ENUMERATION OF TREES ASSOCIATED WITH A CLASS OF NESTED STRATEGIES IN GROUP-TESTING

J. W. Moon and M. Sobel*
Univ. of Alberta Univ. of Minnesota

Technical Report No. 208

May, 1973

* Supported by U. S. Army Grant DA-ARO-D-31-124-72-6187 at the University of Minnesota, Minneapolis, Minnesota.

1. Introduction.

The problem of group-testing is concerned with classifying each of N units into one of two disjoint categories which we call satisfactory and defective (or simply, good and bad). The characteristic feature is that any number of units can be tested simultaneously but the results of a single test on x units, without any chance of error, is that either (i) all x are good, or (ii) at least one of the x units is bad, but it is not known how many or which ones. A strategy (or procedure) consists of sequential sampling scheme that leads to the classification of all N units. Each strategy corresponds to a rooted tree with 2^N terminal points corresponding to the 2^N possible states of nature: each of the N units can be good or bad. We are concerned in this paper with the total number of trees (or strategies) of a special type called NWR; NWR standing for nested with recombination. These strategies have two properties: 1. If a subset of size $m \geq 2$ is found to contain at least one defective then we test a proper subset of these m on the very next test; this is the nested property. 2. If $n \leq N$ units are still not classified, a proper subset S of size s contains at least one defective unit and we take from S a proper subset S' of size s' then the two sets of sizes $n - s$ and $s - s'$ are combined to form a single set before the next test; this is the recombination property. Explicit examples of these NWR procedures can be found in [3].

Remark: It is shown in [3] that the above recombination can be accomplished without any loss of information if we assume a binomial probability structure for the original N units, i.e., independent binomials with a common probability p of being defective. To justify property 2 we adopt the same probability structure, although this may also hold for some other probability formulations. For examples, if the original N units were drawn from a larger set containing a known (or unknown) number of bad units, then we can again combine these two sets without loss of information.

In general the strategy determines the size of the subset at each stage (randomization determines the actual units selected) and various papers such as [3] and [4] deal with the problem of determining which sizes will minimize the expected number of tests needed to classify all N units, for given finite N and also for $N = \infty$. Our object is to determine the total number $h(N)$ of NWR group-testing strategies (or trees) for classifying N units.

We develop both exact and asymptotic formulas for $h(N)$, the total number of NWR strategies for N units. The asymptotic result is that $h(N)$ is approximately $(N \to \infty)$ equal to α^{2^N} / c where $\alpha = 1.526753...$ is defined in (3.11) below and c is an arbitrary positive constant; this approximation holds in the sense that $\log(\alpha^{2^N} / c) / \log h(N) \to 1$ as $N \to \infty$. If we take $c = 4$ then we obtain the stranger results that (i) the ratio $R = \alpha^{2^N} / ch(N) \to 1$ as $N \to \infty$, (ii) that R is strictly increasing with N and hence (iii) that $\alpha^{2^N} / 4$ is a lower bound to $h(N)$ for all $N \geq 1$.

2. Recurrence relations.

For $0 \le m \le n \le N$, let $g(m, n)$ denote the total number of NWR group-testing procedures for classifying the current number n of unclassified units when we know that a particular subset of size m contains at least one bad unit; for convenience let $g(0, 0) = 0$. Note that $g(0, n) = h(n)$ for all $n \ge 0$ and this is in accord with previous notation [2] in group-testing.

If $m = 1$ then we have effectively identified one bad unit and there are n-1 units left to be classified, so

$$(2.1) \qquad g(1, n) = g(0, n-1) \qquad (n = 1, 2...).$$

If $m = 0$ and $n \ge 1$ we test a subset of size i $(1 \le i \le n)$. If all i are good then there are $g(0, n-i)$ ways of continuing; if at least one of the i units are bad then we have $g(i,n)$ ways of continuing. It follows that

$$(2.2) \qquad g(0, n) = \sum_{i=1}^{n} g(0, n-i)\, g(i, n) \qquad (n = 1, 2,...).$$

If $2 \le m \le n$ we use property 1 and test a proper subset of the m units that we know includes at least one bad unit, i.e., $1 \le i \le m-1$. If all are good then there are $g(m-i, n-i)$ ways to continue; if at least one of the i units is bad then we use property 2 and combine the $n-m$ and $m-i$ units, so that there are $g(i, n)$ ways of continuing. Hence

$$(2.3) \qquad g(m, n) = \sum_{i=1}^{m-1} g(i, n) \; g(m-i, n-i) \qquad (2 \le m \le n).$$

3. Main results.

We use the fact (cf. Feller [2; p.73] that the numbers

(3.1) $$C_m = \frac{1}{m}\binom{2m-2}{m-1} \qquad (m = 2, 3, \ldots)$$

satisfy the recurrence relation

(3.2) $$C_m = C_1\ C_{m-1} + C_2\ C_{m-2} + \ldots + C_{m-1}\ C_1.$$

Theorem 1: For $1 \le m \le n$

(3.3) $$g(m, n) = C_m\ g(0, n-1)\ g(0, n-2) \ldots g(0, n-m).$$

Proof The result holds for $m = 1$ by (2.1) since $C_1 = 1$. Suppose it holds whenever $1 \le m \le t-1$, where $2 \le t \le n$. Then

(3.4) $$\begin{aligned} g(t, n) &= \sum_{i=1}^{t-1} g(i, n)\ g(t-i, n-i) \\ &= g(0, n-1)\ g(0, n-2) \ldots g(0, n-t)\left(\sum_{i=1}^{t-1} C_i\ C_{t-i}\right) \\ &= C_t\ g(0, n-1)\ g(0, n-2) \ldots g(0, n-t) \end{aligned}$$

by (2.3), (3.2) and the induction hypothesis; the result (3.3) follows by induction.

Theorem 2: For $n \ge 2$

(3.5) $$g(0, n) = C_{n+1}\ g(0, n-1)\ g(0, n-2) \ldots g(0, 1).$$

Proof It is easy to verify that (3.5) holds for $n = 2$ since $C_3 = 2$ and $g(0, 1) = 1$. Suppose it holds whenever $2 \le n \le k-1$, where $k \ge 3$. Then

(3.6) $$g(0, k) = \sum_{i=1}^{k} g(i, k)\ g(0, k-1)$$

$$= g(0,\ k-1)\ \ g(0,\ k-2)\ \ldots\ g(0,\ 1)\left(\sum_{i=1}^{k} c_i\ \ c_{k+1-i}\right)$$

$$= c_{k+1}\ \ g(0,\ k-1)\ \ g(0,\ k-2)\ \ldots\ g(0,\ 1).$$

by (2.2), Theorem 1, the induction hypothesis and (3.2); the result (3.5) follows by induction.

A useful algorithm for computing $g(0,\ N) = h(N)$ is given in

Corollary 1: For $N \geq 1$

$$(3.7)\qquad h\ (N) = \frac{c_{N+1}}{c_N}\ h^2(N-1) = \frac{2(2N-1)}{N+1}\ h^2(N-1).$$

Proof This follows directly from (3.5).

Explicit formulas for $g(0,\ N) = h(N)$ are given in

Corollary 2: For $N \geq 2$

$$(3.8)\qquad h(N) = \frac{2^{2^{N-1}-1}(2N-1)(2N-3)^2\ (2N-5)^4 \ldots\ 5^{2^{N-3}} \cdot 3^{2^{N-2}}}{(N+1)\ N^2(N-1)^4 \ldots\ 4^{2^{N-3}} \cdot 3^{2^{N-2}}}$$

$$= \frac{(2N)!\ (2N-2)!\ [(2N-4)!]^2\ [(2N-6)!]^4 \ldots\ [2!]^{2^{N-2}}}{(N+1)!\ [N!]^2\ [(N-1)!]^3\ [(N-2)!]^6 \ldots\ [2!]^{3 \cdot 2^{N-3}}}\ .$$

Proof The first expression in (3.8) is obtained by iterating the result in (3.7). Inserting the factors $2N$, $(2N-2)^2$, $(2N-4)^4$, etc. in both numerator and denominator, we then make use of the elementary result that $(2N)\ (2N-2)\ldots 2 = 2^N(N!)$ in the denominator and obtain the second expression of (3.8).

Some idea of the size of these numbers can be had by noting that the first few values are: $h(1) = 1$, $h(2) = 2$, $h(3) = 10$, $h(4) = 280$, $h(5) = 235{,}200$ and $h(6) = 173{,}859{,}840{,}000$. For $N = 33$ the order of magnitude of $h(N)$ is $10^{10^{10}}$ and for $N = 300$ it is $10^{10^{100}}$. This confirms the idea that indirect methods are needed for finding the best strategy in group-testing and that an asymptotic result for $H(N)$ would be desirable.

Our asymptotic result is given in the following three corollaries.

Corollary 3: For $N \geq 1$

$$(3.9) \qquad h(N) = 4^{2^{N}-1} \prod_{i=1}^{N} \left\{ 1 - \frac{3}{2(i+1)} \right\}^{2^{N-i}}$$

Proof The formula clearly holds for $N = 1$. For $N \geq 2$ it follows from Theorem 2 and the expression for $C_{N+1}/\,C_N$ in (3.7), i.e., for $N \geq 1$

$$(3.10) \qquad C_{N+1} = 4 \left\{ 1 - \frac{3}{2(N+1)} \right\} C_N.$$

Corollary 4: For any constant $c > 0$.

$$(3.11) \qquad \lim_{N \to \infty} [c\,h(N)]^{2^{-N}} = 4 \prod_{i=1}^{\infty} \left\{ 1 - \frac{3}{2(i+1)} \right\}^{2^{-i}} = 1.526753\ldots$$

Proof This follows from Corollary 3 and some computer-based calculations for the infinite product.

Corollary 5: For $N \to \infty$

$$(3.12) \qquad \log\,[4h(N)] - 2^{N} \log \alpha = \log \left[\frac{4h(N)}{\alpha^{2^{N}}} \right] \to 0$$

and the approach is strictly monotonic from above; here and below log is to the natural base e.

Proof Using (3.9) and the definition of α in (3.11)

$$\log[4h(N)] - 2^N \log \alpha = 2^N \log\left[\prod_{i=N+1}^{\infty} \left\{1 - \frac{3}{2(i+1)}\right\}\right]^{-2^{-i}}$$

$$= - \sum_{j=1}^{\infty} \frac{1}{2^j} \log\left\{1 - \frac{3}{2(N+j+1)}\right\} < \log\left[1 - \frac{3}{2(N+2)}\right] \to 0,$$

where we used the monotonicity of $- \log(1 - x)$ for $0 < x < 1$. Moreover the summation shows that this quantity is strictly decreasing with N; this proves the corollary.

From Corollary 5 it immediately follows that the ratio $R = \alpha^{2^N}/4h(N) \to 1$ as $N \to \infty$ and hence also that the approximation $\alpha^{2^N}/4$ is a lower bound for all $N \geq 1$.

4. Other Expressions for α

It may be of some interest to point out that the logarithm of the constant α can also be written as a complex integral. First we show

Corollary 6:

$$(4.1) \qquad \alpha = \frac{2^{1-1/\sqrt{2}}[f(1/\sqrt{2})]^{1/\sqrt{2}}}{[f(\frac{1}{2})]^{2+1/\sqrt{2}}},$$

where $f(x) = \prod_{j=1}^{\infty} j^{x^j}$ and $0 < x < 1$,

Proof From the definition of α in (3.11) we easily obtain

$$\alpha = \frac{4 \prod_{j=1}^{\infty} (2j-1)^{(1/2)^j}}{\prod_{j=1}^{\infty} \{2(j+1)\}^{(1/2)^j}} = \frac{4\left[\prod_{j=1}^{\infty} (2j-1)^{(1/\sqrt{2})^{2j-1}}\right]^{1/\sqrt{2}}}{2\left[\prod_{j=1}^{\infty} (j+1)^{(1/2)^{j+1}}\right]^2} \cdot \frac{\left[\prod_{j=1}^{\infty} (2j)^{(1/\sqrt{2})^{2j}}\right]^{1/\sqrt{2}}}{\left[\prod_{j=1}^{\infty} (2j)^{(1/2)^j}\right]^{1/\sqrt{2}}}$$

$$= \frac{2^{1-1/\sqrt{2}} \left[\prod_{j=1}^{\infty} j^{(1/\sqrt{2})^j} \right]^{1/\sqrt{2}}}{\left[\prod_{j=1}^{\infty} j^{(1/2)^j} \right]^{2+1/\sqrt{2}}},$$

which is the desired result stated in (4.1).

We note that to evaluate $f(x)$ above it is sufficient to evaluate

$$(4.2) \qquad F(x) = \log f(x) = \sum_{j=1}^{\infty} (\log j)\, x^j \qquad (0 < x < 1).$$

Using functions related to the Riemann zeta function, we define

$$(4.3) \qquad \varphi(x, s) = \sum_{n=1}^{\infty} \frac{x^n}{n^s} = x\, \Phi\, (x, s, 1),$$

where $\Phi(z, s, v)$ is the well-known (cf. [1; pp 27-28]) function

$$(4.4) \qquad \Phi(z, s, v) = \sum_{n=0}^{\infty} \frac{z^n}{(v+n)^s} \qquad (\text{Re } v > 0;\ \text{Re } s > 0;\ |z| < 1)$$

$$= \frac{-\Gamma(1-s)}{2\pi i} \int_{\infty}^{0+} \frac{(-t)^{s-1} e^{-vt}}{1-z\, e^{-t}}\, dt \qquad (\text{Re } v > 0).$$

From (4.3) we note that for $0 < x < 1$

$$-\frac{\partial\varphi}{\partial s}(x, 0) = \sum_{n=1}^{\infty} (\log n)\, x^n = F(x) = -\, x \frac{\partial}{\partial s} \Phi\, (x, 0, 1)$$

Hence by putting x for z in (4.4), differentiating with respect to s and setting $s = 0$, we obtain

$$(4.5) \qquad F(x) = \frac{x}{2\pi i} \int_{\infty}^{0+} \frac{\gamma + \log(-t)}{t(e^t - x)}\, dt$$

where we used the fact that $[-\Gamma'(x)]_{x=1} = \gamma =$ Euler's constant, .577...

Taking logs on both sides of (4.1) and using (4.5), we can now write another expression for α in the form

$$(4.6) \qquad \log \frac{\alpha}{2^{1-1/\sqrt{2}}} = \frac{1}{2\pi i} \int_{\infty}^{0+} \left[\frac{\gamma + \log(-t)}{2\, t\sqrt{2}} \right] \frac{[2 - e^{t}(1 + \sqrt{2})]}{(e^{t} - \frac{1}{2})\,(e^{t} - \frac{1}{\sqrt{2}})}\, dt$$

5. The problem of Distinguishable units

It should be clear that in the above enumeration of strategies all units were considered indistinguishable except for being good or bad. The units can be regarded as put in random order once before any testing is started. Hence if we want to count strategies with all units distinguishable the number of strategies will of course be much larger. One way of getting this answer for the distinguishable-units problem is to let $f^{*}(i, n)$ denote this larger number and replace (2.1), (2.2) and (2.3), respectively, by

$$(5.1) \qquad f^{*}(i, n) = f^{*}(0, n-1) \qquad\qquad (n = 1, 2, \ldots),$$

$$(5.2) \qquad f^{*}(0, n) = \sum_{i=1}^{n} \binom{n}{i} f^{*}(i, n)\, f(0, n-i) \qquad\qquad (n = 1, 2, \ldots),$$

$$(5.3) \qquad f^{*}(m, n) = \sum_{i=1}^{m-1} \binom{m}{i} f^{*}(i, n)\, f^{*}(m-i, n-i) \qquad\qquad (2 \le m \le n).$$

If we now let $g^{*}(i, n) = f^{*}(i, n)/i!$ and $g^{*}(0, n) = f^{*}(0, n)/n!$, then the resulting equations for (5.2) and (5.3) are the same as in (2.2) and (2.3) and for (5.1) we obtain $g^{*}(1, n) = (n-1)!\ g^{*}(0, n-1)$. Then (3.3) becomes for $1 \le m \le n$

$$(5.4) \qquad g^{*}(m, n) = C_{m} \prod_{j=1}^{m} \{(n-j)!\ \ g^{*}(0, n-j)\}$$

and hence

$$(5.5) \qquad f^{*}(m, n) = m!\ C_{m} \prod_{j=1}^{m} f^{*}(0, n-j).$$

In place of (3.5) we obtain for $n \geq 2$ (since $f^*(0, 1) = 1$)

$$(5.6) \qquad g^*(0, n) = C_{n+1} \prod_{j=1}^{n-1} \{ j!\, g^*(0, j) \}$$

and hence, letting $h^*(n) = f^*(0, n)$,

$$(5.7) \qquad h^*(n) = f^*(0, n) = n!\; C_{n+1} \prod_{j=1}^{n-1} f^*(0, j)$$

Some of these numbers are: $h^*(2) = 4$, $h^*(3) = 120$, $h^*(4) = 161{,}280$, $h^*(5) = 390{,}168{,}576{,}000$. Corresponding to (3.7) we have

$$(5.8) \qquad h^*(N) = N \frac{C_{N+1}}{C_N} [h^*(N-1)]^2 = \frac{2N(2N-1)}{N+1} [h^*(N-1)]^2 .$$

and as a result we can write

$$(5.9) \qquad h^*(N) = N! \prod_{j=2}^{N-1} \{j!\}^{2^{N-1-j}} \; h(N).$$

Letting β be defined by

$$(5.10) \qquad \beta = \prod_{j=2}^{\infty} \{j!\}^{2^{-(j+1)}} ,$$

we note that

$$(5.11) \qquad \lim_{N \to \infty} \left\{\frac{4h^*(N)}{N!}\right\}^{(1/2)^N} = \alpha\, \beta$$

so that the asymptotic $(N \to \infty)$ result for $h^*(N)$ is

$$(5.12) \qquad h^*(N) \sim \frac{N!}{4} (\alpha\, \beta)^{2^N} .$$

The value of β can be found by taking logs in (5.10) and rearranging the order of terms in the resulting infinite series; this gives

$$(5.13) \qquad \log \beta = \sum_{j=1}^{\infty} \frac{\log(j!)}{2^{j+1}} = \sum_{j=1}^{\infty} \frac{\log\ j}{2^j} = F(\tfrac{1}{2})$$

where $F(x)$ is given in (4.2) and (4.5). The numerical values for $F(\frac{1}{2})$, β and $\alpha\, \beta$ are .507834, 1.661688 and 2.536988, respectively.

Acknowledgment

We wish to thank Prof. W. Miller of the University of Minnesota for his help in the derivation of equations (4.3), (4.4) and (4.5).

References

[1] Erdelyi, A. et al. Higher Transcendental Functions Vol. 1 McGraw-Hill, New York.

[2] Feller, W. (1957). An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley, New York.

[3] Sobel, M. and Groll, P. A. (1959). Group-testing to eliminate efficiently all defectives in a binomial samples, Bell System Tech. Jour. **38** 1179-1252.

[4] Sobel, M. (1960). Group-testing to classify all defectives in a binomial sample. A contribution in Information and Decision Processes, Ed. R. E. Machol, McGraw-Hill, New York. p. 127-161.

University of Alberta

University of Minnesota