Studia Scientiarum Mathematicarum Hungarica 22 (1987), 257—263

NON-ADAPTIVE HYPERGEOMETRIC GROUP TESTING

F. K. HWANG and V. T. SÓS

Abstract

In a hypergeometric group testing problem we have a set of n items known to contain exactly d defectives. The problem is to identify all the defectives through group testing with a minimal number of tests where a test consists of a specified subset of the items and has the outcome *pure* if none of the items in the subset is defective and the outcome contaminated otherwise. A testing procedure is called *non-adaptive* if all tests have to be specified simultaneously. We translate this group testing problem into an extremal problem for set-systems and give estimates for the size of the extremal systems.

1. Introduction

In a *hypergeometric group testing* (HGT) problem we have a set of n items known to contain d defectives and $n-d$ good items. Any subset of the n items can be pooled for a test with two possible outcomes: the subset is *pure* if it contains no defectives and the subset is *contaminated* otherwise. The objective is to identify all the defectives using a minimal number of tests. In this paper "minimal" is defined as to minimize the maximum number of tests required (the worst-case number of tests).

A group testing algorithm is called *sequential* if the tests are given sequentially, that is, the decision of which subset to test currently may depend on the outcomes of tests already performed. A group testing algorithm is called *non-adaptive* if all tests have to be specified simultaneously, thus banning any possibility of using feedback information from tests. Since any non-adaptive algorithm can also be used sequentially, it is clear that the sequential algorithms should be expected to perform better than non-adaptive algorithms in general. The line between sequential algorithms and non-adaptive algorithms has not been clearly drawn historically; hence the group testing literature consists almost exclusively of sequential algorithms owing to their better performance as far as the number of tests is concerned. However, with parallel processing a possibility in many potential applications of group testing, one should no longer ignore the potential advantage of time saving in non-adaptive group testing. The purpose of this paper is to call attention to this fact and to provide some exploratory results in this direction.

1980 *Mathematics Subject Classification.* Primary 62K99; Secondary 05A05.
Key words and phrases. Group testing, extremal set-systems, d-complete design.

2. The group-testing problem and some extremal problems for set-systems

Suppose S is the set of n items which contains d defective elements. To design a nonadaptive HGT for finding the defective items we need a system $\mathcal{T}=\{T_1, ..., T_t\}$ of tests which satisfies the following:

for any $I=\{i|T_i$ is contaminated$\}$ there is exactly one d-tuple $D=\{u_1, ..., u_d\}\in S$ so that

$$D\cap T_i \neq \emptyset \quad \text{for} \quad i\in I$$

and

$$D\cap T_i = \emptyset \quad \text{for} \quad i\notin I.$$

For given n and d let $\varphi(n, d)$ be the minimum number of t so that there is a system $\mathcal{T}=\{T_1, ..., T_t\}$ which satisfies this condition. Our main result yields

$$c_1(d)\log n < \varphi(n; d) < c_2(d)\log n \tag{1}$$

where $c_1, c_2>0$ depend only on d. The lower bound follows from the simple Proposition 1, the upper bound follows from Theorem 3.

First we reformulate the problem in a dual form. This yields to different extremal problems for set-systems.

Now let $S=\{u_1, ..., u_n\}$ be a set of n elements, $\mathcal{T}=\{T_1, ..., T_t\}$ be a family of subsets of S. Define the *dual family*

$$\mathcal{C}=\{C_1, ..., C_n\}$$

by

$$C_i=\{j|u_i\in T_j\}, \qquad i=1, ..., n.$$

Observe that for given defective elements $u_{i_1}, ..., u_{i_d}$ a test T_j yields "contaminated" iff $j\in \bigcup_{v=1}^{d} C_{i_v}$. So

$$I=\{j|T_j \text{ is "contaminated"}\}=\bigcup_{v=1}^{d} C_{i_v}.$$

This leads to a reformulation of the problem.

Definition 1. Let $\mathcal{C}=\{C_1, ..., C_n\}$ be a family of subsets of S. $\mathcal{C}$ is a *d-Sidon family,* if all the d-term unions are distinct:

$$\bigcup_{k=1}^{d} C_{i_k} \neq \bigcup_{k=1}^{d} C_{j_k} \tag{2}$$

if $\{i_1, ..., i_d\}\neq\{j_1, ..., j_d\}$.

Proposition 1. *$\mathcal{T}=\{T_1, ..., T_t\}$ is a system of tests of a parallel HGT, iff the dual system $\mathcal{C}=\{C_1, ..., C_n\}$ is a d-Sidon system.*

Proof. Suppose

$$\bigcup_{k=1}^{d} C_{i_k} = \bigcup_{k=1}^{d} C_{j_k}$$

but $\{i_1, ..., i_d\}\neq\{j_1, ..., j_d\}$. Then the test outcomes are identical whether $D=\{u_{i_1}, ..., u_{i_d}\}$ or $D=\{u_{j_1}, ..., u_{j_d}\}$. Thus $\mathcal{T}$ fails to identify the defectives.

Now suppose that $\mathscr{C}$ is a d-Sidon family. Then the set

$$\{j\colon\ T_j \text{ has a “contaminated” outcome}\}$$

corresponds to a unique d-element set $\{u_{i_1}, \ldots, u_{i_d}\}$ which is the set of defectives.

COROLLARY 1. *For any parallel HGT-system* $\mathscr{T}=\{T_1, \ldots, T_t\}$ *on n-elements with d defectives*

$$\binom{n}{d} \leqq 2^t. \tag{3}$$

This gives the lower bound in (1).

So we arrive at a dual form of the original problem which is an extremal problem for set-systems.

PROBLEM 1. Let t, d be given positive integers. Let $f(t, d)$ denote the maximum cardinality of a d-Sidon system on a t element set. Determine $f(t, d)$.

Recently Busch et al. [1] studied a stronger version of parallel HGT-systems, the d-complete designs.

DEFINITION. The family $\mathscr{T}=\{T_1, \ldots, T_t\}$ is called a d-complete design if for any d subscripts $\{i_1, \ldots, i_d\}$

$$\bigcup\{T_j | j \notin \bigcup_{v=1}^{d} C_{i_v}\} = S \setminus \{u_{i_1}, \ldots, u_{i_d}\}. \tag{4}$$

REMARK. By Proposition 1 a d-complete design can be used as a non-adaptive HGT when the number of defectives is d. To find the defectives is very simple:

$$D = S - \bigcup\{T_j | T_j \text{ is pure}\}.$$

Proposition 2 gives a characterization of d-complete designs.

PROPOSITION 2. $\mathscr{T}=\{T_1, \ldots, T_t\}$ *is a d-complete design iff for any d subscripts* $\{i_1, \ldots, i_d\}$

$$C_j \not\subseteqq \bigcup_{k=1}^{d} C_{i_k} \quad \textit{if} \quad j \notin \{i_1, \ldots, i_d\}. \tag{5}$$

PROOF. Suppose $C_1 \subseteqq \bigcup_{k=2}^{d-1} C_k$. Then

$$\{T_j | j \notin \bigcup_{k=2}^{d+1} C_k\} = \bigcup\{T_j | j \notin \bigcup_{k=1}^{d+1} C_k\} \subseteqq \{u_{d+2}, \ldots, u_n\} \neq S - \{u_2, \ldots, u_{d+1}\}.$$

Hence T is not a d-complete design.

Now suppose (5) holds. Then for any choice of distinct $i_0, i_1, \ldots, i_d$ there always exists a T_j such that

$$u_{i_0} \in T_j$$

but

$$u_{i_k} \notin T_j \quad \text{for} \quad k = 1, \ldots, d.$$

Thus

$$u_{i_0} \in \bigcup \{T_j | j \notin \bigcup_{k=1}^{d} C_{i_k}\}.$$

Consequently,

$$\{T_j | j \notin \bigcup_{k=1}^{d} C_{i_k}\} = S - \{u_i, \ldots, u_{i_d}\}.$$

The original HGT problem leads to the question what is the minimum number of sets in a d-design on n elements.

By Proposition 2 this yields to the following dual problem:

Problem 2. Let t, d be given integers. Denote by $g(t, d)$ the maximum cardinality of a system $\mathscr{C} = \{C_1, \ldots, C_n\}$ on t elements which satisfies (5). Determine $g(t, d)$.

A system $\mathscr{C} = \{C_1, \ldots, C_n\}$ satisfies (5) if

$$(6) \qquad |C_i \cap C_j| < \frac{1}{d} |C_i| \qquad \forall i \neq j.$$

So we arrive to

Problem 3. Let t, d be given integers. Denote by $h(t, d)$ the maximum cardinality of a system $\mathscr{C} = \{C_1, \ldots, C_n\}$ on t elements which satisfies (6). Determine $h(t, d)$.

Now we have three extremal problems for sets systems. Since the restriction (6) is stronger than (5) and this is stronger than (2), we have

$$(7) \qquad h(t, d) \leqq g(t, d) \leqq f(t, d).$$

In Theorem 3 we prove

$$(8) \qquad h(t, d) > c^t$$

where $c > 1$ depends only on d. This proves in (1) the upper bound for $\varphi(n; d)$:

Theorem 3.

$$h(t, d) > \frac{1}{2} \left(1 + \frac{1}{(4d)^2}\right)^t.$$

Proof. We use a sort of greedy algorithm to prove the theorem.

Let S be a t-element set, $r = \left[\frac{t}{(4d)^2}\right]$. Put $k = 4dr$, $m = \frac{1}{d} k = 4r$ and $[S]^k = \{A | A \subset S, |A| = k\}$.

Choose $A_1 \in [S]^k$ arbitrarily. Delete all k-sets of S which intersect A_1 in at least m elements. I.e., let

$$\mathscr{B}_1 = \{B | B \in [S]^k, |B \cap A_1| \geqq m\}.$$

We define the sets A_i and the families B_i inductively. Suppose we have already

$A_1, \dots, A_\nu, \mathcal{B}_1, \dots, \mathcal{B}_\nu$. Then choose

$$A_{\nu+1} \in [S]^k \setminus \bigcup_{i=1}^{\nu} \mathcal{B}_i$$

arbitrarily. Define

$$\mathcal{B}_{\nu+1} = \{B \mid B \in [S]^k \setminus \bigcup_{i=1}^{\nu} \mathcal{B}_i,\ |B \cap A_{\nu+1}| \geqq m\}.$$

We proceed as long as we can. Suppose $A_1, \dots, A_M$ have been chosen this way. Since

$$|\mathcal{B}_\nu| \leqq \sum_{i=m}^{k} \binom{k}{i} \binom{t-k}{k-i},$$

we certainly can continue unless

$$M \geqq \binom{t}{k} \left(\sum_{i=m}^{k} \binom{k}{i} \binom{t-k}{k-i} \right)^{-1}.$$

Put $b_i = \binom{k}{i} \binom{t-k}{k-i}$. Obviously

$$\binom{t}{k} > b_r \tag{9}$$

and for $3r \leqq i < k = 4dr$

$$\frac{b_i}{b_{i-1}} = \frac{(k-i)^2}{i(t-2k+i)} \leqq \frac{1}{3} \frac{(4d-3)^2}{(4d)^2 - 8d + 3} < \frac{1}{3}.$$

Hence

$$\binom{k}{m} \binom{t-k}{k-m} < b_r \prod_{i=3r}^{4r} \frac{b_{i+1}}{b_i} < b_r 3^{-r}$$

and

$$\sum_{i=m}^{k} \binom{k}{i} \binom{t-k}{k-m} < 2b_r 3^{-r} < 2b_r \left(1 + \frac{1}{(4d)^2}\right)^t. \tag{10}$$

By (9) and (10)

$$\binom{t}{k} \sum_{i=m}^{k} \binom{k}{m} \binom{t-k}{k-m} > \frac{1}{2} \left(1 + \frac{1}{(4d)^2}\right)^t.$$

This means the above algorithm leads to a family $\mathcal{A} = \{A_1, \dots, A_M\}$, which satisfies (6) and

$$M \geqq \frac{1}{2} \left(1 + \frac{1}{(4d)^2}\right)^t.$$

By Corollary 1 and Theorem 3 we have

Corollary 2.

$$(\log(1 + d^{-1}))^{-1} \log n < \varphi(n, d) < 2(\log 1 + (4d)^{-2}))^{-1} \log n.$$

Remark. A more careful computation would give a better constant instead of 4. We do not see how to diminish the gap between d^{-1} and d^{-2}.

Some open problems

1. We considered three different problems for set-systems. We have the trivial inequalities (7). What can be said on

$$\frac{g(t,d)}{h(t,d)} \quad \text{resp.} \quad \frac{f(t,d)}{g(t,d)}\,?$$

2. Is it possible to give an explicit construction for a "large" system which satisfies (2) resp. (5) or (6)?

Historical remarks

Hwang [8] gave a sequential HGT algorithm with $t=d\log_2 \frac{n}{d}$ tests. Since sequential HGT algorithms also obey the inequality (with a different argument) as given in the Corollary of Theorem 3, Hwang's algorithm is asymptotically optimal for fixed d and large n.

For $d=2$, the best sequential HGT algorithm so far was given by Chang, Hwang and Lin [2]: For $t\geqq 4$,

$$n=\begin{cases}\lceil 43\cdot 2^{(t/2)-5}\rceil-1 & \text{for } t \text{ even,}\\ \lceil 31\cdot 2^{(t-1/2)-4}\rceil-1 & \text{for } t \text{ odd.}\end{cases}$$

Freidlina [6] considered the non-adaptive HGT problem but did not require that all defectives be identified. He gave a construction for a non-adaptive algorithm in which each test consists of a set of random items with the probability of each item being included being q. He showed that for $q=1-2^{-1/d}$ and $t\geqq(1+\varepsilon)\log_2\binom{n}{d}$, then the probability of such an algorithm identifying all defectives is at least $1-\lambda$ where ε and λ are arbitrary numbers in $(0, 1)$.

Shapiro [10] and Fine [5] studied a different version of HGT called the "counterfeit coin" problem in which each test reveals the exact number of defectives contained in it and the number of defectives is unknown at the outset (of course it can be found out in one test). Söderberg and Shapiro [11] gave a non-adaptive algorithm with $t=O(n/\log n)$. Erdős and Rényi [4], and Lindström [9] proved that $t=n/\log_4 n)$ in an asymptotically optimal parallel algorithm for the counterfeit coin problem and gave a construction for such an algorithm.

A criticism of the HGT model is that in real applications one usually can determine only an upper bound but not the exact number of defectives. Now we know that a non-adaptive HGT algorithm or a d-complete design can also be applied to the case where only an upper bound d is known for the number of defectives. This is so because (2) implies

$$\bigcup_{\nu=1}^{k} C_{i_\nu} \neq \bigcup_{\nu=1}^{l} C_{j_\nu} \quad \text{if} \quad 1\leqq k\leqq l\leqq d$$

and

$$\{i_1, \dots, i_k\} \neq \{j_1, \dots, j_l\}.$$

ADDED in proof. Theorem 3 was proved independently by P. Erdős, P. Frankl and Z. Füredi [13].

REFERENCES

[1] BUSH, K. A., FEDERER, W. T., PESOTAN, H. and RAGHAVARAO, D., New combinatorial designs and their application to group testing, 1980 (preprint).

[2] CHANG, G. J., HWANG, F. K. and LIN, S., Group testing with two defectives, *Discrete Appl. Math.* **4** (1982), 97—102. *MR* **84e:** 05010.

[3] ERDŐS, P., FRANKL, P. and FÜREDI, Z., Families of finite sets in which no set is covered by the union of two others, *J. Combin. Theory Ser. A* **33** (1982), 158—166. *MR* **84e:** 05002.

[4] ERDŐS, P. and RÉNYI A., On two problems of information theory, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** (1963), 229—243. *MR* **29** #3268.

[5] FINE, N. J., Solution El 399, *Amer. Math. Monthly* **67** (1968), 697.

[6] FREIDLINA, V. L., A certain problem on the design of screening experiments, *Teor. Verojatnost. i Primenen.* **20** (1975), 100—114 (in Russian). *MR* **51**#9388.

[7] HALL, M., *Combinatorial Theory,* Blaisdell Publishing Co., Waltham, 1967. *MR* **37** #80.

[8] HWANG, F. K., A method for detecting all defective members in a population by group testing, *J. Amer. Statist. Assoc.* **67** (1972), 605—608.

[9] LINDSTRÖM, B., On a combinatorial detection problem, I., *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964), 195—207. *MR* **29**#5750.

[10] SHAPIRO, H. S., Problem El 399, *Amer. Math. Monthly* **67** (1960), 82.

[11] SÖDERBERG, S. and SHAPIRO, H. S., A combinatory detection problem, *Amer. Math. Monthly* **10** (1963), 1066—1070.

[12] SPERNER, E., Ein Satz über Untermengen einer endlichen Menge, *Math. Zeitschr.* **27** (1928), 544—548.

[13] ERDŐS, P., FRANKL, P. and FÜREDI, Z., Families of finite sets in which no set is covered by the union of *r* others, *Israel J. Math.* **51** (1985), no. 1-2.

(Received March 25, 1982; in revised form February 25, 1986)

BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NJ 07974
U.S.A.

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR
ANALÍZIS TANSZÉK
MÚZEUM KRT. 6—8
H—1088 BUDAPEST
HUNGARY