Revisiting Nested Group Testing Procedures: New Results, Comparisons, Robustness

Yaakov Malinovsky

University of Maryland, Baltimore County
Joint work with Paul Albert, NCI

Latent Variable Conference
University of South Carolina
Oct. 12-14, 2016

Introduction

Group testing (GT) procedures are cost, and time-saving identification procedures which have broad applications to

- blood screening for infectious diseases: Bilder, Tebbs, and Chen (2010); Tebbs, McMahan, and Bilder (2013).
- quality control: Sobel and Groll (1959)
- DNA screening: Du and Hwang (2006)
- multiaccess communications: Wolf (1985).

Bilder, C. R., Tebbs, J. M., Chen, P. (2010). Informative retesting. *J. Am. Stat. Assoc.* **1**05, 942–955.

Tebbs, J., McMahan, C., and Bilder, C. (2013). Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics* **6**9, 1064–1073.

Sobel, M., Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech. J.* **3**8, 1179–1252.

Du, D., Hwang, F. K. (2006). Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing. *World Scientific, Singapore*.

Wolf J. K. (1985). Born again group testing: multiaccess comunications. *IEEE Transactions on Information Theory* **31,** 185–191.

Dorfman (1943) Procedure: procedure *D*

The motivation was the need to administer syphilis tests to millions of persons drafted into the U.S. army during World War II. The test for syphilis was a blood test called the Wassermann (1906) test. A nice description is given by W. Feller (1950):

" *A large number, N, of people are subject to a blood test. This can be administered in two ways.*

(i) *Each person tested separately. In this case N tests are required.*

(ii) *The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the k people. If the test is positive, each of the k persons must be tested separately, and in all* $k + 1$ *tests are required for the k people.*

Assume the probability p that the test is positive is the same for all and that people are stochastically independent."

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics* **1**4, 436-440.

Procedure D: Expected number of tests

For $k \geq 2$ the total number of tests is 1 with probability q^k and $k+1$ with probability $1-q^k$. Therefore,

$$E_D(k,p) = \begin{cases} 1 - q^k + \dfrac{1}{k} & \text{for} \quad k \geq 2 \\ 1 & \text{for} \quad k = 1. \end{cases}$$

Define $k_D^*(p) = \arg\min_k E_D(k,p)$. Samuels (1978) showed that,

$$k_D^*(p) = \begin{cases} 1 + [p^{-1/2}] \ \text{or} \ 2 + [p^{-1/2}] & \text{if} \quad p < p_D = 1 - 1/3^{1/3} \approx 0.31 \\ 1 & \text{otherwise.} \end{cases}$$

Remark: for the finite population case (of size N), the solution for procedure D obtained by dynamic programming (with computational complexity $O(N^2)$).

Samuels, S. M. (1978). The exact solution to the two-stage group-testing problem. *Technometrics* **2**0, 497-500.

Remarks: Finite vs Infinite Populations, Nested Class

We consider the cases of the finite and infinite populations. In the GT literature, the infinite population means that we are looking for an optimal group size $k_A^*(p)$ under a particular procedure A which minimizes the expected number of tests per person. In contrast, the finite population optimality under a particular procedure A means that we are looking for an optimal partition of the finite population under which the expected total number of tests is minimal.

A nested algorithm has the property that if the infected subset I is identified, the next subset I_1 that we will test is a proper subset of I, i.e. $I_1 \subset I$. This natural and reasonable class of GT procedures was defined by Sobel and Groll (1959) and Sobel (1960).

Sobel, M., Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech. J.* **3**8, 1179–1252.

Sobel, M. (1960). Group testing to classify efficiently all defectives in a binomial sample. *Information and Decision Processes (R. E. Machol, ed.; McGraw-Hill, New York),* pp. 127-161.

Universal Cut-off Point and Procedure D'

Ungar (1960) characterized the optimality of any group testing algorithm and proved that if $p \geq p_U = (3 - 5^{1/2})/2 \approx 0.38$, then there does not exist an algorithm that is better than individual one-by-one testing, i.e.

$$E(N, p) = N, \quad \text{for} \quad p \geq p_U.$$

There is a logical inconsistency in the D procedure. It is clear that any "reasonable" group testing plan should satisfy the following property:

"A test is not performed if its outcome can be inferred from previous test results".

The D procedure does not satisfy this property, The modified Dorfman procedure (Sobel and Groll (1959)) (defined as D') would not test the last individual in this case.

Ungar, P. (1960). Cutoff points in group testing. *Comm. Pure Appl. Math.* **13,** 49-54.

Procedure D': Expected number of tests

For $k \geq 2$ the total number of tests is

- 1 with probability q^k,
- k with probability $q^{k-1}(1-q)$,
- and $k+1$ with probability $1-q^{k-1}$.

Therefore,

$$E_{D'}(k,p) = 1 - q^k + 1/k - (1/k)(1-q)q^{k-1} \quad \text{for} \quad k \geq 1.$$

$E_{D'}(1,p) = 1$

Infinite Population Case: Procedure D'

Lemma 2 in Pfeifer and Enis (1978):
Let $p \in (0, p_U)$. Then (as a function of the continuous variable k) $E_{D'}(k,p)$ has an absolute minimum which is at the smallest zero of $E'_{D'}(k,p) = \frac{\partial E_{D'}(k,p)}{\partial k}$. This zero is unique in that portion of the domain of $E_{D'}$ for which $E_{D'}(k,p) < 1$.

From the above Result it follows that the optimal value $k^*_{D'}$ is the smallest k value which satisfies:

$$E_{D'}(k,p) \leq E_{D'}(k-1,p) \quad \text{and} \quad E_{D'}(k,p) < E_{D'}(k+1,p).$$

Infinite Population Case: Procedure D'

Remark

In Pfeifer and Enis (1978), it was stated that there is no closed-form expression for the optimal group size $k^*_{D'}(p)$. Although we cannot prove it, the following conjecture is empirically verified:
For the values of $0 < p < p_U = \frac{3-\sqrt{5}}{2}$, $k^*_{D'}$ *equals to* $\lfloor p^{-1/2} \rfloor$ *or* $\lceil p^{-1/2} \rceil$.

Sterrett (1957) Procedure: procedure S

An improvement of the D' procedure in the following way:

- If in the first stage of the procedure D' the group is infected, then in the second stage individuals are tested one-by-one until the first infected individual is identified.
- Then, the first stage of procedure D' is applied to the remaining (non-identified) individuals.
- The procedure is repeated until all individuals are identified.

Result

$$E_S(k,p) = \frac{1}{k}\left[2k - (k-2)q - \frac{1-q^{k+1}}{1-q}\right] \quad \textit{for} \quad k \geq 1.$$

It is easy to check that $E_S(1,p) = 1$.

Sterrett, A. (1957). On the detection of defective members of large populations. *The Annals of Mathematical Statistics* **28**, 1033–1036.

Infinite Population Case: Procedure S

Result

Let $p \in (0, p_U)$. Then (as a function of continuous variable k, $k \geq 1$) $E_S(k, p)$ has an absolute minimum at the unique zero of $E_S'(k, p)$.

From the above Result it follows that the optimal value k_S^* is the smallest k value which satisfies:

$$E_S(k, p) \leq E_S(k-1, p) \quad \text{and} \quad E_S(k, p) < E_S(k+1, p).$$

A Comparison of Procedures D, D', and S

The minimal (optimal) expected number of tests per 100 individuals for the procedures D, D', and S.

p	D		D'		S	
	k_D^*	$100E_D$	$k_{D'}^*$	$100E_{D'}$	k_S^*	$100E_S$
0.001	32	6.2759	32	6.2729	45	4.5844
0.005	15	13.91	15	13.879	21	10.535
0.01	11	19.557	10	19.47	15	15.172
0.05	5	42.622	5	41.807	7	35.977
0.10	4	59.39	4	57.567	5	52.288
0.13	3	67.483	3	64.203	4	60.042
0.15	3	71.921	3	68.308	4	64.784
0.20	3	82.133	3	77.867	3	74.933
0.25	3	91.146	2	84.375	3	83.854
0.30	3	99.033	2	90.5	2	90.5
0.32	1	100	2	92.88	2	92.88
0.35	1	100	2	96.375	2	96.375
0.38	1	100	2	99.78	2	99.78

Finite Population (of size N) Case

Set $E_A(k,\, p) = \dfrac{h_A(k)}{k}$, where $h_A(k)$ is the expected total number of tests.

For a given procedure A, we have to find the optimal *partition* $\{n_1, \ldots, n_I\}$ with $n_1 + \ldots + n_I = N$ for some $I \in \{1, \ldots, N\}$ such that $E_A(k,\, p)$ is minimal:

$$H_A(N) = \min_{m_1,\, m_2, \ldots, m_J} \sum_{i=1}^{J} h_A(m_i), \ \text{s.t.}, \ \sum_{i=1}^{J} m_i = N, \ \ J \in \{1, \ldots, N\}. \tag{1}$$

A common method to solve (1) is dynamic programming (DP):

$$h_A(1) = 1, \ \ H_A(0) = 0, \ \ H_A(1) = 1,$$
$$H_A(k) = \min_{0 \le x \le k-1} \{H_A(x) + h_A(k-x)\}, \quad k = 2, \ldots, N.$$

The computation complexity of the above DP algorithm is $O(N^2)$.

Example $N = 13$

If $p = 0.05$, then the optimal group size for an infinite population under procedure D' is $k^*_{D'} = 5$, and the optimal partition is

$$\{n_1, n_2, n_3\} = \{5, 4, 4\}$$

with $H_{D'}(13) = 5.489$.
For the procedure S with the same $p = 0.05$, $k^*_S = 7$ and the optimal partition is

$$\{n_1, n_2\} = \{6, 7\}$$

with $H_S(13) = 4.685$.

Lee and Sobel (1972) conjectured it as a general result for procedure D. Gilstein (1985) investigated procedure D'. We will prove this result for procedure S.

Lee, J.K., and Sobel, M. (1972). Dorfman and R_1-type procedures for a generalized group testing problem. *Mathematical Biosciences* **1**5, 317–340.

Gilstein, C. Z. (1985). Optimal partitions of finite populations for Dorfman-type group testing. *J. Stat. Plan. Inf.* **1**2, 385–394.

Finite Population Case: procedure S

Result

Suppose we apply the group testing algorithm S for a finite population of size N for a given p. Also suppose that $N = sk_S^*(p)$, *i.e. s subgroups of size* k_S^*. *Then the optimal partition is* $\left\{n_i = k_S^*, i = 1, \ldots, s\right\}$, *i.e.,* $I = s$ *and the infinite population optimal solution is the subgroup size of the optimal partition for the finite population.*

Result

Suppose we apply group testing algorithm S for a finite population of size N for a given p ($p \in (0, 1)$*) and we start with some partition* $\{m_1, \ldots, m_J\}$. *There exists a better (with respect to expected number of tests) partition* $\left\{m_1', \ldots, m_J'\right\}$ *with* $|m_j' - m_i'| \leq 1$ *for all* i, j.

The proof is based on convexity property of $h_S(x)$ with respect to x.

Finite Population Case: procedure S

Result

Suppose we apply the group testing algorithm S for a finite population of size N for a given p ($p \in (0,1)$). Denote a to be be an optimal group size under procedure S for infinite population, and $s = \lfloor \frac{N}{a} \rfloor$ (i.e, s groups of size a) and $\theta = N - sa$ (i.e., remainder $0 < \theta < a$). Then the optimal partition is one of the two following partitions.

(i) *Distribute the remainder θ among s groups (with initial size a) in such a way that $|n_i - n_j| \leq 1$ for all $i, j \in \{1, \ldots, s\}$.*

(ii) *Build up an additional group (group $s+1$) by taking the remainder θ and units from the above s groups (with initial size a) in such way that $|n_i - n_j| \leq 1$ for all $i, j \in \{1, \ldots, s, s+1\}$.*

A Comparison of Procedures D' and S

$s \times a$ means s groups of size a

p	D'		S			
	opt. partition	$100E_{D'}$	opt. partition	$100E_S$	$100E_1$	$H(P)$
0.001	$2 \times 33, 1 \times 34$	6.278	2×50	4.605	1.766	1.141
0.005	$5 \times 14, 2 \times 15$	13.884	5×20	10.537	4.749	4.541
0.01	10×10	19.470	$5 \times 14, 2 \times 15$	15.181	8.320	8.079
0.05	20×5	41.807	$5 \times 6, 10 \times 7$	36.018	28.958	28.640
0.10	25×4	57.567	20×5	52.288	47.375	46.900
0.13	$32 \times 3, 1 \times 4$	64.258	25×4	60.042	56.183	55.744
0.15	$32 \times 3, 1 \times 4$	68.396	25×4	64.784	61.485	60.984
0.20	$2 \times 2, 32 \times 3$	77.872	$32 \times 3, 1 \times 4$	74.974	72.875	72.192
0.25	50×2	84.375	$2 \times 2, 32 \times 3$	83.875	82.191	81.128
0.30	50×2	90.5	50×2	90.5	88.889	88.129
0.32	50×2	92.88	50×2	92.88	91.574	90.438
0.35	50×2	96.375	50×2	96.375	95.633	93.407
0.38	50×2	99.78	50×2	99.78	99.730	95.804

Optimal Nested Procedure (Sobel and Groll, 1959)

Define a binomial set to be a random sample of N units, each of which is defective with probability p and good with probability $q = 1 - p$. The nested procedure requires that between any two successive tests:

(i) future tests are concerned only with units not yet classified as good or defective,

(ii) n units not yet classified have to be separated into only (at most) two sets. One set of size $m \geq 0$, called the “defective set", is known to contain at least one defective unit if $m \geq 1$ (it is not known which ones are defective or exactly how many there are). The other set of size $n - m \geq 0$ is called the “binomial set" because we have no knowledge about it other than the original binomial assumption. Either of these two sets can be empty in the course of experimentation; both are empty at termination.

Sobel, M., Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech. J.* **3**8, 1179–1252.

Computational Complexity of the brutal Search

Sobel and Moon (1977) found the explicit formula for the total number $F(N)$ of the nested group testing procedures:

$$F(N) = 4^{2^N-1} \prod_{i=1}^{N} \left(1 - \frac{3}{2(i+1)})^{2^{N-i}}\right).$$

N	1	2	3	4	5	6
$F(N)$	1	2	10	280	235,200	173,859,840,000

Moon, J. W., Sobel, M. (1977). Enumerating a class of nested group testing procedures. *Journal of Combinatorial Theory, Series B* **2**3, 184–188.

Key Lemma and Computational Complexity

Lemma (Sobel and Groll (1959))

Given a defective set of size $m \geq 2$ and given that a proper subset of size x with $1 \leq x \leq m-1$ also proves to contain at least one defective, then the posteriori distribution associated with $m - x$ remaining units is precisely that $m - x$ independent binomial chance variables with common probability q of being good.

There was a large research effort to reduce the computational complexity:

- Sobel and Groll (1959): $O(N^3)$.
- Sobel (1960): $O(N^2)$.
- Kumar and Sobel (1971): reduced the computation complexity by half as compared with Sobel (1960).
- Hwang (1976): using the results for optimal binary trees (Huffman trees) and optimal alphabetic binary trees, reduced the computational complexity to $O(N)$ (not including the sorting effort).

Kumar, S., and Sobel, M. (1971). Finding a single defective in binomial group-testing. *J. Amer. Statist. Assoc.* **6**6, 824–828.

Hwang, F. K. (1976). An optimal nested procedure in binomial group testing. *Biometrics* **3**2, 939–943.

Connecting of GT and Coding Theory

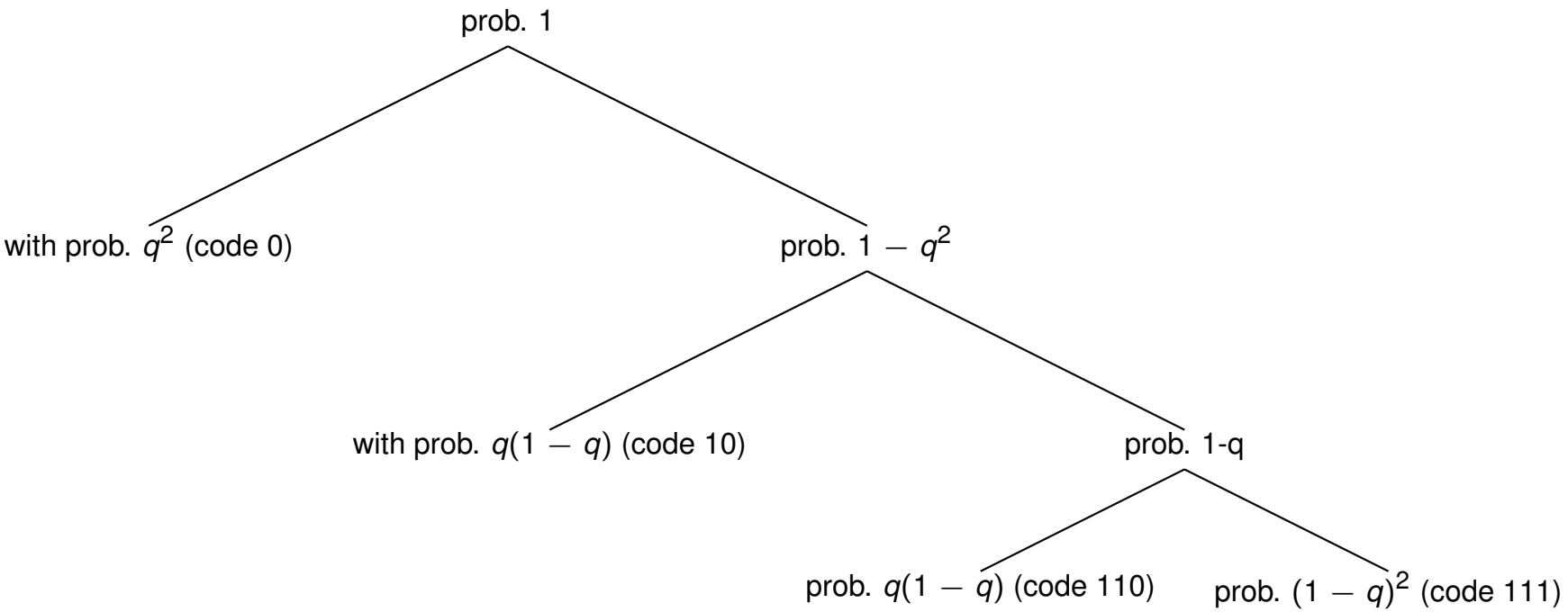
To demonstrate the connection, first consider the case when $N = 2$. There are $M = 2^2 = 4$ possible states of nature:

$$\begin{aligned} -- &\quad \text{with probability} \quad q^2, \\ -+ &\quad \text{with probability} \quad q(1-q), \\ +- &\quad \text{with probability} \quad q(1-q), \\ ++ &\quad \text{with probability} \quad (1-q)^2. \end{aligned}$$

$q = 1 - p$

Optimum Prefix Code: Huffman tree

If $q > (\sqrt{5}-1)/2$ $(p < (3-\sqrt{(5)})/2)$:



$L_4 = 3 - q - q^2$.

Noiseless Coding Theory

In general (for $N \geq 3$) the optimal group testing strategy does not coincide with the optimal prefix code of Huffman.

Therefore, L_M, $M = 2^N$ can serve as a theoretical lower bound which is not attainable in general.

The complexity of calculation of L_M is $O\left(M \log_2(M)\right), M = 2^N$ due to the sorting effort.

A well-known information theory result (*Noiseless Coding Theorem*):

$$H(P) \leq L_M \leq H(P) + 1,$$

where $H(P)$ is the Shannon formula of entropy,

$$H(P) = N\left\{p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}\right\}.$$

Robustness Investigation

In some situations there is only knowledge of an upper bound U of the parameter p.

Under a constant group size setting as in procedures D' and S, we can calculate the minimax group size k_A^{**} for procedures $A = D', S$ as

$$k_A^{**} = \arg \min_{k \in \mathbb{N}^+} \sup_{p \in (0,U]} L_A(k,p),$$

where $L_A(k,p) = E_A(k,p) - E_A(k^*(p),p), \quad A = D', S.$

Table : Robustness of the nested procedure R_1 vs procedures D' and S

	$U = 0.05$			$U = 0.10$			$U = 0.20$			
p	0.001	0.005	0.01	0.001	0.01	0.05	0.001	0.01	0.1	0.2
$100E_{D'}$	10.985	14.841	19.470	13.285	20.109	45.721	14.969	20.945	65.697	92.565
$100E_S$	7.975	11.241	15.185	10.628	16.138	37.760	13.024	17.647	55.928	85.889
$100E_1$	7.468	9.311	11.567	15.287	18.007	30.979	33.233	35.221	53.271	72.875
$100E_1^*$	4.511	6.578	9.194	7.468	11.567	28.958	15.287	18.007	47.375	79.988
$k_{D'}^{**}$	10	10	10	8	8	8	7	7	7	7
k_S^{**}	14	14	14	10	10	10	8	8	8	8

Malinovsky, Y., Albert, P. S. (2015). A note on the minimax solution for the two-stage group testing problem. *The American Statistician* **69**, 45–52.

Thank you!