Supplementary Materials
Revisiting nested group testing procedures: new results, comparisons and robustness

Yaakov Malinovsky*
Department of Mathematics and Statistics
University of Maryland, Baltimore County, Baltimore, MD 21250, USA
and
Paul S. Albert†
Biostatistics Branch, Division of Cancer Epidemiology and Genetics
National Cancer Institute, Rockville, MD 20850, USA

July 17, 2017

*Corresponding author
†The work was supported by the National Cancer Institute Intramural Program.

Web Appendix

F Implementation of Result 5

Remark 1. *(i) From Result 5, it follows that under a partition into s groups of sizes $l = \left\lfloor \frac{N}{s} \right\rfloor$ and $l+1$ (or alternatively $l+1 = a + \left\lfloor \frac{\theta}{s} \right\rfloor$), we have the following relationship:*

$$lg_l + (l+1)g_{l+1} = N, \;\; g_l + g_{l+1} = s,$$

where g_l is the number of groups of size l. The trivial solution of these equations shows that we have $g_l = s(l+1) - N$ and $g_{l+1} = N - sl$.

(ii) From Result 5, it follows that under a partition into $s+1$ groups of sizes $w = \left\lfloor \frac{N}{s+1} \right\rfloor$ and $w+1$ (or alternatively $w+1 = a + \left\lfloor \frac{\theta}{s} \right\rfloor$) we have the following relations:

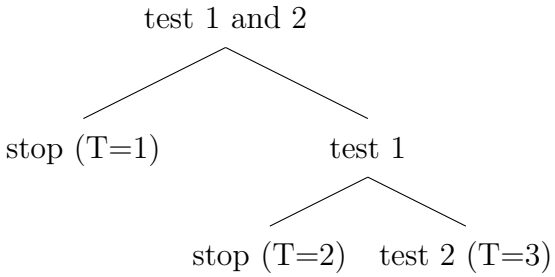
$$wg_w + (w+1)g_{w+1} = N, \;\; g_w + g_{w+1} = s+1,$$

The trivial solution of these equations shows that we have $g_w = (s+1)(w+1) - N$ and $g_{w+1} = N - w(s+1)$.

G Ungar Construction

We recall that Ungar's (1960) fundamental result examines the general problem when group testing is preferable to individual testing.
Result from Ungar (1960): The individual testing is optimal if and only if $p \geq \frac{3-\sqrt{5}}{2}$.

We need to show Ungar's (1960) construction for $N = 2$ to prove future results. The tree below presents a reasonable group testing algorithm for $N = 2$ as it was presented by Ungar (1960) (the left branch of the tree represent the negative test result, and the right branch represents the positive test result):



The expected number of tests per person is $\frac{1}{2}\left(3-q-q^2\right)$. Comparing it with 1, we conclude that this algorithm is better than individual testing when $p < \frac{3-\sqrt{5}}{2}$ or $q > \frac{\sqrt{5}-1}{2}$.

H Connection of group testing and coding theory

The connection of group testing with noiseless-coding theory was presented in group testing literature by Sobel and Groll (1959) and further investigated in Sobel (1960, 1967), Kumar and Sobel (1971), (Hwang, 1974, 1976), Glassey and Karp (1976), Wolf (1985) and Yao and Hwang (1990). For a comprehensive discussion, see Katona (1973).

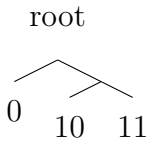
To demonstrate the connection, first consider the case when $N = 2$. There are 4 possible outcomes which correspond to the total number of tests $T = 1, 2, 3$. We use the code $-$ if a person is free from decease and $+$ otherwise. Therefore, there are $M = 2^2 = 4$ possible states of nature:

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

We use a binary (in which only the binary digits 0 and 1 are employed) *prefix code* for these 4 possible states of nature. The code is called a prefix code if no code word is the prefix (contained fully as the beginning) of any other code word. We want to code these 4 states with the binary prefix code with lengths l_1, l_2, l_3, l_4 such that the expected length

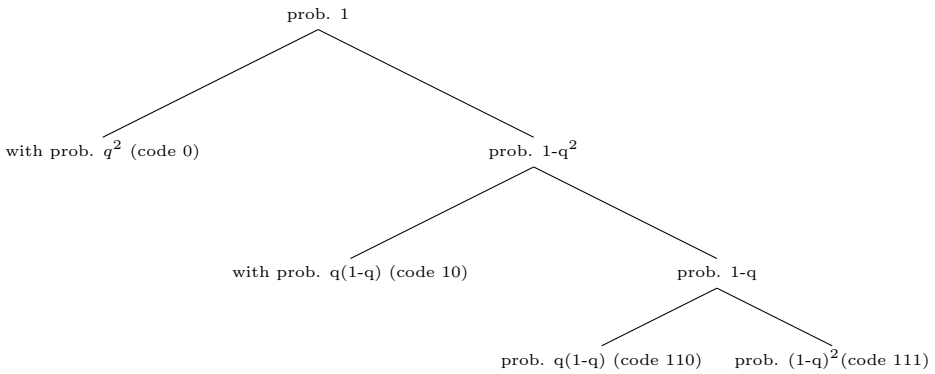
$$p_1 l_1 + p_2 l_2 + p_3 l_3 + p_4 l_4,$$

will be minimal, where $\{p_1, p_2, p_3, p_4\} = \{q^2, q(1-q), q(1-q), (1-q)^2\}$. The solution of this problem is due to Huffman (1952) and the simple explanation here is due to Rényi (1984). Without loss of generality, assume $p_1 \geq p_2 \geq p_3 \geq p_4$. Denote L_4 as the minimal expected length. For $M = 2$ with $p_1 \geq p_2$, the problem is trivial. For $M = 3$, with $p_1 \geq p_2 \geq p_3$, it has to be decided which of these probabilities should be assigned to the point which is one unit away from the root. It is clear that it must be the largest of the three probabilities (i.e., p_1), and that the two smaller ones (p_2 and p_3) should be assigned to the point which is 2 units away. Below are the assignment and the optimal prefix code as a tree presentation for $M = 3$:



As it was noticed, to the largest p_1 is assigned code 0, and to the two others, p_2, and p_3, are assigned codes 10 and 11. Now it is clear how to proceed with $M = 4$ with $p_1 \geq p_2 \geq p_3 \geq p_4$. Denote $p_{34} = p_3 + p_4$. In the previous case, $M = 3$ and we know how to construct a tree. On this tree, we will branch out two new branches from the node with p_{34} and put the numbers p_3 and p_4 at the two terminal nodes.

Now we can apply this optimal construction to the group testing with $N = 2$ ($M = 4$) For $q \geq 1/4$, we have $p_1 = q^2 \geq p_2 = q(1-q) \geq p_3 = q(1-q) \geq p_4 = (1-q)^2$, and for $q \geq (\sqrt{5}-1)/2$ we have $p_1 \geq p_{34} = p_3 + p_4 = 1 - q \geq p_2$.



Therefore, $L_4 = 3 - q - q^2$ and equals to the expected number of tests in the Ungar construction with $N = 2$.

Comment 1. *(i) For $N = 2$, the total expected number of tests under the procedures D' and S equals L_4 and, therefore, is optimal among all group testing algorithms.*

(ii) The optimal group size under procedures D' and S can be 2 (see Tables 1 and 2). Therefore, from part (i) it follows that cut-off points for the procedures D' and S equal UCP p_U.

Comment 2. *(i) Sobel (1967) noticed that 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.*

(ii) In general, the closed-form expression for L_M is not available, except for the case where p and N satisfy some condition (Jakobsson, 1978). It is also well known that the complexity of calculation of L_M is $O\left(M \log_2(M)\right), M = 2^N$ due to the sorting effort. Therefore, even for small N, obtaining the exact value of L_M seems to be impossible.

(iii) A well-known information theory result (Noiseless Coding Theorem; see, e.g., Katona (1973), Cover and Thomas (2006)) is

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

where $H(P)$ is the Shannon formula of entropy, $H(P) = N\left\{p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}\right\}$. It is clear that $H(P)$ is easy to calculate and can serve as a reference information lower bound for any group testing algorithm.

(iv) It is important to note that Yao (1988) obtained improvement over the information lower bound for the values of p close to p_U and Abrahams (1993) obtained it for values of p close to zero.

I Development of the optimal nested procedure

The definition of nested procedure R_1 was presented in Section 3 of the article. We need the following (Sobel, 1960) result and lemma. We will cite them verbatim.

For a defective set of size $m \geq 1$, denote the number of defective units Y and denote by Z the number of defectives presents in the subset of size x (with with $1 \leq x \leq m-1$)

randomly chosen from the defective set, then

$$P\left(Z=0\,|\,Y\geq 1\right)=\frac{q^x(1-q^{m-x})}{1-q^m}. \tag{1}$$

Lemma 1. 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.

Let $G(m,n)$ denote the minimum expected number of tests needed to classify all m units in defective set and the remaining $n-m$ units in a binomial set (G-situation). Define $H_1(n)=G(0,n)$ (H-situation). Then, from (1) and Lemma 1 above, we have the following recursion (dynamic programming) equation which will lead as to the goal $H_1(N)$:

$$\begin{aligned}
&H_1(0)=0,\ H_1(1)=1,\\
&H_1(n)=1+\min_{1\leq x\leq n}\left\{q^xH_1(n-x)+(1-q^x)G(x,n)\right\},\ \ n=2,\ldots,N, \qquad (2)\\
&G(1,n)=H_1(n-1)\ \ n=1,2,\ldots,N,\\
&G(m,n)=1+\min_{1\leq x\leq m-1}\left\{\frac{q^x-q^m}{1-q^m}G(m-x,n-x)+\frac{1-q^x}{1-q^m}G(x,n)\right\},\ 2\leq m\leq n=2,\ldots,N.
\end{aligned}$$

It can be verified that the complexity of the calculation using above DP algorithm (2) is $O(N^3)$.

Sobel (1960) reformulated the problem and found a solution with complexity $O(N^2)$ in the following way. The iteration equations (2) for $n=2,\ldots,N$ will always lead to the "break down" defective set of size $m\geq 2$. In particular, if the unit i is the first defective units in the defective set that we found, then we come to the H-situation with $n-i$ units (follows from Lemma 1 above). This observation allows us to proceed in the following way. Denote $F_1(m)$ as the minimum expected number of tests required to "break down" a defective set of size size $m\geq 2$ and for the first time reach H-situation when q is given. Then $F_1(m)$ does not depend on n and we can write

$$G(m,n)=F_1(m,n)+\sum_{i=1}^{m}\frac{q^{i-1}p}{1-q^m}H(n-i),\ \ 2\leq m\leq n=2,\ldots,N. \tag{3}$$

Denote

$$F_1^*(m)=\frac{1-q^m}{1-q}F_1(m),\ G^*(m,n)=\frac{1-q^m}{1-q}G(m,n).$$

Then combining (2) with (3) leads to the improved dynamic programming algorithm (below) with complexity $O(N^2)$:

$$H_1(0) = 0,\ H_1(1) = 1,$$
$$H_1(n) = 1 + \min_{1\le x\le n}\left\{q^x H_1(n-x) + (1-q)\left[F_1^*(x) + \sum_{i=1}^{x} q^{i-1} H_1(n-i)\right]\right\},\ \ n = 2,\dots,N, \tag{4}$$

$$F_1^*(1) = 0,$$
$$F_1^*(m) = \frac{1-q^m}{1-q} + \min_{1\le x\le m-1}\left\{q^x F_1^*(m-x) + F_1^*(x)\right\},\quad m = 2,\dots,N.$$

Comment 3. *(i) Kumar and Sobel (1971) reduced the computational complexity of* (4) *by half, showing that the value x in the last equation of* (4) *is bounded by $m/2$.*

(ii) For $N = 2$, $H_1(2)$ equals L_4 and, therefore, is optimal among all group testing algorithms. In this case the procedures R_1, S and D' are equivalent (see also Comment 1 part (i)).

We demonstrate the construction of an optimum nested algorithm with the following example.

Example: $p = 0.05,\ N = 13$

n	13	12	11	10	9	8	7	6	5	4	3	2
x_H	13	12	11	10	9	8	7	6	5	4	3	2
x_G	5	4	4	4	4	4	3	2	2	2	2	1

In this table, n is a size of the set C that still is not yet classified, x_H is the size of the subset that we have to check given that the set C is the binomial set (H-situation) and x_G is the size of the subset that we have to check given that set C is the defective set (G-situation).

J Matlab code for the optimum nested procedure

(i) **Matlab function "Fstar"**

#function [Fs Loc]=Fstar(p,n)

$\#q = 1 - p;$

$\#Fs = zeros(n, 1);\ Fs(1, 1) = 0;\ Loc = [];$

#for $m = 2 : 1 : n$

$\#f = floor(m/2);\ l = zeros(f, 1);$

#for $x = 1 : 1 : f$

$\#l(x) = (q^x) * Fs(m - x) + Fs(x);$

#end

$\#Fs(m) = (1 - q^m)/(1 - q) + min(l);$

#%Location of min loc

$\#ll = min(l);\ loc = find(l == ll);\ Loc = [Loc; m \quad loc];$

#end end

(ii) **Matlab function "H1"**

#function $[H_1\ D] = H1(p, n)$

$\#q = 1 - p;\ [Fs\ Loc] = Fstar(p, n);$

$\#H_1 = zeros(n, 1);\ H_1(1, 1) = 1;\ D = [];$

#for $N = 2 : 1 : n$

$\#l = zeros(N, 1);$

$\#qN = ((fliplr(eye(N - 1))) * (q.^[0 : 1 : N - 2]')). * H_1(1 : N - 1, 1);$

$\#S = [];$

#for $x = 1 : 1 : N - 1$

$\#l(x) = (q^x) * H_1(N - x) + (1 - q) * (Fs(x) + sum(qN(((N - 1) - (x - 1)) : (N - 1), 1)));$

$\#S = [S; 1 + l(x)\ N\ x];$

#end

$\#l(N) = (1 - q) * (Fs(N) + sum(qN));$

$\#S = [S; 1 + l(N)\ N\ \ N];\ SS = S(S(:, 1) == min(S(:, 1)), :);$

$\#D = [D; SS];\ H_1(N) = 1 + min(l);$ end

$\#D = D(:, 2 : end);$

References

Abrahams, J. (1993). An improvef lower bound on the minimum expected number of binomial group tests. *Prob. Eng. Inform. Sci.* **7**, 121–124.

Cover, T.M, and Thomas, J. A. (2006). Elements of Information Theory. 2nd Edition. *Wiley.*

Glassey, C. R., and Karp, R. M. (1976). On the optimality of Huffman Trees. *SIAM Journal on Applied Mathematics* **3**1, 368–378.

Huffman, D. A. (1952). A Method for the Construction of Minimum-Redundancy Codes. *P*roceedings of the I.R.E. **4**0, 1098–1101.

Hwang, F. K. (1974). On finding a single defective in binomial group testing. *J. Amer. Statist. Assoc.* **6**9, 146–150.

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

Jakobsson, M. (1978). Huffman coding in bit-vector compression. *Information Processing Letters* **7**, 304–307.

Katona G. O. H. (1973). Combinatorial search problems. *J.N. Srivastava et al., A Survey of combinatorial Theory,* 285–308.

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

Rényi, A. (1984). A diary on information theory. *Alcadémiai Kiadó, Budapest.*

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.

Sobel, M. (1967). Optimal group testing. *Proc. Colloq. on Information Theory, Bolyai Math. Society, Debrecen, Hungary.*

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

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

Yao, Y. C. (1988). An improvement over the information lower bound in binomial group testing. *Prob. Eng. Inform. Sci.* **2**, 313–320.

Yao, Y. C., Hwang, F. K. (1990). On optimal nested group testing algorithms. *J. Stat. Plan. Inf.* **2**4, 167–175.