BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/26

GROUP TESTING TO IDENTIFY THE SAMPLE MINIMUM FROM A DISCRETE UNIFORM DISTRIBUTION

by

Pedro J. Rodríguez
ASA/Census Research Program
Bureau of the Census
Washington, D.C. 20233

Frank K. Hwang
AT & T Bell Laboratories
Murray Hill, NJ 07974

and

Milton Sobel
Department of Mathematics
University of California
Santa Barbara, California 93106

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Kirk M. Wolter

Report completed: August 8, 1984

Report issued: August 8, 1984

Group Testing to Identify the Sample Minimum

From a Discrete Uniform Distribution

Pedro J. Rodríguez-Esquerdo*
ASA-Census Research Program
Statistical Research Division
U.S. Bureau of the Census
Washington D.C., 20233
(301) 763-3846

Frank K. Hwang
AT & T Bell Laboratories
Murray Hill, N.J., 07974

Milton Sobel
Department of Mathematics
University of California
Santa Barbara, California, 93106

May 1984

* Now at the University of Puerto Rico, Department of Mathematics, Rio Piedras, P.R. 00931. This research was supported in part through the ASA/Census Bureau Research Program. The program is jointly funded by the U.S. Census Bureau and the National Science Foundation (grant no. SES 81-2201). These institutions do not necessarily endorse the views expressed here.

ABSTRACT

A group testing procedure for the problem of identifying the sample minimum from a discrete uniform distribution in the minimal expected number of group tests is presented. A relation between the proposed procedure (PP) and other previously studied procedures (Sobel 1971; Hwang 1974, 1980) is shown to exist. Comparisons between the PP and other considered procedures are made and it is shown that the PP is uniformly better. A continuous case procedure is developed and is shown that under certain conditions the PP converges to it.

1. The Basic Problem

1.1 Introduction:

Consider the problem in which we wish to find both the minimum value of a set of stored variables and the position where those variables with value equal to the minimum are stored. The type of sampling allowed is to select a subset of those variables and find out its minimum value by testing simultaneously all the variables in the subset. At this point (unless the subset is of size one) we do not know where this minimum value lies or how many variables are equal to this minimum value, and (unless we test the whole set) we may not know whether the minimum value of the tested subset is equal to the minimum value of the original set. Furthermore, there is a cost of one for every test we perform (on one unit or simultaneously on several units).

We are interested in the goal of finding at least one variable whose value is equal to the minimum value of the original complete set. We wish to do this in the minimal (or near minimal) expected number of group tests. A similar goal applies for the corresponding problem dealing with the maximum value instead of the minimum.

This problem arises in several applications, such as when we are minimizing a discrete function and have stored the values in an array. If the number of values stored in the array is large, it may not be feasible or practical to find the minimum value of the function in a single (simultanoeus) test. So we devise a strategy to test subarrays and find the minimum value of each subarray tested. It is clear that the conditions described above hold for this problem. The problem is to find an efficient or optimal strategy for achieving our goal.

1.2 <u>Formulation of the Problem:</u>

Consider the case when we have N independent identically distributed (i.i.d.) random variables $Y_1,\ldots,Y_N$, each with the discrete uniform distribution over the integers 1,2,3,...,r. Consider the random variable $M=\min(Y_1,Y_2,\ldots,Y_N)$ and let m be the realized value of M for any given point in our sample space S. Let y_k denote the observed value of Y_k for k=1,2,3,...N. Our goal can now be stated as follows: to find, in the minimal (or near minimal) expected number of group tests the value m and any one subscript k, k=1,2,...,N such that $y_k = m$.

In this context a group test is a test on one or more items that furnishes us with the minimum of the item values in the tested subset, i.e., the value of m for the set tested. The subscript k of y_k is sometimes used to refer to a unit. We say that a unit is defective if and only if its value is equal to the overall minimum, i.e., iff $y_k=m$. We say that a set is contaminated if it is known to contain at least one defective unit. The minimum value in a set of size x will be denoted by m_x and the corresponding random variable by M_x.

Before performing any test we do not know the actual value of the overall minimum m, and based on the optimal procedure we may not know this value, even after several tests. The random variable M has a probability distribution which depends on the upper bound r of the uniform distribution and on the number N of random variables that we have at the outset. Thus it is easy to see that for any integer m,

$1 \leq m \leq r$ we have,

$$P(M = m) = \left[\frac{r - m + 1}{r}\right]^N - \left[\frac{r - m}{r}\right]^N \qquad (1.1)$$

In the special case when the minimum value of the whole set is known and it is also known that there is exactly one variable equal to this minimum value, then the Halving Procedure solves this problem optimally (1974). In our case, if the minimum value is known to be m, then our problem becomes analogous to that studied by Kumar and Sobel (1971) and later by Hwang (1974).

2. Finding a Single Unit with Value Equal to the Minimum

2.1 Preliminaries

Under some strategies we find in the very first group test the value m of $M=\min(Y_1,Y_2,\ldots,Y_N)$ by testing simultaneously all the N units. However, the optimal strategy is not necessarily of this type. In fact we start by showing that we can improve on any procedure that starts by testing all the N units.

Suppose that after testing all N units in the first group test and observing the value m of M our procedure requires us to test x units in the second group test. We now compare this with an alternative procedure which tests x units (the same x as before) in the first group test and then the remaining N-x units in the second group test.

In both situations we have used up two group tests and we know with certainty that M = m, and moreover, after two tests we appear to be at the same stage with both of these starting strategies, but, are we really? For "large" values of x, the x units will contain a defective unit with high probability, and in fact, if N is suffiently large, the probability that the overall minimum value is equal to one, Prob(M = 1) will be close to one. Hence, if by taking x units (with $x < N$) in our first group test we find out that the minimum value of the x units is one, we do not have to test the remaining N-x units, and thus we have effectively saved one group test. In any case, the answer to our query above is that we have increased the amount of information about the N units by using the second starting procedure above.

In this section we present several procedures for accomplishing our goal. We propose a procedure (denoted by R_2) that entails the consideration of at most two sets at any given time. One set (possibly empty) will contain the untested units (if any) and the other set V_m (initially empty), will contain v units known to have a minimum value equal to m. This procedure will be described in detail in (2.3) below.

In order to decribe the procedure mentioned above, we need to introduce some definitions and preliminary results. Let V_m denote a set of size v which is known to contain a minimum value equal to m and let XUY denote the pooled set of size x+y of which x units are taken from V_m and y units from the remaining untested units. Then we define $p_v^{(a)}(x,y)$ by

$$p_v^{(a)}(x,y) = \text{Prob [} \min(X \cup Y) = a \mid X \subset V_m \text{]} \tag{2.1}$$

i.e., $p_v^{(a)}(x,y)$ is the probability that if we choose x units from a set V_m of size v known to contain a minimum value equal to m and y units from the remaining (untested) units, the minimum value in the pooled set of these x+y units is equal to a. Clearly for x=v and a>m we have

$p_v^{(a)}(x,y) = 0$.

We also define the function f(m,y) as follows:

$$f(m,y) = (r - m + 1)^y - (r - m)^y = r^y P(\min(Y) = m). \tag{2.2}$$

Lemma 2.1

For $v\neq 0$, $1 \leq a \leq r$, $1 \leq m \leq r$, $0 \leq x \leq v$ and $0 \leq y \leq N$ we have

$$p_v^{(a)}(x,y) = \frac{1}{r^y} \begin{cases} \left[\frac{f(m,v-x)}{f(m,v)}\right] f(a,x+y) & a>m \\ f(m,y) + (r-m)^y\,(r-m+1)^{v-x}\left[\frac{f(m,x)}{f(m,v)}\right] & a=m \\ f(a,y) & a<m \end{cases} \tag{2.3}$$

and for $v=0$, $x=0$ and a, y as above,

$$p_0^{(a)}(0,y) = \left[\frac{r-a+1}{r}\right]^y - \left[\frac{r-a}{r}\right]^y \quad . \tag{2.4}$$

Proof:

See Rodríguez-Esquerdo (1983).

We also define $q_{x,v}^{(m)}$ to be the probability that a subset X of size x, taken from a set V_m of size v contains a minimum value that is greater than m. It is easy to see (Rodríguez-Esquerdo, 1983) that

$$q_{x,v}^{(m)} = P(\,Y_1>m,\ldots,Y_x>m \mid M_v = M = m) = (r-m)^x\left[\frac{f(m,v-x)}{f(m,v)}\right]. \tag{2.5}$$

Lemma 2.2

Let $Y_1,Y_2,\ldots,Y_N$ be as above, i.i.d. uniformly distributed over the integers $1,2,3,\ldots,r$. If the minimum value of the N units is known to be equal to m and if X is a proper random subset consisting of x units from the N units ($x < N$) such that the minimum value in X is also equal to m , then the units in the set N\X are i.i.d. uniformly distributed over the integers $m,m+1,\ldots,r$.

Proof:

Let Y be a randomly chosen unit in the set N\X, then we are interested in the following probability for $k \geq m$:

$$P = P(Y = k \mid \min(N) = m,\ \min(X) = m,\ X \subset N,\ Y \in N\backslash X).$$

We can write this probability P as follows:

$$P = \frac{P(Y = k,\ \min(N) = m,\ \min(X) = m \mid X \subset N,\ Y \in N\backslash X)}{P(\min(N) = m,\ \min(X) = m \mid X \subset N,\ Y \in N\backslash X)}$$

$$= \frac{P(Y = k\ ,\ \min(N\backslash X) > m,\ \min(X) = m \mid Y \in N\backslash X)}{P(\min(N\backslash X) \geq m,\ \min(X) = m)}$$

$$= \frac{P[Y = k,\ \min((N\backslash X)\backslash Y) > m]\ P(\min(X) = m)}{P(\min(N\backslash X) \geq m)\ P(\min(X) = m)}$$

$$= \frac{\frac{1}{r}\ \left[\frac{r - m + 1}{r}\right]^{N-x-1}}{\left[\frac{r - m + 1}{r}\right]^{N-x}} = \frac{1}{r - m + 1}\ .$$

which is what we wanted to show. This result is analogous to a result in binomial group testing (Sobel and Groll, 1959). Note that if a set of size v is contaminated, then the distribution of any unit in the set is given by Lemma 2.1, setting x=1 and y=0.

2.2 The Conditional Procedure $R_c(v \mid M = m)$

Let $H_2(v \mid M_v = M = m)$ be the expected number of tests needed to find a single defective unit when the value of the overall minimum value M is known to be equal to m, we have a contaminated set of size v and we use the conditional procedure R_c, i.e., the procedure when m is known. This case, as we will see, can be interpreted as a binomial group testing problem with $q = (r - m)/(r - m + 1)$.

The (nested) conditional procedure $R_c(v \mid M = m)$ for finding a single unit whose value is equal to m is then described implicitly for v>1 by the following recursive relation , where the actual number of units to test at any stage is given by the minimizing value of x in equation 2.6 below.

$$H_2(v \mid m) = 1 + \min_{1 \le x < m} \{ q^{(m)}_{x,v} H_2(v-x \mid m) + (1 - q^{(m)}_{x,v}) H_2(x \mid m) \} \qquad (2.6)$$

with the following boundary conditions:

$H_2(1 \mid m) = 0$ and

$H_2(v \mid M = r) = 0.$

There is a relation (an analogy) between our conditional procedure and the procedure F(v) for finding a single defective as defined in Sobel and Groll (1959) and also in Kumar and Sobel (1971).

$$F(v) = 1 + \min_{1 \le x < v} \{ \left[\frac{q^x - q^v}{1 - q^v} \right] F(v-x) + \left[\frac{1 - q^x}{1 - q^v} \right] F(x) \} \qquad (2.7)$$

with the single boundary condition

$F(1) = 0.$

If we associate the q in Kumar and Sobel with $(r-m)/(r-m+1)$ then i) $H_2(v \mid m) = F(v)$ and (ii) F(v) will depend on m only through this quantity $q=(r-m)/(r-m+1)$, which is the probability that a unit has a value larger than m given that it is uniformly distributed among the integers $m, m+1, \ldots, r$.

Theorem 2.3

If we let $q=(r-m)/(r-m+1)$ then $H_2(v \mid m) = F(v)$ and the latter depends on m only through q.

Proof:

From the definition of q we inmediately obtain

$r - m + 1 = \frac{1}{1-q}$ and $r - m = \frac{q}{1-q}$ and hence using (2.5) we have

$$q_{x,v}^{(m)} = \left[\frac{q}{1-q} \right]^x \left[\frac{\left[\frac{1}{1-q} \right]^{v-x} - \left[\frac{q}{1-q} \right]^{v-x}}{\left[\frac{1}{1-q} \right]^{v} - \left[\frac{q}{1-q} \right]^{v}} \right] = \frac{q^x - q^v}{1 - q^v} \quad .$$

Therefore, we can write (2.6) in the form

$$H_2(v \mid m) = 1 + \min_{1 \le x < v} \left\{ \frac{q^x - q^v}{1 - q^v} H_2(v-x \mid m) + \frac{1 - q^x}{1 - q^v} H_2(x \mid m) \right\}$$
$$= F(v).$$

Hence, after the value of the overall minimum has been found the conditional procedure $R_c(v \mid m)$ becomes analogous to that of Kumar and Sobel provided we associate the probability q of any one unit being non defective with (r-m)/(r-m+1). This procedure was shown to be optimal by Kumar and Sobel (1971) and later by Hwang (1974).

2.3 The $R_2(N)$ Procedure with At Most Two Sets (AMTS)

We propose a procedure that accomplishes efficiently our goal, it finds the value of the minimum m and a subscript k such that $y_k = m$. At any given point this procedure will never consider more than two distinct sets, i.e., the units within each set do not have to be numbered or otherwise identified. One set (possibly empty) will contain the untested units and another set V_m, (initially empty) is known to contain a minimum value equal to m among the v units in this set. R_2 is unconditional in nature, it does not assume any prior information about the overall minimum value other than the initial conditions of the model which we repeat below explicitly.

The proposed procedure works as follows:

Initial conditions: N i.i.d. units that have the discrete uniform distribution over the integers 1,2,...,r.

Step 1: Choose a random set V of v units and test it. If the minimum value m in this set is equal to 1 or if v=N and m>1 is observed, then use the Conditional Procedure $R_c(v \mid M=m)$ on these v units. Otherwise go to step 2.

Step 2: Choose a subset X from V_m and a subset Y from the remaining untested set N\V. Test the pooled set XUY and let the minimum value observed here be equal to s. If s=1 then perform the Conditional Procedure $R_c(y \mid M=1)$ on the y units. Otherwise go to step 3.

Step 3: If y<N-v (i.e., if there are still untested units) go to step 4, otherwise:

i) if s<m perform the Conditional Procedure $R_c(y \mid M=s)$ on the y units.

ii) if s>m perform the Conditional Procedure $R_c(v-x \mid M=m)$ on the v-x units.

iii) if s=m perform the Conditional Procedure $R_c(\min(x+y,v) \mid M=m)$ on the smaller of the two sets XUY and V.

Step 4: Update N to N-v-y (i.e. for the next test we will have N-v-y untested units).

i) if s<m discard the v units, replace the set V_m by Y_s and go to step 2.

ii) if s>m discard the set $(XUY)_s$, replace the set V_m by $(V\backslash X)_m$ and go to step 2.

iii) if s=m discard the largest of y and v-x, replace the set V_m by the smaller of the two sets $(XUY)_m$ and V_m then go to step 2.

The number of units to be tested at each step (the quantities v,x,y above) are such that they satisfy the recursive relations (2.8) and (2.9) below. This procedure decreases the size of at least one set after each test and thus will eventually find the value of the overall minimum. The boundary conditions (2.11), (2.12) and (2.13) below guarantee that the procedure will terminate succesfully.

In order to write the recursive relations that describe implicitly our proposed procedure we need some notation.

Let H(N) denote the expected number of group tests needed to be performed if at the outset we have N units, we use the AMTS Procedure and the only information available to us initially is that the units are i.i.d. uniformly distributed among the integers 1,2,...,r.

Let $H_1(v,m;n)$ denote the expected number of tests to be performed if we have a set V_m of size v known to contain a minimum value equal to m and we have n untested units.

We can now implicitly define the AMTS Procedure $R_2(N)$ by the following recursive relations. For $N \geq 1$

$$H(N) = 1 + \min_{1 \leq y \leq N} [p_0^{(1)}(0,y)H_2(y|M=1) + \sum_{m>1} p_0^{(m)}(0,y)H_1(y,m;N-y)] \quad (2.8)$$

For $m > 1$ and $n \geq 1$

$$\begin{aligned} H_1(v,m;n) = 1 + \min_{0 \leq x < v} \min_{1 \leq y \leq n} [&p_0^{(1)}(0,y)H_2(y|M=1) \\ &+ \sum_{1<s<m} p_0^{(s)}(0,y)H_1(y,s,\ n-y) \qquad (2.9) \\ &+ \sum_{s>m} p_v^{(s)}(x,y)H_1(v-x,\ m;n-y) \\ &+ p_v^{(m)}(x,y)H_1(\min(x+y,v),\ m;n-y)]. \end{aligned}$$

For $n > 1$ (as a result of boundary condition (2.13) below we need)

$$H_2(n|M=m) = 1 + \min_{1 \le x < n} [q_{x,n}^{(m)} H_2(n-x|M=m) + (1-q_{x,n}^{(m)}) H_2(x|M=m)] \quad (2.10)$$

where the value of $q_{x,n}^{(m)}$ is given by (2.5) above. For $R_2(N)$ we have the following boundary conditions

$$H(1) = 1 \quad (2.11)$$

$$H_2(1|M=m) = 0 \quad (2.12)$$

$$H_2(n|M=r) = 0 \text{ for every } n.$$

$$H_1(v,m;0) = H_2(v|M=m) \text{ for every } m. \quad (2.13)$$

We can define an alternative procedure $R_a(N)$ to the AMTS Procedure for achieving our Goal that starts by testing all of the N units in the first group test. It finds the value of the overall minimum in only one test and thus is optimal for the problem of finding this value. But to achieve our Goal we also have to find the location of the minimum value. The procedure may be described as follows:

$$H_a(N) = 1 + \sum_{m=1}^{r} p_0^{(m)}(0,N) H_2(N|M=m) \quad (2.14)$$

where $H_2(N|M=m)$ is given by (2.10) above. We have shown that we can actually do better (smaller expected number of tests) than by using this procedure.

2.4 The Nested Procedure $R_n(N)$

The nested procedure for this problem is an extension of the idea that as soon as we have a contaminated set we go and search for the defective item in that set. As a result, it can be derived directly from the AMTS procedure by setting x=0 in (2.9), i.e., we do not test units from the contaminated set until we find out the value of the overall minimum.

2.5 The One at a Time Procedure (OAT)

For our Goal, we can test units one by one. It is clear that using the OAT procedure we stop testing units only when the value of the unit tested is equal to one or when we have tested all the units.

The probability that a unit has value one is given by (2.4) above, setting v=0, x=0, a=1 and y=1, namely

$$p_0^{(1)}(0,1) = 1 - \frac{r-1}{r} = \frac{1}{r}. \tag{2.15}$$

Let $H_0(N)$ denote the expected number of tests required to achieve Goal 1 under the OAT procedure, then using (2.15) above we can write

$$H_0(N) = \frac{1}{r} \sum_{i=1}^{N-1} i\left(\frac{r-1}{r}\right)^{i-1} + N\left(\frac{r-1}{r}\right)^{N-1} = r\left[\, 1-\left(\frac{r-1}{r}\right)^{N} \,\right] \tag{2.16}$$

We can also write a recursive equation for $H_0(N)$ as follows:

$$H_0(N) = 1 + \left(\frac{r-1}{r}\right)H_0(N-1) \quad \text{with the boundary} \tag{2.17}$$

condition that $H_0(0) = 0$.

2.6 The Continuous Case Problem

In this section we consider essentially the same problem and Goal 1 as before with one important distinction, we suppose here that we have N i.i.d. random variables $Y_1, Y_2, \ldots, Y_N$ that have the continuous uniform distribution over the interval (0,1). In this case we know, with probability one, that there is exactly one unit with value equal to the overall minimum value.

Since it is well known that the Halving Procedure $R_h(N)$ is optimal for the problem of finding a single defective when it is known that there is exactly one (Sobel, Pasternack and Thomas, 1974) we can be tempted to use $R_h(N)$ for accomplishing our Goal. The difficulty here is that, at the outset, we do not know what a defective unit is, since we do not know the value of the overall minimum. In this section, we describe two procedures that accomplish our Goal and are no worse than the procedure that tests all the N units in the first group test for finding the overall minimum value and then uses the Halving Procedure for finding a single defective unit. These two proposed procedures have the property that they reduce to the Halving Procedure when the value of the overall minimum is known.

Before continuing we need to compute some probabilities. The distribution of the minimum value of y i.i.d. units uniformly distributed in (0,1) is given by

$$P(M \le m) = 1 - (1-m)^y.$$

(2.18)

The distribution of the minimum value in a set Y, given that this minimum value m′ in Y is less than the current minimum value m is given, for $0 \le m' \le m$ by

$$P(\min(Y) \le m' | m' < m) = \frac{P(\min(Y) \le m', \min(Y) < m)}{P(\min(Y) < m}$$

$$= \frac{1 - (1-m')^y}{1 - (1-m)^y} \tag{2.19}$$

And, lastly, we need the probability that if we chose a set X of x units from a set V_m of size v known to contain a minimum value equal to m and y units from the remaining (untested) units, the minimum value is equal to m,

$$P(\min(X \cup Y) = m \mid X \subseteq V_m) = \frac{\binom{v-1}{x-1}}{\binom{v}{x}} (1-m)^y = \frac{x}{v}(1-m)^y. \tag{2.20}$$

Using the same notation as before we describe the first procedure R_{c1} for accomplishing Goal 1 in the continuous case by the following recursive relations: For $N \geq 1$

$$H_{c1}(N) = 1 + \min_{1 \leq y \leq N} [\int_0^1 H_{c1}(y,m;N-y)y(1-m)^{y-1}dm]. \tag{2.21}$$

For $n > 0$

$$\begin{aligned} H_{c1}(v,m;n) = 1 + \min_{0 \leq x < v} \min_{1 \leq y \leq n} [&\int_0^m H_{c1}(y,m';n-y)y(1-m')^{y-1}dm' \\ &+ (\tfrac{v-x}{v})(1-m)^y H_{c1}(v-x,m;n-y) \\ &+ (\tfrac{x}{v})(1-m)^y H_{c1}(\min(x+y,v), m;n-y)]. \end{aligned} \tag{2.22}$$

We need the following boundary conditions,

$$H_{c1}(1) = 1 \tag{2.23}$$

$$H_{c1}(v,m;0) = H_h(v) \tag{2.24}$$

where

$$H_h(v) = K + \frac{2J}{v} \text{ where } v = 2^K + J \text{ and } 0 \leq J < 2^K. \tag{2.25}$$

The procedure R_{c1} requires some explanation. It is an analogous procedure to $R_2(N)$ (see section 2.3), but here, once the value of the overall minimum is known, we use the Halving Procedure (see Pasternack, Sobel and Thomas, 1974). In the H_1 situation (when we have a set size v with current minimum value m and n untested units) we can make some further savings on the expected number of tests if we are willing to allow inference based on probability one to be made. Suppose that we test XUY and that the minimum value contained in XUY is equal to the minimum value contained in V_m (recall that $X \subseteq V_m$), then we may infer with probability one that the minimum value is contained in the set X.

We now describe a recursive procedure, R_{c2} for accomplishing Goal 1 in the continuous case, when inference with probability one is allowed as follows:

For $N \geq 1$

$$H_{c2}(N) = 1 + \min_{1 \leq y \leq N} [\int_0^1 H_{c2}(y,m;N-y)y(1-m)^{y-1}dm]. \quad (2.26)$$

For $n \geq 1$,

$$H_{c2}(v,m;n) = 1 + \min_{0 \leq x < v} \min_{1 \leq y \leq n} [\int_0^m H_{c2}(y,m';n-y)\; y\; (1-m')^{y-1}dm'$$

$$+ (\frac{v-x}{x})(1-m)^y H_{c2}(v-x,m;n-y)$$

$$+ (\frac{x}{v})(1-m)^y H_{c2}(x,m;n-y)], \quad (2.27)$$

where the boundary conditions are the same as for R_{c1}, namely, (2.24) and (2.25) still hold.

Note that R_{c1} and R_{c2} differ only on the approach to the situation when the minimum value in the set XUY is equal to the current minimum value m. R_{c1} keeps the smallest of the two sets XUY and V while R_{c2}

only keeps the set X. We conjecture that R_{c2} is an optimal group testing procedure among all group testing procedures that have probability one of identifying a single defective when the units are uniformly distributed in the interval (0,1]. R_{c2} considers all possible outcomes and by backward optimization chooses the best action to take.

In the following section we discuss some interesting results and comparisons between the different procedures discussed here and in the previous sections.

3. Results On The Procedures For Accomplishing Our Goal

Here we present several results on the procedures described in the previous section. We also present a discussion of the non-optimality of the proposed AMTS procedure.

Theorem 3.1 An Upper Bound for H(N).

Let r, the upper bound of the uniform distribution and N, the initial number of units be fixed, then, for the At Most Two Sets Procedure we have:

$$H(N) \leq 1 + \sum_{m=1}^{r} p_0^{(m)}(0,N)H_2(N|M=m) = H_a(N) \qquad (3.1)$$

Proof:
$H_a(N)$ is a special case of H(N) occurring when we test y=N units in our first test. It is interesting to note that for r=100 we have $H_a(3)=2.9701 > 2.3332=H(3)$. More extensive tables can be found in (Rodríguez Esquerdo, 1983).

Theorem (3.1) above tells us that the expected number of tests, by our proposed AMTs procedure $R_2(N)$ will be no more than the expected number of tests needed if we test all the N units in the first test and then look for a specific unit whose value is equal to the observed minimum.

<u>Theorem 3.2</u> A Lower Bound for H(N).

Let r and N be fixed, then, for the At Most Two Sets Procedure we have:

$$H_a(N) - 1 = \sum_{m=1}^{r} p_0^{(m)}(0,N)H_2(N|M=m) \leq H(N) \qquad (3.2)$$

<u>Proof</u>:

Suppose that we knew the true value m of the overall minimum M without any costs at all to us. Then, since we the conditional Procedure is optimal for finding a single defective unit when the value of the minimum is known, we would use it to find optimally a single defective. As an alternative we can make no use of this knowledge and use the AMTS procedure assuming that the true value of the minimum is unknown, thus for any N and any m, we clearly have the inequality

$$H_2(N|M=m) \leq H(N), \text{ since the left side is optimal.}$$

Since $P(M=m) \geq 0$, $\sum_{m=1}^{r} P(M=m) = 1$ and the above inequality is true for all values of m, m=1, 2,...,r we have from the definition of $H_a(N)$ in (2.14)

$$H_a(N) - 1 = \sum_{m=1}^{r} p_0^{(m)}(0,N)H_2(N|M=m) \leq H(N),$$

which establishes our theorem.

Theorem 3.4 An Upper Bound for $H_n(N)$.

For the nested procedure we have:

$$H_n(N) \leq 1 + \sum_{m=1}^{r} p_0^{(m)}(0,N)H_2(N|M=m) = H_a(N) \tag{3.4}$$

Proof:

The proof is identical to that of Theorem (3.1).

Theorem (3.4) tells us that the Nested Procedure is also no worse than the Alternative Procedure R_a that tests all the N units on the first test and then looks for a specific unit whose value is equal to the observed minimum.

Theorem 3.5

If N and r are fixed, and H(N) denotes the expected number of tests for the AMTS Procedure and $H_n(N)$ the expected number of tests for the Nested Procedure, then,

$$H(N) \leq H_n(N) \tag{3.5}$$

We have a corollary that follows immediately from Theorems (3.1), (3.2), (3.4) and (3.5) above.

Corollary 3.6

$$H_a(N) - 1 \leq H(N) \leq H_n(N) \leq H_a(N) \tag{3.6}$$

We now consider the OAT procedure and show that this procedure is no better than the AMTS procedure. In order to do this we first r a Lemma.

. 3.7

r, n be fixed and m any integer in $\{1,2,\ldots,r\}$, then

$$H_1(1,m;n) \leq H_0(n) \quad . \qquad (3.7)$$

oof:

We prove (3.7) above by induction on n. For n=1 we have

$$H_1(1,m;1) = 1 = H_0(1).$$

For n=2 we have

$$H_1(1,m;2) = 1 + \min_{1\leq y\leq 2} \{p_0^{(1)}(0,y)H\ (y|1) + \sum_{1<s<m} p_0^{(s)}(0,y)H_1(y,s;2-y)$$

$$+ \sum_{s>m} p_0^{(s)}(0,y)H_1(1,m;2-y) + p_0^{(m)}(0,y)H_1(1,m;2-y)\}$$

$$\leq 1 + p_0^{(1)}(0,1)H_2(1|1) + \sum_{1<s<m} p_0^{(s)}(0,1)H_1(1,s;1)$$

$$+ \sum_{s>m} p_0^{(s)}(0,1)\ H_1(1,m;1) + p_0^{(m)}(0,1)H_1(1,m;1) \quad . \qquad (3.8)$$

Using the fact that $H_1(1,m;1) = 1$, $H_2(1|1)=0$ and that $p_0^{(a)}(0,y)=1/r$ for all a, a=1,2,3,...,r we get for inequality (3.8) above

$$H_1(1,m;2) \leq 1 + \sum_{1<s\leq r} \frac{1}{r} = 1 + \frac{r-1}{r} = H_0(2) \quad . \qquad (3.9)$$

Suppose now that (3.7) is true for all integers k, k=1,2,3,...,n-1, then

$$H_1(1,m;n) \le 1 + \frac{1}{r} H_0(1|1) + \frac{1}{r} \sum_{1<s<m} H_1(1,s;n-1)$$

$$+ \frac{1}{r} \sum_{s>m} H_1(1,m;n-1) + \frac{1}{r} H_1(1,m;n-1) . \qquad (3.10)$$

Using the induction hypothesis and the fact that $H_2(1|1) = 0$ we get for (3.10)

$$H_1(1,m;n) \le 1 + \frac{1}{r} (m-2) H_0(n-1) + \frac{r - m + 1}{r} H_0(n-1)$$

$$= 1 + H_0(n-1) \left(\frac{r - 1}{r}\right) = H_0(n), \qquad (3.11)$$

using (2.17) above. This proves our lemma.

We can now prove

Theorem 3.8

Let N,r be fixed, then $H(N) \le H_0(N)$. (3.12)

Proof:

From (2.8) we have

$$H(N) = 1 + \min \{p_0^{(1)}(0,y)H_2(y|M=1) + \sum_{m>1} p_0^{(1)}(0,y) H_1(y,m;N-y) \} ,$$

setting y=1 in the minimization we get

$$H(N) \le 1 + \sum_{m>1} \frac{1}{r} H_1(1,m;N-1) \le 1 + \frac{1}{r} \sum_{m>1} H_0(n-1) = H_0(N). \qquad (3.13)$$

which proves our theorem.

We now prove a convergence result for the AMTS procedure.

Theorem 3.11

For a fixed value of N, as the upper limit r of the discrete uniform distribution increases without bound, the AMTS procedure R_2 approaches the Continuous case procedure R_{c1}, where inference based on probability one is not allowed.

Proof:

The problem of finding the minimum value and a unit whose value is equal to it, when the units are i.i.d. uniformly distributed over the integers 1,2,3,...,r is equivalent to the problem of accomplishing this goal when the units are i.i.d. uniformly distributed over the rational numbers $1/r, 2/r, \ldots, 1$.

The probability $p_v^{(a)}(x,y)$ in (2.3) when m and a are in the set $\{1,2,\ldots,r\}$ can be rewritten for m and c in the set $\{1/r, 2/r, \ldots, 1\}$ as follows

Define $g(m,x) = (1-m+1/r)^x - (1-m)^x$

then

$$p_v^{(c)}(x,y) = \begin{cases} \left[\frac{g(m,v-x)}{g(m,v)}\right] g(c,x+y) & , c>m \\ g(m,y) + (1-m)^y (1-m+\frac{1}{r})^{v-x} \left[\frac{g(m,x)}{g(m,v)}\right], & c=m \\ g(c,y) & , c<m \end{cases} \tag{3.14}$$

It is easy to see, using L'Hopital's rule that

(i) $\lim_{r\to\infty} r p_v^{(c)}(x,y) = y(1-c)^{y-1}$ for $c < m$,

(ii) $\lim_{r\to\infty} p_v^{(m)}(x,y) = \frac{x}{v}(1-m)^y$ for $c = m$, and that

(iii) $\lim_{r\to\infty} \sum_{c>m} p_v^{(c)}(x,y) = \lim_{r\to\infty} \frac{g(m,v-x)}{g(m,v)} = (1-m)^y \quad (\frac{v-x}{v})$

Using the expression for $p_v^{(c)}(x,y)$ in (3.14) we consider (2.9) and show that the terms in brackets converge one by one to the respective terms in brackets in (2.22).

The first term is $p_0^{(1/r)}(0,y)H_2(y|1/r)$ and it is easy to see that

$$\lim_{r\to\infty} p_0^{(1/r)}(0,y)H_2(y|\frac{1}{r}) = 0.$$

Consider now $\sum_{\frac{1}{r}<s\leq m} p_0^{(s)}(0,y)H_1(y,s;n-y)$ where s

ranges over the rational numbers of the form i/r, i=1,2,...,mr-1.

Rewrite this term as:

$$\sum_{\frac{1}{r}<s<m} rp_0^{(s)}(0,y)H_1(y,s;n-y)\,\frac{1}{r}\,,$$

Since $H_1(y,s;n-y)$ is a bounded function and using (i) above, this term converges to the integral

$$\int_0^m y(1-s)^{y-1}H_1(y,s;n-y)ds.$$

Using (iii) above and the same reasoning, we see that

$$\sum_{s>m} p_v^{(s)}(x,y)H_1(v-x,m;n-y) \rightarrow (1-m)^y(\frac{v-x}{v})H_1(v-x,m;n-y).$$

Using (ii) above we have that as $r \rightarrow \infty$,

$$p_v^{(m)}(x,y)H_1(\min(x+y,v),m;n-y) \quad \rightarrow \quad (1-m)^y \quad (\frac{x}{v})\ H_1(\min(x+y),m;n-y).$$

Therefore $H_1(v,m;n)$ converges to

$$1 + \min_{0 \le x < v} \min_{1 \le y \le n} [\int_0^m y(1-s)^{y-1} H_1(y,s,;n-y)ds$$

$$+ (1-m)^y (\frac{v-x}{v}) H_1(v-x,m;n-y)$$

$$+ (1-m)^y (\frac{x}{v}) H_1(\min(x+y),m;n-y)] \quad (3.15)$$

this is just expression (2.22) with H_{c1} replaced by H_1. Now we need to show that

$$p_0^{(1/r)}(0,y)H_2(y|M=1/r) + \sum_{m>1} p_0^{(m)} (0,y)H_1(y,m;N-y)$$

$$= \sum_{m=1/r}^{1} p_0^{(m)}(0,y)H_1(y,m;N-y) \quad (3.16)$$

coverges to

$$\int_0^1 y(1-m)^{y-1} H_{c1}(y,m;N-y)dm.$$

Rewriting the left hand side of (3.16) above as

$$\sum_{m=1/r}^{1} rp_0^{(m)}(0,y)H_1(y,m;N-y)1/r,$$ using (i) above and by

the definition of the integral, the left hand side of (3.16) converges to the desired integral. It is easy to see that boundary condition (2.13) converges to the Halving Procedure.

Therefore, as r increases without bound $H_1(v,m;n) \to H_{c1}(v,m;n)$ and $H(N) \to H_{c1}(N)$.

3.3 Optimality Discussion

The question about the optimality of $R_2(N)$ remains to be answered. We present the reason why $R_2(N)$ is not optimal when considered in the set of all possible group testing procedures. An optimal procedure keeps track of all the possible states and situations arising in the problem, uses all of the available information and then makes the best choice of action to take (group to test) without any restriction. We show that our proposed procedure actually throws away some information.

Initially, the only state possible is the one where we have a set of units and have no knowledge of the overall minimum value (except for the distribution of the units). We then test a non empty subset V of N (for the following discussion we assume that V is a proper subset of N) and find the value m of the minimum M in this set V (which then becomes V_m). Now we have a new state with two sets, N\V and V_m. Our proposed procedure now proceeds to choose a subset X of V_m (possibly empty) and a nonempty subset Y of N\V, pool them in a set XUY, and then test this set. Here we have three possible outcomes:

(i) $\min(XUY) < m$

(ii) $\min(XUY) > m$, and

(iii) $\min(XUY) = m$.

No loss of optimality or difficulties arise in the first two cases or when the set X is empty. Consider the last case (iii), when X is nonempty, $R_2(N)$ chooses the smallest of the two sets $(XUY)_m$ and V_m and calls it the new V_m. No mention is made about the set X. We know that X is the intersection of two sets whose minimum value is equal to m. The probability that X contains a minimum value m has increased with the test from

$$p_V^{(m)}(x,0) = [\frac{f(m,x)}{f(m,v)}](r-m+1)^{v-x} \qquad \text{to}$$

$$\frac{f(m,x)(r-m+1)^{v-x+y}}{(r-m+1)^{v-x}(r-m)^y f(m,x) + (r-m+1)^x f(m,v-x)f(m,y)}$$

after the test, where f(m,x) is given by (2.2) above. Our proposed procedure does not take this latter probability into consideration since it does not keep the set X separated at all times from the set Y (or V\X), rather it mixes the x units with the y units (or v-x units) to get a new set V_m and losses this information.

In Rodriguez-Esquerdo (1983) a specific counterexample is given, where another procedure that takes this information into consideration is presented and shown to achieve the goal in a smaller expected number of tests than the AMTS. The proposed procedure, however, is practical and simple enough to be carried out manually. It is also conjectured that the proposed procedure is optimal in the set of all procedures that allow information to be kept on at most two different sets. Tables for the AMTS and the other discussed procedures are also given here.

Among the topics that need to be considered further is the finding of an overall optimal policy for achieving our goal. A possible model to follow is Friedman's formulation of the Binomial group testing problem. However, this is computationally expensive and amounts to exhaustive search among all possible alternative procedures. A new approach must be found, perhaps using the relationship between the AMTS and the continuous case procedures. We also need to prove the conjectures that were presented in this paper.

Minimum Expected Number of Tests
r=100

N	H(N) AMTS	$H_n(N)$ Nested	$H_a(N)$ Alternative	$H_0(N)$ OAT
1	1.0000 x=1	1.0000 x=1	1.0000 x=1	1.0000 x=1
2	1.9900 x=1	1.9900 x=1	1.9999 x=2	1.9900 x=1
3	2.3332 x=2	2.6419 x=2	2.6616 x=3	2.9701 x=1
4	2.7447 x=2	2.9800 x=2	3.0000 x=4	3.9404 x=1
5	3.0645 x=3	3.3632 x=3	3.3925 x=5	4.9010 x=1
6	3.3179 x=3	3.6201 x=4	3.6586 x=6	5.8520 x=1
7	3.5135 x=4	3.8130 x=4	3.8521 x=7	6.7935 x=1
8	3.6897 x=4	3.9604 x=4	4.0000 x=8	7.7255 x=1
9	3.8623 x=5	4.1652 x=5	4.2136 x=9	8.6483 x=1
10	4.0211 x=5	4.3297 x=6	4.3867 x=10	9.5618 x=1
11	4.1555 x=6	4.4648 x=7	4.5304 x=11	10.4662 x=1
12	4.2772 x=7	4.5778 x=8	4.6520 x=12	11.3615 x=1
13	4.3862 x=8	4.6816 x=8	4.7566 x=13	12.2479 x=1
14	4.4849 x=8	4.7718 x=8	4.8478 x=14	13.1254 x=1
15	4.5749 x=8	4.8513 x=8	4.9282 x=15	13.9942 x=1

x denotes the number of units to test.

REFERENCES

[1] Friedman, Shimon (1982). The Optimal Solution of the Group Testing Problem. Manuscript.

[2] Garey, M.R. and F.K. Hwang (1974). Isolating a Single Defective Using Group Testing. Journal of the American Statistical Association.

[3] Hwang, F.K. (1974). On Finding a Single Defective in Binomial Group Testing. Journal of the American Statistical Association, Vol. 69, No. 345, pp.146-150.

[4] Hwang, F.K. (1980). Optimal Group Testing procedures in Identifying a Single Defective from a Finite Population. Bulletin of the Institute of Mathematics, Academia Sinica, Vol. 8, No. 2/3, pp. 129, 140.

[5] Kumar, S. and M. Sobel (1971). Finding a Single Defective in Binomial Group Testing. Journal of the American Statistical Association, Vol. 66, No. 334, pp. 824-828.

[6] Pasternack, B., M. Sobel and J. Thomas (1974). Group Testing, Halving Procedures and Binary Search. University of Minnesota, School of Statistical Technical Report No. 227.

[7] Rodríguez Esquerdo, P.J. (1983). Group Testing to Identify the Sample Minimum from a Discrete Uniform Distribution. PhD Thesis, University of California, Santa Barbara.

[8] Sandelius, M. (1961). On an Optimal search procedure. American Mathematical Monthly, Vol. 68, pp. 133-134.

[9] Sobel, M. and P.A. Groll (1959). Group Testing to Eliminate Efficiently all Defectives in a Binomial Sample. The Bell System Journal, Vol. 38, pp. 1179-1252.