GROUP-TESTING WITH A NEW GOAL: ESTIMATION*

by

Milton Sobel, University of Minnesota,
Minneapolis, Minnesota

and

R. Elashoff, University of California Medical
School, San Francisco, California

Technical Report No. 165

December 1971

University of Minnesota
Minneapolis, Minnesota

*This paper was supported partially by NSF grant GP-28922X at the University of Minnesota and partially by the National Cancer Institute Contract NIH-71-2180 at the University of California in San Francisco.

1. Introductory Remarks.

The ideas of group-testing have heretofore mainly been used for classifying units into disjoint categories (such as good and defective). In this paper we apply these ideas to the problem of estimation. This paper was motivated by the appearance of a manuscript on this topic by C. Cannings [1] and the simultaneous appearance of applications of these ideas in the field of public health. A previous paper by K. Thompson [4] on this topic was subsequently brought to our attention. Our treatment is asymptotic (the number of tests $\to \infty$) as in [4] but brings into account the cost of the test versus the cost of units; the case treated in [4] corresponds to zero cost for each unit. In this sense the present paper generalizes the results in [4].

In both [4] and the present paper there is an apparently fundamental difficulty in that we use tables based on the value of p (the probability of a unit being defective), although p is unknown and the underlying goal is to estimate p. We shall assume that p is (somehow) estimated at the outset and after every test (say, by maximum likelihood or by an a posterior distribution) and that the convergence to the true value will eventually take place. Since our treatment is asymptotic in the number of tests t this 'fundamental difficulty' will not affect the results of this paper, provided the convergence mentioned above takes place with probability one. From the asymptotic point of view the need for a precise initial or early estimate of p is not essential since this contributes a negligible amount to the overall cost when $t \to \infty$. The study of explicit procedures for approaching the asymptotic situation has not been carried out in this paper. Corresponding small sample results will be published in a separate paper (see also [1] and [4] in this regard). Our criteria (cf. (3.8) below) is to minimize the total cost of testing per unit accuracy obtained in the estimate of p. The two costs considered are the cost C_u of obtaining and preparing each unit for the group-test and the cost C_T of conducting a single group-test. The ratio of these costs (C_u/C_T) is denoted by r and our procedure depends only on r, i.e., we can take C_T to be one. The case $r = 0$ corresponds to that treated in [2], where costs are ignored.

In [4] there is much discussion on the practical difficulties of applying asymptotic results (with the above r equal to zero) to cases in which the number of tests (his n) or the number of units (his m) are limited to small values. Obviously similar cautionings are applicable to our paper. If the above criterion is not the one desired or if the number of tests (or units) is constrained or if the binomial model is questionable for large numbers of units per test (the so-called dilution effect), then an alternative approach may be desirable. Furthermore since our procedure (like that of [4])is one of asymptotic sequential estimation it would be desirable to have some small sample studies on the approach to the asymptotic situation. Further comparisons with [4] will be made below and we also refer the reader to the references in [4] for more related problems.

It is also of interest to note that the value of $q = 1 - p$ that separates the one-at-a-time region (where we use batches of size 1) from its compliment (where we use initial batches of size 2 or more) in the case of no breakdown into subsets as in [4] is as low as $q_1 = 1/3$ for $r = 0$ and increases to only .6 as r increases to 1; its value being

$(1+2r)/(3+2r)$ for any r. This compares with the so-called "cut-off point" in group testing $q_1 = (\sqrt{5} - 1)/2 = .618\ldots$ which also separates these two corresponding strategies in the classification or search problem. It is also of interest that if we allow a single breakdown into subsets when the initial batch size $A = 2$, then the above comparison leads to the familiar dividing point $q_1 = .618\ldots$ for all values of r. Since we use batches of size two or more in estimation for the same range or a wider range of q-values and for most of the r-values of practical interest, we can conclude that group-testing methods and ideas are even more useful for estimation than for classification or search problems.

In one application that arose, rats were collected from the harbor of a large city and, after being killed, dissected, etc., they were carefully examined under the microscope for the presence of some specific type of bacteria. To save costs the same procedure could be carried out by combining small portions of the viscera (say, the liver)of different animal specimens. The goal was to estimate the proportion of rats in this locale that carried this particular germ. If the cost of catching rats and preparing them is negligible compared to the cost of the microscopic search (as was stated by the experimenter in this application) then we would set $r = 0$. In this case the number of rats and the number of tests were large enough to be taken as infinite. In this application we assumed no errors on the part of the technician and no errors due to the dilution effect, i.e., due to combining portions of the viscera of several different animal specimens. Other applications are discussed in [4].

2. Nature of the Problem and the Proposed Solution.

An unlimited number of units are available and we intend to make a large number of tests. The units are all from a common binomial population with unknown probability p of being defective (or having a specific disease D) and $q = 1 - p$ of being good (or free of the disease D). The basic characteristic of group-test sampling is available as a strategy as in [2] and [3], i.e., we can form batches of any size $A \geq 1$ and test the whole batch simultaneously with only the two possible results:

i) all the A units are good, or

ii) at least one of the A units is defective. (We don't know which one(s) or how many.)

For $r = 0$ the cost of a new unit is negligible compared to the cost of a test and in this case it is intuitive that we should never break down any defective batches since we can do 'better' by going to a new batch. (Cf. Remark in Section 3). We therefore have to define what we mean by better and try to find an efficient procedure for any $r \geq 0$. It should be remembered that the optimal initial value of A will depend on the unknown p (cf. Table II) and that we have in mind a sequential procedure that estimates p as we go along.

In an asymptotic analysis a crucial initial consideration is to decide whether we want to make comparisons between procedures on the basis of the same fixed total number of units (u_0) tested or on the basis of the same fixed total number of tests (T_0). If the test cost is zero then we test each unit separately; if the cost of each unit is zero (corresponding to $r = 0$) then we use group-testing methods. By bringing in these two costs into the formulation we avoid both extremes and make both the problem and the answers more realistic. We consider the asymptotic mean square error (MSE) of our estimate on a per test basis multiplied by the total cost of both the testing and the units; we interpret this as a cost per unit information in the Fisherian sense (or as a cost per unit accuracy) and try to minimize this product V over a class of procedures which we now define. Let $R(s)$ depending also on certain constants A, $A_1^{(1)}$, $A_{11}^{(2)}$, $A_{10}^{(2)}$, $A_{111}^{(3)}$, $A_{110}^{(3)}$, $A_{101}^{(3)}$, $A_{100}^{(3)}, \ldots, A_{00\ldots0}^{(s)}$ denote the procedure in which we are prepared to break down (or partition into 2 sets) the batch of original size A at most s times. For example, if A is defective we test a subset of A of size $A_1^{(1)}$; if that is bad we take a subset of it of size $A_{11}^{(2)}$ and if it is good we take a subset of size $A_{10}^{(2)}$ from the remainder set of size $A - A_1^{(1)}$, etc. Let $R_H(s)$ $(s = 0,1,2,\ldots; \infty)$ denote the particular family of such procedures in which the subset size is always exactly half when s is finite and the largest integer $\leq$ half when $s = \infty$; in this family with s finite we assume that A is a multiple of 2^s, without any serious loss of efficiency. In fact, we contend that the halving-procedure family $R_H(s)$ is both practical to use and efficient in the sense of being close to optimal. The values of A and s have to be determined as a function of q and the ratio r of the costs discussed above. For some cases we shall also compute the optimal procedure in $R(s)$ and compare the results with that obtained from the best procedure in the halving family $R_H(s)$.

3. Details of Procedure $R_H(s)$.

We state (without proof) and make use of the following fundamental lemma that appears as Lemma 1 in [2] and also in several other papers on group-testing. A defective set is a set that contains at least one defective unit. If a subset of size A_1 is taken from a defective set of size A and is also defective, then the remainder set of size $A - A_1$ is binomially distributed as at the outset. If this set of size $A - A_1$ is put back into the stockpile of binomial units for future sampling then we refer to this as recombination. (If the stockpile has a queue interpretation then we put these units at the frontof the queue if we want our procedure to have the so-called "first come-first served" property; cf. [3].) The procedures $R_H(s)$ and $R(s)$ are all recombination procedures.

For the procedure $R_H(s)$ with finite s we refer to the batch of size $A/2^s = a$ as a unit batch or U-batch; a batch of size $2a$ is also called two U-batches. When we have a defective set of size greater than a we refer to a G-situation; when we have no defective set or one of size a then we say we are in an H-situation.

The expected total number of U-batches classified between two successive H-situations is given for finite s by

$$(3.1) \qquad E\{B_U|(H, H)\} = P \sum_{j=0}^{S} jQ^{j-1} + SQ^S = \frac{1 - Q^S}{P} = \frac{1 - q^A}{1 - q^a} ,$$

where $Q = q^a$, $P = 1 - Q$ and $S = 2^s = A/a$. The expected number of tests between two successive H-situations is given for finite s by

$$(3.2) \qquad E\{T|(H, H)\} = 1 \cdot q^A + (s+1)(1-q^A) = 1 + s(1-q^A).$$

Moreover, using the independence of successive intervals between H-situations, the product of T_0 and the reciprocal of the result in (3.2) represents the number of such (H, H) intervals we can expect within a fixed large number of tests, T_0. It follows that the asymptotic result for the expected number of U-batches in a fixed large number T_0 of tests is given for finite s by

$$(3.3) \qquad E\{B_U|T_0\} = \frac{(1-q^A)T_0}{(1-q^a)[1 + s(1-q^A)]} ,$$

i.e., by T_0 times the ratio of the results in (3.1) and (3.2). Moreover the expected number of units (or unit batches) per test is a/T_0 (or $1/T_0$) times the result in (3.3). If each test has unit cost and r is the ratio of the cost per unit to the cost per test, then using (3.3) the expected total cost W of T_0 tests is given for finite s by

$$(3.4) \qquad W = T_0 \left\{ 1 + \frac{ra(1-q^A)}{(1-q^a)[1 + s(1-q^A)]} \right\} .$$

Let Z denote the proportion of classified U-batches that are good. For $s = \infty$ we examine every unit and each unit is then a U-batch. Our estimate $\hat{p}$ of p for any s under $R_H(s)$ is the maximum likelihood (m.l.) estimate

$$(3.5) \qquad \hat{p} = 1 - Z^{1/a}.$$

To determine the MSE of $\hat{p}$, we expand $\hat{p}$ in (3.5) by a Taylor expansion about $Z = Q$ and obtain for a large number of tests and for finite s

$$(3.6) \qquad 1 - Z^{1/a} = p - \frac{q^{1-a}}{a} (Z-Q) + \frac{(a-1)q^{1-2a}}{a^2} (Z-Q)^2 + \ldots .$$

The asymptotic expected value of the left member of (3.6) is p and the asymptotic variance based on T_0 tests (for large T_0) and using (3.3) again is for finite s

$$(3.7) \qquad E(1-Z^{1/a}-p)^2 \sim \frac{q^{2-2a}}{a^2} E(Z-Q)^2 = \left(\frac{q^{2-2a}}{a^2}\right) \frac{q^a(1-q^a)}{E\{B_U|T_0\}}$$

$$= \frac{q^{2-a}(1-q^a)^2}{T_0 a^2(1-q^A)} [1 + s(1-q^A)].$$

It is easily seen from (3.6) that the square of the asymptotic bias involves ${T_0}^2$ in the denominator and hence for $T_0 \to \infty$ is of smaller order of magnitude than the asymptotic variance given in (3.7). Thus (3.7) is also the asymptotic MSE of $\hat{p}$.

The product $V = V(A, s)$ of (3.7) and the total cost W in (3.4) gives the asymptotic cost per unit information in the Fisherian sense (or cost per unit accuracy) attained in the estimation of $\hat{p}$ and is independent of T_0, i.e.,

$$(3.8) \qquad V(A, s) = \frac{q^{2-a}(1-q^a)}{a^2(1-q^A)} \{(1-q^a)[1 + s(1-q^A)] + ra(1-q^A)\}.$$

For fixed r and q we want to find the pair (A, s) that minimizes V; recall that $a = A/2^s$.

It can be shown mathematically that for the special case $r = 0$ and any fixed q we can always find a smaller minimum in (3.8) with $s = 0$ than we can with $s \geq 1$. This is apparent in Table II below but has not been shown analytically for all s. We now show this result for $s = 1$ by letting $q^a = y$ in (3.8) and minimizing $V(A, s)/(q \log_e q)^2$ for both $s = 0$ and $s = 1$. For $s = 0$ we obtain a minimum at $y_0 = .2032$ and the minimum value $V(0)$ of $V(A, 0)$ is

$$(3.9) \qquad V(0) = \frac{(q \log q)^2(1-y_0)}{y_0(\log y_0)^2} = 1.5426(q \log_e q)^2.$$

For $s = 1$ we obtain a minimum at $y_0 = .36543$ and the minimum value $V(1)$ of $V(A, 1)$ is

$$(3.10) \qquad V(1) = \frac{(q \log q)^2(1-y_0)(2-y_0{}^2)}{y_0(1+y_0)(\log y_0)^2} = 2.3425(q \log_e q)^2.$$

(For $s = 0$ the value A_0 of A used is $-1.5936/\log_e q$ and for $s = 1$ the value A_1 of A used is $2(\log .365343)/\log q = -2.0138/\log_e q$.) If we disregard the problems arising from the fact that A_1 has to be an even integer and A_0 has to be an integer, then it follows from (3.9) and (3.10) that for $r = 0$ and any q we can get a smaller minimum for $s = 0$ than for $s = 1$ and indeed with a smaller value of A, i.e., $A_0 < A_1$. The above integer constraints disappear if we randomize between two successive integers (or between successive even integers for $s = 1$) to get A exactly equal to the desired value.

To calculate the A-values in Table IIA we first obtained approximations by differentiating (3.8) with respect to A for fixed s and setting the result equal to zero. Letting $I = 1 - q^A$ and $J = 1 - q^a$, we can write the result $A = A_s(r, q)$ for finite s (after simplification) as

$$(3.11) \qquad A = \frac{2^s IJ[2J(1+sI) + raI]}{(-\log_e q)[IJ(2-J)(1+sI) - 2^s(1-I)J^2 + raI^2]} \; ;$$

we omit the algebraic details. This method has the difficulty that the right side of (3.11) also depends on A, so that an iteration of (3.11) and a check on the convergence of this iteration is needed.

By doing this simultaneously for fixed q and different s-values we found the cross-over points on the r-axis. For example, $s = 0$ for $r \leq 1/242$ and $s \geq 1$ for $r \geq 1/241$. Indeed it appears from Table IIA that the minimizing s-value for fixed q is a non-decreasing function of r, although this has not been proved.

These A-values and also missing A-values (due to lack of convergence) were then confirmed and/or corrected and/or obtained by a search algorithm using (3.8) and the fact (which follows from the unique result in (3.11)) that there is a single minimum of (3.8) with respect to A, where the derivative is zero.

Other formulas for A appear below in (5.5) for $s = \infty$ and in (4.2) for the special case $r = s = 0$.

It should be pointed out that since p is not known we will be using a variety of batch sizes until the estimator converges. Then, if we use all the information at hand, the m.l. estimator is the root of a polynomial in q as described in Section X of [2] and also in Section 7 below. Thus (3.7) and the variances in Table IIB are only approximate for small samples.

4. Discussion of the Special Case $r = 0$ and Comments on [4].

As in the previous section the method used in [4] for finding A-values for $r = s = 0$ is differentiation. The differentiation method gives the correct A-values for most (but not all) values of q. A more exact analysis is to use differences to find the exact dividing points and these in turn give the exact A-values in every case. In this section we also develop an iterative scheme to get th$\overset{e}{\text{i}}\overset{e}{\text{s}}$ dividing points and the results are given in Tables IA and IB. Finally some other comments on [4] are included.

We can use differences to get the asymptotic $(T_0 \to \infty)$ answers for A when $r = s = 0$ by solving for the q-value which separates any given A from $A + 1$; this q-value (or so-called dividing point) is the root of

$$(4.1) \qquad \frac{q(1-q^A)}{A^2} = \frac{1 - q^{A+1}}{(A+1)^2} \quad .$$

The answer by differentiation, as mentioned in section 3, is

$$(4.2) \qquad A = -b/\log_e q$$

where b is the unique root of $y = 2(1-e^{-y})$; this constant b is incorrectly given after (9) in [4]; its value to 9 decimals is easily shown to be 1.593624260. Corresponding approximations q' to the exact dividing points q in (4.1) can be obtained by computing $q' = \exp\{-b/(A+.5)\}$. Then $q' - q$ converges to zero quite rapidly as $A \to \infty$. Thus for $A > 28$ (or $q > .946$) the difference $q' - q$ doesn't even show up in the fifth decimal place. Table I gives values of q and q' for increasing A and it also gives A-values for some selected values of q. It should perhaps be mentioned that the author in [4] fails to note that (4.2) above is the exact answer to his equation (8) and he only gives the approximation $A = b/p$ in his equation (10). The last three entries under (8) in his Table 4 have also been corrected in our Tables I and II.

Another useful form of (4.1) is

$$(4.3)\qquad \frac{(A+1)^2}{2A+1} = \frac{1-q^{A+1}}{1-q} = \frac{1-(1-\varepsilon)^{A+1}}{\varepsilon} = A + 1 - \binom{A+1}{2}\varepsilon + \ldots,$$

where we have set $q = 1 - \varepsilon$ (with $\varepsilon > 0$ small) in order to consider the case of q close to one. By dropping q^{A+1} in (4.3) we get a lower bound LB for q (the dividing point between A and $A + 1$) and by dropping terms of order ε^2, ε^3, etc., in (4.3) we get an upper bound UB; these are

$$(4.4)\qquad \text{LB} = \left(\frac{A}{A+1}\right)^2 < q < \frac{2A-1}{2A+1} = \text{UB}.$$

We now develop an iterative scheme for solving (4.1) which was used to compute the exact q-values in Table IA. Solving the two left members of (4.3) for q, we write the iterative scheme as

$$(4.5)\qquad q_{i+1} = \frac{(2A+1)q_i^{A+1} + A^2}{(A+1)^2} \quad (i = 0,1,\ldots),$$

where q_0 is some initial estimate of the root, say an average of the two bounds in (4.4). Although the two bounds in (4.4) both approach 1, their difference goes to zero like $1/A$ and hence does not determine the exact root very quickly; hence the need for the iterative scheme in (4.5).

5. The Procedures $R_H(s)$ for $s = \infty$ and R_{04}.

In this section we consider the halving procedure R_{04} that was introduced in section 6 of [3]. This procedure which classifies every unit as good or defective is the one we use for $s = \infty$. New, useful approximations are obtained here for R_{04} that are valuable to help define the family $R_H(s)$ for $s = \infty$ and make it more explicit. We plan to use R_{04} for $R_H(\infty)$ whenever it gives a result for W in (3.4) that is as good or better than that of $R_H(s)$ for any finite s. In applying $R_H(s)$ for $s = \infty$ it should be noted that we do not assume that the batch sizes are even integers. Table I in [3] gives results for R_{04} for $q = .90$, $.95$ and $.99$ that we use; furthermore it shows numerically how far R_{04} is from an optimal procedure (see R_{01} and R_{21} in the same table) that classifies every unit as good or defective.

In order to get further numerical results for R_{04} we return to the criterion that was used for it, namely to maximize the number of units classified per test in a fixed large number T_0 of tests or, equivalently, to minimize the expected number of tests per unit classified in a fixed large number T_0 of tests. We proceed to consider what happens between two successive H-situations as in section 3 above, considering only 'nested' procedures that continue to search for a single defective in smaller and smaller batches.

The expected number of units classified in an (H, H) interval for any nested procedure R that starts with A units in the H-situation is given by

$$(5.1) \qquad E\{U|(H, H)\} = Aq^A + p\sum_{j=1}^{A} jq^{j-1} = \frac{1 - q^A}{p} .$$

The expected number of tests in any (H, H) interval for our halving procedure R_{04} is approximately given by

$$(5.2) \qquad E\{T|(H, H)\} \sim q^A \cdot 1 + (1-q^A)(1 + \log_2 A) = 1 + (1-q^A)\log_2 A,$$

where A is the starting batch size for the H-situation. This result is exact if A is a power of 2 and is a good approximation otherwise. The closeness of our approximation will be seen below by comparing our results with those in Table I of [3] that were computed by exact recursion methods (cf. Appendix C in [2]).

Using the independence of the (H, H) intervals the reciprocal of the product of (5.2) and T_0 is the approximate number of (H, H) intervals in a fixed large number T_0 of tests. Also T_0 times (5.1) is the total expected number of units classified in these tests. Hence we wish to maximize the ratio of (5.1) to (5.2) or minimize its inverse

$$(5.3) \qquad C(q; R_{04}) = \frac{p}{1 - q^A} + p \log_2 A$$

with respect to (the integer) A; we then find the minimum value of $C(q; R_{04})$ or the maximum value of its inverse. Straightforward differentiation gives a minimum at the root of the transcendental equation

$$(5.4) \qquad -Q(\log_e Q)\log_e 2 = (1-Q)^2,$$

where $Q = q^A$. Thus we find that the continuous solution for A is

$$(5.5) \qquad A = \frac{\log Q}{\log q} = \frac{\log .51276}{\log q} = \frac{.66794}{-\log_e q}$$

and we use the integer closest to this solution. For $q = .90$, $.95$, and $.99$ this gives $A = 6$, 13 and 67 respectively as compared to the exact results 7, 15 and 65 from table I of [3]. For $q = .995$ and $.999$

we obtain from (5.5) $A = 133$ and 668, respectively. The A-values for any given q can also be obtained by solving for the dividing point (say $q_{A,A+1}$) that separates A from $A + 1$ in (5.3); for each $A(A = 1,2,\ldots)$ it is the root of

$$(5.6) \qquad \frac{1}{1 - q^{A}} - \frac{1}{1 - q^{A+1}} = \log_2\left(\frac{A+1}{A}\right) .$$

If the actual value of q lies between $q_{A-1,A}$ and $q_{A,A+1}$, then we use A as the binomial batch size. This method of determining A does not change any of the above five values.

The minimum values of $C(q; R_{04})$ for these five values of q (namely, .90, .95, .99, .995, and .999) are .4719, (.4725); .2878, (.2885); .0811, (.0811); .0456 and .0114 respectively, where the numbers in parentheses are the values from table I of [3] for $q = .90, .95$ and .99. The value of $C\ (q; R_{04})$ can be interpreted as the average number of tests per unit classified.Using (5.6) with $A = 1$ we note that $A = 1$ is to be used for any $q < (\sqrt{5} - 1)/2 = .618\ldots$; the latter is known (cf. [2], [3] and references therein) to be the optimal dividing point between $A = 0$ and $A = 1$ in search problems.

The corresponding value of $V(A, \infty)$ for a fixed large number T_0 of tests under R_{04} is

$$(5.7) \qquad V(A, \infty) = \frac{pq\ E\{T|(H,\ H)\}}{T_0\ E\{U|(H,\ H)\}}\ T_0\left[1 + r\ \frac{E\{U|(H,\ H)\}}{E\{T|(H,\ H)\}}\ \right] = pq[r + \ C(q; R_{04})],$$

where the value of $C(q; R_{04})$ is given above for 5 values of q. As already mentioned,we use $R_{04} = R_H(\infty)$ whenever the value of (5.7) is smaller than that of $R_H(s)$ in (3.8) for a finite value of s.

It has been noted (but not proved) that this occurs only in an interval of the form $s \geq s_0$. In addition we note that for r sufficiently large the value of (5.7) is smaller than (3.8) for any finite value of s since we have for any q and any $A \geq 1$

$$(5.8) \qquad V(A, \infty) \approx rpq < \frac{r(1-q^a)}{aq^{a-2}} \approx V(A, s),$$

and the inequality in (5.8) is easily proved. Thus for any given q we will classify every unit if r is sufficiently large.

The procedure $R_{04} = R_H(\infty)$ is thus seen to be compatible with our family $R_H(s)$ and it gives us a reasonable criterion for deciding when to classify each and every one of our units.

6. The r-Values that Separate s from $s + 1$ in Table II.

In this section we wish to show that for any given q the r-value $r(q)$ that separates (say) the $s = 0$ region from the $s = 1$ region is given by

$$(6.1) \qquad r_{0,1}(q) = C_{0,1} \log_e q$$

where $C_{0,1}$ does not depend on q. With a set of such constants $C_{s,s+1}$ the need for an extensive table like Table II becomes questionable. However we do not have these constants and it is felt that Table II does make the procedure more explicit. In any case the existence of such constants (i.e., that do not depend on q) is of considerable interest and we now prove this.

Let $q^a = y$ and $r' = r/\log_e q$ in (3.8) and let $V_s(y, r')$ denote the function thus obtained; we also write y_s and r'_s to indicate dependence on s. Then for any s we have from (3.8)

$$(6.2) \qquad \frac{V_s(y, r')}{(q \log_e q)^2} = \frac{(1-y)}{(\log_e y)^2 y(1-y^S)} [(1-y)(s+1-sy^S) + r'(1-y^S)(\log_e y)]$$

and we denote the left member of (6.2) by $U_s(y, r')$. Differentiation with respect to y and setting the result equal to zero to find the minimum (and writing r_s' for r') gives for any s

$$(6.3) \qquad \frac{-\log_e y}{(1-y)(1-y^s)} = \frac{2(1-y)(s+1-sy^s) + r_s'(1-y^s)}{(1+y)F(y_s) - sy^s(1-y)^2 + r_s'(\log_e y)(1-y^s)^2} ,$$

where $F(y_s) = (s+1-sy_s^s)(1-y_s^s)(1-y_s)$. Letting y_s denote the value of y at the minimum we can solve (6.3) for $r_s' = r_s'(y_s)$ obtaining the relation

$$(6.4) \qquad r_s' = \frac{sy_s^s(1-y_s)^2(\log_e y_s) - F(y_s)\{(1+y_s)(\log_e y_s) + 2(1-y_s)\}}{(\log_e y_s)(1-y_s^s)^2(1-y_s+\log_e y_s)}$$

For a common q and r we have a common r' at the cross-over point. Hence we equate the expressions (6.4) for two successive values of s. We do this for $s = 0$ and $s = 1$; this gives a relation between y_0 and y_1. If we substitute (6.4) (for $s = 0$ and $s = 1$) back in (6.2) then we obtain the minimum values $V_0(y_0)$ and $V_1(y_1)$ as functions of y_0 alone and y_1 alone, respectively. At the value of r where cross-over occurs these two minima are equal. Hence we have another relation, and both y_0 and y_1 are thus determined. Then (6.4) determines $r_0' = r_1' = r/\log_e q = r'$ (say) and if we let $c_{0,1}$ denote this value then our desired result is proved.

For the special case $s = 0$ and $s = 1$ we obtain from (6.2)

$$(6.5) \qquad U_0(y, r') = \frac{(1-y)}{y(\log_e y)^2}(1+r'),$$

$$(6.6) \qquad U_1(y, r') = \frac{(1-y)}{(\log_e y)^2 y(1+y)}[2 - y^2 + r'(1+y)].$$

The first relation between y_0 and y_1 from (6.4) is obtained by writing $r'_0 = r'_1$; this gives

$$(6.7) \qquad \frac{2(1-y_0) + \log_e y_0}{(\log_e y_0)(1-y_0+\log y_0)} = \frac{2(1-y_1^2)(2-y_1^2) + (\log_e y_1)(2+4y_1-y_1^2-2y_1^3-y_1^4)}{(\log_e y_1)(1+y_1)^2(1-y_1+\log y_1)} .$$

Putting r'_0 and r'_1 into (6.2), we obtain the second relation as

$$(6.8) \qquad \frac{(1-y_0)^2}{y_0(\log_e y_0)^2(1-y_0+\log_e y_0)} = \frac{(1-y_1)^2\{(1+y_1)(2-y_1^2) + (\log_e y_1)y_1(2+2y_1+y_1^2)\}}{(1+y_1)^2 y_1(\log_e y_1)^2(1-y_1+\log_e y_1)} .$$

It should be pointed out that the values of y_0 and y_1 are also useful in the solution since at the cross-over point we can select either of these two s-values and $q^a = y_0$ yields the A-value for $s = 0$ while $q^a = y_1$ yields the A-value for $s = 1$. (Recall that $a = A/2^s$.)

The constant $C_{0,1}$ can also be found numerically by fixing q and searching for the common r'-value at which the two minima are equal. The result obtained in this way for $C_{0,1}$ and checked by (6.7) and (6.8) is $C_{0,1} = -4.1461$ so that for any q

$$(6.9) \qquad r_{0,1}(q) = -4.1461 \log_e q .$$

Thus for $q = .999$ the cross-over value $r_{0,1}$ is between 1/242 and 1/241.

7. Some Comparisons with the Optimal Family R(s).

In this section we study by means of examples how far the halving family $R_H(s)$ is from being optimal with respect to our criterion (3.8). For $s = 0$ the halving family $R_H(s)$ and the optimal family $R(s)$ are clearly identical so that we need not make comparisons for $s = 0$; hence we start with $s = 1$.

Let A denote the binomial batch size as before and let $B = B(q)$ (with $1 \leq B < A$) denote the size of the subset to be taken from a defective set of size A. We consider an optimal procedure for the special case $s = 1$, so that we have an H-situation again after 1 or at most 2 tests. The expected number of tests T between two successive H-situations is

$$(7.1) \qquad E\{T|(H, H)\} = 1 \cdot q^A + 2(1-q^A) = 2 - q^A.$$

The likelihood associated with the results observed is an (H, H) interval is given by

$$(7.2) \qquad L = (q^A)^\alpha (q^B - q^A)^\beta (1-q^B)^\gamma ,$$

where α, β and γ are each 0 or 1 and their sum is 1. Hence

$$(7.3) \qquad \frac{d \log L}{dq} = \frac{\alpha A}{q} + \frac{\beta(Bq^{B-1} - Aq^{A-1})}{q^B - q^A} - \frac{\gamma Bq^{B-1}}{1 - q^B} .$$

Treating α, β, γ as the values from a trinomial distribution with probabilities indicated in (7.2), we find the variance of the maximum likelihood estimate by first computing the inverse of the variance of $\hat{q}$ using the Fisher information method (as was done in Section X of [2]). Using the independence of the (H, H) intervals we obtain

$$(7.4) \qquad (\sigma_{\hat{q}}^2)^{-1} = E\{M(\frac{d \log L}{dq})^2\} = (EM)E\{(\frac{d \log L}{dq})^2\},$$

where M is the number of such (H, H) intervals in a fixed large number T_0 of tests. After much simplification the last factor in (7.4) reduces to

$$(7.5) \qquad E\{(\frac{d \log L}{dq})^2\} = \frac{q^{B-2}\{(A-B)^2 q^A(1-q^B) + B^2(q^B - q^A)\}}{(1-q^B)(q^B - q^A)} ;$$

we omit the algebraic details. From (7.1) and the independence of the

(H, H) intervals, we find that in a fixed large number T_0 of tests we have asymptotically

$$(7.6) \qquad E\{M\} = \frac{T_0}{E\{T|(H, H)\}} = \frac{T_0}{2 - q^A} .$$

To compute the cost of units on a per test basis we say that a unit is charged to a test if (as a result of that test) the procedure does not use that unit in any further tests; the unit need not have been used in the test it is charged to. The only purpose of this definition is to avoid duplicate charges for the same unit. The expected number U of units charged (to any test) in an (H, H) interval is

$$(7.7) \qquad E\{U|(H, H)\} = Aq^A + A(q^B - q^A) + B(1-q^B) = B + (A-B)q^B;$$

this is correct since we use A units if and only if the first B are all good. Hence the expected total cost W of T_0 tests, analogous to (3.4) above, is

$$(7.8) \qquad W = T_0\{1 + \frac{r[B + (A-B)q^B]}{2 - q^A}\} .$$

Combining (7.4), (7.5), (7.6) and (7.8) to form the expected total cost per unit information $V_1^*(A, B, r)$ (as was done in (3.8) for $R_H(s)$), we obtain

$$(7.9) \qquad V_1^*(A, B, r) = \frac{(1-q^B)(q^B - q^A)\{2-q^A + r[B + (A-B)q^B]\}}{q^{B-2}[(A-B)^2q^A(1-q^B) + B^2(q^B - q^A)]} .$$

For given r and any fixed q we wish to find the ordered pair (B, A) with $1 \leq B < A$ which minimizes the value of $V_1^*(A, B, r)$ in (7.9). We consider in Table III several examples in which $s = 1$ is optimal for the $R_H(s)$ family and $s = 1$ also appears to be optimal for the $R(s)$ family. It should be noted that the results (especially the A-value) for the halving procedure $R_H(s)$ can also be used as a good first approximation to the corresponding A-value for the optimal procedure $R(s)$. Hence if we had V_s-formulas as in (7.9) for $s > 2$, it would not be difficult to find the optimal batch sizes by a search algorithm; this has not been done because the V_s-formulas for $s > 2$ appear to be too unwieldy.

We note in Table III that the optimal A-value is close to but not necessarily above or below that of the halving procedure. On the other hand, the optimal B-value appears to be always less than half (and greater than 40 percent) of the optimal A-value. The V-criterion appears to be unaffected in the fifth decimal in all of the cases above if we replace the optimal by the halving procedure. On the other hand, if we test each unit separately (call this procedure R^0) then the cost per unit information for R^0 is

$$V(R^0) = pq(1+r) \tag{7.10}$$

and this represents an increased cost per unit information by a factor of more than 100 in the first row of Table III down to an increase of 40 percent in the last row.

In the optimal case the form of the maximum likelihood (m.l.) estimator $\hat{p}$ (or $\hat{q}$) in terms of B and A and the observed frequencies X, Y and Z for the three factors, respectively, in the trinomial (7.2) for a fixed large number T_0 of tests will now be explicitly derived. In fact we find it useful to derive it by two different methods, thus adding to the desirability of using the m.l. estimator in this application.

In (7.3) we set the result equal to zero, multiply by q and replace X by using the relation

$$(7.11) \qquad X + Y + Z = \frac{T_0}{2 - q^A} ,$$

which follows from (7.6) and the definitions of X, Y, Z. This gives the result

$$(7.12) \qquad \frac{AT_0}{2 - q^A} = \frac{Y(A-B)}{1 - q^{A-B}} + \frac{Z\{Bq^B + A(1-q^B)\}}{1 - q^B} ,$$

and $\hat{q}$ is the root of this equation if Y and Z are not both zero; if $Y = Z = 0$ we take $\hat{q} = 1$. This root must exist since for $q = 0$ the left side is of the form $A(X + Y + Z)$ and the right side is of the form $Y(A-B) + ZA < (Y+Z)A$ and this is smaller than the left side; for $q = 1$ the right side $\to \infty$ if either Y or $Z \neq 0$. Hence at least one root exists if Y and Z are not both zero.

The other method of derivation is to set $\hat{p}$ (the notation will be justified) equal to the proportion of defective units that we expect to have in our defective sets in relation to the total number of units used. The total number of units used is $A(X+Y) + BZ$ and hence, using (7.11), we write

$$(7.13) \qquad \hat{p} = \frac{\dfrac{Y(A-B)\hat{p}}{1 - \hat{q}^{A-B}} + \dfrac{ZB\hat{p}}{1 - \hat{q}^B}}{A(\dfrac{T_0}{2-\hat{q}^A} - Z) + BZ} .$$

Here we used the (easily-shown) fact that for a defective set of size S, i.e., one containing at least one defective unit, the expected number of defective units in it is $Sp/(1-q^S)$. If we cancel the $\hat{p}$ in (7.13) and consider the root $\hat{q}$ of the resulting polynomial it is easy to see that this gives exactly the same polynomial as was obtained in (7.12). This justifies the notation $\hat{q} = 1 - \hat{p}$ and makes the m.l. estimate a natural one to use in such problems.

8. Acknowledgement.

The authors wish to thank Ms. Elaine Frankowski for her assistance with the programming of most of the tables in this paper.

TABLE IA

Exact Dividing Points (q-Values) Between A and A+1

for the special case r = 0

A	LB	Exact q-Value	q'	UB
1	.2500	.3333	.3456	.3333
2	.4444	.5247	.5286	.6000
3	.5625	.6325	.6342	.7143
4	.6400	.7009	.7018	.7778
5	.6944	.7479	.7485	.8182
6	.7347	.7822	.7826	.8462
7	.7656	.8084	.8086	.8667
8	.7901	.8289	.8290	.8824
9	.8100	.8455	.8456	.8947
10	.8264	.8591	.8592	.9048
20	.9070	.9252	.9252	.9512
30	.9365	.9491	.9491	.9672
40	.9518	.9614	.9614	.9753
50	.9612	.9689	.9689	.9802
100	.9803	.9843	.9843	.9901
200	.9901	.9921	.9921	.9950
300	.9934	.9947	.9947	.9967
400	.9950	.9960	.9960	.9975
500	.9960	.9968	.9968	.9980
1000	.9980	.9984	.9984	.9990
2000	.9990	.9992	.9992	.9995

TABLE IB

Exact Values of A for Selected Values of q

q	A	q	A
.35	2	.90	15
.50	2	.95	31
.60	3	.99	159
.70	4	.995	318
.80	7	.999	1593

TABLE IIA

Asymptotic ($T_0 \to \infty$) Halving Procedures for Estimating p

(or q = 1-p) as a function of the ratio r of
the cost[#] of one unit to the cost of one test.

(A is the original binomial batch size, s is the maximum number of times we split the A units into halves, T_0 is the fixed but large number of tests and V is the cost per unit information criterion (3.8).)

r↓ \ q→	.5	.6	.7	.8	.9	.95	.99	.995	.999
.0000	2(0)[*]	3(0)	4(0)	7(0)	15(0)	31(0)	159(0)	318(0)	1593(0)
.0010	2(0)	3(0)	4(0)	7(0)	15(0)	30(0)	142(0)	261(0)	910(0)
.0020	2(0)	3(0)	4(0)	7(0)	15(0)	30(0)	130(0)	230(0)	727(0)
.0025	2(0)	3(0)	4(0)	7(0)	15(0)	29(0)	126(0)	219(0)	672(0)
.0040	2(0)	3(0)	4(0)	7(0)	14(0)	28(0)	115(0)	194(0)	563(0)
.0050	2(0)	3(0)	4(0)	7(0)	14(0)	28(0)	109(0)	182(0)	752(1)
1/150	2(0)	3(0)	4(0)	7(0)	14(0)	27(0)	102(0)	166(0)	668(1)
.0100	2(0)	3(0)	4(0)	7(0)	14(0)	26(0)	91(0)	145(0)	562(1)
.0200	2(0)	3(0)	4(0)	6(0)	13(0)	23(0)	72(0)	112(0)	592(2)
.0250	2(0)	3(0)	4(0)	6(0)	12(0)	21(0)	67(0)	150(1)	752(3)
.0400	2(0)	3(0)	4(0)	6(0)	11(0)	19(0)	56(0)	124(1)	600(3)
.0500	2(0)	3(0)	4(0)	6(0)	11(0)	18(0)	76(1)	112(1)	768(4)
.1000	2(0)	3(0)	3(0)	5(0)	9(0)	14(0)	56(1)	120(2)	768(5)
.2000	2(0)	2(0)	3(0)	4(0)	7(0)	11(0)	60(2)	120(3)	768(6)
.2500	2(0)	2(0)	3(0)	4(0)	6(0)	14(1)	72(3)	160(4)	704(6)
.5000	1(∞)[§]	2(0)	2(0)	3(0)	8(1)	12(1)	80(4)	160(5)	640(7)
1.0000	1(∞)	2(0)	2(0)	4(1)	6(1)	12(2)	64(5)	128(6)	768(8)
2.0000	1(∞)	1(∞)	2(∞)	3(∞)	7(∞)	15(∞)	65(∞)	132(∞)	668(∞)
3.0000	1(∞)	1(∞)	2(∞)	3(∞)	7(∞)	15(∞)	65(∞)	132(∞)	668(∞)
4.0000	1(∞)	1(∞)	2(∞)	3(∞)	7(∞)	15(∞)	65(∞)	132(∞)	668(∞)
5.0000	1(∞)	1(∞)	2(∞)	3(∞)	7(∞)	15(∞)	65(∞)	132(∞)	668(∞)
∞	1(∞)	1(∞)	2(∞)	3(∞)	7(∞)	15(∞)	65(∞)	132(∞)	668(∞)

[*] In each cell the value of A is given first and the value of s is in parentheses.

[#] This cost includes that of obtaining and processing a unit for testing.

[§] The results for $s = 0$ and $s = \infty$ coincide when $A = 1$ and indeed we replace s by ∞ whenever $A = 2^s$ since we then classify every unit.

TABLE IIB

Asymptotic ($T_0 \to \infty$) Halving Procedures for Estimating p

(The first entry# in each cell is the minimum cost per unit information attained in the corresponding cell of Table IIA and the second entry§, when divided by the number of tests T_0, gives the variance of the m.l. estimate.)

r↓ q→	.5	.9	.95	.99	.995	.999
.0000	1.875 E*-1 1.875 E-1	1.389 E-2 1.389 E-2	3.667 E-3 3.667 E-3	1.529 E-4 1.529 E-4	3.841 E-5 3.841 E-5	1.543 E-6 1.543 E-6
.0010	1.879 E-1 1.875 E-1	1.409 E-2 1.389 E-2	3.779 E-3 3.669 E-3	1.758 E-4 1.539 E-4	4.948 E-5 3.924 E-5	3.419 E-6 1.790 E-6
.0020	1.883 E-1 1.875 E-1	1.431 E-2 1.389 E-2	3.889 E-3 3.669 E-3	1.968 E-4 1.562 E-4	5.922 E-5 4.056 E-5	4.956 E-6 2.020 E-6
.0025	1.884 E-1 1.875 E-1	1.441 E-2 1.389 E-2	3.943 E-3 3.677 E-3	2.068 E-4 1.573 E-4	6.381 E-5 4.123 E-5	5.679 E-6 2.119 E-6
.0040	1.890 E-1 1.875 E-1	1.471 E-2 1.393 E-2	4.102 E-3 3.689 E-3	2.355 E-4 1.613 E-4	7.682 E-5 4.326 E-5	7.745 E-6 2.382 E-6
.0050	1.894 E-1 1.875 E-1	1.491 E-2 1.393 E-2	4.206 E-3 3.689 E-3	2.537 E-4 1.642 E-4	8.506 E-5 4.453 E-5	8.984 E-6 2.923 E-6
1/150	1.900 E-1 1.875 E-1	1.523 E-2 1.393 E-2	4.375 E-3 3.707 E-3	2.829 E-4 1.684 E-4	9.825 E-5 4.664 E-5	1.098 E-5 3.077 E-6
.0100	1.913 E-1 1.875 E-1	1.588 E-2 1.393 E-2	4.701 E-3 3.731 E-3	3.381 E-4 1.770 E-4	1.233 E-4 5.031 E-5	1.487 E-5 3.344 E-6
.0200	1.950 E-1 1.875 E-1	1.772 E-2 1.406 E-2	5.613 E-3 3.845 E-3	4.899 E-4 2.008 E-4	1.926 E-4 5.944 E-5	2.577 E-5 4.241 E-6
.0250	1.969 E-1 1.875 E-1	1.858 E-2 1.429 E-2	6.043 E-3 3.963 E-3	5.612 E-4 2.098 E-4	2.234 E-4 7.279 E-5	3.106 E-5 4.890 E-6
.0400	2.025 E-1 1.875 E-1	2.108 E-2 1.464 E-2	7.260 E-3 4.125 E-3	7.651 E-4 2.362 E-4	3.121 E-4 7.925 E-5	4.669 E-5 5.213 E-6
.0500	2.063 E-1 1.875 E-1	2.269 E-2 1.464 E-2	8.032 E-3 4.227 E-3	8.876 E-4 2.878 E-4	3.698 E-4 8.333 E-5	5.700 E-5 5.860 E-6
.1000	2.250 E-1 1.875 E-1	3.004 E-2 1.581 E-2	1.161 E-2 4.837 E-3	1.469 E-3 3.312 E-4	6.405 E-4 1.050 E-4	1.079 E-4 6.858 E-6
.2000	2.625 E-1 1.875 E-1	4.327 E-2 1.803 E-2	1.809 E-2 5.654 E-3	2.544 E-3 4.174 E-4	1.160 E-3 1.297 E-4	2.088 E-4 7.857 E-6
.2500	2.813 E-1 1.875 E-1	4.959 E-2 1.984 E-2	2.102 E-2 7.085 E-3	3.067 E-3 4.896 E-4	1.417 E-3 1.446 E-4	2.590 E-4 7.970 E-6
.5000	3.750 E-1 2.500 E-1	7.822 E-2 2.511 E-2	3.470 E-2 7.600 E-3	5.626 E-3 5.753 E-4	2.682 E-3 1.695 E-4	5.096 E-4 9.106 E-6
1.0000	5.000 E-1 2.500 E-1	1.288 E-1 2.842 E-2	5.998 E-2 9.934 E-3	1.065 E-2 7.037 E-4	5.189 E-3 2.018 E-4	1.010 E-3 9.855 E-6
2.0000	7.500 E-1 2.500 E-1	2.225 E-1 4.253 E-2	1.087 E-1 1.370 E-2	2.060 E-2 8.026 E-4	1.018 E-2 2.268 E-4	2.009 E-3 1.140 E-5
3.0000	1.000 E-1 2.500 E-1	3.125 E-1 4.253 E-2	1.562 E-1 1.370 E-2	3.050 E-2 8.026 E-4	1.515 E-2 2.268 E-4	3.008 E-3 1.140 E-5
4.0000	1.250 E-1 2.500 E-1	5.504 E-1 4.253 E-2	2.037 E-1 1.370 E-2	4.040 E-2 8.026 E-4	2.013 E-2 2.268 E-4	4.007 E-3 1.140 E-5
5.0000	1.500 E-1 2.500 E-1	6.404 E-1 4.253 E-2	2.512 E-1 1.370 E-2	5.030 E-2 8.026 E-4	2.510 E-2 2.268 E-4	5.006 E-3 1.140 E-5
∞	∞ 2.500 E-1	∞ 4.253 E-2	∞ 1.370 E-2	∞ 8.026 E-4	∞ 2.268 E-4	∞ 1.140 E-5

#Based on (3.8) for s finite and on (5.7) for $s = \infty$.

§Based on (3.8) with $r = 0$ for s finite and on (5.7) with $r = 0$ for $s = \infty$.

*E-x denotes 10^{-x}.

TABLE III

Comparison of Halving[§] and Optimal Procedures (s=1)

	q	r	Halving Procedure		Optimal Procedure		
			A	V(A, r)	A	B	$V_1^*(A, B, r)$
1	.999	.005	758	8.9839 E-6	760	348	8.9738 E-6
2	.999	1/150	674	1.0981 E-5	674	311	1.0972 E-5
3	.999	.010	568	1.4874 E-5	566	265	1.4866 E-5
4	.995	.025	150	2.2339 E-4	152	70	2.2315 E-4
5	.995	.040	124	3.1205 E-4	125	58	3.1184 E-4
6	.995	.050	112	3.6978 E-4	113	53	3.6958 E-4
7	.990	.050	76	8.8760 E-4	76	35	8.8658 E-4
8	.990	.100	56	1.4688 E-3	56	26	1.4680 E-3
9	.950	.250	14	2.1008 E-2	15	7	2.0974 E-2
10[#]	.950	.500	12	3.4703 E-2	11	5	3.4660 E-2
11	.900	.500	8	7.8219 E-2	7	3	7.8095 E-2
12	.900	1.000	6	1.2879 E-1	6	3	1.2879 E-1
13	.800	1.000	4	2.6728 E-1	4	2	2.6728 E-2

[§] for all of the 13 cases of Table II where s = 1 is preferred. In each case s = 1 is preferred by both procedures.

[#] Note that the maximum difference of the V-values above occurring in the tenth row is 4.3 E-5. This illustrates the order of magnitude of the additional savings that are possible with more extended computations of the optimal procedure.

BIBLIOGRAPHY

[1] Cannings, C. (1971). The improvement of estimates by the use of pooled samples when the number of tests is limited. Unpublished Report of Laboratorio di Genetica, ed. Evoluzionistica del C.N.R., c/o Instituto di Genetica Dell'Universita' di Pavia, Pavia, Italy.

[2] 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.

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

[4] Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. Biometrics **18** 568-578.