FINITE AND INFINITE MODELS FOR GENERALIZED GROUP-TESTING WITH UNEQUAL PROBABILITIES OF SUCCESS FOR EACH ITEM*

by

Elliott Nebenzahl
California State Univ. at Hayward
Hayward, California

and Milton Sobel
Univ. of Minnesota
Minneapolis, Minn.

Technical Report No. 199

February 1973

University of Minnesota
Minneapolis, Minnesota

*Sponsored partially by U.S. Army Grant DA-ARO-D-31-124-72-6187 at the University of Minnesota, Minneapolis, Minnesota and partially by contract NIH-E-71-2180 at the University of California Medical School in San Francisco, California.

1. Introduction.

In group-testing we are allowed to test any number of units from the same or different sources but each test of a batch of (say, x) units gives us one of two possible results: i) either all x units are good or ii) at least one of the x units is defective and we don't know which one(s). In this paper our goal is again to classify units efficiently in the sense of minimizing the number of tests required. This paper is a generalization of previous work in group-testing [3], [4] in the sense that we allow units to have different probabilities $q_i (i = 1,2,\ldots, k)$ of being good whereas previous work assumed a common value of q for all units. Each value of i $(i = 1,2,\ldots, k)$ corresponds to a different source or stream of units and each stream represents an assembly line type of operation with an unending number of units coming forth to be tested. Each unit from the i^{th} stream can only be good (with probability q_i) or defective (with probability $p_i = 1 - q_i$). All units from the same or different streams represent independent binomial chance variables.

The one restriction that we put on the plan or strategy for testing units is that no stream should be held up indefinitely, i.e., any unit in any of the k streams will be classified in some finite number of tests.

For convenience we sometimes group the units into sets of size k, where the j^{th} set consists of the j^{th} unit from each of the k sources. Let c denote the total number of such sets; we consider both the case $c = \infty$ and the case in which c is large but finite. For convenience and in order to make meaningful comparisons we rephrase our goal as the minimization of the expected number of tests per set of units classified, subject to the restriction mentioned above.

For any set of q_i-values we can always write these as powers of one q (say, the smallest q). We do not treat the most general case of unequal

q_i-values since we assume that the q_i-values are integer powers of the smallest q. In this case we can interpret the unit with probability (say) q^3 as being a set of 3 units, each with probability q, and we only want to know whether this set contains all good units or at least 1 defective unit. The advantage of this interpretation is that we can bring to bear on the problem information from other papers such as [3]. Below we refer to this idea as group-testing with groups or the G-point of view as opposed to the individual or I-point of view.

2. Vertical vs. Horizontal Procedures.

A vertical procedure is one for which the probability that a unit is not classified until after m units from subsequent sets have been classified tends to zero as $m \to \infty$. This insures us that no one source will be held up indefinitely. The problem of finding the optimal vertical procedure is not easy and it turns out to be useful to find the optimal procedure in another class of procedures that we call 'horizontal'. One of the main results of this paper is to point out that for every horizontal procedure we can find a vertical procedure which is equivalent in the sense of having the same expected number of tests per set classified.

To explain the horizontal procedure we assume a large finite c and later let $c \to \infty$. Rather than give a formal definition of a horizontal procedure, we illustrate it by an example. Suppose $k = 4$ and let x_i denote a unit from the i^{th} source which has probability q^i of being good ($i = 1,2,3,4$). To be specific, we take $q = .9$. A horizontal procedure S_h gives us an ordered list of preferred batch structures for testing and a plan (or tree) for testing each of these batches. Suppose a particular horizontal procedure $S_h^{(4)}$ gives us the list

$$
(2.1) \qquad S_h^{(4)} = \begin{array}{ll} \underline{\text{Stage 1}}: & (x_3, x_4) \\ \underline{\text{Stage 2}}: & (x_1, x_2, x_4) \\ \underline{\text{Stage 3}}: & (x_1, x_2, x_2, x_2) \\ \underline{\text{Stage 4}}: & (x_1, x_1, x_1, x_1, x_1, x_1, x_1) \ . \end{array}
$$

Then in Stage 1 we continue to test pairs (x_3, x_4) according to a specific given plan specified by the procedure until the units x_3 are all classified. A particular plan (later shown to be optimal for the example above) is the following , where arrows to the left (right) indicate success (failure):

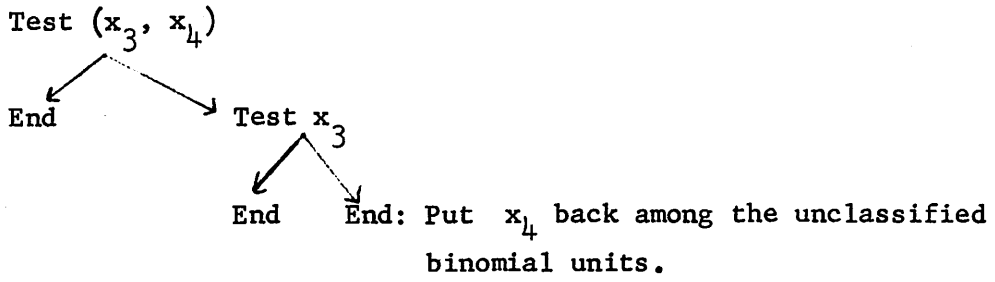


Figure 1. Plan (or Tree) for Testing (x_3, x_4).

Since we start with the same number of x's of each type it is easy to see that the x_3's will be depleted (i.e., classified) before the x_4's under the above scheme.

In the group-testing terminology we only use plans that take us from one H-situation (where the units are all binomially distributed) to the very next H-situation.

In Stage 2 of (2.1) we continue testing triples (x_1, x_2, x_4) according to a specific plan given by the procedure $S_h^{(4)}$, until another type of unit is depleted; in this case the x_4 is depleted earlier. In Stage 3 we continue until the x_2's are depleted and finally in Stage 4 we test seven units of type x_1 according to a specific plan until all the remaining units are classified.

It should be carefully noted that when we let $c \to \infty$ the limit of the horizontal procedure is not a procedure at all. However we are interested in the limit L_∞ as $c \to \infty$ of the expected number of tests per set classified since, as we shall see later, there is a vertical procedure with the same value $L = L_\infty$ where L is the expected number of tests per set classified.

Furtheremore we can find the horizontal procedure with the smallest L_∞-value. We have some reasons to believe that these equivalent vertical procedures are optimal in the entire class of vertical procedures.

A vertical procedure is defined to be equivalent to a horizontal procedure if it has an expected number of tests per set classified that is equal to the limiting L-value of the horizontal procedure.

3. An Example of a Vertical Procedure Equivalent to a Horizontal Procedure.

Using group-testing terminology we define a G-situation to be one in which we have information that some set of units contains at least one defective. The restriction to nested procedures in which we try to locate 1 defective unit in the smallest possible number of tests then gives us an R_1-type procedure (cf. [3]) in which the G and H-situations are the only two that are possible.

For the example illustrating equivalence suppose $k = 3$ and for the three types of units x_1, x_2, x_3 we take $q_1 = .9$ and $q_2 = (.9)^2$ and $q_3 = (.9)^4$ respectively. From the G-point of view we have an equal number of groups of sizes 1, 2 and 4; all units have the common probability .9 of being good and we are only interested in knowing whether or not each entire group is good. From the I-point of view, in the sequal we shall refer to the unit that has probability q^i as type i or simply as the i-unit ($i = 1,2,\ldots$) when the i-values are different; hence for our example above a test on (1, 2, 4) is equivalent to a test on (x_1, x_2, x_3). We now define a horizontal procedure for the above problem and find its L_∞- value; we then derive the equivalent vertical procedure and prove the equivalence by showing that $L = L_\infty$.

A tree is a rule or plan, as in Figure 1 above, which specifies what units to test i) initially in the H-situation and ii) in each succeeding G-situation until the very next H-situation is reached. For convenience we refer to a particular tree by the units tested initially, e.g., the particular tree for testing (x_1, x_2, x_3) can be referred to as the (1, 2, 4)-tree.

Let the horizontal procedure S_h be defined by

$$S_h = \left\{ \begin{matrix} (1, 2, 4) \\ (2, 4) \\ (4, 4) \end{matrix} \right\}, \tag{3.1}$$

where the (1, 2, 4), (2, 4) and (4, 4) trees are given by

$$x = (1, 2, 4) \xrightarrow{\text{End}} G(1, 2, 4)_{x=(1,2)} \xrightarrow{\text{End}} G(1, 2)_{x=(1)} \xrightarrow{H(4)} H(2, 4) , \tag{3.2}$$

$$x = (2, 4) \xrightarrow{\text{End}} G(2, 4)_{x=(2)} \xrightarrow{\text{End}} H(4) , \tag{3.3}$$

$$x = (4, 4) \xrightarrow{\text{End}} G(4, 4)_{x=(4)} \xrightarrow{\text{End}} H(4) . \tag{3.4}$$

In each of the above trees the horizontal (resp., slanted) arrows corresponds to a failure (resp., success) on that particular trial. The expression "H(4)" (say) in (3.2)(say) means that from the (1, 2, 4) grouping that we tested on the first step of the tree, a single 4-unit has been left unclassified at the end of the tree and returns to the binomial state; the word "End" in (3.2) means that the entire (1, 2, 4) grouping has been classified. We start testing with c sets of (1, 2, 4)'s and work (1, 2, 4)-trees until we exhaust the 1-units. The number of such trees is, of course, c and the number of tests involved is $cE\{T|T_{124}\}$, where $E\{T|T_{124}\}$ is the expected number of tests per (1,2,4)-tree. For large c ($c \to \infty$), the number of 4's (resp., 2's) remaining after the 1's have been exhausted is $c(1-q^3)$ (resp., $c(1-q)$). A single 2-unit is used up on each

(2, 4)-tree and thus the number of such trees until the 2's are exhausted is $c(1-q)$; since q^2 4's are classified per (2, 4)-tree, the number of 4-units remaining is $c(1-q^3) - c(1-q)q^2 = c(1-q^2)$. Also, $(1+q^4)$ 4's are classified per (4, 4) tree and thus it takes $\frac{c(1-q^2)}{1+q^4}$ such trees to exhaust the 4's, with the number of tests involved equal to $\frac{c(1-q^2)}{1+q^4} E\{T|T_{44}\}$. It is easy to verify that

$$
\begin{aligned}
E\{T|T_{124}\} &= 3 - q^3 - q^7 \\
E\{T|T_{24}\} &= 2 - q^6 \qquad (3.5) \\
E\{T|T_{44}\} &= 2 - q^8 \quad .
\end{aligned}
$$

Therefore,

$$
L_\infty = \frac{c(3-q^3-q^7) + c(1-q)(2-q^6) + \frac{c(1-q^2)}{1+q^4}(2-q^8)}{c} \approx 2.1196 \qquad (3.6)
$$

where L_∞ is written both as a ratio of polynomials for q close to .9 and also numerically at $q = .9$.

We now define the equivalent vertical procedure. Let the S_i $(i = 1,2,\ldots, 5)$ which represent the five possible H-situations be defined by

$$
\begin{aligned}
&S_1 = H(\ldots.), \; S_2 = H(4,\ldots), \; S_3 = H(4,4,\ldots), \; S_4 = H(2,4,\ldots), \\
&S_5 = H(2,4,4,\ldots),
\end{aligned} \qquad (3.7)
$$

where $H(\ldots.)$ represents a binomial state with an equal number of 1's, 2's, and 4's; $H(4,\ldots.)$ represents one where there is an extra 4, etc. Our vertical procedure can then be described by the trees

$$
\begin{array}{ccccc}
 & & S_i & & S_{i+1} \\
S_i & \nearrow\!\!\rightarrow & G(1,2,4) & \nearrow\!\!\rightarrow & S_{i+3} \\
x=(1,2,4) & & x=(1,2) & & x=(1)
\end{array}
\qquad (i = 1,2), \qquad (3.8)
$$

$$
\begin{array}{ccccc}
 & & S_1 & & S_1 \\
S_3 & \nearrow\!\!\rightarrow & G(4,4) & \nearrow\!\!\rightarrow & S_2 \\
x=(4,4) & & x=(4) & &
\end{array}
\quad , \qquad (3.9)
$$

$$
\begin{array}{ccccc}
 & & S_{i-3} & & S_{i-3} \\
S_i & \nearrow\!\!\rightarrow & G(2,4) & \nearrow\!\!\rightarrow & S_{i-2} \\
x=(2,4) & & x=(2) & &
\end{array}
\qquad (i = 4,5). \qquad (3.10)
$$

It is useful to characterize our system as a Markov chain with state space $\{S_1, S_2, S_3, S_4, S_5\}$. If $P = (p_{ij})$ is defined as the matrix of transition probabilities, i.e., the matrix of probabilities of switching in a single tree from state S_i to state S_j $(i, j = 1,2,\ldots, 5)$, then its value for q near .9 is easily seen to be equal to

$$(3.11) \qquad P = \begin{pmatrix} q^3, & q(1-q^2), & 0, & 1-q, & 0 \\ 0, & q^3, & q(1-q^2), & 0, & 1-q \\ q^4, & 1-q^4, & 0, & 0, & 0 \\ q^2, & 1-q^2, & 0, & 0, & 0 \\ 0, & q^2, & 1-q^2, & 0, & 0 \end{pmatrix} .$$

Using (3.11) and the equations

$$(3.12) \qquad \pi_j = \sum_{i=1}^{5} \pi_i p_{ij} \qquad (j = 1,\ldots, 5)$$

to solve for the π_j, we find that

$$(3.13) \qquad \pi_1 = \frac{q^4}{D}, \quad \pi_2 = \frac{1}{D}, \quad \pi_3 = \frac{1-q^2}{D}, \quad \pi_4 = \frac{q^4(1-q)}{D}, \quad \pi_5 = \frac{1-q}{D},$$

where

$$(3.14) \qquad D = 3 - q - q^2 + 2q^4 - q^5.$$

Next, let ET_i be the expected number of tests and EN_i be the expected number of units classified in the tree starting in state S_i $(i = 1,2,\ldots, 5)$. L, the expected number of tests per set of units classified, evaluated for the vertical procedure is then given by

$$(3.15) \qquad L = 3 \frac{\sum_{i=1}^{5} \pi_i ET_i}{\sum_{i=1}^{5} \pi_i EN_i} ,$$

i.e., L is equal to 3 times the expected numbers of tests per unit classified. Evaluating ET_i and EN_i $(i = 1,2,\ldots, 5)$ for q close to .9 yields

$$\begin{aligned} &ET_i = 3 - q^3 - q^7, \quad EN_i = 1 + q + q^3 \ (i = 1,2), \\ &ET_3 = 2 - q^8, \quad EN_3 = 1 + q^4, \\ &ET_i = 2 - q^6, \quad EN_i = 1 + q^2 \quad (i = 4, 5). \end{aligned} \tag{3.16}$$

Using the above together with (3.13), (3.14) and (3.15) results in (3.6), i.e.,

$$L = \frac{7-2q-2q^2-q^3+5q^4-2q^5-q^6-q^7-q^8}{1+q^4} \approx 2.1196 \tag{3.17}$$

at $q = .9$.

Notice that the expression in (3.17), is the same as in (3.6). This result is not surprising since it is easy in this case to see the equivalence of the horizontal and vertical approaches. In Section 8, it is proved that every horizontal procedure is equivalent to a vertical one.

4. Finding the Optimal Procedure: Trial and Error Method.

In searching for the optimal horizontal procedure we use the G-point of view which helps us to bring to bear information from previous work on group-testing. It was seen in Section 6 of [4] that in the case of an infinite number of units with the same q-value we get an 'optimal' tree i) by maximizing the (Shannon) information in each H-situation and ii) by using the R_1-procedure (cf. [3]) for each G-situation. We use these guidelines as a first step to eliminate lots of possibilities in finding the optimal horizontal procedure. For example, if $k = 4$ and let $q_i = q^i$ $(i = 1,2,3,4)$ where $q = .9$. By [3] we know (using the G-point of view) that we need 7 'units' to maximize the information; thus we can use (3, 4), (1, 2, 4), (1,2,2,2), (1,1,1,1,1,1,1) etc. Furthermore if the collection of 7 'units' is defective then we would like to select 3'units' for test in this G-situation, e.g., from a (1, 2, 4) defective set we can test the combination (1, 2) on the very next test. Generalizing

the above example, let $k = 2, 3, 4$ and 5 with $q_i = q^i (i = 1,2,\ldots, k)$ and $q = .9$. Using the above-mentioned guidelines the optimal horizontal procedures $S_h^{(k)}$, found by trial and error, are given by

$$S_h^{(2)} = \left\{ \begin{matrix} (1,2,2,2) \\ (1,1,1,1,1,1,1) \end{matrix} \right\}, \quad S_h^{(3)} = \left\{ \begin{matrix} (3,2,2) \\ (3,1,1,1,1) \\ (1,1,1,1,1,1,1) \end{matrix} \right\},$$

(4.1)

$$S_h^{(4)} = \left\{ \begin{matrix} (3,4) \\ (1,2,4) \\ (1,2,2,2) \\ (1,1,1,1,1,1,1) \end{matrix} \right\}, \quad S_h^{(5)} = \left\{ \begin{matrix} (3,4) \\ (1,2,4) \\ (1,2,2,2) \\ (1,1,1,1,1,1,1) \\ (5) \end{matrix} \right\}.$$

The last row of $S_h^{(5)}$ has an extremely simple tree, the (5)-tree, since it tests the units one-at-a-time. We find a theoretical basis for these results in the next section.

5. Theoretical Results with Special Reference to $k = 2$.

A group of units $(x_1, x_2,\ldots, x_i)$ will be called coarser than another group $(y_1, y_2,\ldots, y_j)$ if $i < j$ and there exist ordered, unequal integers $\ell_1, \ell_2,\ldots, \ell_{i-1}$ such that

$$(5.1) \qquad x_1 = \sum_{\alpha=1}^{\ell_1} y_\alpha; \quad x_2 = \sum_{\alpha=\ell_1+1}^{\ell_2} y_\alpha;\ldots, x_i = \sum_{\alpha=\ell_{i-1}+1}^{j} y_\alpha$$

where at least one sum contains 2 or more elements.

Unless stated otherwise we assume that all of the trees used in this paper utilize in addition to properties i) and ii) of Section 4 the assumption iii): In a G-situation if there is more than one combination that satisfies ii) we take the coarsest of these combinations. For example, suppose $q_1 = .9$, $q_2 = (.9)^2$ and we use (1,1,1,2,2) in the first test. Then, if it turned out to be defective, we could test (1,1,1) or (1,2) in the resulting G-situation. According to

assumption iii) we prefer (1, 2) since it is coarser than (1,1,1). This assumption has not been proved but has proved to be the right attack in a number of numerical examples. We have proved that if this plan is used in the G-situation then the same plan of using the coarsest combination of those that satisfy property ii) should also be used in the H-situation.

A tree T_1 is said to be coarser than another tree T_2 if from the G-point of view the total number of 'units' at each step of T_1 is the same, i.e., the probability of a positive result is the same and from the I-point of view the group tested under T_1 is coarser than the corresponding group tested in T_2; clearly, the finer tree T_2 can continue beyond T_1. If a family of trees exhibits the property that any pair of trees (in the family) can be compared as to coarseness, then we refer to it as a linear (coarse-fine) family of trees.

We emphasize the case $k = 2$, in which there are only 2 types of units. The two types are ℓ_i-units with probability q^{ℓ_i} of being good, where $\ell_2 = d\ell_1 > \ell_1$ and d is a positive integer. Let $\mathfrak{F}$ be a linear family of trees containing trees which range in coarseness from some coarsest tree in $\mathfrak{F}$ to the finest tree T_{ℓ_1}, where $T_{\ell_1} = (\ell_1, \ell_1, \ldots, \ell_1)$ is also included in $\mathfrak{F}$.

For a given tree T_0, let $E\{T|T_0\}$ denote the expected number of tests under T_0 and let $E\{N(\ell_i)|T_0\}$ denote the expected number of ℓ_i-units classified under T_0 $(i = 1,2)$. As above, we use T_{ℓ_2} to denote the tree $(\ell_2, \ell_2, \ldots, \ell_2)$ consisting only of ℓ_2-units. Let $T(d, \ell_1)$ denote the tree in $\mathfrak{F}$ which starts by testing d of the ℓ_2-units; recall that $\ell_2 = d\ell_1$.

We now state our theorem and illustrate it before going through the proof.

Theorem 1.

A necessary and sufficient condition that a horizontal procedure restricted to a linear family $\mathfrak{F}$ does best by choosing the coarsest possible group of units

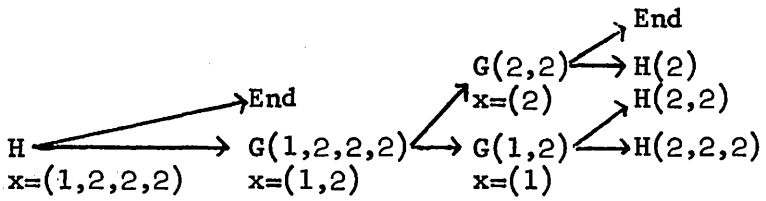
is that

$$(5.2)\qquad \frac{E\{T|T_{\ell_2}\}}{E\{N(\ell_2)|T_{\ell_2}\}} \geq E\{N(\ell_1)|T(d,\ell_1)\}\,\frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}} - (E\{T|T(d,\ell_1)\}-1).$$

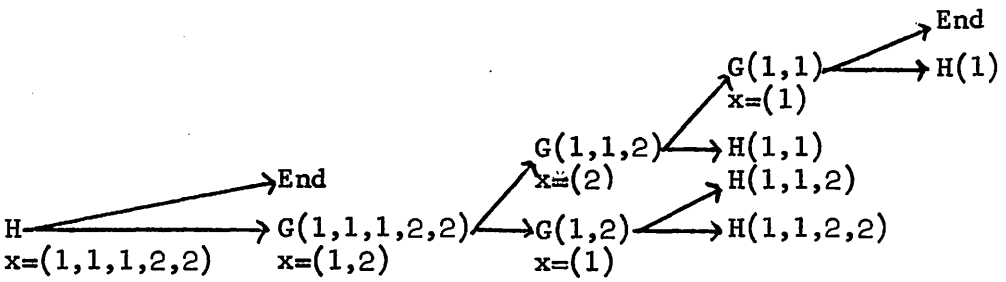
Illustration of the Result:

Let $\ell_1 = 1$ and $\ell_2 = 2\ell_1 = 2$ and suppose x_i has probability q^i of being good ($i = 1,2$) where $q = .9$. The linear family $\mathfrak{F}$ of trees consists of the four trees (1,2,2,2), (1,1,1,2,2), (1,1,1,1,1,2) and (1,1,1,1,1,1,1), which are listed in the order of coarsest to finest. Below we give each of the four trees in detail:

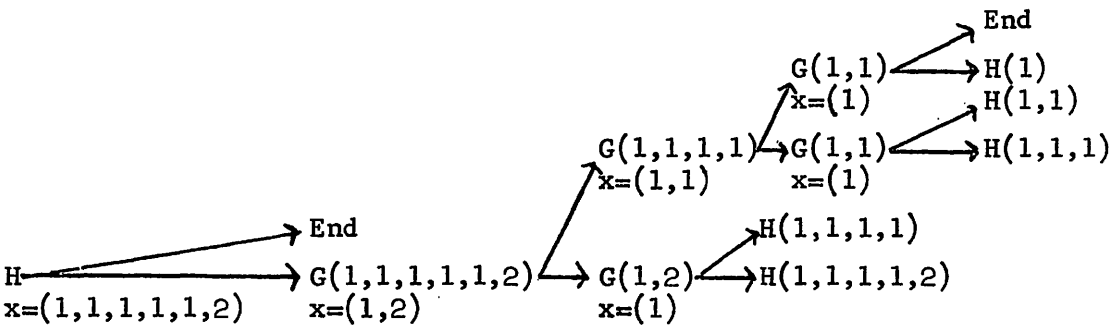
(5.3)

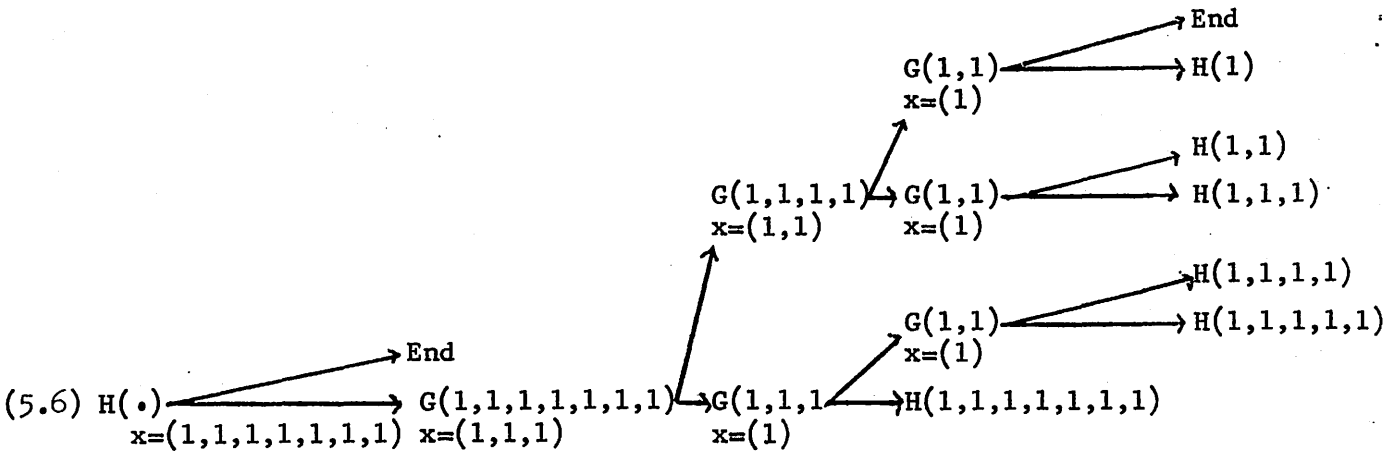


(5.4)

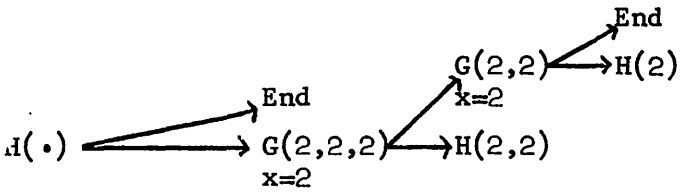


(5.5)

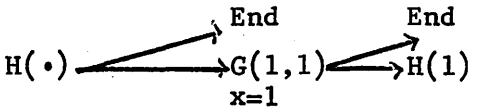




The tree T_{ℓ_1} (or T_1) is given in (5.6); the tree T_{ℓ_2} (or T_2) and the tree $T(d, \ell_1)$ (or $T(2, 1)$ are respectively given by



(5.7) $x = (2,2,2)$,



(5.8) $x = (1,1)$

By elementary calculations we obtain from (5.6), (5.7) and (5.8) for $q = .9$

$$(5.9) \qquad \frac{E\{T|T_1\}}{E\{N(1)|T_1)\}} = \frac{3+q-3q^7}{1+q+q^2+q^4+q^5+q^6} = \frac{2.4651}{5.2170} = .4725,$$

$$(5.10) \qquad \frac{E\{T|T_2\}}{E\{N(2)|T_2\}} = \frac{2+q^2-2q^6}{1+q^2+q^4} = \frac{1.7472}{2.4661} = .7085,$$

$$(5.11) \qquad E\{T|T(2,1)\} = 2-q^2 = 1.19; \; E\{N(1)|T(2,1)\} = 1+q = 1.9.$$

The condition (5.2) of Theorem 1 is satisfied when $q = .9$, $\ell_1 = 1$ and $\ell_2 = 2$ since

$$(5.12) \qquad .7085 \geq (1.9)(.4725) - (1.19-1) = .7078.$$

Hence according to Theorem 1 we conclude that the best horizontal procedure S_h^* is

$$(5.13) \qquad S_h^* = S_h^{(2)} = \left\{ \begin{matrix} (1,2,2,2) \\ (1,1,1,1,1,1,1) \end{matrix} \right\}.$$

Thus the result obtained in Section 4 by trial and error (cf. (4.1) can be obtained from Theorem 1.

Remark.

We note in this example that the same result holds regardless of the relative proportion of 1-units to 2-units at the outset.

Proof of Theorem 1.

Consider 2 successive trees in the linear family $\mathfrak{F}$, which we write as T_c for the coarser and T_F for the finer. If the coarser tree has a units of type ℓ_1 then we can write the trees (in terms of the starting groups) as

$$(5.14) \qquad T_c = (\underbrace{\ell_1, \ell_1, \ldots, \ell_1}_{a}, \underbrace{\ell_2, \ell_2, \ldots, \ell_2}_{b})$$

$$T_F = (\underbrace{\ell_1, \ell_1, \ldots, \ell_1}_{a}, \underbrace{\ell_1, \ell_1, \ldots, \ell_1}_{d}, \underbrace{\ell_2, \ell_2, \ldots, \ell_2}_{b-1})$$

where $a \geq 0$, $b \geq 1$ and $d \geq 2$. We now show that T_c gives better results than T_F if and only if (5.2) holds. Our proof is in three parts; in part A we assume that the ratio $p > 0$ of ℓ_1-type to ℓ_2-type units is sufficiently small so that the coarsest tree in $\mathfrak{F}$ will exhaust the ℓ_1-type unit first, if it is repeatedly applied to the units. It follows that all the trees of $\mathfrak{F}$, and in particular T_c and T_F, will have the same property. In comparing T_c and T_F we will be interested in the events (with reference to the T_c tree):

$$(5.15) \qquad B_1 = \{\text{all } \ell_2\text{-units are classified and the last one classified is good}\},$$

$$(5.16) \qquad B_2 = \{\text{all } \ell_2\text{-units are classified and the last one classified is defective}\}.$$

We write Q $(0 \leq Q \leq 1)$ for the probability that all ℓ_2-units are classified and this enables us to write for the events B_1 and B_2

(5.17) $P\{B_1\} = Qq^{\ell_2}$; $P\{B_2\} = Q(1-q^{\ell_2})$.

The T_C tree differs from the T_F tree only when all the ℓ_2-units are classified and only if the last ℓ_2- unit is defective (cf. (5.3)-(5.6)). This observation leads to the relation

(5.18) $E\{T|T_F\} = E\{T|T_C\} + Q(1-q^{\ell_2})\Sigma i\theta_i$

where θ_i is the conditional probability, given a $G(\ell_1, \ell_1, \ldots, \ell_1)$ situation as a starting point, that it will take i tests to return to the very next H-situation. From the $T(d, \ell_1)$ tree (cf. (5.8)) we note that

(5.19) $E\{T|T(d, \ell_1)\} = 1 + (1-q^{\ell_2})\Sigma i\theta_i$

and hence we can write (5.17) in the form

(5.20) $E\{T|T_F\} = E\{T|T_C\} + Q[E\{T|T(d, \ell_1)\} - 1]$.

Similarly, comparing the units classified, we use (5.16) and obtain

(5.21) $E\{N(\ell_1)|T_F\} = E\{N(\ell_1)|T_C\} + dQq^{\ell_2} + Q(1-q^{\ell_2})\Sigma j\theta'_j$

where θ'_j is the conditional probability, given a $G(\ell_1, \ell_1, \ldots, \ell_1)$ as a starting point, that j units of type ℓ_1 are classified by the time we reach the very next H-situation. From the $T(d, \ell_1)$ tree (cf. (5.8)) we note that

(5.22) $E\{N(\ell_1)|T(d, \ell_1)\} = dq^{\ell_2} + (1-q^{\ell_2})\Sigma j\theta'_j$

and hence we can write (5.20) in the form

(5.23) $E\{N(\ell_1)|T_F\} = E\{N(\ell_1)|T_C\} + QE\{N(\ell_1)|T(d, \ell_1)\}$.

The two results (5.20) and (5.23), will be used later.

To compare the trees T_C and T_F we compare the two horizontal procedures or strategies S_1 and S_2 defined by

(5.24) $S_1 = \begin{Bmatrix} T_C \\ T_{\ell_2} \end{Bmatrix}$, $S_2 = \begin{Bmatrix} T_F \\ T_{\ell_2} \end{Bmatrix}$

both based on the same numbers of units, M_i of type ℓ_i $(i = 1, 2)$. Under the assumption that $p = M_1/M_2$ is small, the units of type ℓ_1 will be depleted first and the expected number of tests under $S_i (i = 1, 2)$ needed to classify all these units is given by

$$E\{T|S_1\} = \frac{M_1}{E\{N(\ell_1)|T_C\}} E\{T|T_C\} + \left[\frac{M_2 - \frac{M_1 E\{N(\ell_2)|T_C\}}{E\{N(\ell_1)|T_C\}}}{E\{N(\ell_2)|T_{\ell_2}\}} \right] E\{T|T_{\ell_2}\}, \tag{5.25}$$

and the same result holds for S_2 if we replace T_C by T_F. To show that $E\{T|S_1\} \leq E\{T|S_2\}$ for p sufficiently small is now equivalent to showing that $\Delta_C \leq \Delta_F$ where

$$\Delta_C = \frac{1}{E\{N(\ell_1)|T_C\}} \left[E\{T|T_C\} - \frac{E\{N(\ell_2)|T_C\}}{E\{N(\ell_2)|T_{\ell_2}\}} E\{T|T_{\ell_2}\} \right], \tag{5.26}$$

and the same holds for Δ_F if T_C is replaced by T_F.

Using (5.20) and (5.23) and the additional fact that

$$E\{N(\ell_2)|T_F\} = E\{N(\ell_2)|T_C\} - Q, \tag{5.27}$$

we obtain for Δ_F the result (in terms of T_C only)

$$\Delta_F = \frac{1}{E\{N(\ell_1)|T_C\} + QE\{N(\ell_1)|T(d,\ell_1)\}} \left[E\{T|T_C\} + Q[E\{T|T(d,\ell_1)\}-1] - \frac{[E\{N(\ell_2)|T_C\} - Q]}{E\{N(\ell_2)|T_{\ell_2}\}} E\{T|T_{\ell_2}\} \right]. \tag{5.28}$$

For convenience we use F, G, H, I, J to denote as follows

$$F = \frac{E\{T|T_{\ell_2}\}}{E\{N(\ell_2)|T_{\ell_2}\}} \; ; \; G = E\{T|T_C\} - E\{N(\ell_2)|T_C\}F;$$
$$H = E\{N(\ell_1)|T_C\}; \; I = E\{T|T(d, \ell_1)\} - 1 + F; \; J = E\{N(\ell_1)|T(d, \ell_1). \tag{5.29}$$

Then by straightforward algebra the inequality $\Delta_C \leq \Delta_F$ becomes

(5.30) $\frac{G}{H} \leq \frac{G + QI}{H + QJ}$ or $GJ \leq HI$.

If we solve the latter inequality for F, then we obtain

$$(5.31) \qquad F \geq \frac{E\{N(\ell_1)|T(d, \ell_1)\}E\{T|T_C\} - E\{N(\ell_1)|T_C\}[E\{T|T(d, \ell_1)\} - 1]}{E\{N(\ell_1)|T_C\} + E\{N(\ell_1)|T(d, \ell_1)\}E\{N(\ell_2)|T_C\}} .$$

To simplify the above result we consider the numerator in (5.31) (call it $N(T_r)$) and the denominator in (5.31) (call it $D(T_r)$) separately with the tree T_C replaced by an arbitrary tree T_r in $\mathfrak{F}$. It is easily seen by using (5.20), (5.23) and (5.27) that

$$(5.32) \qquad N(T_C) = N(T_F) = N(T_r) \quad \text{and} \quad D(T_C) = D(T_F) = D(T_r)$$

for any pair T_C and T_F (contiguous or not) in $\mathfrak{F}$. Hence for any T_C in $\mathfrak{F}$

$$(5.33) \qquad N(T_C) = N(T_{\ell_1}) \quad \text{and} \quad D(T_C) = D(T_{\ell_1}).$$

Using these results in (5.31) and noting that $E\{N(\ell_2)|T_{\ell_1}\} = 0$ we obtain the final result

$$(5.34) \qquad F \geq \frac{N(T_{\ell_1})}{D(T_{\ell_1})} = \frac{E\{N(\ell_1)|T(d, \ell_1)\}E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}} - (E\{T|T(d, \ell_1)\} - 1).$$

This proves Theorem 1 for small p.

Part B of Theorem 1.

For given trees T_C and T_F as in (5.14) we define p^* as the supremum of values of p for which both trees exhaust the ℓ_1-type unit first. We define p^{**} as the infimum value of p for which both tests exhaust the ℓ_2-type first; then $p^* \leq p^{**}$ and we can write

$$(5.35) \qquad p^* = \frac{E\{N(\ell_1)|T_C\}}{E\{N(\ell_2)|T_C\}} , \quad p^{**} = \frac{E\{N(\ell_1)|T_F\}}{E\{N(\ell_2)|T_F\}} .$$

In Part B of our proof we assume that $p^* < p^{**}$ and consider this interval $p^* < p \leq p^{**}$, where the tree T_C exhausts the ℓ_2-type unit first and T_F

exhausts the ℓ_1-type unit first. We consider the two horizontal procedures or strategies S_1' and S_2' defined by

$$(5.36) \qquad S_1' = \begin{Bmatrix} T_C \\ T_{\ell_1} \end{Bmatrix}, \quad S_2' = \begin{Bmatrix} T_F \\ T_{\ell_2} \end{Bmatrix}.$$

For $p^* < p \leq p^{**}$ and M_i $(i = 1, 2)$ as defined above we have as in (5.25)

$$(5.37) \qquad E\{T|S_1'\} = \frac{M_2 E\{T|T_C\}}{E\{N(\ell_2)|T_C\}} + \left[M_1 - \frac{M_2 E\{N(\ell_1)|T_C\}}{EN(\ell_2)|T_C} \right] \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}},$$

$$(5.38) \qquad E\{T|S_2'\} = E\{T|S_2\},$$

where the latter is the analogue of (5.25). Since our result clearly depends only on the ratio p, we can without loss of generality set $M_1 = p$ and $M_2 = 1$. Then for $p^* < p \leq p^{**}$ we can write (5.37) and (5.38), respectively, as

$$(5.39) \qquad E(T|S_1') = \frac{p^*}{E\{N(\ell_1)|T_C\}} E\{T|T_C\} + (p-p^*) \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}}.$$

$$(5.40) \qquad E\{T|S_2'\} = \frac{p}{E\{N(\ell_1)|T_F\}} E\{T|T_F\} + \left[1 - p \frac{E\{N(\ell_2)|T_F\}}{E\{N(\ell_1)|T_F\}} \right] \frac{E\{T|T_{\ell_2}\}}{E\{N(\ell_2)|T_{\ell_2}\}}.$$

We now regard $E\{T|S_i'\}$ and $E\{T|S_i\}$ as functions of p and write $E_p\{T|S_i'\}$ and $E_p\{T|S_i\}$ $(i = 1, 2)$. For $p = p^*$ the second part of (5.37) (in brackets) vanishes and we can write for $p^* < p \leq p^{**}$

$$(5.41) \qquad E_p\{T|S_1'\} = E_{p^*}\{T|S_1\} + (p-p^*) \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}},$$

$$(5.42) \qquad E_p\{T|S_2'\} = E_{p^*}\{T|S_2\} + (p-p^*) \left[\frac{E\{T|T_F\}}{EN(\ell_1)|T_F} - \frac{E\{N(\ell_2)|T_F\}}{E\{N(\ell_1)|T_F\}} \frac{E\{T|T_{\ell_2}\}}{E\{N(\ell_2)|T_{\ell_2}\}} \right].$$

Since we have already shown that $E_{p^*}\{T|S_1\} \leq E_{p^*}\{T|S_2\}$ in Part A of the proof and because of the linearity in p it is sufficient for us to show that

$$(5.43)\qquad E_{p^{**}}\{T|S_1'\} = p^* \frac{E\{T|T_C\}}{E\{N(\ell_1)|T_C\}} + (p^{**} - p^*) \frac{E\{T|T_C\}}{E\{N(\ell_1)|T_{\ell_1}\}} \le p^{**} \frac{E\{T|T_F\}}{E\{N(\ell_1)|T_F\}}$$

$$= E_{p^{**}}\{T|S\},$$

in order to show the inequality for $p^* < p \le p^{**}$. Using (5.35) to replace p^* and p^{**} and (5.20), (5.23) and (5.27), to get it all in terms of the T_C tree, we can write (5.43) as

$$(5.44)\qquad E_{p^{**}}\{T|S_1'\} = \frac{G'}{H_2} + \left[\frac{H_1 + QJ}{H_2 - Q} - \frac{H_1}{H_2}\right] F' \le \frac{G' + QI'}{H_2 - Q} = E_{p^{**}}\{T|S_2'\}$$

where

$$(5.45)\qquad F' = \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}}, \quad G' = E\{T|T_C\}, \quad H_i = E\{N(\ell_i)|T_C\} \ (i = 1, 2)$$

$$I' = E\{T|T(d, \ell_1)\} - 1, \quad J = E\{N(\ell_1)|T(d, \ell_1)\}.$$

Straightforward algebra now shows that (5.44) holds for $p^* < p \le p^{**}$ if and only if

$$(5.46)\qquad \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}} = F' \le \frac{I'H_2 + G'}{JH_2 + H_1}.$$

As in the proof of (5.32) we can easily show that the numerator $N'(T)$ in (5.46), regarded as a function of the tree T, is the same for $T = T_C$ as for $T = T_F$ and the same holds for the denominator $D'(T)$ in (5.46). Thus we can take for T any tree in the family $\mathfrak{F}$ and we can replace T_C in (5.46) by T_{ℓ_1}, obtaining (since H_2 is replaced by 0)

$$(5.47)\qquad I'H_2 + G' = E\{T|T_{\ell_1}\}; \quad JH_2 + H_1 = E\{N(\ell_1)|T_{\ell_1}\}.$$

Thus we have equality in (5.46) and Part B is proved, i.e., Theorem 1 holds for $p^* < p \le p^{**}$. Moreover we attain equality in (5.2) when $p = p^{**}$.

In Part C of the proof we consider the range $p > p^{**}$ where T_C and T_F both exhaust the ℓ_2-type unit first. Defining the horizontal procedures or strategies S_1'' and S_2'' by

(5.48) $$S_1'' = \begin{Bmatrix} T_C \\ T_{\ell_1} \end{Bmatrix}, \quad S_2'' = \begin{Bmatrix} T_F \\ T_{\ell_1} \end{Bmatrix},$$

we find, as in (5.41), and (5.42) that

(5.49) $$E_p\{T|S_i''\} = E_{p^{**}}\{T|S_i'\} + (p-p^{**}) \frac{E\{T|T_{\ell_1}\}}{E\{N(\ell_1)|T_{\ell_1}\}} \quad (i = 1, 2).$$

Since we have already that $E_{p^{**}}\{T|S_1'\} = E_{p^{**}}\{T|S_2'\}$ it follows from (5.49) that for all $p > p^{**}$

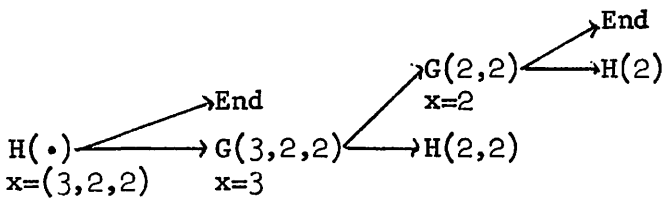
(5.50) $$E_p\{T|S_1''\} = E_p\{T|S_2''\}$$

and this completes the proof of Theorem 1.

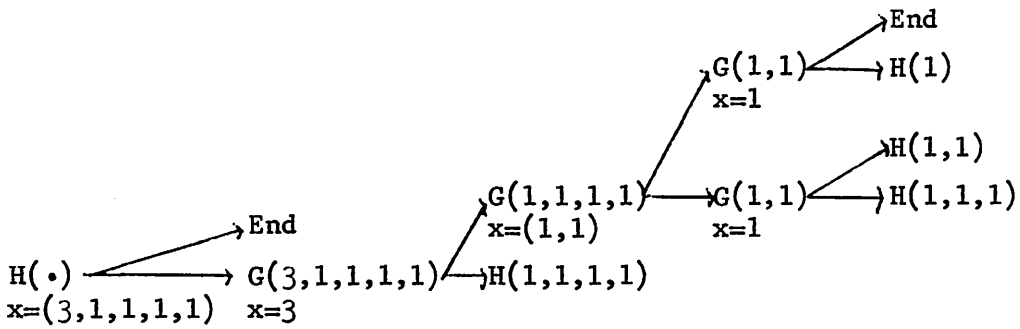
Remarks on Desirable Extensions of Theorem 1.

It would be desirable to find a condition as in Theorem 1 and show that it holds for the case of three types of units, say ℓ_1, ℓ_2, and ℓ_3. Consider the case $\ell_1 = 1$, $\ell_2 = 2$, and $\ell_3 = 3$ as an example; we assume that the three types each have proportion 1/3. The trees of interest no longer form a linear coarse-fine set; in addition to the coarse-fine set for the 2-type case with $\ell_1 = 1$ and $\ell_2 = 2$, we now wish to consider trees for (3, 2, 2), for (3, 2, 1, 1) and for (3, 1, 1, 1, 1). Tree structures for (3, 2, 2) and (3, 1, 1, 1, 1) for $q = .9$ would be

(5.51)



(5.52)



The (3,3,1)-tree (with a total of 7) also maximizes the information on the first step when $q = .9$ but it is eliminated from consideration because any strategy beginning with the (3,3,1)-tree always does worse than the horizontal procedure or strategy

$$(5.53) \qquad \left\{\begin{matrix}(3,3)\\(1,1,1,1,1,1,1)\end{matrix}\right\} = \left\{\begin{matrix}(1,1,1,1,1,1,1)\\(3,3)\end{matrix}\right\} .$$

In other words (3,3,1) is not a good combination for $q = .9$ because it is better to test the 1's and 3's separately, in either order. The optimal strategy for $q = .9$ is $S_h^{(3)}$ given in (4.1).

Similar extensions to the case of 4 types and to any number of types of units would, of course, also be desirable and the trial and error results of Section 4 will then become special applications of the general theorem.

6. Using Recursive Equations to Find the Optimal Horizontal Procedure.

In this section, we illustrate a recursive equation technique for finding the optimal horizontal procedure. It is done for the case when there are $k = 5$ streams of units but the technique is valid for any k.

We begin testing with c units from each stream where c is large but finite. Let $H(p_1, p_2, p_3, p_4, p_5)$ denote the expected number of tests needed to classify all the remaining units if we start from an H-situation in which cp_i units from the i^{th} stream ($i = 1,2,\ldots, 5$) are still unclassified. We assume that at least one of the p_i ($i = 1,2,\ldots, 5$) is positive. Then $H(p_1, p_2, p_3, p_4, p_5)$ for the optimal horizontal procedure satisfies the following equations:

$$(6.1) \qquad H(p_1, p_2, p_3, p_4, p_5) = \min_{T_0 \in \mathfrak{I}(p_1,p_2,\ldots,p_5)} \{b^*(T_0)E\{T|T_0\} + H(c_1(T_0),c_2(T_0),\ldots,c_5(T_0))\},$$

$$(6.2) \qquad b_i(T_0) = \frac{cp_i}{a_i(T_0)}, \quad b^*(T_0) = \min_{i=1,2,\ldots,5} \{b_i(T_0)\},$$

$$c_i(T_0) = p_i - \frac{a_i(T_0)b^*(T_0)}{c},$$

where $a_i(T_0)$ is the expected number of units from the i^{th} stream classified per T_0-tree. For given $(p_1, p_2, \ldots, p_5)$, we restrict $\mathfrak{T}(p_1, p_2, \ldots, p_5)$ to the class of all possible trees which maximize the information obtained from their very first step. This restriction can be justified by reasoning similar to that given at the beginning of Section 4. We note that since for some i

$$c_i(T_0) = 0, \tag{6.3}$$

the H expression on the right-hand-side of the equations in (6.1) contains at least one more zero than the one on the left-hand-side; we thus define

$$H(0,0,0,0,0) = 0. \tag{6.4}$$

as a boundary condition for the set of equations in (6.1). Let us also note that $c_i(T_0)$ and $b_i(T_0)$ are functions of p_i as well as of T_0; this functional dependence is not explicitly shown in these expressions because of their resulting cumbersomeness, but should be understood.

We illustrate the above technique in finding the optimal horizontal procedure for the case when q_i, the probability of the i^{th} unit being good, is equal to q^i $(i = 1,2,\ldots, 5)$, where q is close to .9. From the recursive equations, we find that

$$H(1,1,1,1,1) = b^*(T_1)E\{T|T_1\} + H(c_1(T_1), c_2(T_1), \ldots, c_5(T_1)), \tag{6.5}$$

where T_1 is the (3,4)-tree, where

$$b^*(T_1) = b_3(T_1) = c, \quad E\{T|T_1\} = 2-q^7$$

$$c_1(T_1) = c_2(T_1) = c_5(T_1) = 1, \quad c_3(T_1) = 0, \quad c_4(T_1) = 1-q^3,$$

and where $\mathfrak{T}(1,1,1,1,1)$ consists of the trees

$$\left\{\begin{array}{l}(1,2,4),\ (1,1,1,4),\ (3,4),\ (3,2,2),\ (3,3,1),\\ (3,2,1,1),\ (3,1,1,1,1),(1,1,5),\ (1,1,1,2,2),\ (1,1,1,1,1,2),\\ (1,2,2,2),\ (1,1,1,1,1,1,1),\ (2,5)\end{array}\right\}. \tag{6.7}$$

In finding (6.5), we also find that

$$H(c_1(T_1), c_2(T_1), \ldots, c_5(T_1)) = b^*(T_2)E\{T|T_2\} + H(c_1(T_2), \ldots, c_5(T_2)), \tag{6.8}$$

where T_2 is the (1,2,4) tree, where

$$b^*(T_2) = b_4(T_2) = \frac{c(1-q^3)}{q^3}\ ,\ E\{T|T_2\} = 3-q^3-q^7,$$

$$c_3(T_2) = c_4(T_2) = 0,\ c_1(T_2) = 1 - \frac{1-q^3}{q^3}\ ,\ c_2(T_2) = 1 - \frac{(1-q^3)q}{q^3}\ ,\ c_5(T_2) = 1 \tag{6.9}$$

and where $\mathfrak{F}(c_1(T_1),\ldots, c_5(T_1))$ contains the trees

$$\left\{\begin{matrix}(1,2,4),\ (1,1,1,4),\ (1,1,5),\ (1,1,1,2,2),\ (1,1,1,1,1,2),\\ (1,2,2,2),\ (1,1,1,1,1,1,1),\ (2,5)\end{matrix}\right\}; \tag{6.10}$$

$$H(c_1(T_2),\ldots, c_5(T_2)) = b^*(T_3)E\{T|T_3\} + H(c_1(T_3),\ldots, c_5(T_3)), \tag{6.11}$$

where T_3 is the (1,2,2,2)-tree, where

$$b^*(T_3) = b_2(T_3) = \frac{c\left(1 - \frac{(1-q^3)q}{q^3}\right)}{q+q^3+q^5}\ ,\ E\{T|T_3\} = 3-2q^7,$$

$$c_2(T_3) = c_3(T_3) = c_4(T_3) = 0,\ c_1(T_3) = c_1(T_2) - \frac{b^*(T_3)}{c}\ ,\ c_5(T_3) = 1, \tag{6.12}$$

and where $\mathfrak{F}(c_1(T_2),\ c_2(T_2),\ldots,\ c_5(T_2))$ contains the trees

$$\left\{\begin{matrix}(1,1,5),\ (1,1,1,2,2),\ (1,1,1,1,1,2),\ (1,2,2,2),\\ (1,1,1,1,1,1,1),\ (2.5)\end{matrix}\right\}; \tag{6.13}$$

$$H(c_1(T_3),\ldots, c_5(T_3)) = b^*(T_4)E\{T|T_4\} + H(c_1(T_4),\ldots, c_5(T_4)), \tag{6.14}$$

where T_4 is the (1,1,1,1,1,1,1)-tree, where

$$b^*(T_4) = b_1(T_4) = \frac{c[c_1(T_3)]}{1+q+q^2+q^3+\ldots+ q^6}\ ,\ E\{T|T_4\} = 3+q-3q^7 \tag{6.15}$$

$$c_1(T_4) = c_2(T_4) = c_3(T_4) = c_4(T_4) = 0,\ c_5(T_4) = 1,$$

and where $\mathfrak{F}(c_1(T_3),\ c_2(T_3),\ldots,\ c_5(T_3))$ contains the trees

$$\left\{(1,1,5),\ (1,1,1,1,1,1,1)\right\}; \tag{6.16}$$

and finally

(6.17) $H(0,0,0,0,1) = cE\{T|T_5\}$,

where T_5 is the (5)-tree and

(6.18) $E\{T|T_5\} = 1.$

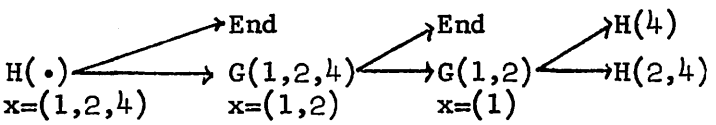
After finding the optimal strategy at each step we conclude that the optimal horizontal procedure for q in the neighborhood of .9 is given by

$$(6.19) \qquad S_h^{(5)} = \left\{ \begin{matrix} (3,4) \\ (1,2,4) \\ (1,2,2,2) \\ (1,1,1,1,1,1,1) \\ (5) \end{matrix} \right\},$$

as stated at the end of Section 4; the expected number of tests per set of units classified, [from (6.5), (6.8), (6.11), (6.14), (6.17)] for $q = .9$ is given to at least 4 decimals by

$$\begin{aligned} (6.20) \qquad \frac{H(1,1,1,1,1)}{c} &= 1.521703 + (.371742)(1.792703) + (.299813)(2.043406) \\ &\quad + (.062956)(2.465109) + 1 \\ &= 1.521703 + .666423 + .612640 + .155193 + 1 \\ &= 3.955959 = 3.9560 \text{ (to four decimal places).} \end{aligned}$$

The (3,4), (5), (1,2,2,2) and (1,1,1,1,1,1,1)-trees are given at the beginning of Section 2, at the end of Section 4 and at the beginning of Section 5, respectively. The (1,2,4)-tree is given by

(6.21)


We can similarly show that for $k = 2,3,4$ and $q_i = q^i$ for $i = 1,2,\ldots, k$, that the optimal horizontal procedures is given by $S_h^{(k)}$ $(k = 2,3,4)$ defined at the end of Section 4, and the expected number of tests per set of units classified for these procedures is

$$\frac{H(1,1,0,0,0)}{c} \approx 1.1803,$$

$$\frac{H(1,1,1,0,0)}{c} \approx 2.0188, \tag{6.22}$$

$$\frac{H(1,1,1,1,0)}{c} \approx 2.9560,$$

respectively. Note that the last result is exactly one less than the result in (6.20) where the 5-units are tested one-at-a-time.

The above calculations were done by hand. The recursive equation technique is particularly important because it enables us to use the computer in more complicated situations.

7. Comparison with Finite Models.

This work on finite models was motivated by an unpublished table of R. R. Coveyou [1]. His table can be used for the batch group-testing problem if we knew a priori that there is exactly one defective unit contained in one of several batches. Then the problem can be handled by known techniques in information theory, coding theory, and/or search theory.

For example, we might have 10 objects grouped into batches of size 1,2,3,4 and we want to know which of these four batches is the one that contains a particular unit which we call defective. All the 10 objects (or units) have the same chance of being the defective unit; thus we can call this a homogeneous model. The problem is to find in the smallest expected number of steps (i.e., tests) which of these four batches (or groups) contains the bad unit. Here we never break up the batches and at each step we select one or more batches and see if all the units contained therein are good or if the bad unit is among them. In this problem the optimal procedure is unique and starts by testing the single batch of size 4 (not two batches adding to five or four). The expected number of tests is 1.9 and the lower bound by information theory is 1.846 according to the table [1] of Coveyou. One of Coveryou's main interests was in systematically arranging a large class of such problems.

Of course, the above model has little in common with our problem since we do not know at the outset whether any units or how many are defective. In our case all the units have a common probability q of being good and $p = 1 - q$ of being defective and we assume mutual independence at the outset. Nevertheless some numerical comparisons between our problem and that of Coveyou (in addition to the fact that he provided a strong motivation) may also be of interest. We apply our general method to the same problem as above with four batches of sizes 1, 2, 3 and 4; special attention is given to the case $q = .9$. We distinguish an ordered and an unordered problem; in the former we apply the restriction that some ordering is given for the four batches (say 1, 2, 3, 4) and we cannot classify any batch strictly before one that precedes it; this is usually called the "firstcome-first served" property in queueing theory. Clearly the unordered problem must give the best results but a basic property of group-testing holds for the fixed-ordering problem and not for the general unordered case.

The derivation of these procedures (due to space problems) will be published separately. To illustrate the results let $H(1,2,3,4)$ denote the total expected number of tests required in the unordered problem for different values of q. The value of $H(1,2,3,4)$ is given below by seven polynomials strung together to form a continuous nonincreasing function, running constant at 4 for q close to zero and decreasing to 1 as $q \to 1$; the value at $q = .9$ is 3.0202.

R_1-Procedure Results for Unordered $H(1,2,3,4)$-Problem

(7.1) $H(1,2,3,4) =$

Polynomials	Range	Initial Strategy
4	$0 < q < .6823$	1 or 2 or 3 or 4
$5-q-q^3$	$.6823 \leq q < .8087$	(1,2) or 3 or 4
$6-2q-q^2-q^5+q^6$	$.8087 \leq q < .8518$	(1,2) or 4
$7-2q-q^2-q^3-q^5$	$.8518 \leq q < .8679$	(1,2,3) or 4
$8-2q-q^2-3q^3-q^6+q^{10}$	$.8679 \leq q < .9057$	(1,2,3)
$8-2q-q^2-2q^3-q^6-q^{10}$	$.9057 \leq q < .9566$	(1,2,3,4)
$9-2q-q^2-3q^3-2q^6$	$.9566 \leq q < 1$	(1,2,3,4)

In this unordered case the optimal strategy in the G-situation does not depend only on the defective set but also on the binomial set. Thus a basic result that holds for batches of size 1 (cf. Theorem 1 of [3]) does not hold here. In the corresponding fixed ordering problem the optimal strategy in the G-situation does depend only on the defective set and this simplifies the theory. Actually we give the numerical answers for several problems related to those used in other sections of this paper. For the unordered problem with $q = .9$ some optimal results in the class of NAR procedures (nested procedures that recombine binomial sets) are

$$\begin{aligned} &H(1,2,3,4,5) = 3.9588 \\ &H(1,2,3,4) = 3.0202 \\ &H(1,2,4) = 2.1296 \end{aligned} \tag{7.2}$$

For the fixed order problem we use the notation H_f and put the arguments in the given fixed order. Then we obtain

$$\begin{aligned} &H_f(1,2,3,4,5) = 3.9936, \\ &H_f(1,2,4,3,5) = 3.9644 \end{aligned} \tag{7.3}$$

and the latter is the best result among the $5! = 120$ possible fixed-orderings of $(1,2,3,4,5)$. For the fixed-ordering problem with batch sizes $(1,2,3,4)$ we obtain

$$\begin{aligned} &H_f(1,2,3,4) = 3.0311, \\ &H_f(3,1,2,4) = 3.0202; \end{aligned} \tag{7.4}$$

the latter result is the same as for the unordered problem and hence is the best result among the $4! = 24$ possible orderings of $(1,2,3,4)$. For the fixed-order problem with batches of size $(1,2,4)$ we obtain

$$H_f(1,2,4) = 2.1296 \tag{7.5}$$

and this is already the minimum of the $3! = 6$ possible orderings of $(1,2,4)$.

8. The Asymptotic Equivalence of Horizontal and Vertical Procedures.

Let c denote the number of sets of units; the sets all have a similar composition, e.g., $(1,2,3,4)$ is a set of size 4 with respective probabilities q, q^2, q^3, and q^4 of being good. We note that for c finite the horizontal type schemes are bona-fide procedures but when $c = \infty$ and the strategy contains more than one type of tree, they are not bona-fide procedures. Hence it is desirable to show for $c = \infty$ that for every horizontal procedure there is an equivalent vertical procedure (i.e., one that will eventually classify any given unit in the infinite collection of sets) which has the same "limiting efficiency" as the horizontal procedure it is associated with.

Before we prove the above (as Theorem 3) let us give a formal definition of the "limiting efficiency" of both a horizontal and a vertical procedure. For the horizontal procedure, we let $T_H^{[c]}$ be the number of tests used in classifying c sets of units and we write its limiting efficiency (denoted by Eff_h) as

$$(8.1) \qquad Eff_h = \lim_{c \to \infty} \frac{ET_h^{[c]}}{c} .$$

For the vertical procedure, we let $T_v^{[N]}$ (resp., $S_v^{[N]}$) be the number of tests used (resp., sets classified) in working through N trees and we write its limiting efficiency (denoted by Eff_v) as

$$(8.2) \qquad Eff_v = \lim_{N \to \infty} \frac{ET_v^{[N]}}{ES_v^{[N]}} .$$

Theorem 3.

For every horizontal 'procedure' S_h with $c = \infty$ there is an equivalent vertical procedure S_v whose group-testing efficiency can be obtained by analyzing the corresponding efficiency of S_h.

Proof:

The basic idea of the proof is that we can crystallize certain structural properties of the horizontal 'procedure' S_h that determine the asymptotic proportion of the time (i.e., of the total number of trees used) that we use the trees of each type. In the definition of the horizontal procedure the various trees are listed in preferential order to ensure that with probability one there is no infinite delay in the classification of any units, i.e., no unit gets classified after an infinite number of units with positive probability. We then argue that the associated vertical procedure must have the same asymptotic structure and hence it is a bona-fide procedure with the same efficiency as S_h.

In the course of the proof we carry along the so-called (1,2,3,4) example which starts with equal proportions of units of type i which have probability q^i (i = 1,2,3,4) of being good and, in particular, for q = .9. The general nature of our proof, however, can be carried over to any other such case. In the above case we start with the horizontal procedure

$$(8.3) \qquad S_h^{(4)} = \left\{ \begin{matrix} (3,4) \\ (1,2,4) \\ (1,2,2,2) \\ (1,1,1,1,1,1,1) \end{matrix} \right\} = \left\{ \begin{matrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{matrix} \right\}$$

which has 4 trees, three of which are given in Figure 1, (5.3) and (5.6) and the fourth is, as in (6.21),

$$(8.4) \qquad \begin{matrix} H(\cdot) \\ x=(1,2,4) \end{matrix} \rightarrow \begin{matrix} \text{End} \\ G(1,2,4) \\ x=(1,2) \end{matrix} \rightarrow \begin{matrix} \text{End} \\ G(1,2) \\ x=1 \end{matrix} \rightarrow \begin{matrix} H(4) \\ H(2,4) \end{matrix} .$$

The order of these 4 trees in (8.1) has been purposely arranged so that for any (large) finite number c of sets with probability approaching one as $c \to \infty$, i) the first tree if repeated over and over again will eliminate the

3-units (before the 4-units), ii) the second tree, if repeated, will eliminate the 4-units (before the 1-units and before the 2-units), iii) the third tree, if repeated, will eliminate the 2-units (before the 1-units, and hence iv) the fourth tree will be used infinitely often in the limiting case ($c \to \infty$).

For any horizontal procedure S_h we can calculate the proportion of trees of each type that are used (in the limit as $c \to \infty$) and thus calculate its limiting ($c \to \infty$) efficiency. The particular calculations for $S_h^{(4)}$ are deferred until later when we use them to illustrate this fact.

Let the limiting proportion for the T_i tree be denoted by p_i ($i = 1,2,3,4$). We first define a vertical procedure $S_v^{(4)}$, then show it has these same limiting proportions and from this fact show that it has the same limiting efficiency as $S_h^{(4)}$.

Recall that a set (1,2,3,4) consists one of each of the types $\ell_i = i$ ($i = 1,2,3,4$). A broken set is one for which at least one unit has been classified. Since the 3-unit always gets used up when we break a set, this is equivalent to saying that at least one unit has been involved in a group-test. Any unit in a broken set is a free unit; if the set is not broken then it is a bound unit.

We define a vertical procedure S_v associated with S_h by establishing a preference scheme among the different H-(i.e., binomial) situations that can arise with respect to free and bound units. We denote a unit of type i as an i-unit ($i = 1,2,3,4$) and a tree that starts by testing (3,4) as a (3,4)-tree. For $S_h^{(4)}$ this preference scheme is

(8.5a) Preference 1: If none of the following 3 preference situations arise then test (3,4)(i.e., use the (3,4)-tree) by breaking up a new set.

(8.5b) Preference 2: If the free units include a 1-unit, a 2-unit and a 4-unit, then use the (1,2,4)-tree.

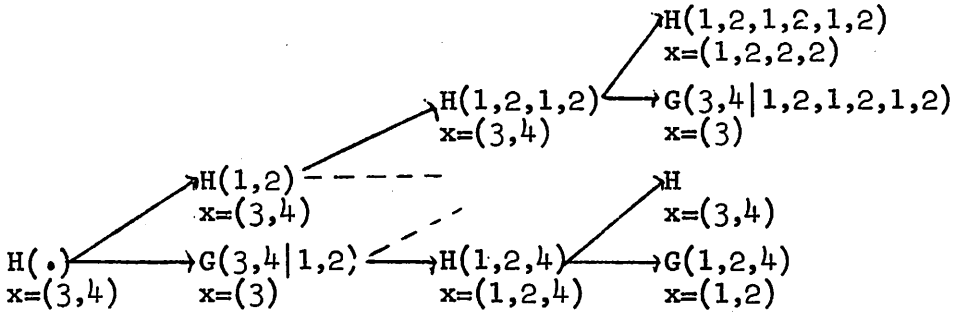
(8.5c) Preference 3: If the free units include a 1-unit and three 2-units, but no 4-units, then use the (1,2,2,2)-tree.

(8.5d) Preference 4: If the free units include seven 1-units, but (no 4-units and at most two 2-units) or (one 4-unit and no 2-units), then use the (1,1,1,1,1,1,1)-tree.

The event in Preference 4 dealing with one 4-unit and no 2-units can be neglected asymptotically since we are eliminating 4-units before 2-units with probability approaching one as $c \to \infty$; in fact we claim that this event cannot occur at all.

These four preference rules implicitly define a vertical procedure S_v; we can describe a small part of the procedure $S_v^{(4)}$ by

(8.6)



where H(•)indicates a binomial situation with no free units; H(1,2)[say] indicates that a 1- and 2-unit are free, etc., and where the dashed lines lead to states not filled in.

We now note that both the vertical procedure $S_v^{(4)}$ defined by (8.3) and the horizontal 'procedure' $S_h^{(4)}$ defined by (8.1) have certain properties in common that characterize the asymptotic structure. These are

1) Every 3-unit is tested with a 4-unit.

2) Every 4-unit that is not classified by 1) is tested simultaneously with both a 1-unit and a 2-unit.

3) Every 2-unit that is not classified by 2) is tested simultaneously with a 1-unit and two other 2-units.

4) Every 1-unit that is not classified by 2) or 3) is tested simultaneously with six other 1-units.

Implicit in these properties is the fact that with probability $\to 1$ as $c \to \infty$ the 3-units get depleted first, then the 4-units, then the 2-units and finally the 1-units. These four properties determine the relative proportion of the trees that are (3,4)-trees, (1,2,4)-trees, (1,2,2,2)-trees and (1,1,1,1,1,1,1)-trees which must of course add to one. Since $S_v^{(4)}$ and $S_h^{(4)}$ have the same asymptotic structure and since $S_h^{(4)}$ has been constructed so that with probability $\to 1$ as $c \to \infty$ there is no infinite delay in classifying any one unit, the same property must also hold for $S_v^{(4)}$. This proves that $S_v^{(4)}$ is indeed a vertical procedure and justifies our notation for it. Moreover the above four asymptotic ($c \to \infty$) relative proportions for $S_h^{(4)}$ must also hold for $S_v^{(4)}$. In particular, with probability $\to 1$ as $c \to \infty$ the (1,1,1,1,1,1,1)-tree will be used infinitely often in the vertical procedure $S_v^{(4)}$.

We claim that the above result holds for any associated pair of horizontal procedure S_h and vertical procedure S_v. The properties that we want for the vertical procedure S_v, such as no infinite delay, are built beforehand into the horizontal procedure S_h. The horizontal procedure S_h then becomes simply a mathematical convenience for calculating the asymptotic ($c \to \infty$) properties of the vertical procedure S_v.

Illustration.

After studying the efficiency of various horizontal schemes (which are easier to work with than vertical schemes) by trial and error methods as in Section 4 we decided for $q = .9$ to first use the (3, 4)-tree and finally arrived at the horizontal procedure $S_h^{(4)}$ in (8.3). Not only can we say that the 3-units are depleted but the proportion of 4-units remaining (assuming a large number c of sets of units) is

(8.18) Prop. (4's after Step I) = $1 - q^3$.

A simple analysis of the (1,2,4)-tree in (8.4), which makes up our Step II, shows that for $q = .9$ the 4-units are depleted before the 1-units or the 2-units. Since the probability of classifying the 4-unit is q^3 it follows that we use the (1,2,4)-tree $N(1,2,4) = c(1-q^3)/q^3$ times compared to the (3,4)-tree being used $N(3,4) = c$ times. Since the 2-unit is classified in Step II with probability q it follows that the proportion of 2-units after Step II is

$$(8.19) \qquad \text{Prop. (2's after Step II)} = \left(1 - \frac{1-q^3}{q^2}\right) = \frac{q^2 + q^3 - 1}{q^2} .$$

In Step III we analyze the (1,2,2,2)-tree given in (5.3) and find that for $q = .9$ the 2-units get depleted before the 1-units. Also the expected number of 2-units classified by the (1,2,2,2)-tree is $q + q^3 + q^5$ and hence the number of these trees used is

$$(8.20) \qquad N(1,2,2,2) = \frac{c(q^2+q^3-1)}{q^3(1+q^2+q^4)} = \frac{c(q^2+q^3-1)(1-q^2)}{q^3(1-q^6)} .$$

Note that we definitely use up one 1-unit per tree in Steps II and III. Hence the proportion of 1-units after Step II is $1-(1-q^3)/q^3 = (2q^3-1)/q^3$ and after Step III it is, using (8.20),

$$(8.21) \qquad \text{Prop.(1's after Step III)} = \frac{2q^3-1}{q^3} - \frac{(q^2+q^3-1)(1-q^2)}{q^3(1-q^6)} = \frac{q+q^2+q^3+q^4-2q^7-2}{q(1-q^6)} .$$

Since each tree in Step IV classifies $(1-q^7)/(1-q)$ units of Type 1, it follows that the number of trees used in Step IV is

$$(8.22) \qquad N(1,1,1,1,1,1,1) = \frac{cp(-2+q+q^2+q^3+q^4-2q^7)}{q(1-q^6)(1-q^7)} = \frac{c(-2+3q-q^5-2q^7+2q^8)}{q(1-q^6)(1-q^7)} .$$

The sum of these four N-values (divided by c) is found to be

$$(8.23) \qquad \frac{1}{c}\Sigma N = \frac{4-q-q^2-q^3-q^4}{1-q^7} = D(\text{say})$$

and hence the four desired relative proportions are

$$(8.24) \qquad \frac{1}{D}\ ,\ \frac{1-q^3}{q^3D}\ ,\ \frac{(q^2+q^3-1)(1-q^2)}{q^3(1-q^6)D}\ ,\ \frac{-2+3q-q^5-2q^7+2q^8}{q(1-q^6)(1-q^7)D}$$

for the trees used in Step I, II, III, and IV, respectively. For $q = .9$ the numerical values for these four proportions are .576531, .214321, .172852, and .036295, respectively.

An interesting result about the computation in (8.23) and (8.24) is that $D = \frac{1}{c}\Sigma N$ represents the asymptotic $(c \to \infty)$ expected number of trees required to classify one set. Since there are no infinite delays, $D/4$ is the asymptotic $(c \to \infty)$ expected number of trees per unit classified. Hence $4/D$ is the asymptotic $(c \to \infty)$ expected number of units classified per tree used; this holds for both procedures $S_h^{(4)}$ and $S_v^{(4)}$.

Hence the use of D as a normalizing constant puts our calculations on a per set basis and if we multiply through by 4 in (8.24) the proportions will be on a per unit basis. In other words, for the number of units U that are classified per tree used we can write the identity, using (8.24)

$$(8.25) \qquad \lim_{c\to\infty} E\{U\} = \frac{1}{D}\{(1+q^3)1 + (1+q+q^3)(\frac{1-q^3}{q^3}) + (1+q+q^3+q^5)\frac{(q^2+q^3-1)}{q^3(1+q^2+q^4)}$$

$$+ (\frac{1-q^7}{1-q})\frac{(-2+3q-q^5-2q^7+2q^8)}{q(1-q^6)(1-q^7)}\} = \frac{4}{D}\ ;$$

the identity being obtained by cancelling D in the last two parts of (8.25). The four quantities $(1+q^3)$, $(1+q+q^3)$, $(1+q+q^3+q^5)$ and $(1+q+q^2+q^3+q^4+q^5+q^6)$ used in (8.25) represent the expected number of units classified in each of the four trees that make up our strategy $S_h^{(4)}$ in (8.3) and each is easily obtained by a simple analysis of the appropriate tree.

The asymptotic $(c \to \infty)$ efficiency is the ratio of the asymptotic expected number of tests to the asymptotic expected number of units classified. For the (1,2,3,4)-problem using the result (8.25), we can write the asymptotic

efficiency (on a per unit classified basis) of any strategy S with no infinite delays as

$$(8.26) \qquad \mathrm{Eff}(S) = \frac{\lim_{c\to\infty} E\{T|S\}}{\lim_{c\to\infty} E\{U|S\}} = \frac{D}{4} \lim_{c\to\infty} E\{T|S\}$$

and we multiply this by 4 to put it on a per set basis. For the strategy $S_h^{(4)}$ the 'per-set' efficiency Eff' reduces to

$$(8.27) \qquad \mathrm{Eff}'(S_h^{(4)}) = (2-q^7) + (3-q^3-q^7)\,\frac{(1-q^3)}{q^3} + \frac{(3-2q^7)(q^2+q^3-1)}{q^3(1+q^2+q^4)}$$

$$+ \frac{(3+q-3q^7)(-2+3q-q^5-2q^7+2q^8)}{q(1-q^6)(1-q^7)} \quad .$$

This has some nice properties. Suppose we consider the corresponding $(1,2,3,4,5)$-problem using the same procedure $S_h^{(4)}$ and testing each 5-unit separately; call this $S_h^{(5)}$. For the vertical procedure $S_v^{(5)}$ the 5-units are tested when they are free and again there is no infinite delay. Then the efficiency on a set basis for $S_h^{(5)}$ or $S_v^{(5)}$ is obtained by simply adding 1 to the result in (8.27).

For $q = .9$ the numerical value for $S_h^{(4)}$ and $S_v^{(4)}$ in(8.14) is 2.9560 and, as mentioned above in Section 4,this was found to be optimal. For the $(1,2,3,4,5)$-problem with $q = .9$, the optimal procedure is the one described above, i.e., the procedure $S_v^{(5)}$ associated with $S_h^{(5)}$, and hence the optimal efficiency is 3.9560 for $q = .9$.

9. Lower Bounds.

For each of these problems one can obtain lower bounds exactly as was done in previous papers [3], [4], [5]; we omit their derivations. One lower bound is based on information lower bound (we denote it by ILB) and it is given by

$$\text{ILB} = -\Sigma Q_i \log_2 Q_i \tag{9.1}$$

where the sum is over all possible states of nature for each set, Q_i is the probability of the i^{th} possible state and $\Sigma Q_i = 1$. For example, in the H(1, 2, 3, 4) problem (finite or infinite model) we consider the 16 possible states of nature: all four batches are good with probability $Q_1 = q^{10}$, the 4-unit is good and the others are bad with probability $Q_2 = q^4(1-q)(1-q^2)(1-q^3)$, etc., so that i runs up to 16 in (9.1).

For $q = .9$ we obtain for the three problems H(1,2,3,4,5), H(1,2,3,4) and H(1,2,4), respectively,

$$\text{ILB}(1,2,3,4,5) = 3.9181;\ \text{ILB}(1,2,3,4) = 2.9419;\ \text{ILB}(1,2,4) = 2.0990. \tag{9.2}$$

Recall that we obtained for the infinite model

$$H_\infty(1,2,3,4,5) = 3.9560;\ H_\infty(1,2,3,4) = 2.9560;\ H_\infty(1,2,4) = 2.1196 \tag{9.3}$$

and for the finite model we obtained

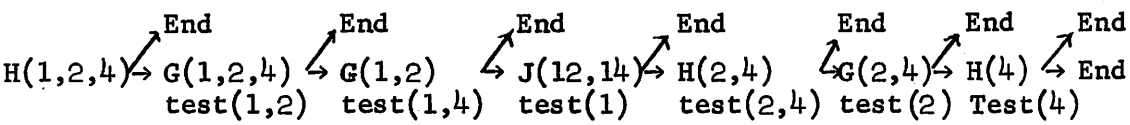
$$H(1,2,3,4,5) = 3.9588;\ H(1,2,3,4) = 3.0202;\ H(1,2,4) = 2.1296. \tag{9.4}$$

The other lower bound is always superior to, i.e., larger than, the ILB. It is the expected length of the most economical code due to Huffman for the given probabilities Q_i; we denote it by HLB. The procedure of deriving it is described in [4]; there is no explicit formula for this lower bound. For $q = .9$ we obtain in the above three problems

$$\text{HLB}(1,2,3,4,5) = 3.9469;\ \text{HLB}(1,2,3,4) = 3.0035;\ \text{HLB}(1,2,4) = 2.1174. \tag{9.5}$$

It is interesting to note that while the ILB is a lower bound for both models the HLB is a lower bound only for the finite model; this further justifies the consideration of both bounds. The HLB (ILB) is generally not attainable by any nested group-testing procedure in the finite (in either) model.

In the finite model version of the (1,2,4)-problem it is interesting to note that a non-nested procedure does exist that attains the HLB at the right end of (9.5). The following non-nested procedure R^*

(9.6)


consisting of a single chain of length 7 and terminating arrows elsewhere, does attain the HLB for $q = .9$. The fact that we tested (14) after '(12) was bad' (and had to write something other than G or H, namely, J(12, 14) in (9.6)) shows that the procedure R^* is not a nested one. In fact, the expected number of tests is easily shown to be (compare with the HLB in (9.5))

$$(9.7) \qquad E\{T|R^*\} = 7-3q-q^2-q^3-q^5-q^6+q^7 = 1+12p-8p^2-4p^3+15p^4-14p^5+6p^6-p^7$$

and the latter is 2.1174 for $p = .1$. The same procedure can be applied in an obvious manner to the infinite model, simply by repeating the operation on successive sets, and the HLB is the resulting measure of efficiency.

Some discussion of when the HLB is attained for the case of a common (known) q is given in [5] but the general case of unequal q's has not been thoroughly investigated. In general, if we start with four or more batches the HLB is not attainable; we conjecture that it is not attainable in the other two problems, (1,2,3,4) and (1,2,3,4,5), discussed above for non-trivial q. In particular, for $q = .9$ we conjecture that these two proposed nested procedures are

optimal among all group-testing procedures.

For the (1,2,3,4)- problem it should be clear from (9.3) and (9.5) that even if there existed a procedure R that attained the HLB = 3.0035 in the finite model, the analogue or repetition of R for the infinite model would still not compete with the proposed nested procedure $S_V^{(4)}$ (cf. (8.3) through (8.6)) which has efficiency $H_\infty(1,2,3,4)=2.9560 < \text{HLB}=3.0035$. This whole question of when there exists a non-nested procedure that is better than the proposed nested procedure will have to be treated in a separate paper.

Acknowledgement.

The authors wish to acknowledge a number of conversations with Professor S. Kumar of the Canadian Government, Department of Indian and Northern Affairs. Thanks are also due to Professor R. A. Elashoff since the basic ideas were developed while working on contract NIH-E-71-2180 at the University of California Medical School in San Francisco, California. The authors also wish to acknowledge with thanks the help of Dr. W. E. Lever of the Oak Ridge National Laboratory, Oak Ridge, Tenn. in deriving the exact form of the procedure R_1 for the (1,2,3,4)-problem as given in (7.1).

Appendix to Section 5.

In this appendix, we give a partial proof (indicating where it is not complete) that if we restrict our attention to trees which maximize the information on their very first step (in the H-situation), then Condition (5.2) for the application of Theorem 1 is always satisfied. (We also assume, as usual, that all trees under consideration satisfy (ii) of Section 4 above.) Since we are often interested in trees which maximize this information, Theorem 1 provides us with a valuable tool in our search for the optimal procedure.

Let $q_{a,a+1}$ denote the root of $1-x^a-x^{a+1} = 0$; it is easily seen that these roots increase with a and converge to 1 as $a \to \infty$. For any given q, this sequence determines an integer $m = m(q)$ such that $q_{m-1,m} < q < q_{m,m+1}$ (with equality on either side, say $q \leq q_{m,m+1}$). It is shown on p.1215 of [3] that for trees which only use one type of unit, each with probability q of being good, the choice of(the above) $m = m(q)$ such units to test on the initial H-situation maximizes the information. The integer m in turn determines two other integers when we write

$$(5.54) \qquad m = 2^{\alpha} + \beta \quad (0 \leq \beta < 2^{\alpha});$$

thus for $m = 4$, we have $\alpha = 2$ and $\beta = 0$. Let $H^*(m)$ denote the expected number of tests to get from one H-situation to the next, if we use m units in our first test.Now (5.54)is in accord with equation (20a) on p. 1190 of [ ] and from (23) of [3], we easily obtain

$$(5.55) \qquad H^*(m) = (\alpha+1)(1-q^m) + q^{m-2\beta},$$

where m is chosen to maximize the information from the first test in the initial H-situation. The result given in (5.55) was shown to hold in an interval ending at $q = 1$ but the fact that this interval always starts to the left of $q_{m-1,m}$ has been verified up to $m = 16$ (and beyond) but not shown to hold for all m; all the calculations to date support this conjecture. This conjecture has been proved by F. K. Hwang in [2].

Suppose we are working with two types of units, say $e_1 = 1$ and $e_2 = d$ (d is a positive integer), i.e., q and q^d are the probabilities of being good for the two types of units. Let $m = 2^\alpha + \beta$ be the maximizing-information value for all q in the interval $q_{m-1,m} < q \leq q_{m,m+1}$ and define T_1 to be the tree that initially tests m of the 1-units. It follows that $E\{T|T_1\} = H^*(m)$ as given in (5.55). It is also easily seen that

$$(5.56) \qquad E\{N(1)|T_1\} = mq^m + mq^{m-1}p + (m-1)q^{m-2}p + \ldots + 1q^0p = \frac{1-q^m}{1-q}.$$

Let us first consider the case when $d = 2$. Then $T(2, 1)$ starts with $x = (1,1)$ and

$$(5.57) \qquad E\{T|T(2,1)\} - 1 = (1-q^2); \; E\{N(1)|T(2,1)\} = 1 + q.$$

Hence for the right side of (5.2), we have

$$(5.58) \qquad (1+q)\,\frac{[(\alpha+1)(1-q^m) + q^{m-2\beta}]}{(1-q^m)/(1-q)} - (1-q^2) = \frac{1-q^2}{1-q^m}\,[\alpha(1-q^m) + q^{m-2\beta}].$$

Let T_2 be the tree which uses only the 2-units and maximizes the information on its first step. Suppose m is odd and hence (for $m > 1$) cannot be a power of two. At this point we are not sure whether $\frac{m-1}{2}$ or $\frac{m+1}{2}$ two's maximize the above information. Consider T_2 first with $\frac{m-1}{2}$ two's and then (later) with $\frac{m+1}{2}$ two's. We use the same formula as in (5.55) above with q replaced by q^2 and (m, α, β) replaced by $(\frac{m-1}{2}, \alpha-1, \frac{\beta-1}{2})$, respectively. This gives for procedure T_2 starting with $\frac{m-1}{2}$ two's

$$(5.59) \qquad E\{T|T_2\} = \alpha(1-q^{m-1}) + q^{m-1-2(\beta-1)}.$$

Similarly, using (5.56) with (q, m) replaced by $(q^2, \frac{m-1}{2})$, respectively,

$$(5.60) \qquad E\{N(2)|T(2)\} = \frac{1-q^{m-1}}{1-q^2}$$

and hence the left side of (5.2) is given by

$$(5.61)\qquad \frac{E\{T|T_2\}}{E\{N(2)|T_2\}} = \frac{1-q^2}{1-q^{m-1}}\,[\alpha(1-q^{m-1}) + q^{m+1-2\beta}].$$

To show the inequality in (5.2) is equivalent to showing that

$$(5.62)\qquad \frac{q^{m+1-2\beta}}{1-q^{m-1}} \geq \frac{q^{m-2\beta}}{1-q^{m}}$$

and this in turn is equivalent to showing that

$$(5.63)\qquad q(1-q^{m}) \geq (1-q^{m-1})$$

or that

$$(5.64)\qquad 1-q^{m-1}-q^{m} \leq 0.$$

The latter inequality holds for $q > q_{m-1,m}$.

We now consider T_2 with $(\frac{m+1}{2})$ two's and break things up into two cases according as (i) $0 < \beta < 2^{\alpha}-1$ or (ii) $\beta = 2^{\alpha}-1$. If $0 < \beta < 2^{\alpha}-1$, we have from (5.55) and (5.56) with (q, m, α, β) replaced by $(q^2, \frac{m+1}{2}, \alpha-1, \frac{\beta+1}{2})$, respectively,

$$(5.65)\qquad E\{T|T_2\} = \alpha(1-q^{m+1}) + q^{m+1-2(\beta+1)},$$

$$(5.66)\qquad E\{N(2)|T_2\} = \frac{1-q^{m+1}}{1-q^2}\,.$$

Then the left side of (5.2) becomes

$$(5.67)\qquad \frac{E\{T|T_2\}}{E\{N(2)|T_2\}} = \frac{1-q^2}{1-q^{m+1}}\,[\alpha(1-q^{m+1}) + q^{m-1-2\beta}].$$

Hence, comparing with (5.58), we have to show that

$$(5.68)\qquad \frac{q^{m-1-2\beta}}{1-q^{m+1}} \geq \frac{q^{m-2\beta}}{1-q^{m}}$$

and this is easily seen to be equivalent to

$$(5.69)\qquad 1-q^{m}-q^{m+1} \geq 0.$$

The latter holds for $q \leq q_{m,m+1}$. If $\beta = 2^{\alpha}-1$, then from (5.55) and (5.56), with (q, m, α, β) replaced by $(q^2, \frac{m+1}{2}, \alpha, 0)$, it follows that

$$E\{T|T_2\} = \alpha(1-q^{m+1}) + 1, \tag{5.70}$$

$$E\{N(2)|T_2\} = \frac{1-q^{m+1}}{1-q^2} \quad . \tag{5.71}$$

Again in this case (namely when $\beta = 2^{\alpha}- 1$), (5.2) is true if and only if

$$1-q^m- q^{m+1} \geq 0. \tag{5.72}$$

Hence the inequalities (5.64), (5.69) and (5.72) all hold for $q_{m-1,m} \leq q \leq q_{m,m+1}$, which is exactly the interval we assumed at the outset. The case m even must also be considered but it is easy to see that this leads to equality in (5.2). We have thus proved for the case $d = 2$ the following theorem:

Theorem 2.

Let the two types of units be the e_1 = 1-unit and e_2 = d-unit. If we restrict ourselves to trees which maximize the information on their initial step (and also satisfy (ii) of Section 4) then condition (5.2) always holds.

It is our conjecture, though, that Theorem 2 is true for all $d \geq 2$. In line with this conjecture, we extend the proof of Theorem 2 by virtue of the following lemma:

Lemma 1.

For any two divisors of m, say $a' > a$,

$$\frac{E\{T|T_a\}}{E\{N(a)|T_a\}} \leq \frac{E\{T|T_{a'}\}}{E\{N(a')|T_{a'}\}} \quad , \tag{5.73}$$

where T_a (resp., $T_{a'}$) is the tree which maximizes the information on its initial step and works only with a-units (resp., a'-units), i.e., with units having probability q^a (resp., $q^{a'}$). We prove Lemma 1 for $a = 1$ and 2 and conjecture that it is true for all a. The case $a = 2$ is the one needed

to show that Theorem 2 holds for all values of $d = a' > a = 2$, which are multiples of 2 and divisors of m. This is so because by proving Theorem 2 for $d = 2$, we have shown that

$$\frac{E\{T|T_2\}}{E\{N(2)|T_2\}} \geq \text{right side of (5.2)}, \tag{5.74}$$

and, together with (5.73), this implies that

$$\frac{E\{T|T_{a'}\}}{E\{N(a')|T_{a'}\}} \geq \text{right side of (5.2)}, \tag{5.75}$$

which in turn proves Theorem 2 for $d = a'$.

We now prove Lemma 1 for $a = 1$ and 2. Let $m = 2^\alpha + \beta$ $(0 \leq \beta < 2^\alpha)$, where we are writing α for $\alpha(m)$ and β for $\beta(m)$. Let m be chosen so that $q_{m-1,m} < q \leq q_{m,m+1}$. The number of units we test on the first step of the $T_{a'}$ tree is thus m/a' (recall that a' is a divisor of m). Replacing (q, m, α, β) by $(q^{a'}, m/a', \alpha(m/a'), \beta(m/a'))$ in (5.55) and (5.56) gives us

$$\frac{E\{T|T_{a'}\}}{E\{N(a')|T_{a'}\}} = (1-q^{a'})\{\alpha(\tfrac{m}{a'}) + 1 + \frac{q^{m-2a'\beta(m/a')}}{1-q^m}\}. \tag{5.76}$$

Let $\alpha(m/a) = \alpha^*$; for $a = 1$ or 2 or any power of 2, $\beta(m/a) = \beta/a$. It thus follows that

$$\frac{E\{T|T_a\}}{E\{N(a)|T_a\}} = (1-q^a)\{\alpha^* + 1 + \frac{q^{m-2\beta}}{1-q^m}\}. \tag{5.77}$$

We say x and y belong to the same power cycles if they have the same α-value, i.e., $\alpha(x) = \alpha(y)$.

Part 1: Assume m/a and m/a' are in the same power cycle.

Suppose first that $a = 1$ and the above assumption holds; then to prove Lemma 1, we have to show (by (5.76) and (5.77)) that

$$(1-q)\{\alpha^* + 1 + \frac{q^{m-2\beta}}{1-q^m}\} \leq (1-q^{a'})\{\alpha^* + 1 + \frac{q^{m-2a'\beta(m/a')}}{1-q^m}\}. \tag{5.78}$$

Since $a' > 1$, $1-q < 1-q^{a'}$ and since $\beta(m/a') \geq 0$ it suffices to show that

$$(5.79) \qquad 1 \leq q^{2\beta}(1+q+\ldots+q^{a'-1}) .$$

Since $a' \geq 2$ we have at least two terms on the right side of (5.79) and we can disregard the remaining terms. Since $\beta < m/2$ it follows that $2\beta \leq m-1$ and hence the result follows since

$$(5.80) \qquad 1 - q^{2\beta} - q^{2\beta+1} \leq 1 - q^{m-1} - q^{m} \leq 0.$$

Suppose $a = 2$ and the assumption of Part 1 holds. We have to show that

$$(5.81) \qquad (1-q^2)\{\alpha^* + \frac{q^{m-2\beta}}{1-q^m}\} \leq (1-q^{d'})\{\alpha^* + \frac{q^{m-2a'\beta(m/a')}}{1-q^m}\}$$

or that

$$(5.82) \qquad \alpha^* q^2(1-q^{a'-2})(1-q^m) + (1-q^{a'})q^{m-2a'\beta(m/a')} \geq (1-q^2)q^{m-2\beta} .$$

Since $\alpha(m/a) = \alpha - 1$ for $a = 2$, we have by the assumption of Part 1

$$(5.83) \qquad \frac{m}{a'} = \frac{2}{a'} 2^{\alpha-1} + \frac{\beta}{a'} \geq 2^{\alpha-1} = \alpha(m/a')$$

and hence, since $a' > 2$,

$$(5.84) \qquad \beta(m/a') = \frac{m}{a'} - 2^{\alpha-1} = \frac{\beta}{a'} - (1 - \frac{2}{a'})2^{\alpha-1} > \frac{\beta}{a'} .$$

Thus we can replace $m - 2a'\beta(m/a')$ in (5.82) by $m-2\beta$ and it suffices to prove that

$$(5.85) \qquad \alpha^* q^2(1-q^{a'-2})(1-q^m) + (1-q^{a'-2})q^{m-2\beta+2} \geq 0,$$

which obviously holds; this completes the proof under the Part 1 assumption.

Part 2: Assume m/a and m/a' are not in the same power cycle.

We will show that it suffices to assume adjoining power cycles, i.e., $\alpha(m/a) = \alpha^* = 1+\alpha(m/a')$.

Suppose $a = 1$ and that m/a and m/a' are in adjoining power cycles. Using (5.76), (5.77) and the fact that $\beta(m/a') \geq 0$, it is sufficient to show that

$$\alpha^* q^a(1-q^{a'-a})(1-q^m) + (1-q^{a'})q^m \geq q^{m-2\beta}(1-q^a) + (1-q^m)(1-q^a). \tag{5.86}$$

For $m = 2$ we take $a = 1$ and $a' = 2$ and the inequality reduces to $1 - q - q^2 \leq 0$, which holds for $m = 2$; hence we can assume $m \geq 3$. For $a = 1$, $\alpha^* \geq 1$ we first treat

Case 1: $a' \geq 2a = 2$ and $m > 2\beta + 1$.

Dividing (5.86) by $1-q^a$, and using the facts that $1-q^{a'-a} \geq 1-q^a$ and

$$1-q^{a'} \geq 1-q^{2a} = (1-q^a)(1+q^a) \tag{5.87}$$

it suffices to show that

$$q^a(1-q^m) + q^m(1+q^a) \geq q^{m-2\beta} + (1-q^m). \tag{5.88}$$

Since $1-q^m \leq q^{m-1}$, it suffices to show that

$$q^a - q^{m-2\beta} \geq q^{m-1}(1-q) \tag{5.89}$$

or equivalently that

$$1 + q + \ldots + q^{m-2\beta-2} \geq q^{m-2} . \tag{5.90}$$

Here we use the fact that $m > 2\beta + 1$ or $m - 2\beta - 2 \geq 0$.

For the case $m = 2\beta + 1$ we disregard the possibility that $a' = 2$ since m is now odd and cannot be a multiple of 2.

Case 2: $m = 2\beta + 1$, $a' \geq 3$.

From (5.86) by substitution and straightforward algebra we now have to show that

$$q^2(1-q^{a'-2}) \geq (1-2q^{2\beta+1})(1-q) \tag{5.91}$$

or equivalently that

$$(5.92) \qquad q^2 + q^3 + \ldots + q^{a'-1} \geq 1 - 2q^{2\beta+1}.$$

Since there is at least 1 term on the left side of (5.92) we have to show that

$$(5.93) \qquad 1 - q^2 - q^{2\beta+1} - q^{2\beta+1} \leq 0$$

which holds for $m = 2\beta+1 \geq 3$ since the first three terms alone are nonpositive.

If m/a and m/a' are $(k+1)$ cycles apart $(k = 1,2,\ldots)$ then $\alpha^* \geq (k+1)$, $a' \geq 3$ and $m \geq 3$. Using (5.76), (5.77) and the fact that $\beta(m/a') \geq 0$, it is sufficient to prove that

$$(5.94) \qquad (k+1)q^a(1-q^{a'-a})(1-q^m) + (1-q^{a'})q^m \geq q^{m-2\beta}(1-q^a) + (k+1)(1-q^m)(1-q^a).$$

Using the fact that (5.86) is true for $\alpha^* = 1$, proving (5.94) amounts to proving that

$$(5.95) \qquad 1 - 2q^a + q^{a'} \leq 1 - 2q + q^3 \leq 0$$

or that

$$(5.96) \qquad (1-q)(1-q-q^2) \leq 0;$$

this holds even for $m \geq 2$.

Since a' cannot be less than 2 for $a = 1$, we have completed the proof for $a = 1$.

Now we consider the case $a = 2$ with m/a and m/a' in adjoining power cycles, i.e., under Part 2. Again we wish to show (5.86). Since $a \leq m/2$, it follows that $m/a \geq 2$ and hence $\alpha^* \geq 1$; it suffices to prove (5.86) with $\alpha^* = 1$ as before.

Case 1: $a' \geq 2a = 4$ and $m \geq 4$.

From (5.86) we have to show that

$$(5.97) \qquad q^a(1-q^{a'-a}) \geq (q^{m-2\beta} + 1 - 2q^m)(1-q^d).$$

Since $1 - q^{a'-a} \geq 1-q^{a}$ for $a' \geq 2a$, it suffices to show that

(5.98) $\quad q^2 \geq 1 + q^{m-2\beta} - 2q^{m}$

or, replacing $1-q^{m}$ by q^{m-1}, that

(5.99) $\quad q^2(1-q^{m-2\beta-2}) \geq q^{m-1}(1-q)$

or equivalently that

(5.100) $\quad q^2 + q^3 + \ldots + q^{m-2\beta-1} \geq q^{m-1}$.

For $m \geq 2\beta + 3$ the result follows from (5.100); hence we need only consider the two remaining possibilities $m = 2\beta + 2$ and $m = 2\beta + 1$. Since $a = 2$ we rule out $m = 2\beta + 1$ since m must be even to be a multiple of a. For $m = 2\beta + 2$ we need only consider $\beta \geq 1$ since $m \geq 4$.

[Remark: We note that the smallest m of this type is $m = 6$ with $a' = 6$ and after that we have $m = 14$ with $a' = 7$ or 14; we note that $\beta(m/a') = 0$ in these cases.]

Since $m = 2^{\alpha} + \beta = 2\beta + 2$ it follows that β is an even integer and hence $m = 2\beta + 2$ contains a factor of 2 but not 4. Hence $a' > a$ must contain a factor other than 2 and since $a' \geq 4$ we can assume that $a' \geq 5$ or that $a' - a \geq 3$. It now follows from (5.97) that it is sufficient to show that

(5.101) $\quad q^2(1+q+q^2) \geq (q^2+1-2q^{m})(1+q)$

or, using the fact that $1-q^{m} \leq q^{m-1}$, that

(5.102) $\quad q^4 \geq q^{m-1}(1-q^2)$

and this clearly holds for $m \geq 5$. If $a' < 2a = 4$ then we need only consider $a' = 3$ and $m \geq 6$.

Case 2: $a' = 3$, $a = 2$ and $m \geq 6$.

For $\beta = 0$ the inequality (5.86) with $\alpha^* = 1$ reduces to $1 - q - q^m \leq 0$ which clearly holds for $m \geq 2$; hence we can assume $\beta \geq 1$. In this case ($\beta \geq 1$) we need the exponent $2a'\beta(m/a')$ in (5.82). Since $a' < 2a$ (strictly)

$$(5.103) \qquad 2a'\beta(m/a') = 2(m-a'2^{\alpha^*-1}) = 2m - \frac{3}{2}(m-\beta) = \frac{m+3\beta}{2}$$

and hence $m - 2a'\beta(m/a') \leq (m-3\beta)/2$. Using (5.100), (5.101) and the above result (5.103) and then setting $\alpha^* = 1$, it suffices to show that

$$(5.104) \qquad q^{\frac{m-3\beta}{2}}(1+q+q^2) \geq (1+q)q^{m-2\beta} + (1-q^m)(1+q-q^2).$$

Replacing $1-q^m$ by q^{m-1} and dividing by $q^{m-2\beta}$, it suffices to show that

$$(5.105) \qquad q^{\frac{\beta-m}{2}}(1+q+q^2) \geq 1 + q + q^{2\beta-1} + q^{2\beta} - q^{2\beta+1}.$$

Since $\beta \leq (m-1)/2$, and $m \geq 6$, the powers on the left side of (5.105) are at most $(-m-1)/4 + 2 \leq 1/4$ and since $q^{1/4} > q$, $q^{-7/4} \geq 1$ and $q^{-3/4} \geq 1$, it suffices to show that

$$(5.106) \qquad 1 - q^{2\beta-1} \geq q^{2\beta}(1-q)$$

or equivalently that

$$(5.107) \qquad 1 + q + \ldots + q^{2\beta-2} \geq q^{2\beta}$$

and this clearly holds for $\beta \geq 1$.

[Remark: Although we assumed $m \geq 6$ in the above, the assumption that m/a and m/a' are in adjoining cycles actually requires that m be at least 18, in which case $m/a = 9$ and $m/a' = 6$. If m/a and m/a' are more than one cycle apart, then the same argument as in (5.94), (5.95) and (5.96) shows that the result holds a fortiori.]

This completes the proof of Lemma 1 for $a = 1$ and 2.

References

[1] Coveyou, R. R. (1971). Tables for an incomplete search problem. Oak Ridge National Laboratory (personal communication).

[2] Hwang, F. K. (1973). On finding a single defective in binomial group-testing. (To appear.)

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

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

[5] Sobel, M. (1967). Optimal Group-Testing. Proceedings of the Colloquium on Information Theory Organized by the Bolyai Mathematical Society. Debrecen, Hungary, 411-488.