ix, 68

To my dear parents and my love

Abstract

Group testing is a procedure that splits samples into multiple groups based on some specific grouping criterion and then tests each group. It is usually used in identifying affected individuals or estimating the population proportion of affected individuals. Improving precision of group testing and saving cost of experiment are two crucial tasks for investigators. Cost-efficiency is a ratio of precision to cost; hence improving cost-efficiency is as crucial as improvement of precision and cost saving. In this thesis, retesting will be considered as a method to improve precision and cost-efficiency, and save cost. Retesting is an extension of group testing. It uses two or more group testing stages, and testing original samples in all of the stages. Hepworth and Watson (2015) proposed a two-stage group testing procedure where two stages have equal group sizes, and the number of groups of the second stage is based on the number of positive groups in the first stage. In this thesis, our main goal is estimating a proportion p under the circumstance of unequal group sizes in two stages, and discovering the most cost-efficient experiment design. Analytical solutions of precision will be provided; we will use these analytical solutions with simulations to analyse some experimental designs, and discover whether doing one group testing only is precise enough or not and if it is worth retesting for each design. In the end, we will combine all these analyses and identify the optimal experiment design.

Acknowledgement

First of all, I would like to show my deepest gratitude to my supervisor, Dr. Stephen Walter, for his patient and persistent guidance throughout my Master program. He not only provided me with a new perspective of knowledge, but also encouraged me to persist when facing difficulties. The goal of this thesis would not have been accomplished without his help.

Second, I would like to extend my thanks to Dr. Graham Hepworth, for his valuable advice and full support for this thesis.

Furthermore, I would like to thank Dr. Shui Feng and Dr. Ben Bolker for being members of my defence committee. I appreciate their time for reviewing my thesis and providing valuable feedback.

Last but not least, I would especially like to thank my dear parents and Mr. Yexin Cui, for their continued support and encouragement throughout this entire process.

Notation

n_1	Number of groups at the first stage
n_2	Number of groups at the second stage
k_1	Number of group sizes at the first stage
k_2	Number of group sizes at the second stage
a_1	Correction at the first stage
a_2	Correction at the second stage
p	True population proportion of affected individuals
π	True population proportion of affected individuals at the second stage
X	Number of positive groups at the first stage
Y	Number of positive groups at the second stage
SE	Standard Error
RSE	Relative Standard Error
RE	Relative Cost-Efficiency

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Group Testing

Group testing (or pooled testing) was first introduced by Dorfman in 1943 to screen U.S. soldiers for syphilis during World War II. It is a procedure to identify affected (positive) individuals or estimate the true proportion of affected individuals in a sample population. An affected or positive individual means a research object is affected by the research of interest. For example, if the research of interest is studying infectious disease, then an individual is defined as affected or positive if he/she has such infectious disease, and an individual is defined as negative if he/she does not have such disease. The true proportion of affected individuals means the percentage of individuals who were affected by the research variable of interest in the sampled population. Group testing is a powerful theory that has broad applications in a great many areas. Definitely, it has helped a lot with blood screening since the original intention of group testing is solving blood testing problems, for example, detecting phenylketonuria, hepatitis B virus and other diseases (Guthrie, 1961; Comanor and Holland, 2006; Bilder, Tebbs and Chen, 2010). Group testing has also been applied to many other fields; for instance, solving some network security problems such as denial-of-service and jamming attacks (Thai 2012; Xuan et al. 2010); encoding the transform coefficients of an image from the wavelet packet and the discrete cosine

transform (Hong, Ladner and Riskin 2003); designing an algorithms for random multiple-access communication channels (Berger et al. 1984); DNA library screening (Ngo and Du, 2000; Schliep, Torney and Rahmann, 2003); and screening individuals for drug use (Gastwirth and Johnson, 1994).

Generally speaking, the statistical study of group testing can be broadly classified into two categories: classification and estimation. Classification will be discussed in Chapter 4. Estimation is the main goal of this thesis; it targets estimating a proportion p of positive individuals, such as evaluating the prevalence of disease of interest in a population (Sobel and Elashoff, 1975; Chen and Swallow, 1990). Gastwirth and Hammick (1989) proposed group testing to estimate the prevalence of AIDS antibodies in blood donors; they used estimation rather than identification because they wanted to protect individuals' civil liberties. Walter, Hildreth and Beaty (1980) used group testing of unequal group sizes within a stage to estimate the infection rates of yellow fever virus in a mosquito population.

1.2 Retesting

When estimating a true proportion, there exists uncertainty about how well an estimate represents the true population. To conceptualize this uncertainty, we can consider how an estimate changes if we repeat the experiment many times, with different samples each time. The closeness of an estimate between different samples is called precision. The precision is closely related to the variance. Suppose μ represents the average value of an estimate in different samples; then the variance

measures the closeness of this estimate to μ. In group testing, investigators will split the whole sample into n groups first, each group having k individuals. If there exist one or more affected individuals in a group, then it will be defined as a positive group. Then investigators will estimate the population proportion of affected individuals by using the total number of positive groups. Nevertheless, the estimate of population proportion is sometimes not precise enough. Therefore, a method that is able to improve precision is needed. Retesting within positive groups and testing additional individuals are two common methods to do this. Retesting is a method that extracts all the positive groups at the first stage, randomly regroups all their individuals, and then retests new groups in the second stage. Rather than testing original samples, testing additional individuals requires to collect more new samples. In statistical studies, testing additional individuals is more popular than retesting because retesting usually gains less precision than testing additional individuals. Nevertheless, in addition to consider precision, promoting cost-efficiency is also a crucial task for investigators. Cost-efficiency is defined as a ratio of precision to cost. With respect to cost-efficiency, testing additional individuals costs more, on the other hand, it is sometimes impractical in an experiment. Instead, retesting is cost-saving and therefore it might have better cost-efficiency than testing additional individuals. When the number of positive groups is 0 or 1 at the first stage, there is no need to go to the second stage, therefore we need not do retesting, and the precision and cost-efficiency will not change; if the number of positive groups is more than 1 at the first stage, then retesting increases precision. In this thesis, we will focus on studying

if it is worth retesting at the second stage based on each sample population's cost-efficiency, precision, and cost. Walter and Hepworth (2019) derived two analytical solutions for the variance of the estimate of true population proportion of affected individuals at the second stage, then compared the two methods by doing simulations to determine the optimal one; Hepworth and Watson (2015) proposed two two-stage procedures and compared each of them with other two methods which are proposed by Hammick and Gastwirth (1994) and Brookmeyer (1999) to determine the most efficient method of retesting.

The other focus of this thesis is studying the effect of adding a correction or not when we estimating the population proportion. The correction was proposed by Burrow (1987) and denoted as a. Its function is eliminating bias and decreasing mean squared error when estimating the population proportion.

1.3 Thesis Structure

In chapter 2, we will first estimate the population proportion without any corrections at first and second stage separately; we will then add a correction to the estimate to compare the estimates with and without correction; next, we will present the estimation of population proportion of affected individuals using overall two testing stages. Also, in order to evaluate whether doing the first stage only is precise enough or not and if it is worth retesting, the variance of the estimate of population proportion at the first stage only and overall two stages will be estimated. In chapter 3, a simulation based on the estimates in chapter 2 will be performed. Then, the simulation

results such as cost-efficiency, precision and bias will be compared using two alternative cost functions. The most cost-efficient combination will be found and some recommendations will be offered to investigators. Some extension of group testing, future work and challenges will be concluded in chapter 4.

Chapter 2 Estimation

2.1 Estimation of proportion of affected individuals in the first and second stage

Suppose there are n_1 groups and each group with k_1 group sizes in the first stage. After a series of experiments, the results show that there are $X = x$ positive groups. Assume p is the true probability of an individual being affected and q is the true probability that an individual is not affected; then X follows a binomial distribution with parameters n_1 and $g(p)$, where $g(p) = 1-(1-p)^{k_1} = 1-q^{k_1}$ and q^{k_1} is the probability that none of k_1 individual is infected. Then the expectation of X is

$$E(X) = n_1 \cdot g(p).$$

Define $E(X) = n_1 \cdot \left[1-(1-p)^{k_1}\right] = n_1 \cdot (1-q^{k_1})$, and let $\hat{p}_1$ be the estimator of p in the first stage. Then the estimator of p at the first stage can be expressed as

$$X = n_1 \cdot (1-\hat{q}^{k_1})$$

$$\Rightarrow \hat{q} = [1-\frac{X}{n_1}]^{\frac{1}{k_1}}$$

$$\Rightarrow \hat{p}_1 = \hat{p} = 1-[1-\frac{X}{n_1}]^{\frac{1}{k_1}} \quad (1)$$

Suppose there are k_2 individuals in each group in the second stage where k_2 is equal to or smaller than k_1. Now, according to the information given above, there

will be $n_2 = \frac{Xk_1}{k_2}$ groups at the second stage. The true prevalence at the second stage is no longer p since it depends on the results of the first stage, therefore we define a new prevalence of the second stage as π, which can be expressed as

$$\text{prevalence at the second stage} = \frac{\text{total number of positive individuals}}{\text{total number of individuals sampled}}$$

$$\Rightarrow \hat{\pi} = \frac{n_1 k_1 \hat{p}}{Xk_1} = \frac{n_1 \hat{p}}{X} \tag{2}$$

After a series of experiments, the result shows that there are $Y = y$ positive groups, where Y follows a binomial distribution conditional on $X = x$ with parameter n_2 and $g(\pi) = 1-(1-\pi)^{k_2}$. Define $E(Y) = n_2 \cdot g(\pi) = n_2 \cdot \left[1-(1-\pi)^{k_2}\right]$, and let $\hat{p}_2$ be the estimator of p in the second stage. Then the estimator of p in the second stage can be expressed as

$$Y = \frac{Xk_1}{k_2} \cdot \left[1-(1-\hat{\pi})^{k_2}\right]$$

$$\Rightarrow \hat{\pi} = 1-(1-\frac{Y}{Xk_1 / k_2})^{\frac{1}{k_2}}$$

$$\Rightarrow \hat{p}_2 = \frac{X}{n_1}\left[1-(1-\frac{Y}{Xk_1 / k_2})^{\frac{1}{k_2}}\right] \tag{3}$$

Burrows (1987) proposed an alternative estimator $\tilde{p}$ to improve the estimator's properties. $\tilde{p}$ has similar steps of calculation with $\hat{p}$ but with correction a where $a = \frac{1}{2}(\frac{k-1}{k})$ to eliminate bias and decrease mean squared error. Set the correction at first stage and second stage as a_1 and a_2 separately, where $a_1 = \frac{1}{2}(\frac{k_1-1}{k_1})$ and

$a_2 = \frac{1}{2}(\frac{k_2 - 1}{k_2})$. Now, the number of positive groups at the first stage X and the number of positive groups at the second stage Y given $X = x$ still follow binomial distribution, but modified maximum likelihood estimate (MLE) $\tilde{p}_1$ becomes $g(\tilde{p}_1) = \frac{X}{n_1 + a_1}$ and $\tilde{\pi}$ becomes $g(\tilde{\pi}) = \frac{Y}{n_2 + a_2}$. Therefore, the alternative estimator $\tilde{p}_1$ and $\tilde{p}_2$ can be expressed as

$$1 - (1 - \tilde{p}_1)^{k_1} = \frac{X}{n_1 + a_1}$$

$$\Rightarrow \tilde{p}_1 = 1 - (1 - \frac{X}{n_1 + a_1})^{\frac{1}{k_1}} \tag{4}$$

$$1 - (1 - \tilde{\pi})^{k_2} = \frac{Y}{Xk_1 / k_2 + a_2}$$

$$\Rightarrow \tilde{\pi} = 1 - (1 - \frac{Y}{Xk_1 / k_2 + a_2})^{\frac{1}{k_2}}$$

$$\Rightarrow \tilde{p}_2 = \frac{X}{n_1}\left[1 - (1 - \frac{Y}{Xk_1 / k_2 + a_2})^{\frac{1}{k_2}}\right] \tag{5}$$

Although equations (1) and (4), (3) and (5) look very similar, and a_1, a_2 do not exceed 0.5, correction strongly influences the results. Some results may look abnormal if we do not add correction in the equation of estimators. In the next chapter, we will compare differences between estimators with and without correction by using real data.

2.2 Estimation of proportion of affected individuals using overall two testing stages

We will use a weight function to evaluate the estimator of p for both first stage and second stage combined. Define the estimator of overall p (i.e. for both first stage and second stage combined) as $\tilde{p}_{1+2}$. The weight of the estimate of p at the first (second) stage is defined as an inverse proportion to its variance, which is able to minimize the variance of $\tilde{p}_{1+2}$. By using the property of weight function, the estimator of overall p can be expressed as

$$\tilde{p}_{1+2} = \frac{w_1\tilde{p}_1 + w_2\tilde{p}_2}{w_1 + w_2} \tag{6}$$

where $w_1 = \frac{1}{Var(\hat{p}_1)}$ and $w_2 = \frac{1}{Var(\hat{p}_2)}$. After simplifying the expression,

$$\tilde{p}_{1+2} = \frac{\tilde{p}_1 Var(\hat{p}_2) + \tilde{p}_2 Var(\hat{p}_1)}{Var(\hat{p}_1) + Var(\hat{p}_2)}.$$

Note that the derivation of $Var(\tilde{p}_2)$ is very complicated since the equation of $\tilde{p}_2$ contains correction, hence we will use $Var(\hat{p}_2)$ instead, and therefore standardize by using $Var(\hat{p}_1)$ instead of $Var(\tilde{p}_1)$.

By using equation (6), variance of estimator of overall p can be expressed as

$$Var(\tilde{p}_{1+2}) = \left(\frac{1}{w_1 + w_2}\right)^2 \cdot \left[w_1^2 Var(\tilde{p}_1) + w_2^2 Var(\tilde{p}_2) + 2Cov(\tilde{p}_1, \tilde{p}_2)\right]$$

Note that covariance of $\tilde{p}_1$ and $\tilde{p}_2$ is unknown here since its derivation is complicated, therefore in chapter 3 we will use simulation to obtain the variance of $\tilde{p}_{1+2}$ instead of attempting to derive an analytic expression for it.

2.3 Variance of estimators

Following the process of the variance estimation by Hepworth and Walter (2019). In section 2.1, we have assumed that the number of positive groups at the first stage X follows a binomial distribution with parameter n_1 and $g(p)$. Therefore variance of X can be expressed as

$$\text{var}(X) = n_1 g(p_1)(1 - g(p_1)) \tag{7}$$

The modified maximum likelihood estimate (MLE) $\tilde{p}_1$ is $g(\tilde{p}_1) = \frac{X}{n_1 + a_1}$. Now, we can express the variance of $g(\tilde{p}_1)$ in two ways. The first is obtained based on the property of the derivative

$$Var(g(\tilde{p}_1)) = g'(p_1)^2 Var(\tilde{p}_1) \tag{8}$$

The second is obtained based on the MLE $\tilde{p}_1$

$$Var(g(\tilde{p}_1)) = \frac{1}{(n_1 + a_1)^2} Var(X) \xrightarrow{\text{using eq.(7)}} \frac{1}{(n_1 + a_1)^2} n_1 g(p_1)(1 - g(p_1)) \tag{9}$$

Therefore equation (8) and (9) are equal, and new equation can be expressed as

$$g'(p_1)^2 Var(\tilde{p}_1) = \frac{1}{(n_1 + a_1)^2} n_1 g(p_1)(1 - g(p_1)) \tag{10}$$

Next, in order to get the variance of modified MLE $\tilde{p}_1$, we need to plug $g(p_1) = 1 - (1 - p_1)^{k_1}$ and $g'(p_1) = k_1(1 - p_1)^{k_1 - 1}$ into equation (10) and simplify it,

$$\left[k_1(1 - p_1)^{k_1 - 1}\right]^2 Var(\tilde{p}_1) = \frac{1}{(n_1 + a_1)^2} n_1 \left[1 - (1 - p_1)^{k_1}\right](1 - p_1)^{k_1}$$

$$\Rightarrow Var(\tilde{p}_1) = \frac{1 - (1 - p_1)^{k_1}}{(1 + \frac{a_1}{n_1})^2 k_1^{\,2} n_1 (1 - p_1)^{k_1 - 2}} \tag{11}$$

The variance of the modified MLE $\tilde{p}_1$ is asymptotically equal to the variance of

MLE $\hat{p}_1$

$$Var(\hat{p}_1) = \frac{1-(1-p_1)^{k_1}}{k_1{}^2 n_1 (1-p_1)^{k_1-2}}$$

The variance of the modified MLE $\tilde{p}_2$ is calculated very similarly with $Var(\tilde{p}_1)$,

$$g'(\pi)Var(\tilde{\pi}) = \frac{1}{(n_2+a_2)^2} n_1 g(\pi)(1-g(\pi))$$

$$\Rightarrow Var(\tilde{\pi}) = \frac{1-(1-\pi)^{k_2}}{(1-\pi)^{k_2-2} X k_1 k_2 (1+\frac{a_2}{Xk_1 / k_2})^2}$$

$$\Rightarrow Var(\tilde{p}_2) = \frac{X}{n_1{}^2}\left[\frac{1-(1-\pi)^{k_2}}{(1-\pi)^{k_2-2} k_1 k_2 (1+\frac{a_2}{Xk_1 / k_2})^2}\right] \qquad (12)$$

The variance of $\hat{p}_2$ can be obtained from two approaches. The first approach is assuming that the estimator of p at the second stage is conditional on the number of positive groups at the first stage X. Note that if the number of positive groups at the first stage is zero, then there is no need to do retesting, therefore $\hat{p}_{1+2} = \hat{p}_1 = \hat{p}_2 = 0$ and $Var(\hat{p}_2 \mid X) = 0$; if the number of positive groups at the first stage is one, then doing retesting will be meaningless and wasting money. Therefore the estimator of overall p will be equal to the estimator of p in the first stage, and $Var(\hat{p}_2 \mid X) = 0$. Now, by using the equation (12), we can express the variance of MLE $\hat{p}_2$ conditional on X as

$$Var(\hat{p}_2 \mid X) = \begin{cases} \frac{X}{n_1{}^2}\left[\frac{1-(1-\pi)^{k_2}}{(1-\pi)^{k_2-2} k_1 k_2 (1+\frac{a_2}{Xk_1 / k_2})^2}\right] \approx \frac{X}{n_1{}^2}\left[\frac{1-(1-\pi)^{k_2}}{(1-\pi)^{k_2-2} k_1 k_2}\right], & \text{if } X > 1 \\ 0, \quad \text{if } X = 0 \text{ or } 1 \end{cases}.$$

Now consider $X>1$, if π is small, then we can using the Taylor's first order expansion to rewrite the expression of $Var(\hat{p}_2 \mid X)$ as

$$Var(\hat{p}_2 \mid X>1) \approx \frac{X}{n_1^{\,2}} \frac{\left[1-(1-k_2)\pi\right]}{k_1 k_2}\left[1+(k_2-2)\pi\right]$$

$$\Rightarrow Var(\hat{p}_2 \mid X>1) \approx \frac{X\pi}{n_1^{\,2} k_1}\left[1+(k_2-2)\pi\right] \tag{13}$$

In section 2.1, we have already known $\hat{\pi} = \frac{n_1 p}{X}$, therefore

$$Var(\hat{p}_2 \mid X>1) \approx \frac{p_1}{n_1 k_1}\left[1+(k_2-2)\frac{n_1 p_1}{X}\right] \tag{14}$$

Now, in order to obtain the variance of $\hat{p}_2$, we will apply the law of total variance

$$Var(\hat{p}_2) = E\left[Var(\hat{p}_2 \mid X)\right] + Var(E\left[\hat{p}_2 \mid X\right]) \tag{15}$$

By using equation (3) and the Taylor's first order expansion,

$$\begin{aligned} E\left[\hat{p}_2 \mid X\right] &= E\left[\frac{X}{n_1}\left[1-(1-\frac{Y}{Xk_1/k_2})^{\frac{1}{k_2}}\right]\right] \\ &\approx \frac{X}{n_1}\left(\frac{1}{k_2}\cdot\frac{E[Y]}{Xk_1/k_2}\right) \end{aligned} \tag{16}$$

Y given X follows a binomial distribution with parameter Xk_1/k_2 and $g(\pi)$, if π is small, then

$$E[Y] = \frac{Xk_1}{k_2}\left[1-(1-\pi)^{k_2}\right] \approx Xk_1\pi$$

$$\approx Xk_1\frac{n_1 p_1}{X}$$

Therefore, equation (16) can be simplified as

$$E[\hat{p}_2 \mid X] \approx \frac{X}{n_1}\cdot\frac{1}{k_2}\cdot\frac{Xk_1\, n_1 p_1/X}{Xk_1/k_2} \approx p_1$$

As a result, $Var(E[\hat{p}_2 \mid X]) \approx Var(p_1) \approx 0$. Hence, we only need to consider $E[Var(\hat{p}_2 \mid X)]$ in equation (15).

Now, based on equation (14) we can obtain that

$$E\left[Var(\hat{p}_2 \mid X)\right] \approx \frac{p_1}{n_1 k_1}\left[1 + (k_2 - 2) n_1 p_1 \cdot E\left[\frac{1}{X}\right]\right] \text{ if } X > 1 \tag{17}$$

Assume $\theta = g(p_1) = 1 - (1 - p_1)^{k_1}$. Johnson, Kotz and Kemp (1992) derived the approximate expectation for inverse of the number of positive groups at the first stage X when X is larger than zero,

$$E\left[\frac{1}{X} \mid X > 0\right] \approx \left(\frac{n_1 - 2}{n_1}\right)\left[(n_1 + 1)\theta - 1\right]^{-1} \tag{18}$$

Meanwhile, $E\left[\frac{1}{X} \mid X > 0\right]$ can be expressed as

$$\begin{aligned} E\left[\frac{1}{X} \mid X > 0\right] &= \frac{\sum_{X=1}^{n} \frac{1}{X} P[X = x]}{P[X > 0]} \\ &= \frac{P[X = 1] + \sum_{x=2}^{n} \frac{1}{X} P[X = x]}{P[X > 0]} \end{aligned} \tag{19}$$

However, we want to know $E\left[\frac{1}{X} \mid X > 1\right]$. By applying the property of conditional expectation and equation (19),

$$\begin{aligned} E\left[\frac{1}{X} \mid X > 1\right] &= \frac{\sum_{X=2}^{n} \frac{1}{X} P[X = x]}{P[X > 1]} \\ &= \frac{E\left[\frac{1}{X} \mid X > 0\right] P[X > 0] - P[X = 1]}{P[X > 1]} \end{aligned}$$

The number of positive groups at the first stage X follows a binomial distribution with parameter n and θ, therefore $P[X > 0] = 1 - (1 - \theta)^{n_1}$, $P[X = 1] = n_1 \theta (1 - \theta)^{n_1 - 1}$

and $P[X>1]=1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}$. Hence,

$$E\left[\frac{1}{X}\mid X>1\right]=\frac{\left(\frac{n_1-2}{n_1}\right)\left[(n_1+1)\theta-1\right]^{-1}\left[1-(1-\theta)^{n_1}\right]-n_1\theta(1-\theta)^{n_1-1}}{1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}} \qquad (20)$$

Now, plugging equation (20) into equation (17), we can obtain that

$$E[Var(\hat{p}_2\mid X>1)]\approx\frac{p_1}{n_1k_1}\left\{1+\frac{(k_2-2)n_1p_1\left[\frac{n_1-2}{n_1}\left[(n_1+1)\theta-1\right]^{-1}\left[1-(1-\theta)\right]^{n_1}-n_1\theta(1-\theta)^{n_1-1}\right]}{1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}}\right\}$$

By using equation (15), the variance of the estimator of p at the second stage can be expressed as

$$Var(\hat{p}_2)=E[Var(\hat{p}_2\mid X)]+0$$
$$\Rightarrow Var(\hat{p}_2)=E[Var(\hat{p}_2\mid X>1)]P(X>1)+E[Var(\hat{p}_2\mid X=0,1)]P[X=0,1]$$

Note that $Var(\hat{p}_2\mid X)=0$ when $X=0\text{ or }1$, hence the expectation of $Var(\hat{p}_2\mid X=0,1)$ will be zero. Therefore,

$$Var(\hat{p}_2)=\frac{p_1}{n_1k_1}\left\{\left[1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}\right]+(k_2-2)n_1p_1\left[\frac{n_1-2}{n_1}\left[(n_1+1)\theta-1\right]^{-1}\left[1-(1-\theta)\right]^{n_1}-n_1\theta(1-\theta)^{n_1-1}\right]\right\} \qquad (21)$$

The second approach assumes that the estimator of p at the second stage is conditional on the joint distribution of X and π. Suppose we have n_1 groups at the first stage, and denote $m_1,...,m_n$ as the number of positive individuals in each group. Assume group i is positive; then m_i will follow a positive binomial distribution with parameter k_1 and p_1 where $1\le m_i\le k_1$. Now, the number of positive individuals at the first stage changes from $n_1k_1p_1$ to $\sum_{m_i>0}m_i$. Hence, the estimated prevalence at the second stage will be

$$\hat{\pi} = \frac{\sum_{m_i>0} m_i}{Xk_1} \tag{22}$$

If the true probability at the first stage is small, then we expect $m_i = 1$, therefore $\hat{\pi} \approx \frac{X}{Xk_1} \approx \frac{1}{k_1}$. According to the approximate value of $\hat{\pi}$, assume that $E(\hat{\pi}) \approx \frac{1}{k_1}$ and $E(\hat{\pi}^2) \approx \frac{1}{k_1^2}$. By using equation (13), the variance of the estimator of p at the second stage when $X > 1$ can be expressed as

$$\begin{aligned} Var(\hat{p}_2 \mid X>1) &= E_\pi\left[Var(\hat{p}_2 \mid X>1, \pi)\right] \\ &= E_\pi\left[\frac{X\pi}{n_1^2 k_1}(1+(k_2-2)\pi)\right] \\ \Rightarrow Var(\hat{p}_2 \mid X>1) &= \frac{X}{n_1^2 k_1^2}(1+\frac{k_2-2}{k_1}) \end{aligned} \tag{23}$$

From equation (15), we have already known that we only need to consider $E\left[Var(\hat{p}_2 \mid X)\right]$, therefore we need to get $E\left[X \mid X>1\right]$. Knowing that the number of positive groups at the first stage follows a binomial distribution with parameter n_1 and θ, hence

$$\begin{aligned} E\left[X \mid X>0\right] &= \frac{n_1\theta}{1-(1-\theta)^{n_1}} \\ &= \frac{\sum_{X=1}^{n} XP\left[X=x\right]}{P\left[X>0\right]} \\ &= \frac{P\left[X=1\right]+\sum_{X=2}^{n} XP\left[X=x\right]}{P\left[X>0\right]} \end{aligned} \tag{24}$$

$$E\left[X \mid X>1\right] = \frac{\sum_{X=2}^{n} XP\left[X=x\right]}{P\left[X>1\right]} \tag{25}$$

Plugging equation (24) into equation (25), we can obtain that

$$E[X \mid X>1]=\frac{E[X \mid X>0]P[X>0]-P[X=1]}{P[X>1]}$$

$$\Rightarrow E[X \mid X>1]=\frac{\frac{n_1\theta}{1-(1-\theta)^{n_1}}\left[1-(1-\theta)^{n_1}\right]-n_1\theta(1-\theta)^{n_1-1}}{1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}}$$

$$\Rightarrow E[X \mid X>1]=\frac{n_1\theta\left[1-(1-\theta)^{n_1-1}\right]}{1-n_1\theta(1-\theta)^{n_1-1}-(1-\theta)^{n_1}} \tag{26}$$

If p_1 is small, then $\theta=1-(1-p_1)^{k_1}\approx k_1p_1$. Therefore, the expectation of equation (23) can be expressed as

$$E[Var(\hat{p}_2 \mid X>1)]=\frac{1+\frac{k_2-2}{k_1}}{n_1^{\,2}k_1^{\,2}}E[X \mid X>1]$$

$$\Rightarrow E[Var(\hat{p}_2 \mid X>1)]\approx\frac{p_1\left[1-(1-\theta)^{n_1-1}\right]}{n_1k_1P[X>1]}\left(1+\frac{k_2-2}{k_1}\right) \tag{27}$$

Hence, according to equation (15) and (27), the variance of the estimator of p at the second stage can be expressed as

$$Var(\hat{p}_2)=E[Var(\hat{p}_2 \mid X>1)]P[X>1]+E[Var(\hat{p}_2 \mid X=0,1)]P[X=0,1]+0$$

Note that $Var(\hat{p}_2 \mid X)=0$ when $X=0$ or 1, therefore

$$\Rightarrow Var(\hat{p}_2)\approx\frac{p_1}{n_1k_1}\left[1-(1-\theta)^{n_1-1}\right]\left(1+\frac{k_2-2}{k_1}\right) \tag{28}$$

Chapter 3 Simulation Results

3.1 Simulation

In the simulations, we will choose some typical values of the number of groups at the first stage (n_1), the number of individuals in each group at the first stage (k_1) and at the second stage (k_2), and the true prevalence of affected individuals p . The values of n_1 , k_1 and p are set as in Hepworth and Watson (2015), and Hepworth and Walter (2019). We will group each number of n_1 , k_1 and k_2 , and select five values of p which correspond to each group. Note that if all groups are positive or negative, then there is no need to do retesting. Therefore, in order to avoid such a situation as far as possible, Hepworth and Watson (2015) stated that the five values of p are selected in order that the probability of a positive group in a group testing is not 0 or 1. Note that the number of groups at the second stage $\frac{Xk_1}{k_2}$ might be non-integral, so we will round these numbers to the next highest integer (e.g. $X=3$, $k_1=6$, $k_2=12$, $\frac{Xk_1}{k_2}=\frac{18}{12}\approx 2$). Following is the data set:

$$n_1=(10,30,50,100)$$

$$k_1=(6,12,20,50,100)\ k_2=(6,12,20,50,100)$$

$$p\subset(0.001,0.002,0.005,0.01,0.02,0.05,0.1,0.2,0.3)$$

Generally speaking, there are supposed to be $4\times5\times5\times5=500$ combinations. But

we have to pay attention to the size of k_1 and k_2 when the group sizes are different in the two stages. According to equation (3) and (5), there is no doubt that $\frac{Y}{Xk_1/k_2}$ and $\frac{Y}{Xk_1/k_2+a_2}$ must be no greater than 1 if the group sizes are the same in the two stages; but in the case of different group sizes, $\frac{Y}{Xk_1/k_2}$ and $\frac{Y}{Xk_1/k_2+a_2}$ might be larger than 1 if k_2 is larger than k_1 (e.g. $n_1=10$, $k_1=6$, $X=3$, $k_2=12$, $Y=2$). Hence, we will drop all groups that have k_2 larger than k_1 in the simulation.

In the simulations, we will perform $N=100000$ runs on each combination. The total number of simulations N and combinations are set as in Hepworth and Walter (2019), but the estimation equations used and the simulation process are somewhat different since there are many special conditions if group sizes in the two stages are different. The same simulation process setting with Hepworth and Walter (2019) is that if the number of positive groups at the first stage X is 0 or 1, then the number of positive groups at the second stage will always be 0 or 1 which is expensive and meaningless, therefore there is no need to go to the second stage if $X=0\text{ or }1$ and first and second stage estimation will be same. Due to different group sizes in two stages, there are two main special conditions that are different with Hepworth and Walter (2019): first, k_2 needs to be smaller or equal to k_1; second, if the number of groups at the second stage $n_2=\frac{Xk_1}{k_2}$ is indivisible, then the result will be rounded to the next highest integer.

In the next few sections, we will find out one or more most cost-efficient

combinations by fixing some parameters and then comparing each combination's properties including relative cost-efficiency, precision, bias and so on.

3.2 Simulation Comparison

In statistics, simulation is a method that can generate random numbers based on models rather than collecting a real data set, and it is a fast tool to approximate the results of a true data set. Our first goal of the simulation is going to evaluate which analytical solution is better by comparing the variance derived analytically and the variance of the second stage obtained by simulation. The second goal of the simulation is to identify one or more most valuable combinations which are composed by n_1, k_1, k_2 and p. The 'value' of a combination can be evaluated in several ways, including cost-efficiency, precision, if it is worth retesting and so on.

3.2.1 Comparison of two analytical solutions with and without correction

To evaluate how correct the equation (21) and equation (28) are, we can calculate the ratio of the variances derived analytically (i.e. equation (21) and (28)) to the variance of the second stage obtained by simulation. Note that at the end of section 2.1, we mentioned that the equation with and without correction looks very similar but actually 'correction' can be very important. The denominator of the ratio is the variance of the second stage obtained by simulation, in other words, the denominator of the ratio is obtained from the variance of 100000 results of equation (3) (i.e. without correction) or equation (5) (i.e. with correction), where n_1, k_1 and k_2 are

chosen from the data set in section 3.1 and X, Y are determined by simulation. Let us first take a look at equation (3) and (5) in detail. The only difference between equation (3) and (5) is that equation (5) has a correction a_2 which equation (3) does not have. The correction a_2 can be expressed as

$$a_2 = \frac{1}{2}\left(\frac{k_2 - 1}{k_2}\right), \quad k_2 > 0$$

where a_2 must be smaller than 0.5 since $\frac{k_2 - 1}{k_2}$ is smaller than 1. It looks negligible since it is very small, for example, if we have a combination $n_1 = 10$, $k_1 = 12$, $k_2 = 6$, $X = 2$, $Y = 3$, then equation (3) will equal to 0.04126 and equation (5) will equal to 0.03453 which are very close. The effect of correction might not look very significant if we look at only one case, but if we have 100000 cases and compute the variance, the difference will become clearer. Table 1 indicates the ratio of the variance derived analytically (i.e. the first approach means using equation (21), the second approach means using equation (28)) to the variance of the second stage obtained by simulation, where upper value is derived by applying equation (5) (i.e. with correction) and lower value is derived by applying equation (3) (i.e. without correction) in simulation. The ratio needs to be as close to 1 as possible since it is a measurement of how different are the variances derived analytically and the variance of the second stage obtained by simulation. Hence, a ratio is said to be acceptable if it is close to 1.

Overall, the ratio derived by applying equation (5) (i.e. with correction) will be more recommended than the ratio derived by applying equation (3) (i.e. without

correction). Let us look at the results without a correction first. There are some obvious observations for the results without correction that can be found in Table 1: most ratios will be close to 1 (which is acceptable) if the values of k_1 and k_2 have large differences; when $k_1 = k_2$, the ratio is always acceptable if the value of p is in the middle (e.g. when $k_1 = k_2 = 100$, the middle value of p is 0.005; when $k_1 = k_2 = 50$, the middle value of p is 0.01) and $n_1 = 100$, the ratio is totally unacceptable if p is very small or large (e.g. when $k_1 = k_2 = 100$, $p = 0.001$ is very small, $p = 0.02$ is very large); when k_2 is approximately half of k_1, the ratio will be acceptable for the most of the time when n_1 is 30 or more, but not acceptable for the most of the time when $n_1 = 10$. Note that the values of k_1 and k_2 will be the same or k_2 is approximately half of k_1 for the case shown in Table 2.

	same					k_2 is approximately half of k_1			
k_1	100	50	20	12	6	100	50	20	12
k_2	100	50	20	12	6	50	20	12	6

Table 2. Combinations of same k_1 and k_2 / k_2 is approximately half of k_1

From Table 1, we can observe that most ratios that without correction (i.e. lower values) are acceptable and close to the ratio with correction when k_1 and k_2 have large differences, sometimes the ratio without correction is more acceptable than the ratio with correction, sometimes not. Nevertheless, some ratios without correction are totally unacceptable when the values of k_1 and k_2 are the same or k_2 is approximately half of k_1 in either the first or second approach. For example, when $k_1 = 100$ and $k_2 = 100$, ratios are extremely low for most values of n_1 and p

except when $n_1 = 100$ and $p = 0.005, 0.01$. In contrast, the results of simulation obtained by applying equation (5) (i.e. upper values, with correction) looks much better, and only a few ratios with very small n_1 (i.e. $n_1 = 10$) and p (i.e. $p = 0.001$) are extremely large. Hence, from Table 1, we can conclude that if we want to avoid an extremely low ratio of the variance derived analytically to the variance of the second stage derived by simulation, it is necessary to add a correction when we evaluate the estimator of p at the second stage.

Concerning the correction, a further question arises: the numerator of the ratio (i.e. equation (21) or (28)) is derived by not adding any corrections, whereas for the denominator of the ratio we have confirmed that one should apply the equation with correction (i.e. equation (5)). A guess here is if we unify the numerator and the denominator by adding a correction to both sides, the ratio would be closer to 1 than the ratio (i.e. upper value) in Table 1. Deriving the variance of the estimator of p at the second stage with correction is a big challenge, but it might be worth to do it in the future.

Overall, the first approach will be recommended if n_1 is very small, or $n_1 \geq 30$ when k_1 and k_2 have large differences; the second approach will be recommended if $n_1 \geq 30$ when $k_1 = k_2$ or the value of k_2 is approximately half of k_1 . To compare the results of the ratio with correction for the first approach and second approaches, let us take a look at the results that are presented in the form of figure and table. Figure 1 shows the ratio of the variance derived analytically to the variance of the second stage obtained by simulation for each approach for each combination of

$n_1 = 10$, p, k_1 and some appropriate values of k_2 (which are depended on the value of k_1). Figure 2 and Figure 3 are similar to Figure 1 but with $n_1 = 30$ and $n_1 = 50$ separately. We will choose one or two k_2 that is same or approximately half of k_1 for all k_1, and one k_2 that has large difference with k_1 for $k_1 = (20,50,100)$: $k_1 = 100$, $k_2 = (6,50,100)$; $k_1 = 50$, $k_2 = (6,20,50)$; $k_1 = 20$, $k_2 = (6,12,20)$; $k_1 = 12$, $k_2 = (6,12)$; $k_1 = 6$, $k_2 = 6$.

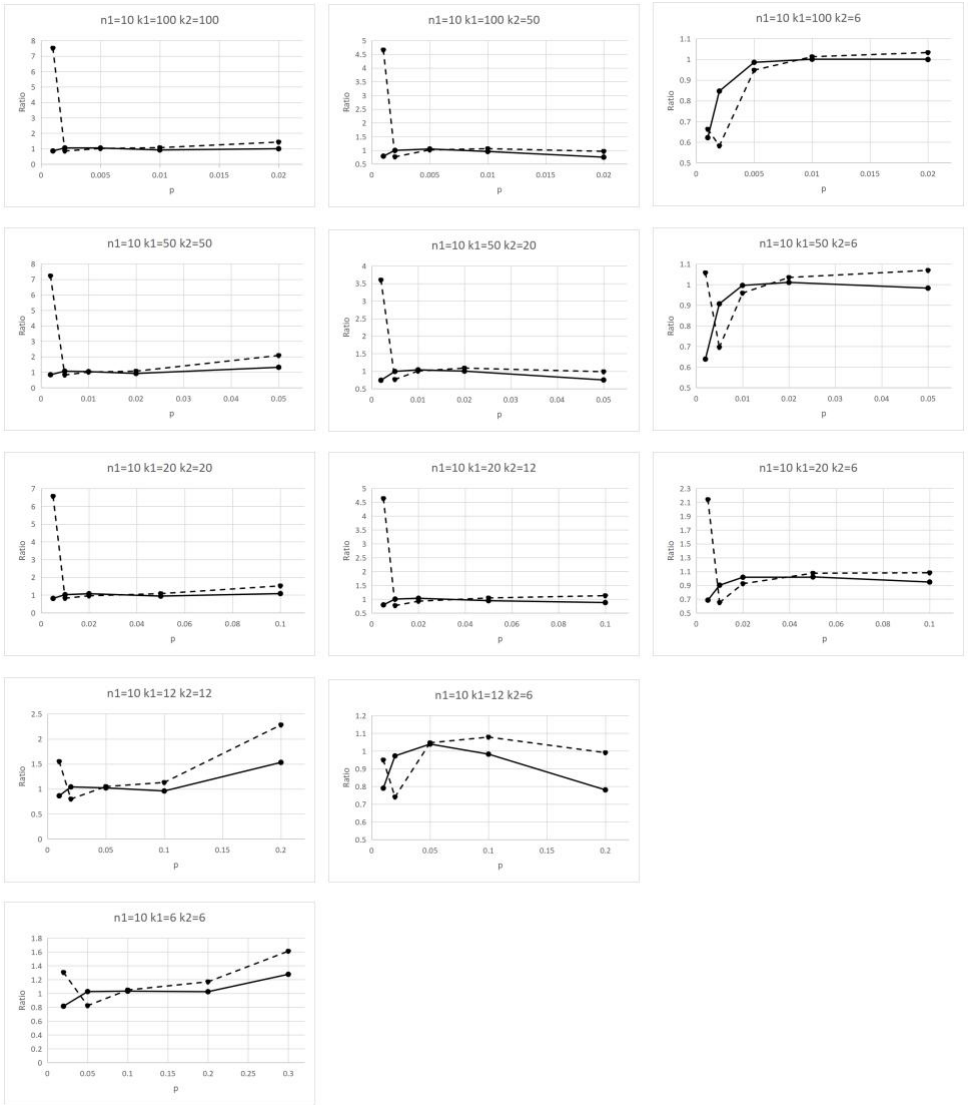

Figure 1. Ratio of the variance derived by $\begin{cases} \text{first approach(---)} \\ \text{second approach(—)} \end{cases}$ to the variance derived by simulation when $n_1 = 10$. x-axis is the value of true prevalence p; y-axis is the value of the ratio.

Note that x-axis represents the true prevalence; y-axis represents the ratios; the

dashed line represents the trend of the ratio of the variance derived by the first approach to the variance derived by simulation, and the solid line represents trend of the ratio of the variance derived by the second approach to the variance derived by simulation.

From Figure 1, we can observe that the trend of the ratio for the first approach and the second approach will have a similar pattern if p is not very small. Simultaneously, no matter the value of k_2 is same, approximately half or has large difference with k_1, the ratio for the first approach is always larger than the second approach if p is very small or very large; some ratios for the first approach are extremely high and totally unacceptable if p is very small; if p is not very small, most ratios for the first and second approach are very close, but the trend of the ratio for the second approach is more smoothly than the first approach, and more ratios for the second approach are closer to 1; the first approach will have more acceptable ratios as the value of p getting larger, but it has smaller number of acceptable ratios than the second approach in total. Therefore, the second approach will be recommended if n_1 is very small.

From Figure 2 and Figure 3, we can observe that the trend of the ratio for the first approach and second approaches have very similar patterns. When $k_1 = k_2$ (i.e. five plots in the first column), the ratio for the first approach is always larger than the ratio for the second approach; the difference between the ratio for the first approach and second approach will get larger as the value of p gets larger; the first approach has more acceptable ratios (i.e. close to 1) than the second approach. Hence, the first

approach will be recommended if $k_1 = k_2$ when n_1 is 30 or more. When the value of k_2 is approximately half of k_1 (i.e. four plots in the second column), most ratios for the second approach are acceptable if p is not very large, all ratios for the first approach are acceptable and the trend of the ratio for the first approach is more smoothly than the second approach. Hence, the first approach will be recommended

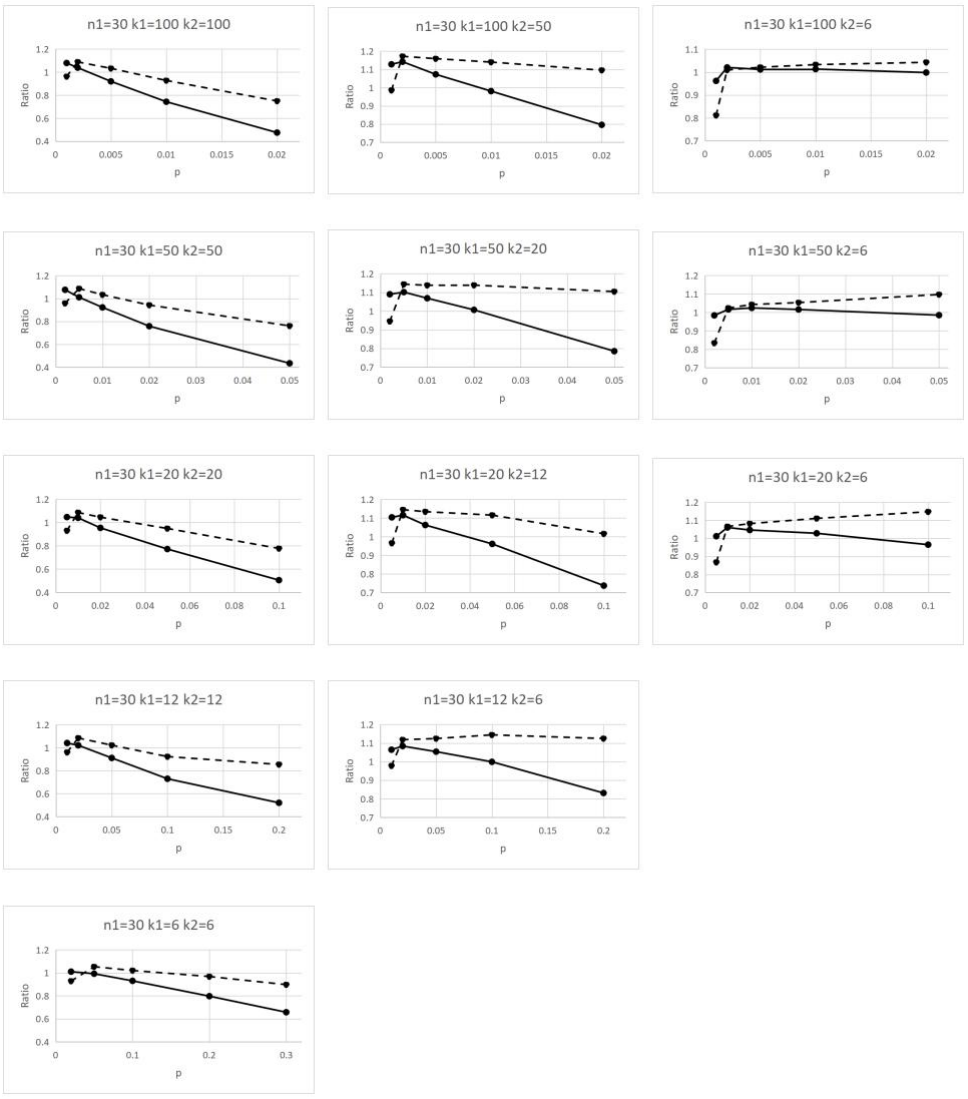

Figure 2. Ratio of the variance derived by $\begin{cases} \text{first approach(- - -)} \\ \text{second approach(—)} \end{cases}$ to the variance derived by simulation when $n_1 = 30$. x-axis is the value of true prevalence p ; y-axis is the value of the ratio.

if the value of k_2 is approximately half of k_1 when n_1 is 30 or more. When the

values of k_1 and k_2 have large differences (i.e. three plots in the third column), the ratio for the first approach is always smaller than the ratio for the second approach if p is very small; the difference between the ratio for the first and second approaches will get larger as the value of p gets larger; all ratios for the first and second approaches are very close and acceptable, but the trend of the ratio for the second

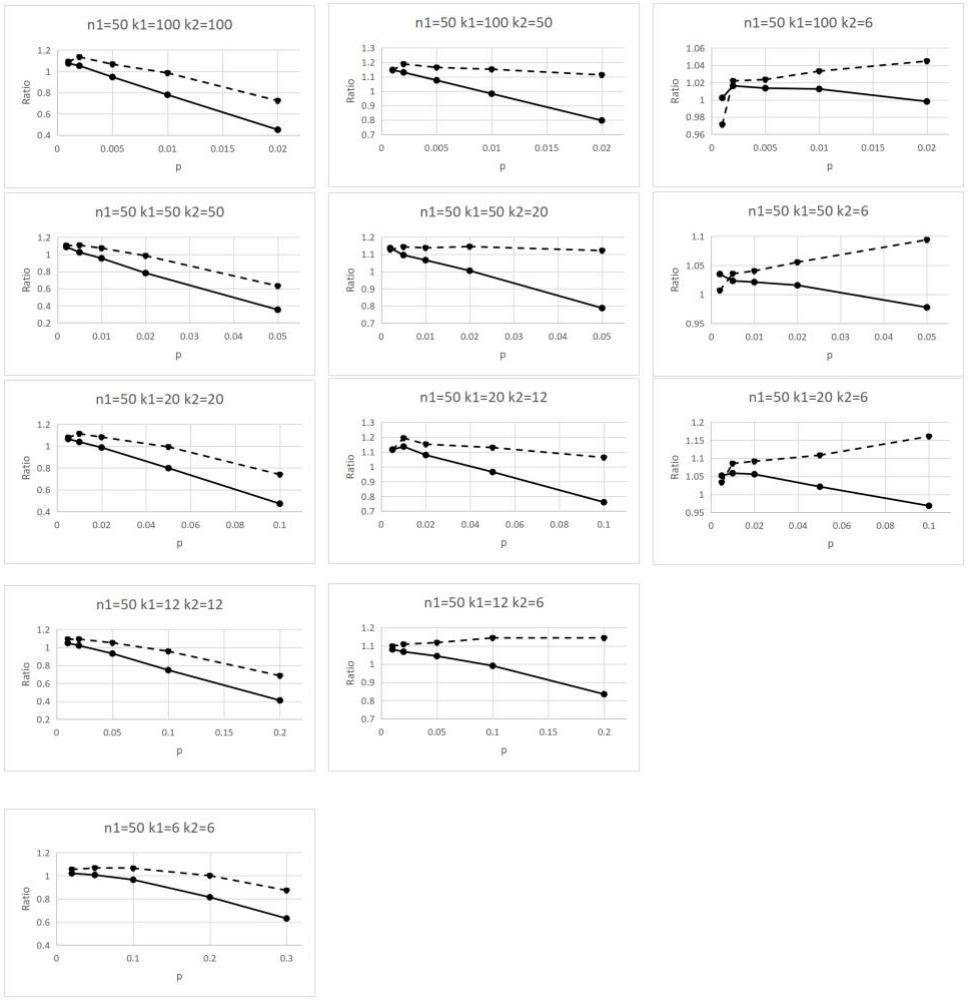

Figure 3. Ratio of the variance derived by $\begin{cases}\text{first approach}(---)\\ \text{second approach}(—)\end{cases}$ to the variance derived by simulation when $n_1 = 50$. x-axis is the value of true prevalence p; y-axis is the value of the ratio.

approach is smoother than the first approach, and the ratio for the second approach is closer to 1 than the first approach, which means that the ratio for the second approach is more acceptable than the first approach. Hence, the second approach will be

recommended if the values of k_1 and k_2 have large differences when n_1 is 30 or more.

In summary, when n_1 is very small, the second approach will be recommended; when $n_1 \geq 30$, the first approach will be preferred if $k_1 = k_2$ or the value of k_2 is approximately half of k_1, and the second approach will be preferred when the values of k_1 and k_2 have large differences.

Now, we will move to look at the results that are presented in the form of a table. We will look at the second approach first. When the values of k_1 and k_2 have large differences and n_1 is large enough (i.e. larger or equal to 30), the ratio of variance is always between 0.9 to 1.1 for all values of p. The ratio in such range is approaching to 1, in other words, the variance derived analytically by using the second approach is close to the real result when k_1 and k_2 vary greatly with 30 or more n_1. However, when the values of k_1 and k_2 have large differences but with $n_1 = 10$, most ratios of variance are between 0.84 to 1.04 for the largest four values of p, but some ratios are around 0.65 when the value of p is smallest. For the range of 0.84 to 1.04, it is certain that the ratio between 0.9 to 1.04 is an nearly ideal ratio, and the ratio between 0.84 to 0.9 is still narrowly acceptable although it is not that perfect. When the values of k_1 and k_2 are the same and $n_1 \geq 30$, the variance derived analytically by using the second approach is close to the real result for the three smallest values of p; when k_2 is approximately half of k_1 and n_1 is 30 or more, the second analytical solution is acceptable for the four smallest values of p. When n_1 is the smallest and k_1 and k_2 are the same or k_2 is approximately half of k_1, the second analytical

solution is always between 0.9 to 1.08 for the middle three values of p, and sometimes acceptable for the largest value of p (i.e. $k_1 = k_2 = 100$; $k_1 = k_2 = 20$).

The first approach has a similar overall pattern to the second approach, but it still has some minor changes. First, when the values of k_1 and k_2 have large differences, the acceptable ratio changes towards larger n_1 (i.e. 50 or more) for all value of p and towards larger p (largest three value of p instead of largest four) when $n_1 = 10$; the first analytical solution is appropriate when $n_1 = 30$ for the largest four values of p. Second, when the values of k_1 and k_2 are same or k_2 is approximately half of k_1, the acceptable ratio changes towards larger p for all n_1.

3.2.2 Determine the most valuable combination

In an experiment, in addition to get a good experimental result, investigators are also interested in discovering the most cost-efficient combination which is able to accomplish an experiment with a minimum cost and get the most precise result. Cost-efficiency is closely related to precision and cost, it can be expressed as

$$\text{Cost - Efficiency} = \frac{\text{Precision}}{\text{Cost}} = \frac{1/\text{Variance}}{\text{Cost}} \tag{29}$$

We will use different cost functions in the next two subsections. In section 3.2.2.1, we will define 'cost' as the total cost using a testing function, in other words, it is equal to the number of groups at the first stage n_1 if we are only interested at the cost-efficiency of the first stage; it is equal to the total number of groups at first and second stage $n_1 + n_2$ if we are interested in the cost-efficiency of two stages overall.

Instead, in section 3.2.2.2, we will define 'cost' as the total cost using a sampling function, equal to the total number of individuals being tested $n_1 k_1$. Note that when applying the cost function in section 3.2.2.2, the cost is still equal to $n_1 k_1$ if we are going to do retesting, because the individuals being tested in the second stage are the same as in the first stage.

In the next two subsections, we will determine whether using the first stage only is precise enough or not and if it is worth to go on to the second stage. By looking at the cost-efficiency of overall two stages, all the most efficient combinations will be chosen by fixing n_1, k_1 and p or by fixing $n_1 k_1$ and p, then, the optimal combination will be determined by looking at these 'the most efficient' combinations' criterion, including standard error, relative cost-efficiency, precision, and bias.

3.2.2.1 Simulation Comparison by fixing n_1, k_1 and p

The first goal of this subsection is finding out the value of k_2 which would form the most efficient combination with given n_1, k_1 and p. Table 3 displays the cost-efficiency of each combination when applying the cost function that is based on the total number of groups tested. In each combination, the first value represents the cost-efficiency of testing overall two stages, and the numerator of equation (29) is the variance of overall p which is derived by using the first approach (i.e. equation (21)); the second value represents the similar thing with the first value, but the numerator of equation (29) is derived by using the second approach (i.e. equation (28)); the third value represents the cost-efficiency of testing the first stage only. From

this table, we can observe that the cost-efficiency of testing the first stage only is always greater than the cost-efficiency of testing overall two stages, which means that retesting will reduce the cost-efficiency. In fact, it is not a surprising observation because retesting requires additional cost, and some retesting will probably increase a little precision which is not worth the extra cost. Nevertheless, some investigators might still want to go to the second stage, for example, the precision of testing the first stage only is not enough, and then investigators decide to spend a little more money to go to the second stage to improve precision. Note that the cost-efficiency of testing overall two stages by using the first and second approaches are both shown in Table 3, but we will only look at one of them based on the conclusions in section 3.2.1. The choice of approach for cost-efficiency is shown below:

The value of n_1	The values of k_1 and k_2	The choice of approach for cost-efficiency
n_1 is very small	Any value of k_1 and k_2	The second approach
$n_1 \geq 30$	$k_1 = k_2$	The first approach
$n_1 \geq 30$	k_2 is approximately half of k_1	The first approach
$n_1 \geq 30$	k_1 and k_2 have large differences	The second approach

Table 4. The choice of approach for cost-efficiency

Overall, in the case where n_1, k_1 and p are fixed, the value of k_2 which would form the most efficient combination with given n_1, k_1 and p is always equal to k_1 or approximately half of k_1. This can be proved by observing the results

in Table 3. From Table 3, we can observe that when $k_1 = (12,50,100)$, the value of k_2 which would form the most cost-efficient combination with given n_1, k_1 and p is half of k_1 if $n_1 = 10$ and p is relative small; when $k_1 = (12,20,50)$, the value of k_2 which would form the most cost-efficient combination with given n_1, k_1 and p is half of k_1 if $n_1 = (50,100)$ and p is relative large; otherwise, the value of k_2 which would form the most cost-efficient combination with given n_1, k_1 and p is always equal to k_1. Thus, we can conclude that the most cost-efficient combination is consisted by using $k_1 = k_2$ when k_1 is very small all the time and k_1 is not very small (i.e. $k_1 = (12,20,50,100)$) for most of the time; however, when k_1 is not very small, some of the most cost-efficient combinations will be consisted by a given k_1 and a value of k_2 that is approximately half of k_1.

Next, we will evaluate the 'value' of those 'most cost-efficient' combinations and determine the optimal combination based on some criterion. Table 5 displays the most cost-efficient combination for each given n_1, k_1 and p, and their standard error, bias and relative cost-efficiency. Note that the standard error, bias, and relative cost-efficiency of testing overall two stages are all shown in Table 5, but we will only look at the criterion of one approach based on the conclusions in Table 3.

First, the standard error will be discussed. The standard error is a square root of the variance, it measures how precise an estimate is, as the standard error getting smaller, an estimate will be more precise. Column 5 and 9 in Table 5 represent the standard error of overall p which is derived by using the first approach and second approach separately. From these two columns, we can observe that it is very hard to

determine which combination has a better estimate of overall p because each estimate is corresponding with different true prevalence. Therefore, we will use the Relative Standard Error (RSE) to evaluate the standard error of these combinations. RSE is defined as a fraction of the standard error and the true prevalence, and it can be expressed as

$$\text{Relative Standard Error} = \frac{(\text{estimate})\ \text{Standard Error of overall } p}{\text{True prevalence}} = \frac{SE(\tilde{p}_{1+2})}{p} \quad (30)$$

Usually, RSE is displayed as a percentage. The combination with a high percentage of RSE represents that there is more relative variation in the estimates, which means that such combination will subject to high estimation error and it needs to be careful when using such design. If a combination has a low percentage of RSE, then it represents that there is less relative variation in the estimates, which means that this combination is acceptable and it is good enough for using the first stage only. In this thesis, a relative standard error is defined as acceptable if it is below 20%; in other words, if the RSE is below 20%, then testing the first stage only is precise enough.

Overall, we will recommend investigators to use the combination that has as large n_1 as possible. This can be proved by comparing the RSE of combinations in Table 5. Column 6 and 10 in Table 5 represent the relative standard error where the overall p is derived by using the first approach and second approach separately. Based on the conclusions in Table 3, we will look at column 6 (i.e. the RSE of the first approach) when $n_1 \geq 30$, and we will look at column 10 (i.e. the RSE of the second approach) when $n_1 = 10$. From these two columns, we can observe that the relative standard

error will be more acceptable when n_1 and p are getting larger, and extremely unacceptable when n_1 and p are very small no matter what the values of k_1 and k_2 are. However, different kinds of experiments will have a different range of true prevalence. For example, rat-bit fever is a rare infectious disease with only several cases in the world each year, so it has extremely low true prevalence; instead, malaria is a common disease and it has high true prevalence. Therefore, by looking at (relative) standard error only, if the research variable of interest is common and has high true prevalence, then the combination of large enough number of groups n_1 with any value of k_1 and a value of k_2 that is the same or approximately half of k_1 will be recommended. Under this condition, the estimate of proportion is good enough for using the first stage only, thus one recommendation for next step is investigators can decide to accept the experiment result and not go on to the second stage; the other recommendation is they can decide to go on to the second stage if they have enough cost and interested at seeing if the second stage will give them a more precise result. However, if the research variable of interest is rare and has low true prevalence, the difficulty of the experiment will increase and it is very hard to get an estimate of the proportion which has a very small relative standard error. Even so, the recommended combination is still to have a large enough number of groups n_1 with any value of k_1 and a value of k_2 that is the same or approximately half of k_1, which is able to minimize the error in the case of low true prevalence. Under this condition, the estimate of the proportion is not so bad but not that ideal for using the first stage only, so one recommendation for next step is investigators can decide to stop the

experiment and change a different design, the other recommendation is they may decide to continue the experiment and go on to the second stage to see whether the result has improved or not.

Overall, if the cost function is based on the total number of groups tested, then the estimate of p using two testing stages combined is close to the true prevalence. Column 7 and 11 in Table 5 represent the bias between the estimate of p using two testing stages combined and the true prevalence, where column 7 represents the estimate of p using two stages combined is derived by using the first approach and column 11 represents the estimate of p using two stages combined is derived by using the second approach. Note that both two columns are multiplied by 100000 to make it easier to read. Based on the conclusions in Table 3, we will look at column 7 (i.e. the bias of the first approach) when $n_1 \geq 30$, and we will look at column 11 (i.e. the bias of the second approach) when $n_1 = 10$. A small bias represents that the difference between the estimation and the true result are small; in other words, the estimation is close to the true result. Therefore, bias needs to be as small as possible. From these two columns, we can observe that all bias are very small. The largest bias (times 100000) is -2078.97 when $n_1 = 10$, $p = 0.2$ and $k_1 = k_2 = 12$, but since it is 100000 times bias, so the estimate of p using two testing stages combined is $\widetilde{p}_{1+2} = p + \text{bias} = 0.2 + \frac{-2078.97}{100000} = 0.1792103$, which is close to the true prevalence $p = 0.2$. Therefore, we can conclude that all estimates of p using two testing stages combined are close to their true prevalence.

Column 8 and 12 in Table 5 represent the relative cost-efficiency which is derived

by using the first approach and second approach separately. Based on the conclusions in Table 3, we will look at column 8 (i.e. the relative cost-efficiency of the first approach) when $n_1 \geq 30$, and we will look at column 12 (i.e. the relative cost-efficiency of the second approach) when $n_1 = 10$. Relative cost-efficiency is one of the important criteria to evaluate the 'value' of a combination. It is a fraction of the cost-efficiency of two stages combined and the first stage only, it can be expressed as

$$\text{Relative Cost - Efficiency} = \frac{\text{Cost - Efficiency}_{\text{overall}}}{\text{Cost - Efficiency}_{\text{first stage}}}$$

$$= \frac{Var(\tilde{p}_1) \cdot Cost_1}{Var(\tilde{p}_{1+2}) \cdot Cost_{1+2}} \tag{31}$$

where the cost function here is based on the total number of groups tested, $Cost_1 = n_1$ and $Cost_{1+2} = n_1 + n_2$. Therefore,

$$\text{Relative Cost - Efficiency}_{\text{(the total cost using testing function)}} = \frac{Var(\tilde{p}_1) \cdot n_1}{Var(\tilde{p}_{1+2}) \cdot (n_1 + n_2)} \tag{32}$$

The relative standard error tells investigators whether doing the first stage only is precise enough or not, instead, the relative cost-efficiency tells investigators if it is worth to go on to the second stage. Usually, a combination is optimal if its relative cost-efficiency is larger or equal to 1, which means that the cost-efficiency of retesting is larger or equal to the cost-efficiency of doing the first stage only. But note that in Table 3, we can observe that the cost-efficiency of using the first stage only is always larger than the cost-efficiency of retesting. Even so, some investigators might still want to do retesting if it is worth to go on to the second stage, where 'worth' means spending more funds to get a more precise result. A relative cost-efficiency that is

close to 1 represents that cost relative to precision of testing two stages combined is approximately same with doing the first stage only, which means that if investigators are testing overall two stages by using the cost as doing the first stage only, then they will get an approximately same precision with doing the first stage only. For example, if the relative cost-efficiency is 0.9, then it indicates that testing two stages combined is 90% as efficient as testing the first stage only. Since the cost-efficiency of using the first stage only is always larger than the cost-efficiency of retesting when the cost function is based on the total number of groups tested, therefore, in table 5, a relative cost-efficiency is defined as acceptable if it is above 0.8 (below 1 but close to 1), and we may recommend investigators to go on to the second stage.

Let us look at the RE when $n_1 = 10$ (i.e. column 12, the second approach) first. From column 12, we can observe that if p is very small (0.5% or less), then the relative cost-efficiency is always acceptable when $k_1 = (20,50)$ and acceptable for most of the time when $k_1 = 100$; if p is small (larger than 0.5%, less or equal to 1%), then the relative cost-efficiency is acceptable when $k_1 = (12,20)$, and entirely not acceptable when $k_1 = (50,100)$; if p is large (larger than 1%, less or equal to 5%), then the relative cost-efficiency is always acceptable when $k_1 = (6,100)$, sometimes acceptable when $k_1 = (12,50)$, and entirely not acceptable when $k_1 = 20$; if p is very large (larger than 5%), then the relative cost-efficiency is always acceptable when $k_1 = 20$, sometimes acceptable when $k_1 = (6,12)$. Next, let us look at the RE when $n_1 \geq 30$ (i.e. column 8, the first approach). From column 8, we can observe that all findings when $n_1 \geq 30$ are the same as the findings when $n_1 = 10$,

except the following conditions: when $0.01 < p \le 0.05$, the relative cost-efficiency is sometimes acceptable when $k_1 = 6$ instead of acceptable for all the time, and not acceptable when $k_1 = 100$ instead of acceptable for all the time; when $p > 0.05$, the relative cost-efficiency is totally not acceptable when $k_1 = 6$ instead of acceptable for some time, and not acceptable when $k_1 = 20$ instead of acceptable for all the time. The summary of the observations from column 8 and 12 are listed in Table 6.

		acceptable or not	
p	**k1**	**n1=10**	**n1>=30**
p<=0.005	100	acceptable for most of the time	acceptable for most of the time
	50	always acceptable	always acceptable
	20	always acceptable	always acceptable
0.005<p<=0.01	100	not acceptable	not acceptable
	50	not acceptable	not acceptable
	20	always acceptable	always acceptable
	12	always acceptable	always acceptable
0.01<p<=0.05	100	always acceptable	not acceptable
	50	sometimes acceptable	sometimes acceptable
	20	not acceptable	not acceptable
	12	sometimes acceptable	sometimes acceptable
	6	always acceptable	sometimes acceptable
p>0.05	20	always acceptable	not acceptable
	12	sometimes acceptable	sometimes acceptable
	6	sometimes acceptable	not acceptable

Table 6. The summary of the observations of relative cost-efficiency

Note that the final recommendation will be given based on the true prevalence, therefore, since $p \le 0.005$ and $0.005 < p \le 0.01$ are close, these two ranges will be combined ($p \le 1\%$) and defined as relatively low true prevalence or rare; $0.01 < p \le 0.05$ and $p > 0.05$ will be combined ($p > 1\%$) and defined as relatively high true prevalence or common. When we analyzing the RSE, we have concluded that the combination that has as large n_1 as possible will be recommended. Therefore, it is significant to have an acceptable RE when n_1 is large.

Let us look at Table 6 for relatively low true prevalence first. We can observe that if p is very small (0.5% or less), then the RE is acceptable for all n_1 when

$k_1 = (20,50,100)$; if p is small (larger than 0.5%, less or equal to 1%), the RE is acceptable for all n_1 when $k_1 = (12,20)$, but no longer acceptable when $k_1 = (50,100)$. In summary, if $k_1 = (12,20)$, the RE is acceptable when n_1 is large, which means that it is worth to go on to the second stage if n_1 is large, $k_1 = (12,20)$ and a value of k_2 that is the same or approximately half of k_1 . Next, let us look at Table 6 for relatively high true prevalence. We can observe that if p is large (larger than 1%, less or equal to 5%), the RE is acceptable for all n_1 when $k_1 = (6,12,50)$; if p is very large (larger than 5%), then the RE is acceptable for all n_1 when $k_1 = 12$, but no longer acceptable when $k_1 = 6$. In summary, if $k_1 = (12,50)$, the RE is acceptable when n_1 is large, which means that it is worth to go on to the second stage if n_1 is large, $k_1 = (12,50)$ and a value of k_2 that is the same or approximately half of k_1 .

Overall, by looking at the relative standard error, the bias and the relative cost-efficiency, if the cost function we applied is based on the total number of groups tested, and the value of n_1 , k_1 and p are fixed, then we will make the following recommendation:

1. If the research variable of interest has relatively high true prevalence ($p > 1\%$) and investigators do not have extra cost or they are not interested at seeing if the second stage can give a better result, then the combination of large n_1 with any value of k_1 will be recommended, and investigators only need to do the first stage if they follow the recommendation.

2. If the research variable of interest has relatively high true prevalence ($p > 1\%$) and

investigators have extra fund and they are interested in seeing if the second stage can give a better result, then the combination of large n_1, $k_1 = (12,50)$ and a value of k_2 that is same or approximately half of k_1 will be recommended.

3. If the research variable of interest has relatively low true prevalence ($p \le 1\%$), then the combination of large n_1, $k_1 = (12,20)$ and a value of k_2 that is same or approximately half of k_1 will be recommended, and investigators can either decide to change the design of combination after doing the first stage or go to the second stage to see if they can get a better result.

3.2.2.2 Simulation comparison by fixing n_1k_1 and p

In this subsection, we will do a similar work with section 3.2.2.2, but the value of n_1k_1 and p will be fixed, and the cost function, the equation of cost-efficiency and relative cost-efficiency will be changed. The cost function we used in this subsection is based on the total number of individuals tested, where $Cost_{(\text{first stage only})} = Cost_{(\text{overall two stages})} = n_1k_1$. Therefore, the new equation for the relative cost-efficiency can be expressed as

$$\text{Relative Cost - Efficiency}_{(\text{the total cost using sampling function})} = \frac{Var(\tilde{p}_1)\cdot(n_1k_1)}{Var(\tilde{p}_{1+2})\cdot(n_1k_1)} \qquad (33)$$

By comparing the equation (32) and (33), if investigators decide to not go to the second stage, then $n_2 = 0$, and the relative cost-efficiency by using the testing based cost function (i.e. equation (32)) will be equal to the relative cost-efficiency by using the sampling based cost function (i.e. equation (33)); if investigators decide to do

retesting, then the relative cost-efficiency by using the sampling based cost function will be larger than the relative cost-efficiency by using the testing based cost function. Therefore, the RE by using the sampling based cost function is always greater or equal to the RE by using the testing based cost function.

We have $n_1 = (10,30,50,100)$ and $k_1 = (6,12,20,50,100)$, so n_1k_1 can be combined as

n_1k_1	n_1=10	n_1=30	n_1=50	n_1=100
k_1=6	60	180	300	600
k_1=12	120	360	600	1200
k_1=20	200	600	1000	2000
k_1=50	500	1500	2500	5000
k_1=100	1000	3000	5000	10000

Table 5. Different combinations of n_1k_1

Note that since we want to compare the cost-efficiency of each combination with given n_1k_1 and p, therefore the n_1k_1 that only appears once in Table 5 will be dropped and we will select common p in each (n_1,k_1) with given n_1k_1. Hence, we will find the proper value of k_2 to form the most cost-efficient combination with the following nk_1 and p:

$$n_1k_1 = 600: \ (n_1,k_1) = \{(100,6),(50,12),(30,20)\}, \ p = (0.02,0.05,0.1)$$

$$n_1k_1 = 1000: \ (n_1,k_1) = \{(50,20),(10,100)\}, \ p = (0.005,0.01,0.02)$$

$$n_1k_1 = 5000: \ (n_1,k_1) = \{(100,50),(50,100)\}, \ p = (0.002,0.005,0.01,0.02).$$

Table 7 displays the cost-efficiency of each combination above when applying the cost function that is based on the total sample size. Column 6 and 7 represent the cost-efficiency of testing overall two stages by using the first approach and second

approaches separately; column 8 represents the cost-efficiency of testing the first stage only. Note that the cost-efficiency of testing overall two stages by using the first and second approaches are both shown in Table 7, but we will only look at one of them based on Table 4. Different with the observations from Table 3, we can observe that in Table 7, most cost-efficiency of testing overall two stages are approximately equal or larger than the cost-efficiency of testing the first stage only, which means that doing additional stage will have similar precision with doing the first stage only, or even increasing the precision of the first stage. This observation makes more sense for investigators to do retesting. From Table 7, we can observe that the value of k_2 which would form the most cost-efficient combination with given n_1k_1 and p is always equal to 6 whatever the value of n_1k_1 and p are. Therefore, $k_2 = 6$ will be recommended if the value of n_1k_1 and p are fixed.

Table 8 displays the standard error, relative standard error, bias, and relative cost-efficiency for 'most cost-efficient' combination we found in Table 7. Based on Table 4, we will look at column 6 to 9 (i.e. the first approach) when $n_1k_1 = 600$ since all three 'most cost-efficient' combinations have large enough n_1 ($n_1 \geq 30$) and the value of k_2 is the same or approximately half of k_1; we will look at column 10 to 13 (i.e. the second approach) when $n_1k_1 = 1000$ since the first two 'most cost-efficient' combinations have large enough n_1 ($n_1 \geq 30$) and k_1 and k_2 have large differences, and the last 'most cost-efficient' combination has very small n_1 ($n_1 = 10$); we will look at column 10 to 13 (i.e. the second approach) when $n_1k_1 = 5000$ since all four 'most cost-efficient' combinations have large enough n_1

($n_1 \geq 30$) and k_1 and k_2 have large differences.

The relative standard error in Table 8 shows that doing the first stage only is precise enough (i.e. below 20%) if the total sample size is small (i.e. 600, 1000) and p is very large ($p \geq 0.05$) or if the total sample size is large (i.e. 5000) and p is relatively large ($p \geq 1\%$); doing the first stage only is not precise enough if the total sample size is small and $p < 0.05$ or if the total sample size is large and p is relatively small ($p < 0.01$). Table 9 concludes the observations above.

total sample size (n_1k_1)	**p**	**doing the first stage only is precise enough or not**
small (i.e. 600,1000)	p<0.05	not precise enough
small (i.e. 600,1000)	p>=0.05	precise enough
large (i.e. 5000)	p<0.01	not precise enough
large (i.e. 5000)	p>=0.01	precise enough

Table 9. The observations of RSE

However, different experiment has different research object, therefore the total number of individuals can be collected is depended on the variety of research object. For example, humans and mosquitoes, it is more possible for investigators to collect over ten thousand mosquitoes than to collect over ten thousand people. Therefore, the total sample size will not be very large if the research object is human, instead, the total sample size can be extremely large if the research object is mosquitoes. In this section, we only analyzed the experiment with a total sample size that is not extremely small or large (i.e. 600, 1000, 5000) because our setting of combination is limited. Nevertheless, we can still make a guess for the experiment with extremely small or large total sample size based on the observations of the relative standard error (Table 9): if the total sample size is extremely small, then it needs a true prevalence

that is much larger than 5% to make the result of doing the first stage only precise enough; if the total sample size is extremely large, then it needs a true prevalence that is much smaller than 1% to make the result of doing the first stage only precise enough; otherwise, doing the first stage only will be not precise enough. Analyzing the experiment with an extremely small or large total sample size is a significant work, and this guessing will be verified or against in future work.

Overall, if the cost function is based on the total sample size, then the estimate of p using two testing stages combined is close to the true prevalence. From column 8 and column 12 in Table 8, we can observe that all bias are very small. The largest bias (times 100000) is -13.91 when $n_1k_1 = 1000$ and $p = 0.02$, but since it is 100000 times bias, so the estimate of p using two testing stages combined is $\widetilde{p}_{1+2} = p + \text{bias} = 0.02 + \frac{-13.91}{100000} = 0.0198609$, which is close to the true prevalence $p = 0.02$. Therefore, we can conclude that all estimates of p using two testing stages combined are close to their true prevalence.

From the relative cost-efficiency in Table 8, we can observe that the relative cost-efficiency is larger than 1 for all the time, which means that it is worth to do retesting whatever the value of n_1k_1 and p are.

Overall, by looking at the relative standard error, the bias and the relative cost-efficiency, if the cost function we applied is based on the total sample size, and the value of n_1k_1 and p are fixed, then we will make the following recommendation:

1. If investigators can collect either small or large sample sizes, and the research

variable of interest has a relative small true prevalence to the total sample size, then a combination with $k_2 = 6$ will be recommended. Under this situation, doing the first stage only is not precise enough, so investigators can either decide to change the design or go to the second stage. However, retesting will be more recommended since testing overall two stages will gain a little bit more precision at most of the time.

2. If investigators can collect either small or large sample sizes, and the research variable of interest has a relative large true prevalence to the total sample size, then a combination with $k_2 = 6$ will be recommended. Under this situation, doing the first stage only is precise enough, so investigators can either decide to stop at the first stage or go to the second stage. However, retesting will be more recommended since testing overall two stages will always gain much more precision, which is worth the extra fund.

Chapter 4 Discussion

4.1 Extension of Group testing: Classification

Our main goal of this thesis is estimating the true proportion of affected units, which is one reason for using group testing. Classification is the other category of group testing. It is Dorfman (1943)'s primary incentive for using group testing, and it is aimed at identifying positive units, or in other words, detecting individuals with the disease of interest (Kim et al., 2007). Nevertheless, there exist some imperfect cases in group testing, such as misclassification. Misclassification occurs when an individual is classified into the wrong population subgroup. For example, suppose there are 50 people in an experiment, 45 of them are healthy and 5 of them have cancer. Now these 50 people are mixed up and tested individually to see if they have cancer, if a healthy person is diagnosed with cancer or a people who have cancer is diagnosed as health, then this person is classified into the wrong subgroup, which is a misclassification. Many statistical studies of group testing assume that the test samples can be analyzed by group testing precisely without any errors. However, when the proportion p is relatively large to the total sample size, we need to be cautious in choosing the number of group size k. If the group size k is too large, it will

result in a high mean squared error (MSE) of the estimation of the true proportion, and test samples will be misclassified into the wrong subgroup. Therefore, it is very important to choose an optimal group size k (Liu et al., 2011; Chen and Swallow, 1990). Graff and Roeloffs (1972) gave an extension of group testing based on Dorfman (1943) under the condition of known test error between outcome and true state. Burns and Mauro (1987) summarized the former's conclusion and proposed group testing with accidental test error .

4.1.1 Dorfman's group testing

Dorfman (1943)'s group testing has a different objective with this thesis. The goal of group testing in this thesis is estimating the proportion of affected individuals, nevertheless, Dorfman's incentive is identifying individuals with the disease of interest. Dorfman (1943) (see also Malinovsky and Albert 2018) proposed a screening procedure intended to decrease the expected number of tests required to identify soldiers with syphilis. He regards n soldiers as a whole group, and then collected their blood samples separately. He began with a test on this group of blood samples. If the result shows negative, then it declares that none of the soldiers have syphilis in this group, therefore no further test is needed; if the result shows positive, it states that there exists at least one soldier has syphilis in this group, then each soldier has to be retested individually, therefore it needs $n+1$ tests in this group, we can also call this as the Dorfman two-stage procedure or retesting. Nevertheless, the total number of tests needed will not exceed $n+1$ in any case. When the population proportion of

affected individuals is small, then it requires a small expected number of tests.

4.1.2 Sterrett's group testing

Sterrett (1957) modified and suggested a more advanced procedure based on Dorfman's screening procedure. He aimed at further reduction in the expected number of tests needed. If the result of group test is positive in the first stage, then each individual is tested one-by-one instead of testing all individuals in the second stage, this process stops after the first appearance of a nonconforming individual. Then group the remaining untested individuals as a new group, and repeat the same procedure in the first stage. If the result is negative, then it states that there exists only one nonconforming individual; if the result is positive, then test each individual one-by-one until the first appearance of a nonconforming individual. The rest can be done in the same manner until all items are tested.

4.2 Future work and Challenge

One extension for our work would be determining an appropriate number of group sizes at the second stage based on the estimate of the true proportion at the first stage, which is called an adaptive group testing scheme. Normally, group testing can be split into two categories: non-adaptive group-testing scheme and adaptive group-testing scheme. The former is more common than the latter since the derivation of the adaptive scheme is more complicated. In this thesis, we used a non-adaptive scheme. The number of group sizes at the first stage and the second stage are fixed at the

beginning. A non-adaptive group-testing scheme tests N groups and each with group size k; if the test result of a certain group is positive, then it means one or more individuals in this group has a trait that conforms to the research variable of interest. An adaptive group-testing scheme tests N_1 groups and each with group size k_1 in the first stage, N_2 groups and each with group size k_2 in the second stage, and so on in a similar manner, the number of group sizes of the next stage will be determined during the experiment and depending on the maximum likelihood estimation of p in the previous stage and the number of tests in the stage to be tested currently. An adaptive scheme refers to a multi-stage scheme, Hughes-Oliver and Swallow (1994) proposed a two-stage adaptive algorithm and derived the number of group sizes in the second stage based on the MLE of p in the first stage.

In addition to the one-stage and two-stage algorithm, the research of three or more schemes is also concerned, which could be an extension of our work. Schultz et al. (1973) proposed multiple-stage procedures for drug screening and gave an example of three-stage designing, then concluded that drugs might be declared active if and only if they pass through all three stages. In statistical research, it is important to choose the most optimal number of stages. The derivation of formulas will be complicated when there are three or more stages, simultaneously, the cost of additional stages must be considered. However, the number of stages needs to be determined on the exact situation, sometimes one- or two-stage would be better, but sometimes it may require more stages.

In section 4.1, we talked about an imperfect case—misclassification—in group

testing. In addition to the imperfect case, there exists a special case in group testing: unequal group sizes. Unequal group sizes can be divided into two varieties: unequal group sizes between stages, and unequal group sizes within a stage. Both of these two varieties might be caused by either deliberate design or unforeseen occurrence. Unequal group sizes between stages are talked in this thesis. An extension of our work would be allowing unequal group sizes within a stage. Furthermore, a more complicated extension work would be allowing unequal group sizes within both two stages, and simultaneously, allowing unequal group sizes between two stages. Statisticians have done some researches on the case of unequal group sizes within a stage. For example, Walter, Hildreth and Beaty (1980) used group testing of unequal group sizes within a stage to estimate the infection rates of yellow fever virus in a mosquito population; Chen and Swallow (1990) designed a set of unequal group sizes within a stage and proposed a grouping test based on the Binomial model with these group sizes; Le (1981)'s research of interest is estimating the infection rates in populations of organisms, but unfortunately, it is impractical to test every unit separately, so instead, he chooses to divide the organisms into multiple groups at random. The derivation of a confidence interval is difficult and complicated when group sizes are unequal. Hepworth (2005) and Hepworth (1996) developed confidence intervals and exact confidence intervals for unequal group sizes.

Overall, there were three challenges in this thesis.

The first is deriving the equation of the variance of the estimate of overall p, $Var(\tilde{p}_{1+2})$, which includes Burrow's correction and was talked in section 2.2. The

difficulty for this challenge is deriving the covariance of modified MLE at the first and second stage, $Cov(\tilde{p}_1, \tilde{p}_2)$.

The second challenge is evaluating whether doing the first stage only is precise enough or not and if it is worth to go to the second stage when the total sample size is extremely small or large, which was talked in section 3.2.2.2. We have this challenge in this thesis because we only set 4 numbers in the figure of n_1 (i.e. $n_1 = (10,30,50,100)$) and 5 numbers in the figure of k_1 (i.e. $k_1 = (6,12,20,50,100)$), our setting of combination is very limited. In the future, we may solve this challenge by examining more n_1 and k_1.

The last challenge is finding out an appropriate cost function. In this thesis, we considered two cost functions: total cost using the testing function (section 3.2.2.1) and total cost using the sampling function (section 3.2.2.2). However, in addition to consider the cost of testing a group or an individual in an experiment, we also need to consider other factors that would cost extra funds. For example, investigators need to spend some costs on collecting data set before testing (Sobel and Elashoff, 1975).

Table 1. The ratio of the variance derived analytically (i.e. first approach means using equation (21), second approach means using equation (28)) to the variance of the second stage obtained by simulation. The upper value is derived by applying equation (5) (i.e. with correction) and the lower value is derived by applying equation (3) (i.e. without correction) in simulation.

				n1(First Approach)				n2(Second Approach)			
k_1	k_2	p	Correction	10	30	50	100	10	30	50	100
100	100	0.001	with	7.5175	0.9622	1.0917	1.1660	0.8537	1.0793	1.0747	1.0944
			without	0.0022	0.0004	0.0008	0.0053	0.0003	0.0005	0.0008	0.0049
		0.002	with	0.8519	1.0888	1.1348	1.1552	1.0505	1.0390	1.0526	1.0847
			without	0.0003	0.0008	0.0030	0.0745	0.0003	0.0008	0.0028	0.0700
		0.005	with	1.0141	1.0329	1.0668	1.1131	1.0425	0.9202	0.9468	0.9852
			without	0.0004	0.0025	0.0292	1.0465	0.0004	0.0022	0.0259	0.9263
		0.01	with	1.0691	0.9278	0.9848	1.0206	0.9186	0.7434	0.7796	0.8007
			without	0.0004	0.0026	0.0447	0.9558	0.0003	0.0021	0.0354	0.7499
		0.02	with	1.4294	0.7498	0.7240	0.7828	0.9980	0.4758	0.4513	0.4817
			without	0.0003	0.0005	0.0017	0.0754	0.0002	0.0003	0.0010	0.0464
	50	0.001	with	4.6584	0.9858	1.1452	1.1984	0.7879	1.1284	1.1471	1.1493
			without	1.2263	0.3377	0.8818	1.1713	0.2074	0.3865	0.8832	1.1233
		0.002	with	0.7631	1.1722	1.1881	1.1884	0.9976	1.1426	1.1305	1.1398
			without	0.0660	0.7029	1.0681	1.1709	0.0863	0.6852	1.0163	1.1231
		0.005	with	1.0164	1.1595	1.1646	1.1662	1.0494	1.0733	1.0753	1.0748
			without	0.0640	1.1184	1.1402	1.1541	0.0661	1.0353	1.0527	1.0636
		0.01	with	1.0601	1.1405	1.1519	1.1629	0.9575	0.9811	0.9824	0.9855
			without	0.1045	1.0994	1.1273	1.1506	0.0944	0.9457	0.9614	0.9752
		0.02	with	0.9632	1.0961	1.1127	1.1272	0.7506	0.7958	0.7972	0.7996
			without	0.0424	1.0381	1.0784	1.1102	0.0330	0.7538	0.7726	0.7875
	20	0.001	with	2.0568	0.8831	1.0446	1.0994	0.6856	1.0322	1.0651	1.0780
			without	1.8922	0.8519	1.0317	1.0958	0.6308	0.9957	1.0519	1.0745
		0.002	with	0.6549	1.0875	1.0997	1.0953	0.9114	1.0834	1.0742	1.0738
			without	0.6178	1.0764	1.0946	1.0927	0.8598	1.0724	1.0692	1.0712
		0.005	with	0.9949	1.0914	1.0945	1.0970	1.0317	1.0517	1.0533	1.0548
			without	0.9704	1.0844	1.0903	1.0949	1.0063	1.0449	1.0493	1.0527
		0.01	with	1.0662	1.1020	1.1060	1.1037	1.0154	1.0237	1.0230	1.0176
			without	1.0442	1.0945	1.1015	1.1015	0.9944	1.0168	1.0188	1.0155
		0.02	with	1.0733	1.1316	1.1264	1.1316	0.9471	0.9611	0.9496	0.9488
			without	1.0451	1.1219	1.1205	1.1287	0.9222	0.9528	0.9447	0.9463
	12	0.001	with	1.2713	0.8439	1.0053	1.0581	0.6504	0.9939	1.0317	1.0459
			without	1.2135	0.8291	0.9995	1.0567	0.6208	0.9765	1.0257	1.0446
		0.002	with	0.6145	1.0437	1.0561	1.0609	0.8755	1.0481	1.0416	1.0484
			without	0.5993	1.0389	1.0540	1.0598	0.8538	1.0432	1.0395	1.0473
		0.005	with	0.9702	1.0564	1.0563	1.0617	1.0077	1.0331	1.0323	1.0370
			without	0.9598	1.0532	1.0544	1.0607	0.9969	1.0300	1.0304	1.0360
		0.01	with	1.0405	1.0663	1.0684	1.0662	1.0107	1.0198	1.0192	1.0150
			without	1.0304	1.0628	1.0664	1.0652	1.0009	1.0165	1.0172	1.0140
		0.02	with	1.0557	1.0875	1.0900	1.0864	0.9780	0.9835	0.9811	0.9745
			without	1.0430	1.0832	1.0874	1.0851	0.9663	0.9796	0.9788	0.9733
	6	0.001	with	0.6628	0.8115	0.9712	1.0214	0.6220	0.9621	1.0023	1.0166
			without	0.6489	0.8070	0.9693	1.0209	0.6089	0.9567	1.0003	1.0162
		0.002	with	0.5824	1.0096	1.0219	1.0302	0.8469	1.0208	1.0162	1.0250
			without	0.5809	1.0081	1.0212	1.0298	0.8447	1.0192	1.0155	1.0247
		0.005	with	0.9479	1.0216	1.0235	1.0287	0.9858	1.0120	1.0135	1.0184
			without	0.9456	1.0204	1.0228	1.0283	0.9835	1.0108	1.0128	1.0181
		0.01	with	1.0129	1.0331	1.0333	1.0309	1.0007	1.0136	1.0126	1.0093
			without	1.0091	1.0318	1.0325	1.0305	0.9970	1.0123	1.0118	1.0090
		0.02	with	1.0330	1.0432	1.0449	1.0371	0.9994	0.9985	0.9981	0.9890
			without	1.0283	1.0416	1.0440	1.0366	0.9949	0.9970	0.9972	0.9886
50	50	0.002	with	7.2277	0.9593	1.1040	1.1636	0.8379	1.0769	1.0875	1.0931
			without	0.0087	0.0016	0.0032	0.0208	0.0010	0.0018	0.0032	0.0195
		0.005	with	0.8240	1.0874	1.1077	1.1279	1.0676	1.0117	1.0255	1.0529
			without	0.0011	0.0042	0.0196	0.6249	0.0014	0.0039	0.0181	0.5834
		0.01	with	1.0108	1.0340	1.0736	1.1027	1.0395	0.9232	0.9550	0.9782
			without	0.0015	0.0098	0.1015	1.0373	0.0016	0.0087	0.0903	0.9202
		0.02	with	1.0679	0.9435	0.9843	1.0289	0.9205	0.7590	0.7823	0.8105
			without	0.0015	0.0102	0.1219	0.9644	0.0013	0.0082	0.0969	0.7597
		0.05	with	2.0840	0.7615	0.6312	0.6425	1.3187	0.4347	0.3534	0.3547
			without	0.0016	0.0011	0.0020	0.0200	0.0010	0.0006	0.0011	0.0111

	20	0.002	with	3.5997	0.9455	1.1276	1.1584	0.7419	1.0902	1.1365	1.1196
			without	2.9706	0.8604	1.0867	1.1437	0.6122	0.9920	1.0953	1.1054
		0.005	with	0.7648	1.1441	1.1432	1.1326	0.9942	1.1018	1.0957	1.0906
			without	0.5854	1.1139	1.1257	1.1239	0.7608	1.0728	1.0790	1.0822
		0.01	with	1.0050	1.1385	1.1378	1.1432	1.0392	1.0692	1.0662	1.0696
			without	0.7684	1.1144	1.1235	1.1360	0.7945	1.0466	1.0528	1.0629
		0.02	with	1.0889	1.1391	1.1450	1.1621	1.0024	1.0068	1.0048	1.0144
			without	1.0087	1.1146	1.1302	1.1547	0.9286	0.9852	0.9918	1.0079
		0.05	with	0.9864	1.1048	1.1218	1.1427	0.7508	0.7856	0.7874	0.7944
			without	0.9295	1.0639	1.0974	1.1304	0.7076	0.7565	0.7703	0.7858
	12	0.002	with	2.2073	0.8930	1.0703	1.1030	0.6880	1.0419	1.0898	1.0797
			without	2.0007	0.8548	1.0543	1.0984	0.6236	0.9974	1.0734	1.0751
		0.005	with	0.7356	1.0834	1.0930	1.0855	0.9575	1.0595	1.0640	1.0598
			without	0.6920	1.0726	1.0869	1.0825	0.9007	1.0490	1.0581	1.0569
		0.01	with	0.9928	1.1002	1.0918	1.0989	1.0290	1.0571	1.0475	1.0533
			without	0.9618	1.0913	1.0864	1.0962	0.9968	1.0485	1.0424	1.0507
		0.02	with	1.0645	1.1066	1.1128	1.1179	1.0099	1.0221	1.0229	1.0240
			without	1.0371	1.0973	1.1071	1.1150	0.9839	1.0134	1.0177	1.0214
		0.05	with	1.0717	1.1413	1.1497	1.1541	0.8950	0.9088	0.9072	0.9044
			without	1.0300	1.1268	1.1411	1.1498	0.8601	0.8973	0.9003	0.9010
	6	0.002	with	1.0576	0.8340	1.0068	1.0402	0.6388	0.9842	1.0351	1.0306
			without	1.0140	0.8210	1.0017	1.0390	0.6125	0.9689	1.0298	1.0294
		0.005	with	0.6958	1.0242	1.0357	1.0313	0.9069	1.0165	1.0233	1.0203
			without	0.6843	1.0210	1.0340	1.0304	0.8919	1.0134	1.0217	1.0194
		0.01	with	0.9591	1.0431	1.0404	1.0456	0.9962	1.0245	1.0212	1.0259
			without	0.9502	1.0404	1.0388	1.0448	0.9870	1.0218	1.0196	1.0250
		0.02	with	1.0349	1.0535	1.0555	1.0565	1.0110	1.0161	1.0158	1.0152
			without	1.0260	1.0505	1.0537	1.0556	1.0023	1.0133	1.0141	1.0143
		0.05	with	1.0693	1.0974	1.0940	1.0966	0.9830	0.9854	0.9778	0.9768
			without	1.0568	1.0932	1.0914	1.0954	0.9715	0.9816	0.9755	0.9757
20	20	0.005	with	6.5689	0.9307	1.0799	1.1462	0.8111	1.0473	1.0660	1.0798
			without	0.0486	0.0100	0.0204	0.1213	0.0060	0.0113	0.0201	0.1143
		0.01	with	0.8208	1.0863	1.1140	1.1434	1.0235	1.0408	1.0386	1.0785
			without	0.0063	0.0211	0.0715	0.7562	0.0078	0.0202	0.0667	0.7133
		0.02	with	0.9580	1.0462	1.0828	1.1065	1.0775	0.9546	0.9876	1.0092
			without	0.0087	0.0460	0.2697	1.0407	0.0098	0.0420	0.2460	0.9492
		0.05	with	1.0826	0.9494	0.9932	1.0318	0.9424	0.7727	0.7989	0.8228
			without	0.0093	0.0574	0.4648	0.9707	0.0081	0.0469	0.3739	0.7741
		0.1	with	1.5195	0.7779	0.7401	0.8044	1.0844	0.5059	0.4731	0.5076
			without	0.0084	0.0108	0.0319	0.4371	0.0060	0.0070	0.0204	0.2758
	12	0.005	with	4.6356	0.9653	1.1151	1.1685	0.7962	1.1041	1.1163	1.1202
			without	2.1196	0.7678	0.9528	1.1325	0.3641	0.8781	0.9538	1.0857
		0.01	with	0.7693	1.1451	1.1943	1.1541	1.0049	1.1159	1.1364	1.1072
			without	0.5242	0.9225	1.1173	1.1320	0.6848	0.8990	1.0632	1.0860
		0.02	with	0.9290	1.1346	1.1525	1.1415	1.0346	1.0628	1.0793	1.0690
			without	0.1393	1.0650	1.1190	1.1252	0.1552	0.9976	1.0479	1.0537
		0.05	with	1.0457	1.1159	1.1301	1.1440	0.9468	0.9613	0.9650	0.9705
			without	0.1178	1.0606	1.0979	1.1281	0.1066	0.9136	0.9375	0.9570
		0.1	with	1.1279	1.0158	1.0630	1.0999	0.8796	0.7370	0.7608	0.7793
			without	0.1274	0.9311	1.0139	1.0756	0.0993	0.6755	0.7256	0.7621
	6	0.005	with	2.1401	0.8674	1.0333	1.0819	0.6864	1.0118	1.0521	1.0593
			without	1.9083	0.8238	1.0145	1.0761	0.6120	0.9610	1.0329	1.0536
		0.01	with	0.6514	1.0658	1.0856	1.0789	0.9019	1.0602	1.0589	1.0565
			without	0.5968	1.0491	1.0775	1.0749	0.8264	1.0435	1.0510	1.0526
		0.02	with	0.9241	1.0817	1.0917	1.0804	1.0172	1.0464	1.0558	1.0450
			without	0.8811	1.0707	1.0850	1.0771	0.9698	1.0358	1.0494	1.0418
		0.05	with	1.0749	1.1107	1.1087	1.1082	1.0218	1.0280	1.0213	1.0173
			without	1.0404	1.0991	1.1018	1.1047	0.9890	1.0173	1.0150	1.0141
		0.1	with	1.0816	1.1475	1.1607	1.1715	0.9478	0.9650	0.9684	0.9716
			without	1.0342	1.1313	1.1509	1.1666	0.9062	0.9514	0.9602	0.9675

12	12	0.01	with	1.5494	0.9613	1.0921	1.1220	0.8640	1.0415	1.0488	1.0632
			without	0.0308	0.0332	0.0748	0.4346	0.0172	0.0360	0.0718	0.4118
		0.02	with	0.7997	1.0854	1.0934	1.1146	1.0417	1.0219	1.0221	1.0506
			without	0.0178	0.0709	0.2467	0.9814	0.0232	0.0667	0.2306	0.9251
		0.05	with	1.0488	1.0228	1.0525	1.0914	1.0222	0.9117	0.9324	0.9625
			without	0.0280	0.1650	0.6476	1.0359	0.0273	0.1471	0.5737	0.9136
		0.1	with	1.1308	0.9235	0.9581	1.0174	0.9599	0.7304	0.7479	0.7866
			without	0.0247	0.1050	0.5279	0.9526	0.0209	0.0831	0.4121	0.7365
		0.2	with	2.2810	0.8550	0.6850	0.6862	1.5314	0.5209	0.4098	0.4050
			without	0.0303	0.0189	0.0288	0.1571	0.0203	0.0115	0.0172	0.0927
	6	0.01	with	0.9499	0.9780	1.0980	1.1116	0.7905	1.0646	1.0808	1.0788
			without	0.7208	0.8710	1.0666	1.0964	0.5999	0.9558	1.0499	1.0640
		0.02	with	0.7407	1.1184	1.1088	1.1086	0.9729	1.0840	1.0678	1.0728
			without	0.5557	1.0771	1.0879	1.0981	0.7298	1.0440	1.0478	1.0626
		0.05	with	1.0469	1.1247	1.1179	1.1259	1.0396	1.0541	1.0440	1.0487
			without	0.8620	1.0961	1.1012	1.1176	0.8560	1.0273	1.0283	1.0409
		0.1	with	1.0790	1.1446	1.1438	1.1663	0.9828	0.9993	0.9907	1.0043
			without	0.9130	1.1114	1.1243	1.1565	0.8316	0.9704	0.9738	0.9958
		0.2	with	0.9911	1.1248	1.1434	1.1687	0.7809	0.8315	0.8349	0.8458
			without	0.4536	1.0648	1.1082	1.1511	0.3574	0.7871	0.8092	0.8330
6	6	0.02	with	1.3052	0.9292	1.0551	1.0814	0.8161	1.0109	1.0215	1.0335
			without	0.0995	0.1230	0.2546	0.7499	0.0622	0.1338	0.2464	0.7167
		0.05	with	0.8246	1.0541	1.0680	1.0956	1.0279	0.9931	1.0072	1.0369
			without	0.0771	0.3016	0.6736	1.0372	0.0961	0.2841	0.6353	0.9817
		0.1	with	1.0508	1.0210	1.0655	1.0973	1.0341	0.9309	0.9659	0.9906
			without	0.1113	0.4321	0.8877	1.0508	0.1095	0.3939	0.8047	0.9487
		0.2	with	1.1693	0.9686	1.0012	1.0486	1.0262	0.7976	0.8146	0.8458
			without	0.0991	0.2976	0.7155	0.9894	0.0870	0.2451	0.5822	0.7980
		0.3	with	1.6117	0.8980	0.8738	0.9172	1.2775	0.6579	0.6307	0.6548
			without	0.1019	0.1198	0.2682	0.7597	0.0808	0.0878	0.1936	0.5424

Table 3. The cost-efficiency of combination with given n, k_1 and p when applying the cost function that is based on the total number of groups tested

k_1	n_1	p	k_2	Cost-efficiency of testing two stages combined by using the first approach	Cost-efficiency of testing two stages combined by using the second approach	Cost-efficiency of testing the first stage only
100	10	0.001	100	90477.58	85966.93	94441.29
			50	88420.55	88166.73	
			20	84600.95	85539.72	
			12	82427.75	83112.36	
			6	80133.99	80199.48	
		0.002	100	37775.78	38377.59	44575.75
			50	36557.15	36825.76	
			20	32392.92	32350.44	
			12	29532.24	29416.65	
			6	26368.76	26206.66	
		0.005	100	11314.99	11331.58	14691.45
			50	9771.42	9770.96	
			20	6730.21	6719.22	
			12	5051.38	5040.97	
			6	3338.39	3330.36	
		0.01	100	4019.76	4020.02	5182.43
			50	3294.70	3314.46	
			20	2043.34	2055.02	
			12	1408.56	1413.83	
			6	798.55	799.85	
		0.02	100	1808.32	1803.73	2171.72
			50	1255.77	1232.80	
			20	823.39	827.36	
			12	567.15	569.66	
			6	318.92	319.67	
	30	0.001	100	81556.98	82528.11	95791.75
			50	80569.21	80987.11	
			20	70025.37	70124.93	
			12	61234.89	61250.01	
			6	48920.69	48885.77	
		0.002	100	36931.00	36759.21	45390.12
			50	34300.66	34275.14	
			20	26161.46	26161.99	
			12	20513.34	20512.21	
			6	13624.21	13621.24	
		0.005	100	11306.08	11186.06	15370.14
			50	9698.56	9695.90	
			20	6262.71	6271.37	
			12	4448.99	4453.75	
			6	2581.04	2582.39	
		0.01	100	3994.55	3911.88	5719.04
			50	3383.60	3403.18	
			20	2053.16	2067.16	
			12	1408.02	1414.58	
			6	788.08	789.79	
		0.02	100	1094.88	1037.74	1382.26
			50	1071.87	1107.37	
			20	698.39	719.48	
			12	483.71	493.17	
			6	273.09	275.47	

	50	0.001	100	81402.93	81271.99	94920.14
			50	79004.16	79009.26	
			20	66711.00	66723.16	
			12	56475.84	56477.05	
			6	41548.19	41541.05	
		0.002	100	36986.66	36727.94	45150.72
			50	33887.88	33836.75	
			20	25484.80	25487.24	
			12	19748.93	19752.33	
			6	12724.36	12725.75	
		0.005	100	11396.56	11281.49	15421.48
			50	9672.44	9669.40	
			20	6207.74	6216.50	
			12	4392.84	4397.59	
			6	2538.50	2539.86	
		0.01	100	4089.04	4010.17	5819.44
			50	3404.02	3422.41	
			20	2056.16	2070.21	
			12	1408.54	1415.18	
			6	787.13	788.86	
		0.02	100	1093.93	1017.53	1448.26
			50	1081.01	1113.93	
			20	696.53	716.71	
			12	486.40	495.73	
			6	274.27	276.64	
	100	0.001	100	82625.24	82116.13	95589.58
			50	79310.50	79174.59	
			20	66319.70	66307.44	
			12	55484.86	55484.43	
			6	39458.50	39459.53	
		0.002	100	37549.98	37336.77	45815.92
			50	34235.22	34189.36	
			20	25435.50	25435.58	
			12	19652.21	19654.61	
			6	12518.39	12519.52	
		0.005	100	11507.81	11402.46	15484.68
			50	9658.01	9653.88	
			20	6183.23	6191.99	
			12	4374.48	4379.33	
			6	2519.21	2520.60	
		0.01	100	4117.46	4041.13	5872.50
			50	3403.78	3422.07	
			20	2043.11	2057.03	
			12	1401.10	1407.75	
			6	782.34	784.08	
		0.02	100	1144.10	1066.91	1529.68
			50	1096.01	1124.86	
			20	703.76	722.96	
			12	488.62	497.48	
			6	274.75	277.01	
50	10	0.002	50	22408.28	21247.26	23414.28
			20	21663.84	21721.91	
			12	21087.94	21315.08	
			6	20536.34	20679.39	
		0.005	50	7092.39	7234.63	8638.41
			20	6421.41	6450.13	
			12	5797.91	5792.77	
			6	4941.92	4922.01	
		0.01	50	2866.65	2871.32	3754.42
			20	2269.54	2268.76	
			12	1831.97	1829.69	
			6	1276.42	1274.00	
		0.02	50	1018.61	1018.20	1323.60
			20	766.37	771.21	
			12	573.04	576.28	
			6	356.92	357.98	
		0.05	50	478.28	463.54	587.95
			20	231.56	220.26	
			12	191.58	187.85	
			6	121.47	120.83	

	30	0.002	50	20398.33	20638.85	23906.49
			20	19560.91	19639.21	
			12	18076.25	18105.08	
			6	15356.13	15353.75	
		0.005	50	6999.33	6951.83	8747.65
			20	5955.01	5952.34	
			12	4958.39	4958.93	
			6	3515.44	3515.92	
		0.01	50	2845.69	2816.67	3861.64
			20	2231.07	2232.60	
			12	1744.07	1746.44	
			6	1117.84	1118.83	
		0.02	50	1010.99	992.63	1431.05
			20	769.15	775.07	
			12	579.12	583.25	
			6	353.99	355.37	
		0.05	50	190.40	175.48	222.40
			20	184.86	190.71	
			12	150.94	155.89	
			6	96.92	98.67	
	50	0.002	50	20679.59	20648.30	24032.76
			20	19435.10	19439.31	
			12	17674.83	17678.19	
			6	14350.27	14349.78	
		0.005	50	7038.22	6987.84	8816.77
			20	5945.64	5941.75	
			12	4944.28	4944.79	
			6	3456.55	3457.29	
		0.01	50	2871.63	2843.77	3889.69
			20	2223.26	2224.42	
			12	1728.04	1730.15	
			6	1104.35	1105.30	
		0.02	50	1025.49	1007.02	1456.25
			20	774.22	779.96	
			12	581.86	586.04	
			6	354.25	355.65	
		0.05	50	163.84	147.18	202.23
			20	177.17	185.29	
			12	144.38	150.48	
			6	92.95	94.99	
	100	0.002	50	20642.78	20519.69	23842.45
			20	19168.43	19148.75	
			12	17347.93	17343.40	
			6	13859.62	13859.70	
		0.005	50	7051.39	7009.04	8821.50
			20	5913.73	5909.94	
			12	4900.69	4901.00	
			6	3398.24	3398.86	
		0.01	50	2888.52	2861.95	3898.13
			20	2224.46	2225.70	
			12	1726.96	1729.19	
			6	1098.19	1099.17	
		0.02	50	1039.33	1022.29	1467.13
			20	776.14	782.23	
			12	581.04	585.30	
			6	352.81	354.23	
		0.05	50	167.85	149.11	212.91
			20	179.61	187.44	
			12	145.41	151.25	
			6	93.52	95.51	

20	10	0.005	20	3614.20	3599.15	3784.02
			12	3523.74	3383.44	
			6	3462.15	3486.08	
		0.01	20	1513.07	1538.56	1792.93
			12	1440.77	1448.39	
			6	1373.12	1375.20	
		0.02	20	635.73	640.44	813.35
			12	568.87	569.92	
			6	475.94	475.24	
		0.05	20	170.15	170.15	221.67
			12	153.00	152.84	
			6	109.36	109.92	
		0.1	20	81.46	81.30	99.19
			12	50.17	48.21	
			6	46.14	46.17	
	30	0.005	20	3233.47	3272.71	3805.64
			12	3215.71	3240.32	
			6	2987.20	2994.85	
		0.01	20	1485.72	1479.83	1818.39
			12	1385.43	1384.04	
			6	1200.04	1200.02	
		0.02	20	622.25	617.03	826.82
			12	566.75	565.81	
			6	443.27	443.50	
		0.05	20	168.25	165.38	240.07
			12	150.59	151.05	
			6	110.57	111.26	
		0.1	20	48.36	46.09	60.77
			12	47.70	48.52	
			6	38.89	40.02	
	50	0.005	20	3263.89	3259.77	3788.95
			12	3132.67	3132.80	
			6	2917.08	2917.91	
		0.01	20	1489.61	1479.93	1823.99
			12	1437.11	1434.50	
			6	1191.46	1191.29	
		0.02	20	631.15	626.25	833.51
			12	572.99	572.11	
			6	445.13	445.36	
		0.05	20	170.29	167.55	242.01
			12	150.76	151.23	
			6	109.99	110.68	
		0.1	20	47.30	44.32	66.07
			12	48.76	49.70	
			6	39.13	40.28	
	100	0.005	20	3291.94	3273.60	3801.81
			12	3207.39	3200.86	
			6	2915.06	2914.08	
		0.01	20	1502.58	1495.23	1824.73
			12	1408.49	1406.15	
			6	1186.61	1186.45	
		0.02	20	630.62	626.04	832.45
			12	568.60	567.63	
			6	440.78	440.98	
		0.05	20	171.40	168.78	244.42
			12	151.23	151.70	
			6	109.86	110.54	
		0.1	20	50.02	46.98	66.07
			12	50.06	50.93	
			6	39.72	40.82	

12	10	0.01	12	1048.57	1013.71	1142.19
			6	1036.57	1033.25	
		0.02	12	440.23	449.23	536.62
			6	419.40	422.24	
		0.05	12	133.84	133.69	177.66
			6	113.86	113.88	
		0.1	12	49.57	49.76	63.26
			6	40.47	40.77	
		0.2	12	35.44	34.65	44.47
			6	18.54	17.60	
	30	0.01	12	957.10	964.70	1136.23
			6	933.42	936.19	
		0.02	12	437.12	434.73	545.03
			6	398.04	397.79	
		0.05	12	133.79	132.52	185.74
			6	113.38	113.47	
		0.1	12	47.29	46.39	67.26
			6	40.81	41.18	
		0.2	12	14.37	13.56	16.62
			6	14.57	14.91	
	50	0.01	12	966.29	962.61	1134.02
			6	924.82	924.35	
		0.02	12	433.31	430.74	539.89
			6	391.28	390.95	
		0.05	12	133.71	132.49	185.50
			6	112.12	112.19	
		0.1	12	47.79	46.86	68.54
			6	40.74	41.09	
		0.2	12	12.17	11.17	14.89
			6	13.82	14.34	
	100	0.01	12	968.18	963.43	1134.59
			6	922.57	921.68	
		0.02	12	436.79	434.64	543.55
			6	392.71	392.41	
		0.05	12	135.34	134.17	187.07
			6	112.44	112.50	
		0.1	12	48.77	47.96	69.35
			6	41.01	41.38	
		0.2	12	12.30	11.18	15.40
			6	14.10	14.62	
6	10	0.02	6	263.61	256.49	288.45
		0.05	6	86.39	87.79	108.83
		0.1	6	36.07	36.04	48.47
		0.2	6	14.16	14.21	18.48
		0.3	6	10.62	10.70	13.30
	30	0.02	6	245.99	248.00	292.35
		0.05	6	86.11	85.68	110.63
		0.1	6	35.51	35.27	49.77
		0.2	6	13.43	13.28	19.03
		0.3	6	6.85	6.73	8.80
	50	0.02	6	247.81	247.07	291.40
		0.05	6	86.53	86.11	110.97
		0.1	6	36.21	35.97	50.34
		0.2	6	13.58	13.42	19.40
		0.3	6	6.91	6.72	9.19
	100	0.02	6	246.81	245.84	288.85
		0.05	6	87.01	86.65	111.06
		0.1	6	36.37	36.14	50.41
		0.2	6	13.81	13.66	19.70
		0.3	6	7.05	6.87	9.46

Table 5. The standard error, relative standard error, bias (x100000) and relative cost-efficiency of each 'most efficient' combination with given n_1, k_1 and p.

'*' indicates the criterion when the second approach is best

'#' indicates the criterion when the first approach is best

				First Approach				Second Approach			
k_1	n_1	p	k_2	Standard error of testing two stages combined	Relative standard error of testing two stages combined	Bias*100000	Relative Cost-Efficiency	Standard error of testing two stages combined	Relative standard error of testing two stages combined	Bias*100000	Relative Cost-Efficiency
100	10	0.001	50	0.0010*	1.02*	-0.32*	0.94*	0.0010	1.02	-0.97	0.93
100	10	0.002	100	0.0015	0.76	-0.76	0.85	0.0015	0.76	-0.69	0.86
100	10	0.005	100	0.0025	0.51	-1.06	0.77	0.0025	0.51	-1.02	0.77
100	10	0.01	100	0.0039	0.39	-6.07	0.78	0.0039	0.39	-6.09	0.78
100	10	0.02	100	0.0055	0.27	-102.12	0.83	0.0055	0.27	-108.90	0.83
100	30	0.001	100	0.0006	0.61	0.70	0.85	0.0006	0.61	-0.66	0.86
100	30	0.002	100	0.0009	0.44	0.27	0.81	0.0009	0.44	0.29	0.81
100	30	0.005	100	0.0015	0.29	0.40	0.74	0.0015	0.29	0.44	0.73
100	30	0.01	100	0.0023	0.23	1.49	0.70	0.0023	0.23	1.62	0.68
100	30	0.02	100	0.0040	0.20	1.65	0.79	0.0041#	0.21#	0.94#	0.750#
100	50	0.001	100	0.0005	0.47	0.54	0.86	0.0005	0.47	0.52	0.86
100	50	0.002	100	0.0007	0.34	0.15	0.82	0.0007	0.34	0.16	0.81
100	50	0.005	100	0.0011	0.22	0.57	0.74	0.0011	0.23	0.59	0.73
100	50	0.01	100	0.0017	0.17	0.46	0.70	0.0017	0.17	0.51	0.69
100	50	0.02	100	0.0031	0.16	2.76	0.76	0.0032#	0.16#	2.89#	0.70#
100	100	0.001	100	0.0003	0.33	0.16	0.86	0.0003	0.33	0.17	0.86
100	100	0.002	100	0.0005	0.24	-0.10	0.82	0.0005	0.24	-0.10	0.81
100	100	0.005	100	0.0008	0.16	-0.06	0.74	0.0008	0.16	-0.06	0.74
100	100	0.01	100	0.0012	0.12	-0.03	0.70	0.0012	0.12	0.02	0.69
100	100	0.02	100	0.0022	0.11	0.86	0.75	0.0022#	0.11#	0.70#	0.70#
50	10	0.002	20	0.0021*	1.03*	-0.51*	0.93*	0.0020	1.02	-1.73	0.93
50	10	0.005	50	0.0035	0.69	-0.04	0.82	0.0034	0.69	0.18	0.84
50	10	0.01	50	0.0050	0.50	-1.96	0.76	0.0050	0.50	-1.87	0.76
50	10	0.02	50	0.0078	0.39	-9.30	0.77	0.0078	0.39	-11.01	0.77
50	10	0.05	50	0.0104	0.21	-526.53	0.81	0.0106	0.21	-546.22	0.80
50	30	0.002	50	0.0012	0.61	0.86	0.85	0.0012	0.61	0.77	0.86
50	30	0.005	50	0.0020	0.40	0.37	0.80	0.0020	0.40	0.43	0.80
50	30	0.01	50	0.0029	0.29	0.49	0.74	0.0029	0.29	0.56	0.73
50	30	0.02	50	0.0045	0.22	2.41	0.71	0.0045	0.23	2.47	0.69
50	30	0.05	50	0.0095	0.19	-37.01	0.86	0.0099#	0.20#	-42.65#	0.80#
50	50	0.002	50	0.0009	0.47	0.47	0.86	0.0009	0.47	0.48	0.86
50	50	0.005	50	0.0015	0.31	-0.21	0.80	0.0015	0.31	-0.18	0.79
50	50	0.01	50	0.0022	0.22	0.04	0.74	0.0022	0.22	0.09	0.73
50	50	0.02	50	0.0035	0.17	0.79	0.70	0.0035	0.17	0.81	0.69
50	50	0.05	20	0.0058	0.12	7.62	0.88	0.0057	0.11	7.72	0.92
50	100	0.002	50	0.0007	0.33	0.13	0.87	0.0007	0.33	0.14	0.86
50	100	0.005	50	0.0011	0.22	0.25	0.80	0.0011	0.22	0.26	0.79
50	100	0.01	50	0.0016	0.16	0.31	0.74	0.0016	0.16	0.33	0.73
50	100	0.02	50	0.0024	0.12	0.23	0.71	0.0024	0.12	0.26	0.70
50	100	0.05	20	0.0041	0.08	5.93	0.84	0.0040	0.08	6.00	0.88
20	10	0.005	20	0.0051	1.03	-1.26	0.96	0.0053	1.06	-0.53	0.90
20	10	0.01	20	0.0076	0.76	2.16	0.84	0.0076	0.76	2.33	0.86
20	10	0.02	20	0.0110	0.55	-4.61	0.78	0.0109	0.55	-3.99	0.79
20	10	0.05	20	0.0190	0.38	-21.60	0.77	0.0190	0.38	-25.15	0.77
20	10	0.1	20	0.0256	0.26	-535.51	0.82	0.0256	0.26	-566.46	0.82
20	30	0.005	20	0.0031	0.62	4.51	0.85	0.0031	0.61	4.22	0.86
20	30	0.01	20	0.0044	0.44	6.04	0.82	0.0044	0.44	6.14	0.81
20	30	0.02	20	0.0064	0.32	7.40	0.75	0.0064	0.32	7.60	0.75
20	30	0.05	20	0.0110	0.22	9.83	0.70	0.0111	0.22	10.44	0.69
20	30	0.1	20	0.0193	0.19	6.06	0.79	0.0197#	0.20#	4.26#	0.75#

20	50	0.005	20	0.0024	0.47	2.60	0.86	0.0024	0.47	2.62	0.86	
20	50	0.01	20	0.0034	0.34	3.30	0.82	0.0038	0.34	3.40	0.81	
20	50	0.02	20	0.0049	0.24	2.43	0.76	0.0049	0.25	2.50	0.75	
20	50	0.05	20	0.0085	0.17	4.05	0.70	0.0085	0.17	4.18	0.69	
20	50	0.1	12	0.0129	0.13	37.36	0.78	0.0128	0.13	40.91	0.79	
20	100	0.005	20	0.0017	0.33	-0.14	0.87	0.0017	0.33	-0.11	0.86	
20	100	0.01	20	0.0024	0.24	-0.18	0.82	0.0024	0.24	0.19	0.82	
20	100	0.02	20	0.0035	0.17	0.32	0.76	0.0035	0.17	0.35	0.75	
20	100	0.05	20	0.0060	0.12	-1.09	0.70	0.0060	0.12	-1.03	0.69	
20	100	0.1	12	0.0090	0.09	11.54	0.76	0.0089	0.09	13.65	0.77	
12	10	0.01	6	0.0093*	0.93*	-10.45*	0.91*	0.0093	0.93	-11.17	0.91	
12	10	0.02	12	0.0139	0.70	-3.01	0.82	0.0138	0.69	-2.42	0.84	
12	10	0.05	12	0.0228	0.46	-3.68	0.75	0.0228	0.46	-4.04	0.75	
12	10	0.1	12	0.0344	0.34	-90.72	0.78	0.0343	0.34	-101.58	0.79	
12	10	0.2	12	0.0383	0.19	-2018.71	0.80	0.0387	0.19	-2078.97	0.80	
12	30	0.01	12	0.0056	0.56	8.81	0.84	0.0056	0.56	8.50	0.85	
12	30	0.02	12	0.0079	0.40	5.22	0.80	0.0080	0.40	5.44	0.80	
12	30	0.05	12	0.0131	0.26	7.06	0.72	0.0132	0.26	7.30	0.71	
12	30	0.1	12	0.0203	0.20	14.65	0.70	0.0205	0.20	15.79	0.69	
12	30	0.2	6	0.0283	0.14	-35.71	0.88	0.0280	0.14	-31.34	0.90	
12	50	0.01	12	0.0043	0.43	5.27	0.85	0.0043	0.43	5.34	0.85	
12	50	0.02	12	0.0062	0.31	4.57	0.80	0.0062	0.31	4.71	0.80	
12	50	0.05	12	0.0101	0.20	3.65	0.72	0.0102	0.20	3.87	0.71	
12	50	0.1	12	0.0156	0.16	6.21	0.70	0.0158	0.16	7.01	0.68	
12	50	0.2	6	0.0225	0.11	0.94	0.93	0.0221	0.11	0.53	0.96	
12	100	0.01	12	0.0030	0.30	2.91	0.85	0.0031	0.31	2.94	0.85	
12	100	0.02	12	0.0043	0.22	1.68	0.80	0.0044	0.22	1.70	0.80	
12	100	0.05	12	0.0071	0.14	4.53	0.72	0.0071	0.14	4.58	0.72	
12	100	0.1	12	0.0109	0.11	0.51	0.70	0.0110	0.11	1.05	0.69	
12	100	0.2	6	0.0157	0.08	1.66	0.92	0.0155	0.08	1.55	0.95	
6	10	0.02	6	0.0189	0.94	2.85	0.91	0.0191	0.96	3.61	0.89	
6	10	0.05	6	0.0307	0.61	3.48	0.79	0.0305	0.61	3.25	0.81	
6	10	0.1	6	0.0438	0.44	-23.99	0.74	0.0438	0.44	-24.28	0.74	
6	10	0.2	6	0.0640	0.32	-174.21	0.77	0.0639	0.32	-189.09	0.77	
6	10	0.3	6	0.0709	0.24	-1217.37	0.80	0.0706	0.24	-1272.17	0.80	
6	30	0.02	6	0.0111	0.55	15.40	0.84	0.0110	0.55	14.62	0.85	
6	30	0.05	6	0.0175	0.35	0.23	0.78	0.0176	0.35	9.59	0.77	
6	30	0.1	6	0.0253	0.25	11.99	0.71	0.0254	0.25	12.74	0.71	
6	30	0.2	6	0.0378	0.19	38.19	0.71	0.0380	0.19	39.08	0.70	
6	30	0.3	6	0.0509	0.17	16.27	0.78	0.0513	0.17	12.52	0.76	
6	50	0.02	6	0.0085	0.43	10.14	0.85	0.0085	0.43	10.32	0.85	
6	50	0.05	6	0.0135	0.27	8.12	0.78	0.0136	0.27	8.32	0.78	
6	50	0.1	6	0.0194	0.19	4.88	0.72	0.0195	0.19	5.36	0.71	
6	50	0.2	6	0.0291	0.15	1.73	0.70	0.0293	0.15	2.11	0.69	
6	50	0.3	6	0.0392	0.13	12.97	0.75	0.0398	0.13	14.11	0.73	
6	100	0.02	6	0.0060	0.30	3.35	0.85	0.0060	0.30	3.42	0.85	
6	100	0.05	6	0.0095	0.19	0.71	0.78	0.0096	0.19	0.73	0.78	
6	100	0.1	6	0.0137	0.14	0.06	0.72	0.0137	0.14	0.30	0.72	
6	100	0.2	6	0.0204	0.10	-8.26	0.70	0.0205	0.10	0.81	0.69	
6	100	0.3	6	0.0275	0.09	23.50	0.75	0.0278	0.09	24.35	0.73	

Table 7. The cost-efficiency of combination with given n_1k_1 and p when applying the cost function that is based on the total sample size

n_1k_1	p	k_1	n_1	k_2	Cost-efficiency of testing two stages combined by using the first approach	Cost-efficiency of testing two stages combined by using the second approach	Cost-efficiency of testing the first stage only
600	0.02	6	100	6	45.80	45.62	48.14
		12	50	12	43.78	43.53	44.99
		12	50	6	46.34	46.30	44.99
		20	30	20	41.30	40.95	41.34
		20	30	12	44.00	43.92	41.34
		20	30	6	46.12	46.14	41.34
	0.05	6	100	6	18.45	18.45	18.44
		12	50	12	16.23	16.08	15.46
		12	50	6	17.84	17.85	15.46
		20	30	20	13.77	13.54	12.00
		20	30	12	15.59	15.64	12.00
		20	30	6	17.26	17.37	12.00
	0.1	6	100	6	8.89	8.84	8.40
		12	50	12	6.83	6.70	5.71
		12	50	6	8.24	8.31	5.71
		20	30	20	4.54	4.32	3.04
		20	30	12	5.89	5.99	3.04
		20	30	6	7.64	7.86	3.04
1000	0.005	20	50	20	178.38	178.15	189.45
		20	50	12	182.75	182.77	189.45
		20	50	6	190.48	190.53	189.45
		100	10	100	155.01	155.24	146.91
		100	10	50	167.37	167.36	146.91
		100	10	20	178.09	177.80	146.91
		100	10	12	181.25	180.88	146.91
		100	10	6	183.74	183.30	146.91
	0.01	20	50	20	87.87	87.30	91.20
		20	50	12	93.58	93.41	91.20
		20	50	6	94.93	94.92	91.20
		100	10	100	65.07	65.07	51.82
		100	10	50	73.23	73.67	51.82
		100	10	20	81.87	82.34	51.82
		100	10	12	84.56	84.88	51.82
		100	10	6	86.75	86.90	51.82
	0.02	20	50	20	41.95	41.62	41.68
		20	50	12	44.03	43.97	41.68
		20	50	6	46.55	46.57	41.68
		100	10	100	33.65	33.57	21.72
		100	10	50	34.12	33.49	21.72
		100	10	20	43.46	43.67	21.72
		100	10	12	46.24	46.44	21.72
		100	10	6	48.65	48.77	21.72

5000	0.002	50	100	50	451.82	449.13	476.85
		50	100	20	473.88	473.39	476.85
		50	100	12	482.31	482.18	476.85
		50	100	6	488.70	488.70	476.85
		100	50	100	436.02	432.97	451.51
		100	50	50	458.85	458.15	451.51
		100	50	20	475.92	475.96	451.51
		100	50	12	480.66	480.74	451.51
		100	50	6	485.05	485.10	451.51
	0.005	50	100	50	172.09	171.06	176.43
		50	100	20	183.29	183.18	176.43
		50	100	12	187.42	187.43	176.43
		50	100	6	190.84	190.88	176.43
		100	50	100	158.52	156.92	154.21
		100	50	50	171.98	171.93	154.21
		100	50	20	181.95	182.21	154.21
		100	50	12	185.11	185.31	154.21
		100	50	6	187.81	187.91	154.21
	0.01	50	100	50	80.49	79.75	77.96
		50	100	20	88.20	88.25	77.96
		50	100	12	90.98	91.10	77.96
		50	100	6	93.48	93.56	77.96
		100	50	100	66.70	65.41	58.19
		100	50	50	76.92	77.34	58.19
		100	50	20	85.16	85.74	58.19
		100	50	12	87.86	88.28	58.19
		100	50	6	90.20	90.40	58.19
	0.02	50	100	50	33.97	33.42	29.34
		50	100	20	40.15	40.46	29.34
		50	100	12	42.32	42.64	29.34
		50	100	6	44.28	44.46	29.34
		100	50	100	20.42	18.99	14.48
		100	50	50	29.53	30.43	14.48
		100	50	20	37.10	38.18	14.48
		100	50	12	39.96	40.73	14.48
		100	50	6	42.30	42.67	14.48

Table 8. The standard error, relative standard error, bias (x100000) and relative cost-efficiency of each 'most efficient' combination with given n_1k_1 and p.

					First Approach				Second Approach			
n1k1	k1	n1	p	k2	Standard error of testing two stages combined	Relative standard error of testing two stages combined	Bias*100000	Relative Cost-Efficiency	SE(overall phat)	Relative SE	Bias*100000	Relative Cost-Efficiency
600	12	50	0.02	6	0.0060	0.30	1.99	1.03	0.0060	0.30	2.01	1.03
600	6	100	0.05	6	0.0095	0.18	0.43	1.00	0.0096	0.18	0.56	1.00
600	6	100	0.1	6	0.0137	0.14	0.06	1.06	0.0137	0.14	0.30	1.05
1000	20	50	0.005	6	0.0023	0.46	0.99	1.01	0.0023	0.46	0.98	1.01
1000	20	50	0.01	6	0.0032	0.32	2.60	1.04	0.0032	0.32	2.62	1.04
1000	100	10	0.02	6	0.0045	0.23	-14.30	2.24	0.0045	0.23	-13.91	2.25
5000	50	100	0.002	6	0.0006	0.32	-0.16	1.02	0.0006	0.32	-0.16	1.02
5000	50	100	0.005	6	0.0010	0.20	0.10	1.08	0.0010	0.20	0.11	1.08
5000	50	100	0.01	6	0.0015	0.15	0.20	1.20	0.0015	0.15	0.20	1.20
5000	50	100	0.02	6	0.0021	0.11	0.37	1.51	0.0021	0.11	0.38	1.52

Bibliography

Berger, T., Mehravari, N., Towsley, D., & Wolf, J. (1984). Random Multiple-Access Communication and Group Testing. *IEEE Transactions on Communications*, **32**(7), 769–779.

Bilder, C. R., Tebbs, J. M., & Chen, P. (2010). Informative Retesting. *Journal of the American Statistical Association*, **105**(491), 942–955.

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

Burns, K. C., & Mauro, C. A. (1987). Group testing with test error as a function of concentration. *Communications in Statistics - Theory and Methods*, **16**(10), 2821–2837.

Burrows, P. M. (1987). Improved Estimation of Pathogen Transmission Rates by Group Testing. *Phytopathology*, **77**(2), 363

Chen, C. L., & Swallow, W. H. (1990). Using Group Testing to Estimate a Proportion, and to Test the Binomial Model. *Biometrics*, **46**(4), 1035.

Comanor, L., & Holland, P. (2006). Hepatitis B virus blood screening: unfinished agendas. *Vox Sanguinis*, **91**(1), 1–12.

Dorfman, R. (1943). The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, **14**(4), 436–440.

Gastwirth, J. L., & Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *Journal of Statistical Planning and Inference*, **22**(1), 15–27.

Gastwirth, J. L., & Johnson, W. O. (1994). Screening with Cost-Effective Quality Control: Potential Applications to HIV and Drug Testing. *Journal of the American Statistical Association*, **89**(427), 972–981.

Graff, L. E., & Roeloffs, R. (1972). Group Testing in the Presence of Test Error; An Extension of the Dorfman Procedure. *Technometrics*, **14**(1), 113–122.

Guthrie, R. (1961). Blood Screening for Phenylketonuria. *Jama*, **178**(8), 863.

Hammick, P. A., & Gastwirth, J. L. (1994). Group Testing for Sensitive Characteristics: Extension to Higher Prevalence Levels. *International Statistical Review / Revue Internationale De Statistique*, **62**(3), 319.

Hepworth, G. (1996). Exact Confidence Intervals for Proportions Estimated by Group Testing. *Biometrics*, **52**(3), 1134.

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.

Hepworth, G., & Watson, R. (2015). Revisiting retesting in the estimation of proportions by group testing. *Communications in Statistics - Simulation and Computation*, **46**(1), 261–274.

Hong, E. S., Ladner, R. E., & Riskin, E. A. (2003). Group testing for image compression using alternative transforms. *Signal Processing: Image Communication*, **18**(7), 561–574.

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.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). Univariate Discrete Distributions. *Wiley Series in Probability and Statistics*.

Kim, H.-Y., 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.

Le, C. T. (1981). A New Estimator For Infection Rates Using Pools Of Variable Size. *American Journal of Epidemiology*, **114**(1), 132–136.

Liu, A., Liu, C., Zhang, Z., & Albert, P. S. (2011). Optimality of group testing in the presence of misclassification. *Biometrika*, **99**(1), 245–251.

Malinovsky, Y., & Albert, P. S. (2018). Revisiting Nested Group Testing Procedures: New Results, Comparisons, and Robustness. *The American Statistician*, **73**(2), 117–125.

Ngo, H., & Du, D.-Z. (2000). A survey on combinatorial group testing algorithms with applications to DNA Library Screening. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science Discrete Mathematical Problems with Medical Applications*, 171–182.

Schliep, A., Torney, D., & Rahmann, S. (2003). Group testing with DNA chips: generating designs and decoding experiments. *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*.

Schultz, J. R., Nichol, F. R., Elfring, G. L., & Weed, S. D. (1973). Multiple-Stage Procedures for Drug Screening. *Biometrics*, **29**(2), 293.

Sobel, M., & Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika*, **62**(1), 181–193.

Sterrett, A. (1957). On the Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, **28**(4), 1033–1036.

Thai, M. T. (2012). *Group Testing Theory in Network Security An Advanced Solution*. New York, NY: Springer New York.

Walter, S. D., & Hepworth, G. (2019). Estimation of proportions by group testing with retesting of positive groups. *Communications in Statistics - Theory and Methods*, 1–11.

Walter, S. D., Hildreth, S. W., & Beaty, B. J. (1980). Estimation Of Infection Rates In Populations Of Organisms Using Pools Of Variable Size. *American Journal of Epidemiology*, **112**(1), 124–128.

Xuan, Y., Shin, I., Thai, M. T., & Znati, T. (2010). Detecting Application Denial-of-Service Attacks: A Group-Testing-Based Approach. *IEEE Transactions on Parallel and Distributed Systems*, **21**(8), 1203–1216.