Group-Testing with Preassigned Confidence Levels for a Correct Classification of all Units

by

Milton Sobel

Technical Report 223

December 1973

University of Minnesota
Minneapolis, Minnesota

* Supported by U.S. Army Grant DA-ARO-D-31-124-72-G187 at the University of Minnesota, Minneapolis, Minnesota.

In this paper we use the simplest model for group-testing, i.e., at the outset we have a binomial model with N independent units, each of which is good with probability q and defective with probability $p = 1 - q$; both N and q are known. Group-testing is characterized by the fact that we can jointly test any integer number of units $1 \leq x \leq N$ with only two possible disjoint outcomes for each test: (i) all the x are good or (ii) at least one of the x is defective and, if $x > 1$, we don't know which one(s) or how many. Strategies for carrying out such a procedure so as to minimize the expected number of tests needed to classify without uncertainty everyone of the N units have been considered in a series of papers [3], [4], [5]. Some of these deal with the case of q unknown, with or without a given prior for q, some deal with $N = \infty$ or N large and unknown, some deal with a known number of defectives and some deal with optimality questions.

In this paper we consider the larger class of group-testing procedures defined by the fact that we can assert at the outset that the probability of correctly classifying all the N units $P(CC)$ is at least P^*(if we use that procedure); here P^* is preassigned and can be regarded as a joint confidence level. Any procedure R in this class is called P^*-admissible. Subject to this condition, $P\{CC|R\} \geq P^*$, we wish to find the procedure R that requires the smallest expected number of tests $E(T)$. It would be desirable to have a lower bound for $E(T)$ that holds for any group testing procedure since we can then guage how close we are to an optimal result;

such a bound is derived in Section 5 .

The basic procedure that we develop here, denoted by $R_{U,D}$, is to look for at most D defective units and then assert that all the units still unclassified, if there are any, are good; the value of D is computed at the outset as a function of N, q and P^*. We then give recursions for the unconditional expected number of tests. [In [6] a related problem is considered where D is given as an upper bound on the number of defectives present among the N units and there we use the conditional expectation of E(T) given that the number of defective units is at most D.]

It is proved in Section 3 that the strategy defined by procedure $R_{U,D}$ is exactly the same as the strategy defined by procedure R_1 studied in [3] and [5]. In fact, for $D \geq N$ at the outset (or if the current values d, n satisfy $d \geq n$ at any point) the procedure $R_{U,D}$ becomes identical with R_1 in strategy, and in the value of E(T), from that point on. Since the strategy of $R_{U,D}$ is the same as that of R_1 there is no necessity of drawing up extensive new tables since tables for procedure R_1 [3] are already in the literature.

Some comparisons with the results of Thomas et al. [7] are made in Section 4. They use a halving procedure (see also R_4 and R_5 in [3]) and employ $D = 1$ in our formulation but do not calculate or control the resulting overall probability of a correct classification; their objective was to emphasize the reduction in radiation exposure attainable by using the halving procedure and assuming that the leaker found, if any, is the only one present. Unlike the comparisons made in their paper, in this paper only procedures with exactly the same P(CC) will be compared by their E(T)-values; the same P(CC) can usually be obtained with the help of randomization.

This paper with its unconditional approach and a related paper [6], which takes a conditional approach assuming we have a known upper bound on the number of defectives present, were both partly motivated by a recent paper by Thomas, Pasternack, Vacirca and Thompson [7]. They develop a halving procedure R_T for locating a single defective, if it exists, in the unconditional problem (with no upper bound on the number of defectives present). Although they compared the expected number of tests for different procedures, comparing one procedure which classifies all the units with another that finds at most one defective, their real interest was in reducing the expected exposure to personnel checking N "sealed" radioactive sources for leakage; here the likelihood of more than one leaker in a group of standard size (say 50) is assumed to be small on empirical grounds. They claim to have recognized that the halving procedure R_T was subject to uncertainty in the correct classification of all units, but they did not examine the numerical value or the full implications of this uncertainty. In a subsequent paper [8] they further investigate the use of group-testing methods for the goal of reducing the total expected radiation exposure; it should be noted that this goal does not necessarily lead to the same results as our present goal of minimizing the expected number of tests, subject to satisfying a lower bound on the P(CC). Thus they have added some new vistas to the applications of group-testing methods.

2. Definition of Procedure $R_{U,D}$.

Let $D = D(P^*,N,q)$ denote the smallest integer such that

(2.1) $$P\{\mathcal{D} \le D|N, q\} \ge P^*$$

where $\mathcal{D}$ denotes the binomially-distributed random number of defectives among N independent units with common known probability q of being good and $P^* < 1$ is preassigned. Then our procedure $R_{U,D}$, defined below by recursive formulas, looks for at most D defectives, i.e., it either classifies all N units or finds D defectives, whichever comes sooner. Thus we get a correct classification if and only if $\mathcal{D} \le D$ and hence by (2.1)

(2.2) $$P\{CC\} = P\{\mathcal{D} \le D\} \cdot 1 + P\{\mathcal{D} > D\} \cdot 0 \ge P^*.$$

Hence this procedure is P^*-admissible and by randomizing between two successive values of D we can make the P{CC} exactly equal to P^*.

Note that, unlike the attitude in [6], where we treat D as an upper bound to the number of defectives present, we now regard D as the maximum number of defectives we will look for in order to satisfy (2.1). In accordance with this point of view, we compute our E(T)-expressions below with unconditional probabilities.

Let d denote the current value of D, just as n denotes the current value of N, the number of unclassified units. Let $H_{u,d}(n|q) = H_d(n)$ denote the expected number of additional tests needed for termination under procedure $R_{U,D}$ when we are in a binomial (or H) situation with n un-

classified units and at most d more defectives to look for $(0 \le d \le D)$. In a so-called G-situation we have two sets to work with: one is known to contain at least one defective unit and is called a defective set (of size m, say), the other is a binomial set as at the outset and is of size $n-m$. Let $G_{u,d}(m,n|q) = G_d(m,n)$ denote the expected number of additional tests needed as above except that we start with the G-situation. It will be noted that for $D \ge N$ ($D = N$ or its equivalent $D = \infty$ is obtained by setting $P^* = 1$), our procedure $R_{U,D}$ is identical with procedure R_1 studied in [3] and [5]; hence we can denote the latter by $R_{U,\infty}$ and its expectation formulas by $H_\infty(n)$ and $G_\infty(m,n)$ to avoid any confusion caused by the single subscript. In all cases we revert to the double subscript when there is danger of confusion.

The recursive formulas below define Procedure $R_{U,D}$ in terms of q, N, D through their current values q,n,d, respectively. For $n \ge 1$

$$(2.3) \qquad H_d(n) = 1 + \min_{1 \le x \le n} \{q^x H_d(n-x) + (1-q^x) G_d(x,n)\} .$$

For $2 \le m \le n$

$$(2.4) \qquad G_d(m,n) = 1 + \min_{1 \le x < m} \{(\frac{q^x - q^m}{1-q^m}) G_d(m-x,n-x) + (\frac{1-q^x}{1-q^m}) G_d(x,n)\} .$$

The boundary conditions are

$$(2.5) \qquad H_d(n) = 0 \qquad \text{if } d = 0 \text{ or } n = 0 ,$$

$$(2.6) \qquad G_d(1,n) = H_{d-1}(n-1) \qquad (n = 1, 2, \dots)$$

A third superfluous boundary condition tells us that for $d = 1$ and $n > m$

$$(2.7) \qquad G_1(m,n) = G_1(m,m) \; .$$

This is superfluous under Procedure $R_{U,D}$ since our procedure is nested and in the $G_1(m,n)$-situation it first looks for one defective unit in the defective set of size m. Hence, using (2.5) with $d = 1$ and (2.6) with $d = 0$, we never need to test the n-m binomial units on the left side of (2.7).

Note that we write d as a subscript on the right side of (2.3) and (2.4) even after obtaining x good units; technically, it should read $\min(d,n-x)$ in these cases. However, if it should happen that $n-x \leq d$ (for the minimizing x), it means that we have to classify <u>all</u> the remaining units individually and, from that point on, our procedure $R_{U,D}$ becomes identical with $R_{U,\infty}$ and no subscript is necessary for H or G. Hence the subscript d can remain as it appears on the right side of (2.3) and (2.4), i.e., it is correct, but it may be superfluous. The argument above that procedures $R_{U,D}$ and $R_{U,\infty}$ became identical is based on the fact that (2.3), (2.4) and (2.6) are the same for both procedures and (2.5) differs only in that there is no subscript d in the definition of $R_{U,\infty}$ in [3]. Hence, if at any point we obtain a situation with $d \geq n$, then the subscript becomes superfluous and the procedure becomes the same as that of $R_{U,\infty}$ from that point on. In particular, if $D \geq N$ then the total procedure, and hence also the resulting E(T)-functions, are the same fcr both procedures.

3. The Strategy of Procedure $R_{U,D}$.

To analyze the procedure $R_{U,D}$ and show that we don't need any new tables, we first simplify our equations (2.3) through (2.6) by introducing a function $F(m)$. Recall that a nested procedure is one that always gives priority to testing a part of the defective set whenever the latter is non-trivial, i.e., when $m \geq 2$; it does this without mixing units from the two types of sets.

Let $F_d(m,n)$ denote the expected number of tests required to reach the next H-situation under procedure $R_{U,D}$ if we start with a defective set of size $m \geq 2$ and a binomial set of size $n-m$; we wish to show that this function depends neither on n nor on d, but only on m, i.e., $F_d(m,n) = F(m)$. For convenience of analysis we assume that the units are ordered and our procedure will have a "first come-first served" property with respect to this ordering. By the nature of the nested procedure if we start with a defective set with $m \geq 2$, we always find exactly one defective unit, namely the first defective unit in that ordering, on the way to the very next H-situation. Given m, the conditional probability that the first defective is in the i^{th} position is $pq^{i-1}/(1-q^m)$. After we find this unit we are back in an H-situation with $n-i$ unclassified units and at most $d-1$ defectives to look for. Hence the relation between $F_d(m,n)$, $G_d(m,n)$ and $H_{d-1}(n)$ is

$$(3.1) \qquad G_d(m,n) = F_d(m,n) + \frac{p}{1-q^m} \sum_{i=1}^{m} q^{i-1} H_{d-1}(n-i)$$

For convenience we define

$$(3.2)\qquad G_d^*(m,n) = (\frac{1-q^m}{1-q})\ G_d(m,n) \quad \text{and} \quad F_d^*(m,n) = (\frac{1-q^m}{1-q})\ F_d(m,n)\ .$$

From (2.6), (3.1) and (3.2) we have the boundary condition

$$(3.3)\qquad F_d^*(1,n) = F_d(1,n) = 0 \qquad\qquad (n = 1,\ 2,\ldots;\ \ d = 0,\ 1,\ldots)\ .$$

<u>Lemma 1</u>: Under Procedure $R_{U,D}$ the strategy in the G-situation does not depend on d or n.

<u>Proof</u>: From (3.1) and (3.2) we have

$$(3.4)\qquad G_d^*(m,n) = F_d^*(m,n) + \sum_{i=1}^{m} q^{i-1}\ H_{d-1}(n-i)\ .$$

If we use (3.1) and (3.2) for all the G-expressions in both sides of (2.4) then we find that the three summations cancel each other and we obtain the simpler recursion, with the same minimizing x-value as (2.4),

$$(3.5)\qquad F_d^*(m,n) = \frac{1-q^m}{1-q} + \min_{1\ x<m}\ \{q^x\ F_d^*(m-x,\ n-x) + F_d^*(x,n)\}\ ,$$

and the boundary condition for this is (3.3). Since (3.5) and (3.3) hold for all d and do not depend on d it follows that the minimizing x-value does not depend on d. Using an induction proof, we assume that $F^*(a,b)$ does not depend on b for $b < n$ and also for $b = n$, $a < m$. Then it follows from (3.5) that $F^*(m,n)$ does not depend on n. Since (3.3) does not depend on n, the induction proof is complete and we can replace $F_d^*(m,n)$ and $F_d(m,n)$ by $F^*(m)$ and $F(m)$, respectively.

Theorem 1: Under Procedure $R_{U,D}$ the strategy in the H-situation does not depend on d.

Proof: Using (3.1) and (3.2) we can write (2.3) in the form

$$(3.6) \qquad H_d(n) = 1 + \min_{1 \le x \le n} \{q^x H_d(n-x) + p F_d^*(x) + p \sum_{i=1}^{m} q^{i-1} H_{d-1}(n-i)\} .$$

Assuming that the minimizing x in (3.6) is less than n, we now iterate the recursion by using the same result for $H_d(n-x)$ on the right side of (3.6). This gives

$$(3.7) \quad H_d^{(2+)}(n) = 1 + \min \{q^x + pF^*(x) + pq^x F^*(y) + q^{x+y} H_d(n-x-y) + p\sum_{i=1}^{x+y} q^{i-1} H_{d-1}(n-i)\}$$

where the minimum is over integer partitions of n $(x+y \le n,\ x \ge 1,\ y \ge 1)$ into three disjoint parts with at most one zero; this zero has to be terminal. Then, allowing x to be n in (3.6),

$$(3.8) \qquad H_d(n) = \min \{H_d^{(1)}(n),\ H_d^{(2+)}(n)\} ,$$

where $H_d^{(1)}(n)$ is the right side of (3.6) with x = n (a 1-part partition). If we continue this iteration we come to a point (say, at the r^{th} step) where all the rest of the units are tested. Then writing x_1 for x, x_2 for y, etc., we have $n = x_1 + x_2 + \dots + x_r$ and, since $H_d(0) = 0$ for all d, we can write

$$(3.9) \quad H_d(n) = 1 + \mathrm{Min}\{q^{x_1}+q^{x_1+x_2}+\dots+q^{n-x_r}+ pF^*(x_1)+pq^{x_1} F^*(x_2)+\dots+pq^{n-x_r} F^*(x_r)\}$$
$$+ p \sum_{i=1}^{n} q^{i-1} H_{d-1}(n-i) ,$$

where the minimum is over $r(r = 1, 2,.., n)$ and over partitions of n into r positive parts. Since the part to be minimized in (3.9) no longer depends on d, the result is proved.

Corollary 1: The strategy for Procedure $R_{U,D}$ is exactly the same as for Procedure $R_{U,\infty}$ for any q-value.

Proof: Since the strategy does not depend on d or D we can take $D = N$ and the procedure then agrees exactly with that of $R_{U,\infty}$ studied in [3] and [5].

It follows that we do not need any new tables to describe the strategy of Procedure $R_{U,D}$. Hence the tables of this paper are devoted only to numerical results and lower bounds for Procedure $R_{U,D}$.

4. Some Comparisons.

In [7] a similar problem is considered. The unconditional expected number of tests is derived as in this paper. However they take $D = 1$ and do not control the $P(CC)$. By taking $D = 1$ we will have the same $P(CC)$ and hence the comparison will be a fair one. This point may seem trivial but it has to be emphasized since procedures with different $P(CC)$ values are compared in [7]. Although for $D=1$ and $N=50$ the attained $P(CC)$ $(=.9106)$ is moderate for $q = .99$, it should be carefully noted that it is very low $(.2794)$ for $q = .95$ and extremely low $(.0338)$ for $q = .90$. Hence for the latter two q-values it is highly desirable to use a larger D-value. However, for the purpose of comparison, we will stick with $D = 1$.

We denote the halving procedure with $D=1$ by R_T; in other contexts it is called binary search. To be specific we test all the N units at the outset and then we take x equal to the largest integer in $N/2$, i.e., $x = [N/2]$. We terminate when 1 defective unit is found. It has the property that one need not know the value of q to carry out the procedure and the value of $E(T)$, or $H_T(n)$, can be written as a single expression that holds for all values of q; these expressions were not given in [7] and we include them here for $N = 15, 30, 50, 60$ and 100. For $D = 1$ and any q

$$H_T(15) = 4(1-q^{15}) + q \tag{4.1}$$

$$H_T(30) = 5(1-q^{30}) + q - pq^{15} \tag{4.2}$$

$$H_T(50) = 6(1-q^{50}) + q - \frac{(1-q^{18}+q^{22}-q^{43})}{1 + q + q^2} \tag{4.3}$$

$$H_T(60) = 6(1-q^{60}) + q - pq^{15}(1 + q^{15} + q^{30}) \tag{4.4}$$

$$H_T(100) = 7(1-q^{100})+q - \frac{q^3}{1+q+q^2} \{1-q^{18}+q^{22}(1-q^{21})(1+q^{25}+q^{50})\} \, . \tag{4.5}$$

The derivation of the above is straightforward and is omitted.

In contrast, the expressions for $R_{U,D}$ with $D = 1$ vary with q and we include some of these for intervals containing $q = .90$, $.95$ and $.99$. Given the strategy in [3] it is straightforward to derive the polynomial for $H_{U,1}(N)$ at the desired q-value; we use Table VA, B, C of [3] which goes up to $N = 100$ at $q = .90$, $.95$, and $.99$ and obtain

$$H_{U,1}(15) = \left\{ \begin{array}{ll} 3(1-q^{15}) + q + q^7 & q \approx .90 \quad (x = 7) \\ 4(1-q^{15}) + q & q \approx .95 \quad (x = 15) \\ 4(1-q^{15}) + q & q \approx .99 \quad (x = 15) \end{array} \right. \tag{4.6}$$

$$H_{U,1}(50) = \left\{ \begin{array}{ll} 3(1-q^{50})+q+q^8+q^{15}+q^{22}+q^{29}+q^{36}+q^{42} & q \approx .90 \ (x = 7) \\ 4(1-q^{50})+q^3+q^{16}+q^{29}+q^{44} & q \approx .95 \ (x = 13) \\ 6(1-q^{50}) + q^{14} & q \approx .99 \ (x = 50) \end{array} \right. \tag{4.7}$$

$$(4.8)\qquad H_{U,1}(100) = \begin{cases} 3(1-q^{100})+q^{79}+q^{85}+q^{91}+q^{98}+\dfrac{q(1-q^{77})}{1-q^{7}} & q \approx .90 \ (x = 7) \\ 4+q^{85}+q^{98}-5q^{100}+\dfrac{q^{2}(1-q^{84})}{1-q^{14}} & q \approx .95 \ (x = 14) \\ 7 + q^{29} + q^{98} - 8q^{100} & q \approx .99 \ (x = 100) \end{cases}$$

In each case the initial value of x is shown and the expression holds in an interval containing the indicated value of q.

Table 1 shows the numerical comparisons that are obtained from these formulae. It also includes a column for the ratio of the two numbers that can be interpreted as the efficiency of the result for procedure R_T relative to that obtained for procedure $R_{U,1}$.

Table 1: Comparison of Procedures $R_{U,1}$ and R_T for $D = 1$

N	q = .90			q = .95			q = .99		
	$H_{U,1}$	H_T	Ratio (Eff.)	$H_{U,1}$	H_T	Ratio (Eff.)	$H_{U,1}$	H_T	Ratio (Eff.)
15	3.7606	4.0764	92.3%	3.0968	3.0968	100.0%	1.5498	1.5498	100.0%
50	4.7010	6.6169	71.0%	5.3203	6.2430	85.2%	3.2387	3.2561	99.5%
100	4.7250	7.6458	61.8%	5.7276	7.6405	75.0%	5.1924	5.2550	98.8%

It is interesting to note that the monotonicity of $H_{U,1}$ as a function of q (for fixed $D = 1$) no longer holds.

For any fixed q and D the asymptotic ($N \to \infty$) value of $H_T(N)$ is approximately $1 + \log_2 N$ and tends to ∞. On the other hand the value of $H_{U,1}(N)$ is approximately $\frac{1}{1-q^x} + F(x)$ for large N where x depends only on q; hence $H_{U,1}(N) \to$ a finite result as $N \to \infty$. It follows that the asymptotic ($N \to \infty$) efficiency of R_P relative to $R_{U,1}$ is zero for any fixed values of q and D. The same argument does not hold for fixed q and P^* but since the value of P^*, or the $P(CC)$, is not computed in [7], we need not make this comparision.

Table 2: Exact Values of the Expected Number of Tests and Lower Bounds for Procedure $R_{U,D}$.

N		q = .90			q = .95			q = .99		
		D = 1	D = 2	D = 3	D = 1	D = 2	D = 3	D = 1	D = 2	D = 3
2	EXACT	1.1900	1.2900	1.2900	1.0975	1.1475	1.1475	1.0199	1.0299	1.0299
	HLB	1.1900	1.2900	---	1.0975	1.1475	---	1.0199	1.0299	---
	ILB	0.8911	0.9380	---	0.5585	0.5836	---	0.1608	0.1615	---
3	EXACT	1.4420	1.6510	1.6610	1.2353	1.3376	1.3401	1.0494	1.0695	1.0696
	HLB	1.4420	1.5880	1.5980	1.2353	1.2973	1.2998	1.0494	1.0599	1.0600
	ILB	1.2710	1.4023	1.4070	0.8169	0.8585	0.8587	0.2400	0.2424	0.2424
4	EXACT	1.6878	2.0201	2.0500	1.3710	1.5300	1.5375	1.0788	1.1092	1.1095
	HLB	1.6878	1.9509	1.9692	1.3710	1.4651	1.4686	1.0788	1.0907	1.0908
	ILB	1.6129	1.8582	1.8755	1.0625	1.1428	1.1435	0.3184	0.3231	0.3232
5	EXACT	1.9575	2.4254	2.4855	1.5360	1.7556	1.7707	1.1173	1.1582	1.1588
	HLB	1.9575	2.3550	2.3985	1.5360	1.6736	1.6808	1.1173	1.1311	1.1313
	ILB	1.9206	2.3026	2.3428	1.2958	1.4628	1.4267	0.3960	0.4039	0.4040
10	EXACT	3.1126	4.4152	4.7865	2.3401	2.9334	3.0283	1.3240	1.4214	1.4251
	HLB	3.1126	4.3690	4.7001	2.3401	2.8461	2.9134	1.3240	1.3606	1.3615
	ILB	3.0547	4.2924	4.6215	2.2984	2.7918	2.8065	0.7725	0.8070	0.8079
15	EXACT	3.7606	5.9278	6.8406	3.0968	4.1588	4.4096	1.5498	1.7119	1.7214
	HLB	3.7606	5.8897	6.7558	3.0968	4.0948	4.3015	1.5498	1.6250	1.6283
	ILB	3.7243	5.8393	6.7025	3.0742	4.0534	4.2720	1.1306	1.2084	1.2118
25	EXACT	4.3874	7.8466	10.0646	4.1659	6.2812	7.0649	2.0430	2.3593	2.3899
	HLB	4.3874	7.7923	9.9585	4.1659	6.1997	6.9410	2.0430	2.2478	2.2631
	ILB	4.3533	7.7713	9.9423	4.1391	6.1875	6.9156	1.7951	2.0032	2.0189
50	EXACT	4.7010	9.2679	13.4709	5.3203	9.4905	12.1739	3.2387	4.0730	4.2215
	HLB	4.7010	9.2254	13.3896	5.3203	9.4371	12.0666	3.2387	3.9502	4.0628
	ILB	4.6658	9.1973	13.3633	5.2872	9.4146	12.0464	3.1913	3.9139	4.0255
100	EXACT	4.7250	9.4486	14.1646	5.7276	11.2773	16.3644	5.1924	7.4067	8.1066
	HLB	4.7250	9.4066	14.0880	5.7276	11.2385	16.2873	5.1924	7.3180	7.9666
	ILB	4.6898	9.3783	14.0591	5.6900	11.2096	16.2601	5.1221	7.2570	7.8983

5. Lower Bounds and Optimality Discussion.

As in other problems of group-testing (see e.g., Section 12 of [5]) we develop two lower bounds: one is the information lower bound (ILB), based on the Shannon-Weiner information concept and the other is based on the concept of the (Huffman) lower bound on the expected length (or Huffman cost) of a binary code with preassigned probabilities for each code word. As in the other problems the Huffman lower bound (HLB) is a sharper bound but usually does not lend itself to any simple expresssion. Moreover it is either not available for $N = \infty$ or else it approaches the ILB in some sense as $N \to \infty$. On the other hand we obtain an explicit expression for the ILB. One method of showing optimality (which we claim for procedure $R_{U,D}$ with $D = 1$) is to show that the attained $H_{U,1}(N)$ value is equal to the HLB for every N and all q.

The ILB is based on the identity

$$(5.1) \qquad \sum_{j=0}^{D-1} \binom{N}{j} p^j q^{N-j} + \sum_{j=D}^{N} \binom{j-1}{D-1} p^D q^{j-D} = 1 ;$$

this holds since the first sum runs through the probabilities for the various possible number of defectives less than D and the second sum exhausts the possible positions of the D^{th} defective; the union of all these events is a disjoint exhaustive set.

It follows from (5.1) that

$$(5.2) \qquad p^D \sum_{j=D}^{N} \binom{j-1}{D-1} q^{j-D} = p^D \sum_{\alpha=0}^{N-D} \frac{\Gamma(\alpha+D)}{\Gamma(D)\alpha!} q^{\alpha} = I_p(D, N-D+1) ,$$

where $I_p(x,y)$ is the usual incomplete beta function, which we define to be 0 for $y = 0 < x$ and to be one for $x = 0 < y$. Note that in the first case of (5.1) we observe a particular one of the $\binom{N}{j}$ events with probability $p^j q^{N-j}$ and in the second case we observe a particular one of the $\binom{j-1}{D-1}$ events with probability $p^D q^{j-D}$. Hence the expected total information $E(I)$ learned under procedure $R_{1,D}$ is

$$(5.3) \qquad E(I) = -\sum_{j=0}^{D-1} \binom{N}{j} p^j q^{N-j} \log_2(p^j q^{N-j}) - \sum_{j=D}^{N} \binom{j-1}{D-1} p^D q^{j-D} \log_2(p^D q^{j-D})$$

$$= - p(\log_2 p)[N\ I_q(N\ I_q(N-D+1,\ D-1) + \frac{D}{p} I_p(D,\ N-D+1)]$$

$$- q(\log_2 q)[N\ I_q(N-D,D) + \frac{D}{p} I_p(D+1,\ N-D)] .$$

Since we can gain at most 1 unit of information per test, the maximum expected information we can gain from all the T tests is $E(T)$ and this must be greater than $E(I)$ in (5.3). Thus $E(I)$ is the ILB we are seeking. For $D = 1$ and $D = N$ it gives

$$(5.4) \qquad E(T|D = 1) \geq (- p \log_2 p - q \log_2 q)(\frac{1-q^N}{p}) = (\frac{1-q^N}{p})\ U(q) ,$$

$$(5.5) \qquad E(T|D = N) \geq N(- p\log_2 p - q\log_2 q) = N\ U(q) ,$$

and the latter agrees with the result for Procedure R_1 (see Section XII of [3]). If we think of D/N as approaching a limit λ as $N \to \infty$ then it is easy to show from (5.3) that

$$(5.6) \qquad E(I) \to E_\infty(I) = \begin{cases} \dfrac{\lambda N U(q)}{p} & \text{for } \lambda < p \\ NU(q) & \text{for } \lambda \geq p \text{ ;} \end{cases}$$

hence $E_\infty(I)$ is a simple approximation to the ILB (for N large and λ not extreme) if we set λN equal to D; for extreme λ or moderate N it may turn out that $E_\infty(I)$ is not a lower bound.

The Huffman lower bound (HLB) is obtained by utilizing the same identity (5.1) and again treating both combinatorial coefficients as "repetition factors", not as part of the probability of a basic event. Then we have a total number of probabilities S equal to

$$(5.7) \qquad S = \sum_{j=0}^{D-1} \binom{N}{j} + \sum_{j=D}^{N} \binom{j-1}{D-1} = 2^N I_{\frac{1}{2}}(N-D+1,\ D) + \binom{N}{D} \leq 2^N ,$$

which reduces to $N + 1$ for $D = 1$ and to 2^N for $D = N$. The Huffman routine combines the 2 smallest numbers and replaces them by a new number. Then we reorder (by magnitude) the set of $S-1$ numbers and repeat the process. As a check, we note that the last new number is equal to 1. The sum of all these new numbers is the desired HLB. Exact values of the expected number of tests under procedure $R_{U,D}$ together with the corresponding ILB and HLB bounds are given in table 2 for $D = 1, 2, 3$, $q = .90, .95, .99$ and selected values of N. Although this table does not cover a large range of P^*-values the illustrative examples in Section 6 show how the exact calculation can be carried out for any P^* using the fact that the strategy is known.

For the special case $D = 1$ the identity (5.1) reduces to

$$(5.8) \qquad q^N + p + qp + \dots + q^{N-1}p = 1$$

and the same identity (and indeed the same problem) has arisen in other contests. In [5] we regarded the basic binomial problem (with $P^* = 1$ and $D = N$) as composed of subproblems in each of which we looked for at most 1 defective unit. It was noted that procedure R_1 attained the HLB for each subproblem and hence was optimal in each subproblem, although not necessarily in the overall problem. In [2] we considered $N = \infty$ and looked for a single defective and the same optimality property was utilized. A proof of this optimality property was given by Hwang [1] who relates the optimal solution for finding at most 1 defective unit with finite N to the problem of finding the best alphabetic code for N states of nature with probabilities given on the left side of (5.8).

6. Comparison with Another Procedure R_S.

One other procedure R_S (and an equivalent variation R'_S of R_S) for the P^*-problem will be defined, briefly discussed, and then compared with procedure $R_{U,D}$. Comparisons will be made only between procedures with exactly the same $P\{CC\}$; this is accomplished by some form of randomization.

Procedure R_S is defined by using the same recursive equations as in (2.3), (2.4) and (2.6) except that the subscripts are all deleted and the boundary condition (2.5) is replaced by

$$(6.1) \qquad H(n) = 0 \qquad \text{if} \quad n = 0 \quad \text{or} \quad P\{CC|n\} \geq P^* .$$

Here $P(CC|n)$ denotes the probability of correctly classifying (by a guess) the n remaining unclassified units (in the H-situation) without any further tests. For $q \geq P^*$ this amounts to stopping (in the H-situation) when n is small enough so that $q^n \geq P^*$. For this procedure we assume that $q \geq P^* \geq 1/2$ and then the question of stopping in a G-situation with a guess does not arise since for any defective set of size $m \geq 2$ the probability of a correct quess is less than $1/2$. Hence we cannot satisfy the P^*-condition by stopping in a G-situation and there is no advantage to adding this to our boundary conditions.

A variation of R_S, denoted by R'_S, is easier to carry out and, because it is equivalent to R_S, shows that the strategy of R_S is exactly the same as that of procedure R_1 of [3] based on a smaller number of units. Let the integer M be defined by the inequalities

$$(6.2) \qquad q^{M+1} < P^* \leq q^M .$$

Then procedure R'_S sets aside at the outset M units (that are not classified) and uses the same recursive equations as R_S, with (6.1) replaced by

(6.3) $\quad H(n) = 0 \qquad \text{if} \quad n = 0$,

but applies them to only N - M units. These equations are the same as those of procedure R_1 in [3] and hence R'_S is the same as R_1 based on N - M units. Moreover, to satisfy the P^*- condition, we need only guess that all the M units set aside are good.

To see that R_S and R'_S are equivalent we note that every stopping point of R_S must be an $H(n_i)$ with $N_i \leq M$. Hence R_S classifies at least as many units as R'_S and, since both are operating in an optimal nested manner, R_P must have an expected number of tests that is at least as large as that of R'_S. On the other hand, in the class of procedures that do not keep track of (or make use of) the current (separate) numbers of units already classified as good and defective, the procedure R_S is an optimal nested procedure and hence must be at least as good (i.e., with an H-function that is at least as small) as R'_S. Hence R_S and R'_S must be equivalent procedures, i.e., they have an identical expected number of tests needed for termination. Hence R_S is equivalent to the application of procedure R_1 of [3] to N - M units.

Procedure $R_{U,D}$ does keep track of the number of units shown to be defective (actually, it records D minus that number) and we claim it is uniformly (for all q, all N and all P^*) better than procedure R_S, but this result has not been proved. We now show that for N large the procedure

$R_{U,D}$ is better than R_S for any fixed values of q and P^*. In the course of the proof we note that the amount by which $R_{U,D}$ is better is of the order $\sqrt{N}$, while the expected number of tests for both procedures has the same leading term of the form CN. Hence we can also assert that $R_{U,D}$ is asymptotically $(N \to \infty)$ equivalent to procedure R_S for any fixed values of q and P^*.

Since the strategy for both procedures is the same as for procedure R_1 of [3], it is sufficient to show for large N (i) that the total number of units classified under $R_{U,D}$ is less than under R_S and (ii) that the number of defectives classified under $R_{U,D}$ is less than under R_S.

To show (i), we have to show that for large N the expected number of units not classified under $R_{U,D}$ is at least M or

$$(6.4) \qquad M \le \sum_{j=D}^{N} (N-j)\binom{j-1}{D-1} p^D\, q^{j-D} = NI_p(D,N-D) - \frac{D}{p}\, I_p(D+1\ ,\ N-D)$$

$$= N \sum_{j=D}^{N-1} \binom{N-1}{j} p^j\, q^{N-1-j} - \frac{D}{p} \sum_{j=D+1}^{N} \binom{N}{j} p^j\, q^{N-j}$$

$$\le \sum_{j=D}^{N-1} \binom{N-1}{j} p^j\, q^{N-1-j} - \frac{D}{p}\,(1-P^*)\ ,$$

where we used (2.1) in the last step and (5.1) earlier. Since $M \ge 1$ depends only on P^*, it suffices to show that in an asymptotic expansion of the right side of (6.4) the coefficient of N vanishes and the coefficient of $\sqrt{N}$ is positive. Using an Edgeworth series for (5.1) (with continuity correction) up to terms of order $1/\sqrt{N}$, we let $w = w(N)$ denote the standardized variable $(D - 1/2 - N_p)/\sqrt{Npq}$ and obtain

$$(6.5) \qquad \Phi(w) - \frac{(q-p)(w^2-1)\varphi(w)}{6\sqrt{Npq}} \approx P^*$$

where $\Phi(w)$ and $\varphi(w)$ denote the standard normal c.d.f. and density, respectively. Let w_0 denote the root of $\Phi(w_0) = P^*$. We write $w = w_0 + \epsilon$ in (6.5) and use a 2-term Taylor expansion for $\Phi(w)$; in the second term the correction is of smaller order of magnitude and is ignored. Then, cancelling P^* on both sides of (6.5) we can solve for ϵ, then for w and finally for D, obtaining

$$(6.6) \qquad D \approx Np + w_0\sqrt{Npq} - \frac{1}{2} + \left(\frac{q-p}{6}\right)(w_0^{\,2} - 1).$$

For the sum on the right side of (6.4) we need $w_1 = w(N-1)$ and by (6.6) we have

$$(6.7) \qquad w_1 = \frac{D - \frac{1}{2} - (N-1)p}{\sqrt{(N-1)pq}} \approx w_0 + \frac{(q-p)(w_0^2 - 1)-q}{\sqrt{(N-1)pq}} .$$

Hence, the sum on the right side of (6.4) is (up to terms of order $1/\sqrt{N}$)

$$(6.8) \qquad N\{1 - \Phi(w_1) + \frac{(q-p)(w_1^{\,2}-1)\varphi(w_1)}{\sqrt{(N-1)pq}}\}$$

$$\approx N\{1 - \Phi(w_0) - \varphi(w_0)[\frac{(q-p)(w_1^{\,2}-1) - q}{\sqrt{(N-1)pq}}] + \frac{(q-p)(w_0^{\,2}-1)\varphi(w_0)}{\sqrt{(N-1)pq}}\}$$

$$= N\{1 - P^* + \frac{\varphi(w_0)}{\sqrt{(N-1)}}\sqrt{\frac{q}{p}}\} .$$

Using (6.8) and the two leading terms of (6.6) we obtain for the right side (RS) of (6.4).

$$(6.9)\qquad RS = w_0\sqrt{Nq/p}\ [\frac{\varphi(w_0)}{w_0} - (1 - P^*)] = w_0\sqrt{Nq/p}\ [\frac{\varphi(w_0)}{w_0} - \Phi(w_0)]\ .$$

By the well-known Feller-Laplace inequality the expression in brackets in (6.9) is positive for all P^* and this proves the result (i).

To prove (ii) we have to show that

$$(6.10)\qquad (N-M)p \ge \sum_{j=0}^{D} j\binom{N}{j}p^j q^{N-j} + D\sum_{j=D+1}^{N}\binom{N}{j}p^j q^{N-j}$$

$$= Np\sum_{j=1}^{D}\binom{N-1}{j-1}p^{j-1}q^{N-j} + D(1 - P^*)$$

$$= Np - p\ \{\sum_{\alpha=D}^{N-1}\binom{N-1}{\alpha}p^{\alpha}\ q^{N-1-\alpha} + \frac{D}{p}(1 - P^*)\}\ ,$$

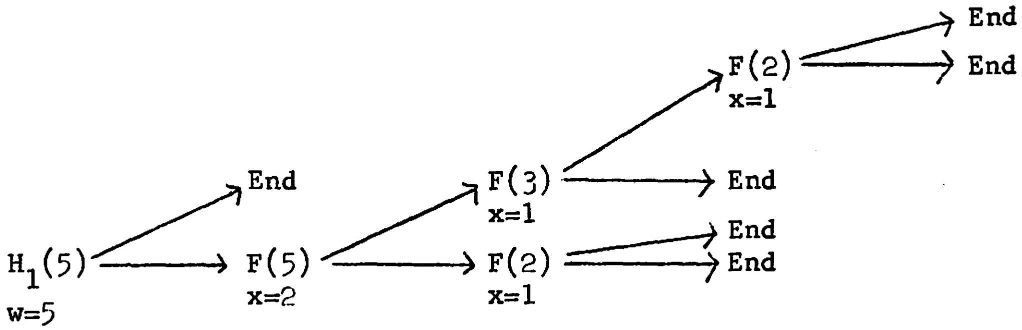
which reduces to the same inequality that was shown in (6.4). It follows that procedure $R_{U,D}$ is better than procedure R_S for sufficiently large N. Since we have cancelled Np in (6.10) and N in order to write (6.4) and the strategies are the same for both procedures, it also follows that they are asymptotically $(N \to \infty)$ equivalent.

To illustrate this result numerically suppose $N = 100$, $q = .95$ and $P^* = .9025 = q^2$, so that $M = 2$ for convenience. Then we can randomize between $D = 7$ $(P\{\mathcal{D} \le 7\} = .8720395)$ and $D = 8$ $(P\{\mathcal{D} \le 8\} = .9369104)$ to obtain a $P(CC)$ exactly equal to .9025; we select $D = 7$ with probability .5305 and $D = 8$ with probability .4695. The expected number of tests is 27.900 for the randomized $R_{U,D}$ and 28.382 for R_S which was procedure R_1

to classify 98 units. The ratio of these expectations is 98.3% and this can be interpreted as the efficiency of procedure R_S relative to procedure $R_{U,D}$.

In a second illustration we fix D, determine P^* and randomize on M. Suppose $N = 5$, $q = .90$ and $D = 1$. Then it is easy to verify, using (2.1), that $P^* \le ,9185$; for convenience, we take P^* equal to this value. Since we know that $R_{U,D}$ uses the same strategy as R_1 it is easy to see from [3] that the tree for $R_{U,D}$ (with $D = 1$, $q = .90$ and $N = 5$) is

(6.11)



where horizontal (slanted) arrows indicate at least 1 bad (all good). From (6.11) we obtain the formula for the expected number of tests

$$(6.12) \qquad H_1(5) = q^5 + 4q^3(1-q^2) + 3[q^2p + 1-q^2] = 3 + q^3 - 3q^5 = 1.9575$$

for $q = .90$. For procedure R_S we randomize and use $M = 0$ with probability .1854 and $M = 1$ with probability .8146 in order to get a P(CC) exactly equal to $P^* = .9185$. Hence

$$(6.13) \qquad H(5|R_S) = (.1854)(2.490) + (.8146)(2.051) = 2.1324$$

where the entries are $H(5|R_1)$ and $H(4|R_1)$, respectively taken from [3].

The efficiency of R_S relative to $R_{U,D}$ in this case is 91.8%. Such a calculation can be carried out for any q, any P^* and any moderate N since we know that the strategy (or x-value) in each situation is the same as for procedure R_1.

We conjecture that procedure $R_{U,D}$ is uniformly better than R_S for all fixed values of q, N and P^*. We also conjecture that $R_{U,D}$ is an optimal nested procedure for the P^*-problem. Finally we conjecture that R_S is an optimal nested procedure for the P^*-problem among those procedures that keep track of the total number of units classified but not of its breakdown into good and defective units.

7. Acknowledgement.

I wish to thank Mr. Fan-Nan O. Lin and Mr. Arthur Roth of the University of Minnesota and the students of my seminar in the Fall of 1973 for the instructive and inspiring discussions that contributed to this paper, especially in regard to the procedure R_S. Thanks are also due to Dr. Bernard S. Pasternack for discussions about this problem. I also wish to thank Mrs. Betty Hinkley for programming the exact and HLB entries in Table 2.

References

[1] Hwang, F. K. (1974). An optimal nested binomial group-testing procedure for identifying all defectives. (To appear).

[2] Kumar, S. and Sobel, M. (1971). Finding a single defective in binomial group-testing, JASA, 66, 824-828.

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

[4] Sobel, M. and Groll, P. A. (1966). Binomial group-testing with an unknown proportion of defectives, Technometrics, 8, 631-656.

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

[6] Sobel, M. (1973). Group-testing with at most D defectives: binomial and hypergeometric models. University of Minnesota School of Statistics, Tech. Report No. 204, (submitted for publication).

[7] Thomas, J., Pasternack, B. S., Vacirca, S. J. and Thompson, D. L. (1973). Application of group-testing procedures in radiological health, Health Physics, 25 259-266.

[8] Thomas, J., Pasternack, B. S., Bohning, D. E. and Vacirca, S. J. (1974). A group-testing method for reducing exposure to personnel leak-testing sealed radium sources. (Submitted).