

Effects of Group Size on the MLE of Proportion in Group Testing

Jie Mi

Department of Mathematics and Statistics
Florida International University
Miami, FL 33199 USA
mi@fiu.edu

ABSTRACT

Under some conditions the proportion p of a certain trait can be estimated based on group testing. The method of group testing has several merits such as saving costs, protecting privacy etc. and has been applied to many fields. Its performance certainly depends on the selected group size. In particular, it was observed from practice that too large group size does not do a good job. There are lots of work on selecting the “optimal” group size. This article studies the effects of group size on the performance of the MLE of p by considering its bias and limiting distribution in the case of non-perfect group testing.

Keywords: Group testing, Specificity, Sensitivity, Overestimation, Convergence in law.

2000 Mathematics Subject Classification: 62F15, 62P10.

1 Introduction

The history of group testing can be traced back to (Dorfman, 1943). A great deal of research work appeared since then. We list just a few of them among others: (Thompson, 1962), (Swallow, 1985), (Gastwirth and Hammick, 1989), (Gastwirth and Johnson, 1994), (Hughes-Oliver and Swallow, 1994), (Tu and Litvak, 1994), (Tu, Litvak and Pagano, 1995), (Lui, 1995), (Gupta and Malina, 1999), (Tebbs and Swallow, 2003), (Hepworth and Watson, 2009), (Pritchard and Tebbs, 2011a), (Pritchard and Tebbs, 2011b), (Hepworth, 2013), (Tebbs, McMahan and Bilder, 2013), (Huang and Mi, 2019), (Mi, 2019). Group testing has a long history of successful applications in various areas. To fix our idea let's just consider a rare disease. Let $\mathbb{T}$ be a screening test. Testing each individual (experimental) unit has a binary response: positive or negative, for example. By group testing, however, the units are classified into groups each of which includes k units, and then the test $\mathbb{T}$ is applied to the groups instead of each unit individually. It is assumed in group testing that a group has a positive response, if and only if, at least one unit in the group would have a positive response if it was tested individually. This assumption holds in many areas such as human infectious disease and drug discovery. In such cases, individual specimens such as blood or urine, can be combined together without loss of test accuracy.

Let ξ be a Bernoulli random variable, $(\xi = 1)$ and $(\xi = 0)$ denote the events that a randomly selected unit has the disease or not, respectively. Our target parameter is the proportion $p \equiv P(\xi = 1)$, the proportion of units that have the rare disease in a population. Let D be the event that a group selected at random includes at least one unit with the disease, and $\bar{D}$ be its complement. Moreover, let $+$ and $-$ be the events that a group selected at random is diagnosed as positive or negative, respectively. Note that the following two conditional probabilities in performing test $\mathbb{T}$ are critical. One is the *sensitivity* of the test, the other is the *specificity*. They are defined as $\pi_1 \equiv P(+|D)$ and $\pi_0 \equiv P(-|\bar{D})$, respectively. Practitioners usually assume that both the sensitivity and specificity remain the same for different group sizes (see, for instance, (Tu et al., 1995), (Hung and Swallow, 1999)). Considering the above, the probability that a group selected at random has positive response with the test $\mathbb{T}$ is given as $\theta \equiv \theta_k \equiv P(+) = \pi_1 - r(1 - p)^k$, where k is the group size and $r \equiv \pi_1 + \pi_0 - 1$. In general, both π_1 and π_0 are close to 1 and consequently $r > 0$. We will keep these assumptions throughout the paper. It is easy to see that a perfect test with $\pi_1 = \pi_0 = 1$ can be characterized by $r = 1$. Most research works on group testing focus on the MLE of p . It is obvious that the performance of the MLE of p depends on the group size in addition to other factors. It sounds reasonable intuitively to believe that the MLE would overestimate p when k is large since it will make θ close to π_1 and thus to 1, so most tested groups would show positive response. The purpose of this article is to explore the effect of group size on the performance of the MLE by considering its bias and limit distribution as $k \rightarrow \infty$. It is hoped that the obtained results could validate the intuitive thoughts.

This article focuses on non-perfect test $\mathbb{T}$ with $0 < r < 1$. (Graff and Roeloffs, 1972) was probably the earliest work considering non-perfect group testing. More works appeared since then. For example, (Rogan and Gladen, 1978), (Gastwirth, 1987), (Lew and Levy, 1989), (Gastwirth, Johnson and Reneau, 1991), (Tu et al., 1995), (Pfeiffer, Rutter, Gail, Struewing and Gastwirth, 2002), (Liu, Liu, Zhang and Albert, 2012) among others. In Section 2 we will study the case of binomial sampling. Ranges of the group size k , within which the MLE p overestimates p , are given; and the limit distribution of the MLE of p is also derived there. The similar results are established for the case of the negative binomial sampling in Section 3. A summary is stated in the last section.

2 Binomial Sampling

Let n and k be two predetermined positive integers. Suppose $N = nk$ experimental units are classified into n groups each of which includes k units. Define binary random variable $X_i, i = 1, 2, \dots, n$ by

$$X_i = 1, \text{ if the test result of the } i\text{th group is } +; \quad X_i = 0, \text{ otherwise.} \quad (2.1)$$

Denote $T \equiv \sum_{i=1}^n X_i$. Clearly, T follows binomial distribution, i.e., $T \sim B(n, \theta)$ where $\theta = \pi_1 - r(1 - p)^k$.

2.1 When does the MLE of p Overestimate p ?

It is known that the MLE of p , denoted as $\hat{p}$, is given as

$$\hat{p} = \hat{p}(T) = 1 - \left[\frac{\pi_1 - \min\{\pi_1, \max(1 - \pi_0, \frac{T}{n})\}}{r} \right]^{1/k}. \quad (2.2)$$

See, for instance, (Tu and Litvak, 1994), (Tu et al., 1995), (Liu et al., 2012), (Ding and Xiong, 2015), (Xiong and Ding, 2015).

Obviously $\hat{p}(t)$ increases with t for any given k, n, π_0, π_1 and p . It means that larger observation of T yields larger estimate $\hat{p}(t)$. It was known in the literature (see, for instance, (Hepworth, 1996)) that $\hat{p}(T)$ overestimates p for any group size $k \geq 1$ if both $\pi_0 = 1$ and $\pi_1 = 1$. We will show in the following that the same result is true for the imperfect test when k is in a certain range.

Definition 2.1. A finite real sequence $\{a_i, i = 0, 1, \dots, n\}$ is a convex sequence if $2a_i \leq a_{i-1} + a_{i+1}, \forall i = 1, 2, \dots, n-1$.

Lemma 2.1. Let $a_i = \hat{p}(i), i = 0, 1, \dots, n$ defined by (2.2). Suppose that $\pi_0 \geq 1 - 1/n$ and

$$\pi_1 \geq 1 - \left(\frac{2^k - 2}{2^k - 1} \right) \frac{1}{n}, \quad (2.3)$$

then $\{a_i, i = 0, 1, \dots, n\}$ is a convex sequence.

Remark 2.1. Note the fraction

$$\frac{2^k - 2}{2^k - 1}$$

increased in k so if

$$\pi_1 \geq 1 - \left(\frac{2}{3} \right) \frac{1}{n},$$

then the assumption (2.3) will be true for all $k \geq 2$.

Proof. From the expression of $\hat{p}(t)$ the following is true

$$\hat{p}(t) = \begin{cases} 0, & \text{if } 0 \leq t < n(1 - \pi_0); \\ 1 - \left[\frac{\pi_1 - t/n}{r} \right]^{1/k}, & \text{if } n(1 - \pi_0) \leq t \leq n\pi_1; \\ 1, & \text{if } n\pi_1 < t \leq n. \end{cases} \quad (2.4)$$

Under the assumption

$$\pi_1 \geq 1 - \left(\frac{2^k - 2}{2^k - 1} \right) \frac{1}{n}$$

we have $\pi_1 > 1 - 1/n$ and so $n\pi_1 > n - 1$. Hence the expression (2.4) can be rewritten as

$$\hat{p}(t) = \begin{cases} 0, & \text{if } 0 \leq t \leq j; \\ 1 - \left[\frac{\pi_1 - t/n}{r} \right]^{1/k}, & \text{if } j + 1 \leq t \leq n - 1; \\ 1, & \text{if } t = n, \end{cases} \quad (2.5)$$

where $j \equiv \lfloor n(1 - \pi_0) \rfloor$ is the largest integer part of $n(1 - \pi_0)$.

Further, we define function $\psi(t)$ on the interval $[j+1, n-1]$ as

$$\psi(t) \equiv 1 - \left[\frac{\pi_1 - t/n}{r} \right]^{1/k}, j+1 \leq t \leq n-1. \quad (2.6)$$

Obviously, $a_i = \hat{p}(i) = \psi(i), i = j+1, \dots, n-1$. The function $\psi(t)$ is clearly increasing in $t \in [j+1, n-1]$. Moreover, the second derivative of $\psi(t)$ is given as

$$\psi''(t) = \frac{1-1/k}{k(nr)^2} \cdot \left[\frac{\pi_1 - t/n}{r} \right]^{1/k-2} > 0,$$

This shows that $\psi(t)$ is a convex function of $t \in [j+1, n-1]$. This immediately implies that the sequence $\{a_i, i = j+1, 1, \dots, n-1\}$ is a convex sequence.

In order to prove the desired result we need to show (i) $2a_{j+1} \leq a_j + a_{j+2}$; and (ii) $2a_{n-1} \leq a_{n-2} + a_n$.

Case (i): Note that $\pi_0 \geq 1 - 1/n$ implies $n(1 - \pi_0) \leq 1$ and so $j = \lfloor n(1 - \pi_0) \rfloor = 0$ or 1 . The inequality $2a_{j+1} \leq a_j + a_{j+2}$ is equivalent to

$$2 \left(1 - \left[\frac{\pi_1 - \frac{j+1}{n}}{r} \right]^{1/k} \right) \leq 1 - \left[\frac{\pi_1 - \frac{j+2}{n}}{r} \right]^{1/k}. \quad (2.7)$$

The inequality is equivalent to

$$1 + \left[\frac{\pi_1 - \frac{j+2}{n}}{r} \right]^{1/k} \leq 2 \left[\frac{\pi_1 - \frac{j+1}{n}}{r} \right]^{1/k}. \quad (2.8)$$

or

$$r^{1/k} + \left(\pi_1 - \frac{j+2}{n} \right)^{1/k} \leq 2 \left(\pi_1 - \frac{j+1}{n} \right)^{1/k}. \quad (2.9)$$

Let function $f(x) \equiv x^{1/k}$. It is easy to see that $f(x)$ is an increasing concave function of $x > 0$.

Taking $x_1 = r$ and $x_2 = \pi_1 - (j+2)/n > 0$, we thus have

$$f(x_1) + f(x_2) \leq 2f\left(\frac{x_1 + x_2}{2}\right) \quad (2.10)$$

or

$$\begin{aligned} r^{1/k} + \left(\pi_1 - \frac{j+2}{n} \right)^{1/k} &\leq 2 \left(\pi_1 - \frac{j+2}{2n} - \frac{1-\pi_0}{2} \right)^{1/k} \\ &\leq 2 \left(\pi_1 - \frac{j+1}{n} \right)^{1/k}, \end{aligned} \quad (2.11)$$

where the last inequality is true because $f(x)$ is an increasing function of $x > 0$, $\pi_0 \geq 1 - 1/n$ and $j = 0, 1$. This shows that (i) is true.

Case (ii): The inequality $2a_{n-1} \leq a_{n-2} + a_n$ is equivalent to

$$2 \left(1 - \left[\frac{\pi_1 - \frac{n-1}{n}}{r} \right]^{1/k} \right) \leq 1 + \left(1 - \left[\frac{\pi_1 - \frac{n-2}{n}}{r} \right]^{1/k} \right), \quad (2.12)$$

or

$$\pi_1 - \frac{n-2}{n} \leq 2^k \left(\pi_1 - \frac{n-1}{n} \right). \quad (2.13)$$

The inequality (2.13) is true if and only if

$$\pi_1 \geq 1 - \frac{2^k - 2}{2^k - 1} \cdot \frac{1}{n} \quad (2.14)$$

and therefore the result (ii) follows. $\square$

Theorem 2.2. Suppose that $\pi_0 > 1 - 1/n$ and $\pi_1 > 1 - (2/3)n^{-1}$.

If further

$$k \leq \frac{\ln(r^{-1}(\pi_1 - (1 - n^{-1})))}{\ln(1 - p)}, \quad (2.15)$$

then $\hat{p}$ overestimates p , i.e., $E(\hat{p}) \geq p$. In particular if

$$k \leq \frac{-\ln(n - 5/3)}{\ln(1 - p)}, \quad (2.16)$$

then the inequalities (2.15) hold.

Proof. We keep using the notation in the proof of Lemma 2.1. Clearly $j = \lfloor n(1 - \pi_0) \rfloor = 0$ and $n\pi_1 > n - 1$. Hence the function $\psi(t)$ given in (2.6) is defined on the interval $[1, n - 1]$. Moreover, the domain of $\psi(t)$ can be extended to $[n(1 - \pi_0), n - 1]$ and $\psi(n(1 - \pi_0)) = 0$. To see this note that $0 < n(1 - \pi_0) < 1$ since $\pi_0 > 1 - 1/n$, hence $\pi_1 - t/n \leq r$ and $\psi(t)$ is well defined for any $t \in [n(1 - \pi_0), n - 1]$. It is easy to see that the function $\psi(t)$ is again convex on the extended domain $[n(1 - \pi_0), n - 1]$. The expression (2.5) can now be rewritten as

$$\hat{p}(t) = \begin{cases} 0, & \text{if } t = 0; \\ 1 - [\frac{\pi_1 - t/n}{r}]^{1/k}, & \text{if } 1 \leq t \leq n - 1 \\ 1, & \text{if } t = n. \end{cases} \quad (2.17)$$

Denote the point with coordinates (t, a_t) , $t = 0, 1, \dots, n$ in (t, a) -plane as A_t , and denote the line segment connecting points A_t and A_{t+1} as $\ell_{t,t+1}$, $t = 0, 1, \dots, n - 1$. Note that according to Remark 2.1 the inequality (2.3) is true. Let $g(t)$ be the function with the graph consisted by $\ell_{0,1}, \ell_{1,2}, \dots, \ell_{n-1,n}$. Then $g(t)$ must be a convex function of $t \in [0, n]$ due to Lemma 2.1. It is also obvious that $g(t) = a_t = \hat{p}(t)$, $\forall t = 0, 1, \dots, n$. Moreover, $g(t) \geq \psi(t)$, $\forall t \in [n(1 - \pi_0), n - 1]$ since $g(t) = \psi(t)$, $\forall t = 1, \dots, n - 1$, $\psi(n(1 - \pi_0)) = 0$, and $\psi(t)$ is convex on the interval $[n(1 - \pi_0), n - 1]$.

By Jensen's Inequality then we have $Eg(T) \geq g(E(T))$. Note that

$$E(\hat{p}(T)) = \sum_{t=0}^n \hat{p}(t)P(T=t) = \sum_{t=0}^n a_t P(T=t) = \sum_{t=0}^n g(t)P(T=t) = E(g(T))$$

since $a_t = g(t)$, $t = 0, 1, \dots, n$. Hence it holds that $E(\hat{p}(T)) = E(g(T)) \geq g(E(T))$.

The mean of T is $E(T) = n\theta = n(\pi_1 - r(1 - p)^k)$. Below we are going to show that $n(1 - \pi_0) < E(T) \leq n - 1$.

The inequality $n(1 - \pi_0) \leq E(T)$ is equivalent to

$$\begin{aligned} n(1 - \pi_0) &< n(\pi_1 - r(1 - p)^k); \\ r(1 - p)^k &< r. \end{aligned}$$

It is certainly true.

On the other hand, $E(T) \leq n - 1$ is equivalent to

$$\begin{aligned} n(\pi_1 - r(1 - p)^k) &\leq n - 1; \\ k &\leq \frac{\ln(r^{-1}(\pi_1 - (1 - n^{-1})))}{\ln(1 - p)}. \end{aligned}$$

Thus from the assumption (2.15) it holds that $n(1 - \pi_0) \leq E(T) \leq n - 1$. Therefore, $E(\hat{p}(T)) = E(g(T)) \geq g(E(T)) \geq \psi(E(T)) = p$, i.e., $\hat{p}$ overestimates p .

In order to show that (2.16) implies (2.15) it is sufficient to show that

$$\frac{-\ln(n - 5/3)}{\ln(1 - p)} \leq \frac{\ln(r^{-1}(\pi_1 - (1 - n^{-1})))}{\ln(1 - p)}. \quad (2.18)$$

It is obvious that $n^{-1} \geq \pi_1 - (1 - n^{-1})$. Moreover, $r = \pi_0 + \pi_1 - 1 \geq (1 - n^{-1}) + (1 - (2/3)n^{-1}) - 1 = 1 - (5/3)n^{-1}$. Thus we obtain

$$r^{-1}(\pi_1 - (1 - n^{-1})) \leq \frac{n^{-1}}{1 - (5/3)n^{-1}} = \frac{1}{n - 5/3}$$

and consequently

$$\frac{\ln(r^{-1}(\pi_1 - (1 - n^{-1})))}{\ln(1 - p)} \geq \frac{-\ln(n - 5/3)}{\ln(1 - p)}.$$

□

Corollary 2.3. Assume that $\pi_0 > 1 - 1/n$ and $\pi_1 = 1$. Then $\hat{p}$ overestimates p for any group size $k \geq 1$.

Proof. Notice that under the assumptions $r = \pi_0$, the domain of the function $\psi(t)$ can be extended from $[n(1 - \pi_0), n - 1]$ to $[n(1 - \pi_0), n]$, and the expression (2.17) can be rewritten as

$$\hat{p}(t) = \begin{cases} 0, & \text{if } t = 0; \\ 1 - [\frac{1-t/n}{\pi_0}]^{1/k}, & \text{if } 1 \leq t \leq n. \end{cases} \quad (2.19)$$

Moreover, $g(t) \geq \psi(t), \forall t \in [n(1 - \pi_0), n]$. Similar to the proof of Theorem 2.2 it can be shown that

$$E(\hat{p}(T)) = E(g(T)) \geq g(E(T)) \geq \psi(E(T)) = p$$

since $n(1 - \pi_0) < E(T) < n$ hold obviously. □

Remark 2.2. The known result that $\hat{p}$ overestimates p for any group size $1 \leq k < \infty$ in the case of perfect group testing is certainly a special case of Corollary 2.3.

Remark 2.3. If $\pi_0 = 1, \pi_1 > 1 - (2/3)n^{-1}$ then $\hat{p}$ overestimates p for any group size satisfying

$$1 \leq k \leq \frac{-\ln n}{\ln(1 - p)}. \quad (2.20)$$

This holds since it is clear this time that $r = \pi_1$ and

$$\frac{-\ln n}{\ln(1 - p)} \leq \frac{\ln[1 - \pi_1^{-1}(1 - n^{-1})]}{\ln(1 - p)}.$$

Remark 2.4. Let $n = 20, \pi_0 = 1, \pi_1 = 0.97 > 0.9667 = 1 - (2/3)(1/n)$. If $p = 0.1$, then $k \leq 28.43316$; $p = 0.01$, then $k \leq 298.0729$; $p = 0.001$, then the bound given in inequality (2.20) is $k \leq 3,980$, if $p = 0.0005$ then $k \leq 7,762$, and if $p = 0.0001$ then $k \leq 38,814$.

upper bound for k	$p = 0.1$	$p = 0.01$	$p = 0.001$	$p = 0.0005$	$p = 0.0001$
$\frac{-\ln n}{\ln(1-p)}$	28	298	3,980	7,762	38,814

 Table 1: Upper Bound (2.20) when $n = 20, \pi_0 = 1, \pi_1 = 0.97$

2.2 Limit of the Estimator when Group Size Goes to Infinity

It is easy to see that the MLE $\hat{p}$ is a strongly consistent estimator of p , i.e., $\lim_{n \rightarrow \infty} \hat{p} = p$. In this subsection we will explore the limit of $\hat{p}$ when the group size $k \rightarrow \infty$.

We will use the notations T_k to emphasize the dependence of T on the group size k . Accordingly, we will use $\hat{p}_k$ to show the dependence of the MLE $\hat{p}$ on k .

Lemma 2.4. Suppose that $\pi_1 = 1$ and $0 < \pi_0 \leq 1$. Then $T_k \rightarrow n$ with probability one as $k \rightarrow \infty$.

Proof. For any $\varepsilon \in (0, 1)$ obviously

$$(n - T_k \geq \varepsilon) = (T_k \leq n - \varepsilon) = (T_k \leq n - 1).$$

Hence,

$$P(n - T_k \geq \varepsilon) = P(T_k \leq n - 1) = 1 - P(T_k = n) = 1 - \theta_k^n \equiv b_k$$

and consequently

$$\sum_{k=1}^{\infty} P(n - T_k \geq \varepsilon) = \sum_{k=1}^{\infty} b_k.$$

We have

$$\frac{b_{k+1}}{b_k} = \frac{1 - \theta_{k+1}^n}{1 - \theta_k^n} = \frac{(1 - \theta_{k+1})(1 + \theta_{k+1} + \cdots + \theta_{k+1}^{n-1})}{(1 - \theta_k)(1 + \theta_k + \cdots + \theta_k^{n-1})}.$$

Note that under the assumption $\pi_1 = 1$ we have $r = \pi_0 + \pi_1 - 1 = \pi_0$,

$$\begin{aligned} 1 - \theta_{k+1} &= 1 - [\pi_1 - r(1 - p)^{k+1}] \\ &= \pi_0(1 - p)^{k+1} \end{aligned}$$

and hence

$$\frac{1 - \theta_{k+1}}{1 - \theta_k} = \frac{\pi_0(1 - p)^{k+1}}{\pi_0(1 - p)^k} = 1 - p.$$

On the other hand, as $k \rightarrow \infty$ it holds that

$$\theta_k = 1 - [1 - \pi_0(1 - p)^k] \rightarrow 0$$

and thus

$$1 + \theta_k + \cdots + \theta_k^{n-1} \rightarrow 1.$$

The above implies that

$$\frac{b_{k+1}}{b_k} \rightarrow 1 - p \in (0, 1)$$

as $k \rightarrow \infty$. By d'Alembert rule it is true that

$$\sum_{k=1}^{\infty} P(n - T_k \geq \varepsilon) = \sum_{k=1}^{\infty} b_k < \infty, \forall \varepsilon > 0.$$

Therefore, $T_k - n \rightarrow 0$ with probability one as $k \rightarrow \infty$ according to the Borel-Cantelli lemma. $\square$

Remark 2.5. In fact it can also be shown that T_k increases in $k \geq 1$ according to the likelihood ration order $\leq_l$.

Corollary 2.5. Let $\pi_1 = 1$ and $0 < \pi_0 \leq 1$. Then it holds almost surely that for each sample point ω there exists $k_0 = k_0(\omega)$ such that $T_k(\omega) = n, \forall k \geq k_0$.

Proof. Let Ω be the underlying sample space and $\Omega_0 = \{\omega \in \Omega : T_k(\omega) \rightarrow n \text{ as } k \rightarrow \infty\}$. Clearly $P(\Omega_0) = 1$ due to Lemma 2.4. The fact that for any $T_k(\omega) \rightarrow n$ for any $\omega \in \Omega_0$ certainly implies the desired result since T_k is integer valued random variable. $\square$

Theorem 2.6. Assume that $\pi_1 = 1$ and $0 < \pi_0 \leq 1$. Then it is true that

$$\lim_{k \rightarrow \infty} \hat{p}_k = 1 \text{ almost surely}$$

Proof. From the assumption for any k we have

$$\hat{p}_k = 1 - \left[\frac{1 - \max(1 - \pi_0, \frac{T_k}{n})}{\pi_0} \right]^{1/k}.$$

According to Corollary 2.5 for any $\omega \in \Omega_0$ and sufficiently large k it holds that $T_k = n$ and thus

$$\hat{p}_k = 1 - \left[\frac{1 - \max(1 - \pi_0, 1)}{\pi_0} \right]^{1/k} = 1.$$

$\square$

Theorem 2.7. Suppose that $0 < \pi_1 < 1$ and $0 < \pi_0 \leq 1$. Then as $k \rightarrow \infty$ it holds that $\hat{p}_k$ converges in distribution to a binary random variable Δ_B with probability distribution

$$P(\Delta_B = 0) = \sum_{i < n\pi_1} \binom{n}{i} \pi_1^i (1 - \pi_1)^{n-i}$$

and $P(\Delta_B = 1) = 1 - P(\Delta_B = 0)$.

Proof. Denote the value of $\hat{p}_k$ when $T_k = i$ as $\hat{p}_k(i)$. From the expression (2.4) the following is true

$$\hat{p}_k(i) = \begin{cases} 0, & \text{if } 0 \leq i \leq n(1 - \pi_0); \\ 1 - \left[\frac{\pi_1 - i/n}{\pi_0} \right]^{1/k}, & \text{if } n(1 - \pi_0) < i < n\pi_1; \\ 1, & \text{if } n\pi_1 \leq i \leq n. \end{cases} \quad (2.21)$$

Clearly for any $0 < \alpha < 1$

$$\begin{aligned}(\hat{p}_k \leq \alpha) &= (\hat{p}_k = 0) \cup (0 < \hat{p}_k \leq \alpha) \\ &= (T_k \leq n(1 - \pi_0)) \cup (n(1 - \pi_0) < T_k \leq n(\pi_1 - r(1 - \alpha)^k)).\end{aligned}$$

Hence

$$\begin{aligned}P(\hat{p}_k \leq \alpha) &= P(0 \leq T_k \leq n(\pi_1 - r(1 - \alpha)^k)) \\ &= \sum_{0 \leq i \leq n(\pi_1 - r(1 - \alpha)^k)} \binom{n}{i} \theta_k^i (1 - \theta_k)^{n-i}\end{aligned}\quad (2.22)$$

We consider two cases below.

Case 1: $n\pi_1$ is an integer. In this case it is true that $n\pi_1 - 1 < n(\pi_1 - r(1 - \alpha)^k) < n\pi_1$ if k is sufficiently large. Thus it implies that for sufficiently large k

$$\sum_{0 \leq i \leq n(\pi_1 - r(1 - \alpha)^k)} \binom{n}{i} \theta_k^i (1 - \theta_k)^{n-i} = \sum_{0 \leq i \leq n\pi_1 - 1} \binom{n}{i} \theta_k^i (1 - \theta_k)^{n-i}.$$

Consequently,

$$\begin{aligned}\lim_{k \rightarrow \infty} \sum_{0 \leq i \leq n(\pi_1 - r(1 - \alpha)^k)} \binom{n}{i} \theta_k^i (1 - \theta_k)^{n-i} &= \lim_{k \rightarrow \infty} \sum_{0 \leq i \leq n\pi_1 - 1} \binom{n}{i} \theta_k^i (1 - \theta_k)^{n-i} \\ &= \sum_{0 \leq i \leq n\pi_1 - 1} \binom{n}{i} \pi_1^i (1 - \pi_1)^{n-i} \\ &= \sum_{0 \leq i < n\pi_1} \binom{n}{i} \pi_1^i (1 - \pi_1)^{n-i}\end{aligned}$$

since $\lim_{k \rightarrow \infty} \theta_k = \lim_{k \rightarrow \infty} [\pi_1 - r(1 - p)^k] = \pi_1$. That is, $\lim_{k \rightarrow \infty} P(\hat{p}_k \leq \alpha) = P(\Delta_B \leq \alpha)$ or $\hat{p}_k$ converges to Δ_B in distribution as $k \rightarrow \infty$.

Case 2: $n\pi_1$ is not an integer. Let $\lfloor n\pi_1 \rfloor$ denote the largest integer part of $n\pi_1$. Similar to Case 1 it is true that $\lfloor n\pi_1 \rfloor < n(\pi_1 - r(1 - \alpha)^k) < n\pi_1$ if k is sufficiently large and therefore $i \leq n(\pi_1 - r(1 - \alpha)^k)$ is equivalent to $i \leq \lfloor n\pi_1 \rfloor$ or $i < n\pi_1$ if k is large enough. The desired result thus follows. $\square$

Remark 2.6. Note that when $\pi_1 = 1$ it follows that $n\pi_1 = n$,

$$\sum_{i < n\pi_1} \binom{n}{i} \pi_1^i (1 - \pi_1)^{n-i} = \sum_{i \leq n-1} \binom{n}{i} \pi_1^i (1 - \pi_1)^{n-i} = 0,$$

and so the result of Theorem 2.7 becomes that of Theorem 2.6.

3 Negative Binomial Sampling

Suppose that there is a sequence of independent groups of size k . Define $X_i, i \geq 1$ as in (2.1). According to the negative binomial model, group testing is ended at the first time when the n th group with positive response is observed. Denote the number of tests until this moment as S , i.e.,

$$S = \inf\{m \geq n : \sum_{i=1}^m X_i = n\}.$$

Clearly, S follows the negative binomial distribution with probability distribution function

$$P(S = s) = \binom{s-1}{n-1} \theta^n (1-\theta)^{s-n}, s = n, n+1, \dots$$

The MLE of p , denoted as $\tilde{p}$, is thus given as

$$\tilde{p} = 1 - \left[\frac{\pi_1 - \max\{1 - \pi_0, \min(\pi_1, \frac{n}{S})\}}{r} \right]^{1/k}$$

(see, for example, (Xiong, 2016)).

Obviously $\tilde{p} = \tilde{p}(s)$ is decreasing in $s \geq n$ and can be rewritten as

$$\tilde{p}(s) = \begin{cases} 0, & \text{if } s \geq n/(1 - \pi_0); \\ 1 - \left[\frac{\pi_1 - n/s}{r} \right]^{1/k}, & \text{if } n/\pi_1 \leq s \leq n/(1 - \pi_0); \\ 1, & \text{if } n \leq s \leq n/\pi_1. \end{cases} \quad (3.1)$$

3.1 Conditions Causing Overestimation of the MLE

(Pritchard and Tebbbs, 2011b) pointed out that $\tilde{p}$ always overestimates p for any group size k when $r = 1$ by the Jensen's Inequality. We will give sufficient conditions under which $\tilde{p}$ also overestimates p when the test is imperfect.

Lemma 3.1. Let $c_i = \tilde{p}(i), i = n, n+1, \dots$. Suppose that

$$\pi_1 \geq \frac{n(3n+7)}{3(n+1)(n+2)} \quad \text{and} \quad \pi_0 \geq \frac{2}{n+2}. \quad (3.2)$$

Then $\{c_i, i \geq n\}$ is a convex sequence.

Proof. Let function $\varphi(s)$ be defined as

$$\varphi(s) \equiv 1 - \left[\frac{\pi_1 - n/s}{r} \right]^{1/k}, \quad \frac{n}{\pi_1} \leq s \leq \frac{n}{1 - \pi_0}. \quad (3.3)$$

Clearly,

$$\varphi'(s) = -\frac{n}{kr} \left[\frac{\pi_1 - n/s}{r} \right]^{1/k-1} \cdot \frac{1}{s^2} < 0,$$

and

$$\varphi''(s) = -\frac{n}{kr} \left(\left(\frac{1}{k} - 1 \right) \left[\frac{\pi_1 - n/s}{r} \right]^{1/k-2} \left(\frac{n}{r} \right) \cdot \frac{1}{s^4} - \left[\frac{\pi_1 - n/s}{r} \right]^{1/k-1} \cdot \frac{2}{s^3} \right) > 0.$$

These imply that $\varphi(s)$ is a decreasing convex function on the interval $[n/\pi_1, n/(1 - \pi_0)]$.

From (3.2) we see that $\pi_1 > n/(n+1)$ and so we have $n/\pi_1 < n+1$. Further, if $n/(1 - \pi_0)$ is integer, we have $\varphi(j) = \varphi(n/(1 - \pi_0)) = 0$ where $j \equiv \lfloor n/(1 - \pi_0) \rfloor$ is the largest integer part of $n/(1 - \pi_0)$. Thus we can rewrite (3.1) as

$$\tilde{p}(s) = \begin{cases} 0, & \text{if } s \geq j+1; \\ \varphi(s), & \text{if } n+1 \leq s \leq j; \\ 1, & \text{if } s = n. \end{cases} \quad (3.4)$$

From $\pi_0 \geq 2/(n+2)$ we see that

$$\begin{aligned} 1 - \pi_0 &\leq \frac{n}{n+2} \\ j = \left\lfloor \frac{n}{1 - \pi_0} \right\rfloor &\geq n+2 > \frac{n}{\pi_1}. \end{aligned} \quad (3.5)$$

Hence, $[n+1, j] \subseteq [n/\pi_1, n/(1 - \pi_0)]$. It indicates that the sequence $\{c_i, n+1 \leq i \leq j\}$ is convex since $c_i = \tilde{p}(i) = \varphi(i)$, $n+1 \leq i \leq j$ and $\varphi(s)$ is convex on $[n/\pi_1, n/(1 - \pi_0)]$.

To show the claimed result we will have to show

$$2c_{n+1} \leq c_n + c_{n+2} \quad (3.6)$$

and

$$2c_j \leq c_{j-1} + c_{j+1}. \quad (3.7)$$

The inequality (3.6) is equivalent to the following

$$\begin{aligned} 2\left(1 - \left[\frac{\pi_1 - n/(n+1)}{r}\right]^{1/k}\right) &\leq 1 + \left(1 - \left[\frac{\pi_1 - n/(n+2)}{r}\right]^{1/k}\right) \\ 2^k\left(\pi_1 - \frac{n}{n+1}\right) &\geq \pi_1 - \frac{n}{n+2} \\ \pi_1 &\geq \frac{2^k(n/(n+1)) - n/(n+2)}{2^k - 1}. \end{aligned} \quad (3.8)$$

It can be verified easily that the right side of (3.8) is a strictly decreasing function of k for any fixed n . This implies that

$$\begin{aligned} \frac{2^k(n/(n+1)) - (n/(n+2))}{2^k - 1} &\leq \frac{4(n/(n+1)) - (n/(n+2))}{3} \\ &= \frac{n(3n+7)}{3(n+1)(n+2)}. \end{aligned}$$

Therefore, the condition (3.2) implies (3.8) and consequently (3.6).

In the case of $\pi_0 = 1$ we have $j = \infty$, $c_j = c_\infty = \tilde{p}(\infty) = 0$, and so (3.7) is certainly true. Below we let $2/(n+2) < \pi_0 < 1$. Under this assumption, $j < \infty$ and due to the fact that $c_{j+1} = \tilde{p}(j+1) = 0$, the needed (3.7) is equivalent to

$$\begin{aligned} 2\left(1 - \left[\frac{\pi_1 - n/j}{r}\right]^{1/k}\right) &\leq 1 - \left[\frac{\pi_1 - n/(j-1)}{r}\right]^{1/k} \\ 1 + \left[\frac{\pi_1 - n/(j-1)}{r}\right]^{1/k} &\leq 2\left[\frac{\pi_1 - n/j}{r}\right]^{1/k}. \end{aligned} \quad (3.9)$$

Let

$$f(s) \equiv 1 - \varphi(s) = \left[\frac{\pi_1 - \frac{n}{s}}{r}\right]^{1/k}.$$

Obviously, $f(s)$ is an increasing concave function of $s \in [n/\pi_1, n/(1 - \pi_0)]$. If we let $s_1 = n/(\pi_1 - r)$, $s_2 = j - 1$. Clearly, $n/\pi_1 < s_1 = n/(1 - \pi_0)$, $s_2 < n/(1 - \pi_0)$, and $s_2 = j - 1 \geq (n+2) - 1 > n/\pi_1$, which is from (3.5). These show $s_1, s_2 \in [n/\pi_1, n/(1 - \pi_0)]$. Due to the concavity of $f(s)$ it follows that

$$f(s_1) + f(s_2) \leq 2f\left(\frac{s_1 + s_2}{2}\right),$$

or

$$1 + \left[\frac{\pi_1 - \frac{n}{j-1}}{r} \right]^{1/k} \leq 2f\left(\frac{s_1 + s_2}{2}\right). \quad (3.10)$$

It is not difficult to verify that

$$\frac{s_1 + s_2}{2} \leq j$$

since $j + 1 > n/(1 - \pi_0) = n/(\pi_1 - r)$. Hence it holds that

$$f\left(\frac{s_1 + s_2}{2}\right) \leq f(j) = \left[\frac{\pi_1 - \frac{n}{j}}{r} \right]^{1/k}. \quad (3.11)$$

Combining (3.10) and (3.11), we obtain (3.9) and complete the proof. $\square$

Remark 3.1. It can be observed that if $n \geq 2$, $\pi_1 \geq 1 - (2n)^{-1}$ and $\pi_0 > 2/(n + 2)$ then the assumption (3.2) holds and so $\{a_i, i \geq n\}$ is a convex sequence. The difference of the two expressions is pretty small. For example when $n = 20, 30$ we have $1 - 2n^{-1} = 0.975, 0.9833$ and (3.2) gives 0.9668, 0.9778 respectively.

Theorem 3.2. Suppose that the inequality (3.2) holds. If the group size k satisfies

$$1 \leq k \leq \frac{\ln((r(n+1))^{-1}((n+1)\pi_1 - n))}{\ln(1-p)} \quad (3.12)$$

then $\tilde{p}$ overestimates p .

Proof. Denote the point with coordinates (s, c_s) , $s = n, n + 1, \dots$ in (s, c) -plane as B_s , and denote the line segment connecting points B_s and B_{s+1} as $\gamma_{s,s+1}$, $s = n, n + 1, \dots$

Let $h(s)$ be the function with the graph consisted by $\gamma_{n,n+1}, \gamma_{n+1,n+2}, \dots$. Then $h(s)$ must be a convex function of $s \in [n, \infty)$ due to Lemma 3.1. It is also obvious that $h(s) = c_s = \tilde{p}(s)$, $\forall s = n, n + 1, \dots$

By Jensen's Inequality then we have $E(h(S)) \geq h(E(S))$. Note that

$$E(\tilde{p}(S)) = \sum_{s=n}^{\infty} \tilde{p}(s)P(S=s) = \sum_{s=n}^{\infty} h(s)P(S=s) = E(h(S))$$

since $c_s = h(s)$, $s = n, n + 1, \dots$. Hence it holds that $E(\tilde{p}(S)) = E(h(S)) \geq h(E(S))$.

It has been shown in the above that $\varphi(s)$ is a decreasing convex function on the interval $[n/\pi_1, n/(1 - \pi_0)]$. Note that $\varphi(n/(1 - \pi_0)) = 0$. Hence, it is easy to see that $h(s) \geq \varphi(s)$, $\forall s \in [n/\pi_1, n/(1 - \pi_0)]$.

Now, if it can be shown that $n + 1 \leq E(S) \leq n/(1 - \pi_0)$, then it follows immediately that $E(\tilde{p}(S)) = E(h(S)) \geq h(E(S)) \geq \varphi(E(S)) = p$.

Below we are going to show that $n + 1 \leq E(S) \leq n/(1 - \pi_0)$. The mean of S is $E(S) = n/\theta = n/(\pi_1 - r(1 - p)^k)$. Hence, the inequality $n + 1 \leq E(S)$ is equivalent to

$$\begin{aligned} n + 1 &\leq \frac{n}{\pi_1 - r(1 - p)^k}; \\ r(n + 1)(1 - p)^k &\geq (n + 1)\pi_1 - n; \\ k \ln(1 - p) &\geq \ln \left\{ \frac{(n + 1)\pi_1 - n}{r(n + 1)} \right\} \end{aligned}$$

which is certainly true due to assumption (3.12). On the other hand, the desire inequality $E(S) \leq n/(1 - \pi_0)$ is equivalent to

$$\begin{aligned} \frac{n}{\pi_1 - r(1-p)^k} &\leq \frac{n}{1 - \pi_0}; \\ r(1-p)^k &\leq \pi_1 + \pi_0 - 1 = r \end{aligned}$$

which is trivial.

Therefore, it must be true that then $\tilde{p}$ overestimates p . $\square$

Corollary 3.3. *Let $\pi_1 = 1$ and $\pi_0 \geq 2/(n+2)$. Then $\tilde{p}$ overestimates p for any group size $k \geq 1$.*

Proof. It is sufficient to notice that this time $\varphi(s)$ is a convex function on $[n, n/(1 - \pi_0)]$ and $\varphi(n) = 0 = h(0)$. $\square$

3.2 Limiting Behavior of the MLE as $k \rightarrow \infty$

To indicate the dependence of S and MLE of p on k we will denote S and $\tilde{p}$ as S_k and $\tilde{p}_k$ if it is necessary.

Lemma 3.4. *Suppose that $\pi_1 = 1$. Then $S_k \rightarrow n$, a.s. as $k \rightarrow \infty$ for any $\pi_0 \in (0, 1]$. Particularly, for all sample point ω except for a set of probability zero there exists an integer $k_0 = k_0(\omega)$ such that $S_k(\omega) = n, \forall k \geq k_0$.*

Proof. We have $\theta_k = \pi_1 - r(1-p)^k = 1 - \pi_0(1-p)^k$ since $\pi_1 = 1$. Denote the number of tests needed until the first positive group is observed as Y_1 and let $Y_i, 1 \leq i \leq n$ be iid random variables. Clearly, Y_i follows geometric distribution with $P(Y_i = y) = \theta_k(1 - \theta_k)^{y-1}$ for any integer $y \geq 1$. Further, it is obvious that $S_k = \sum_{i=1}^n Y_i$ in distribution. We will use $Y_i^{(k)}$ to display the dependence of Y_i on the group size k below.

For any $\epsilon \in (0, 1)$ it holds that

$$\begin{aligned} P(S_k - n > \epsilon) &= P\left(\sum_{i=1}^n Y_i^{(k)} > n + \epsilon\right) \leq P\left(\bigcup_{i=1}^n (Y_i^{(k)} > 1 + \epsilon/n)\right) \\ &\leq \sum_{i=1}^n P(Y_i^{(k)} \geq 2). \end{aligned}$$

From this inequality we have

$$\begin{aligned} \sum_{k=1}^{\infty} P(|S_k - n| > \epsilon) &\leq \sum_{k=1}^{\infty} \sum_{i=1}^n P(Y_i^{(k)} \geq 2) \\ &= \sum_{k=1}^{\infty} n P(Y_1^{(k)} \geq 2) = n \sum_{k=1}^{\infty} [1 - \theta_k] \\ &= n \pi_0 \sum_{k=1}^{\infty} (1-p)^K < \infty. \end{aligned}$$

Therefore, by the Borel-Cantelli lemma S_k converges to n with probability one as $k \rightarrow \infty$. $\square$

Theorem 3.5. *Suppose that $\pi_1 = 1$. Then $\lim_{k \rightarrow \infty} \tilde{p}_k = p$ with probability one.*

Proof. It follows from Lemma 3.4 immediately. $\square$

Theorem 3.6. Assume $0 < \pi_1 < 1$. Then $\tilde{p}_k$ converges in distribution to binary random variable Δ_{NB} that has distribution

$$P(\Delta_{NB} = 1) = \sum_{n \leq s < n/\pi_1} \binom{s-1}{n-1} \pi_1^n (1 - \pi_1)^{s-n}$$

and $P(\Delta_{NB} = 0) = 1 - P(\Delta_{NB} = 1)$.

Proof. It is sufficient to show that for any $\alpha \in (0, 1)$ it holds that

$$\lim_{k \rightarrow \infty} P(\tilde{p}_k \leq \alpha) = \sum_{s \geq n/\pi_1} \binom{s-1}{n-1} \pi_1^n (1 - \pi_1)^{s-n}. \quad (3.13)$$

From (3.1) we have

$$\begin{aligned} (\tilde{p}_k \leq \alpha) &= (\tilde{p}_k = 0) \cup (0 < \tilde{p}_k \leq \alpha) \\ &= (S_k \geq \frac{n}{\pi_1 - r(1 - \alpha)^k}). \end{aligned}$$

Let

$$j \equiv \lfloor n/(1 - \pi_0) \rfloor.$$

It then follows that

$$\begin{aligned} P(\tilde{p}_k \leq \alpha) &= \sum_{s \geq n/[\pi_1 - r(1 - \alpha)^k]} \binom{s-1}{n-1} \theta_k^n (1 - \theta_k)^{s-n} \\ &= 1 - \sum_{n \leq s < n/[\pi_1 - r(1 - \alpha)^k]} \binom{s-1}{n-1} \theta_k^n (1 - \theta_k)^{s-n}. \end{aligned}$$

We consider two cases below.

Case A: n/π_1 is an integer. In this case there does not exist integer between n/π_1 and $n/[\pi_1 - r(1 - \alpha)^k]$ if k is sufficiently large. Thus it implies that for sufficiently large k

$$\sum_{n \leq s < n/[\pi_1 - r(1 - \alpha)^k]} \binom{s-1}{n-1} \theta_k^n (1 - \theta_k)^{s-n} = \sum_{n \leq s < n/\pi_1} \binom{s-1}{n-1} \theta_k^n (1 - \theta_k)^{s-n}$$

and further as $k \rightarrow \infty$

$$\begin{aligned} P(\tilde{p}_k \leq \alpha) &= 1 - \sum_{n \leq s < n/\pi_1} \binom{s-1}{n-1} \theta_k^n (1 - \theta_k)^{s-n} \\ &\longrightarrow 1 - \sum_{n \leq s < n/\pi_1} \binom{s-1}{n-1} \pi_1^n (1 - \pi_1)^{s-n}. \end{aligned} \quad (3.14)$$

Case B: n/π_1 is not an integer. This case we clearly have the same conclusion as (3.14).

Therefore, at any rate the desired result is true. $\square$

4 Summary

The present article studies the effects of group size on the estimators of proportion p based on group testing in two aspects. One is the limit of the estimators when the group size goes to infinity, the other is whether the estimators overestimate p .

In both binomial and negative binomial samplings we have shown that the MLEs $\hat{p}$ and $\tilde{p}$ converge to 1 with probability one as the group size $k \rightarrow \infty$ when the *sensitivity* of the test $\pi_1 = 1$ regardless of the value of the *specificity* π_0 of the test. On the other hand, when $0 < \pi_1 < 1$ then $\hat{p}$ and $\tilde{p}$ converge to a binary random variable in distribution.

As was pointed out, for instance, in (Hepworth, 1996), (Pritchard and Tebbs, 2011b) that if test $\mathbb{T}$ is perfect, i.e., both $\pi_1 = 1$ and $\pi_0 = 1$ hold then both $\hat{p}$ and $\tilde{p}$ overestimate p for any group size $k \geq 1$. In the present paper, for the imperfect test with $\pi_1 = 1$ we can show that if π_0 satisfies certain condition, then both $E(\hat{p}) \geq p$ and $E(\tilde{p}) \geq p$ for any $k \geq 1$. However, when $\pi_1 < 1$ we can only prove that under certain conditions there exists finite bound k^* such that $E(\hat{p}) \geq p$ and $E(\tilde{p}) \geq p$ for any $1 \leq k \leq k^*$. Note that in most applications of group testing the proportion p is very small, and thus k^* is quite large. Consequently, the group size k used in practice are smaller than k^* , and hence both $\hat{p}$ and $\tilde{p}$ overestimate p under these situations. It is unclear at this moment whether $\hat{p}$ or/and $\tilde{p}$ overestimate p for any $k \geq 1$ without making some assumptions on π_0 and π_1 . This is certainly worthy of further study in the future.

References

- Ding, J. and Xiong, W. 2015. Robust group testing for multiple traits with misclassification, *Journal of Applied Statistics* **42**: 2115–2125.
- Dorfman, R. 1943. The detection of defective members of large populations, *Annals of Mathematical Statistics* **14**: 436–440.
- Gastwirth, J. L. 1987. The statistical precision of medical screening procedures: application to polygraph and aids antibodies test data, *Statistical Science* **2**: 213–238.
- Gastwirth, J. L. and Hammick, P. 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**: 15–27.
- Gastwirth, J. L. and Johnson, W. 1994. Screening with cost-effective quality control: Potential applications to hiv and drug testing, *J. Am. Stat. Assoc.* **89**: 972–981.
- Gastwirth, J. L., Johnson, W. O. and Reneau, D. M. 1991. Bayesian analysis of screening data: Application to aids in blood donors, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **19**: 135–150.
- Graff, L. E. and Roeloffs, R. 1972. Group testing in the presence of test error; an extension of the dorfman procedure, *Technometrics* **14**: 113–122.

- Gupta, D. and Malina, R. 1999. Group testing in presence of classification errors, *Statistics in Medicine* **18**: 1049–1068.
- Hepworth, G. 1996. Exact confidence intervals for proportions estimated by group testing, *Biometrics* **52**(33, pages = 1134-1146,).
- Hepworth, G. 2013. Improved estimation of proportions using inverse binomial group testing, *J. Agric., Biol., Environ. Stat.* **18**: 102–119.
- Hepworth, G. and Watson, R. K. 2009. Debiased estimation of proportions in group testing, *J. R. Stat. Soc. Ser. C (Appl. Stat)* **58**: 105–121.
- Huang, K. and Mi, J. 2019. Study on optimal group size in grouping test, *Journal of Probability and Statistical Science* **17**: 97–108.
- Hughes-Oliver, J. M. and Swallow, W. H. 1994. A two-stage adaptive group-testing procedure for estimating small proportions, *J. Am. Stat. Assoc* **89**: 982–993.
- Hung, M. and Swallow, W. H. 1999. Robustness of group testing in the estimation of proportions, *Biometrics* **55**: 231–237.
- Lew, R. A. and Levy, P. S. 1989. Estimation of prevalence on the basis of screening tests, *Statist. Med.* **8**: 1225–1230.
- Liu, A., Liu, C. L., Zhang, Z. W. and Albert, P. S. 2012. Optimality of group testing in the presence of misclassification, *Biometrika* **99**: 245–251.
- Lui, K. J. 1995. Confidence limits for the population prevalence rate based on the negative binomial distribution, *Statistics in Medicine* **14**: 1471–1477.
- Mi, J. 2019. Some limit results in estimation of proportion based on group testing, *Metrika* **82**: 1021–1038.
- Pfeiffer, R. M., Rutter, J. L., Gail, M. H., Struewing, J. and Gastwirth, J. L. 2002. Efficiency of dna pooling to estimate joint allele frequencies and measure linkage disequilibrium, *Genet. Epidemiol.* **22**: 94–102.
- Pritchard, N. A. and Tebbs, J. M. 2011a. Bayesian inference for disease prevalence using negative binomial group testing, *Biomet. J.* **53**: 40–56.
- Pritchard, N. A. and Tebbs, J. M. 2011b. Estimating disease prevalence using inverse binomial pooled testing, *J. Agric., Biol., Environ. Stat.* **16**: 70–87.
- Rogan, W. J. and Gladen, B. 1978. Estimating prevalence from the results of a screening test, *American Journal of Epidemiology* **107**: 71–76.
- Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission, *Phytopathology* **75**: 882–889.

- Tebbs, J. M., McMahan, C. and Bilder, C. R. 2013. Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project, *Biometrics* **69**: 1064–1073.
- Tebbs, J. M. and Swallow, W. H. 2003. Estimating ordered binomial proportions with the use of group testing, *Biometrika* **90**: 471–477.
- Thompson, K. 1962. Estimation of the proportion of vectors in a natural population of insects, *Biometrics* **18**: 568–578.
- Tu, X. M. and Litvak, E. 1994. Studies of aids and hiv surveillance screening tests: can we get more by doing less ?, *Statistics in Medicine* **13**: 1905–1919.
- Tu, X. M., Litvak, E. and Pagano, M. 1995. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to hiv screening, *Biometrika* **82**: 287–289.
- Xiong, W. J. 2016. The optimal group size using inverse binomial group testing considering misclassification, *Communications in statistics - Theorey and Methods* **45**: 4600–4610.
- Xiong, W. J. and Ding, J. 2015. Robust procedures for experimental design in group testing considering misclassification, *Statistics and Probability Letters* **100**: 35–41.