Probability in the Engineering and Informational Sciences, **17**, 2003, 335–350. Printed in the U.S.A.

HYPERGEOMETRIC GROUP TESTING WITH INCOMPLETE INFORMATION

SHAUL K. BAR-LEV
Department of Statistics
University of Haifa
Haifa 31905, Israel
E-mail: barlev@stat.haifa.ac.il

WOLFGANG STADJE
Department of Mathematics and Computer Science
University of Osnabrück
49069 Osnabrück, Germany
E-mail: wolfgang@mathematik.uni-osnabrueck.de

FRANK A. VAN DER DUYN SCHOUTEN
Center for Economic Research
Tilburg University
5000 LE Tilburg, The Netherlands
E-mail: F.A.vdrDuynSchouten@kub.nl

We studied several group testing models with and without processing times. The objective was to choose an optimal group size for pooled screening of a contaminated population so as to collect a prespecified number of good items from it with minimum testing expenditures. The tested groups that were found to be contaminated were used as a new sampling population in later stages of the procedures. Since testing may be time-consuming, we also considered deadlines to be met for the testing process. We derived algorithms and exact results for the underlying distributions, enabling us to find optimal procedures. Several numerical examples are given.

1. INTRODUCTION

The general purpose of group testing is to reduce the number of tests needed to decide for each item in a given "contaminated" population, whether it is good or

© 2003 Cambridge University Press 0269-9648/03 $16.00

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

defective. The basic idea is to pool samples taken from the population, screen them together, and observe, for any such group, one of two possible outcomes: either it is "clean," implying that all items in the group are good, or it is contaminated, implying that at least one item in the group is defective, but without knowing which and how many are defective, so that such a group may need individual rescreening. Intuitively, it seems clear that such an approach can save tests and thus time and money—in particular if the fraction of defective items is rather small.

Dorfman [4] was the first to suggest group testing for blood samples in order to screen a large number of people for a certain disease. The objective was to get an exact result for every individual, and since the disease was rare, great savings could be obtained by sample pooling. Since this pioneering work, group testing procedures have been applied in various areas, not only for analyzing blood or urine samples to detect syphilis, human immunodeficiency virus (HIV) or other diseases as well as for DNA screening, but also in quality control for industrial production systems (Bar-Lev, Boneh, and Perry [1]). One of the key references is the monograph by Ding-Zhu and Hwang [3] in which algorithms for the worst-case analysis of the detection problem for defective items are studied in detail. Applications to HIV screening are given, among others, by Hammick and Gastwirth [5], Litvak, Tu, and Pagano [8], Tu, Litvak, and Pagano [11], Wein and Zenios [12], and Hung [6], who used binomial grouping in hypotheses testing for the classification of quantitative covariables. Combinatorial questions in the context of DNA library screening were recently treated by Macula [9,10].

In the models studied in this article, a total population of size N contains a certain number G of good items, and we assume that G is a random variable following some known distribution. In some situations, G is deterministic and known in advance, and sometimes any item can be considered to be defective independently of the others with a known probability p, so that G has a binomial distribution with parameters N and $q = 1 - p$. However, the distribution of the residual number of good items will, of course, change in the successive stages of the procedures. The objective is to identify a required number D of good items (the *demand*). Samples of size m are taken successively and tested in groups. Without much loss of generality, we assume that N and D are multiples of m. If a group is found to be clean, its items are put aside and collected to meet the demand requirement. If it is found contaminated, there are two submodels: Either it is put aside for potential use in later stages (Model A) or it is immediately returned to the current pool (Model B). In Model A, the items of the contaminated groups are taken as the new sampling population in a second stage after all items of the original population have been tested, and so on. We assume that there is a prespecified upper bound for the number of group tests; testing being costly, this can be considered as a budget restriction. In both models, the group testing continues until the demand is fully satisfied or the maximum number of tests is reached. Since sampling without replacement is a central feature of the procedures, we call them *hypergeometric group testing*.

As a further ingredient to both models, we also introduce processing times. Let the durations of the group tests be independent and identically distributed (i.i.d.)

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

random variables and assume that there is a certain deadline by which the testing process must be finished. Both models can be endowed with this additional feature. Then, the process ends if either the required number of good items is achieved, the number of permitted tests is reached, or the deadline is exceeded.

In Bar-Lev, Stadje, and Van der Duyn Schouten [2], a one-stage model (in which groups found contaminated are immediately scrapped) was studied. Of course, reconsidering "bad" groups and trying to detect nondefective items among the "suspicious" ones by regrouping and new testing makes better use of the available population; on the other hand, multistage procedures may cause higher costs for retesting and are mathematically more complicated. The computational effort to handle them seems to be enormous. Which of the two models A and B is more cost-efficient depends on all the system parameters. It seems reasonable to conjecture that for stochastically small fractions of defective items in the population, Model B is based on the more economical procedure, because immediately returning a contaminated group to the current pool will increase the proportion of good items. However, since the fraction of defective items grows as good items are successively removed from the population, it may be better to change the strategy to putting contaminated groups aside until the end of the current stage (i.e., to use the policy of Model A). For certain values of N and D and stochastically large distributions of G, Model A is preferable.

2. MODEL DESCRIPTION, COST STRUCTURE, AND OBJECTIVE FUNCTIONS

We start with a more formal description of the dynamics of our models.

Model A: At the beginning of any stage, we know the residual population size and the number of good items still needed. Let these numbers be N_i and md_i for stage i, respectively. In particular, $N_1 = N$ and $md_1 = D$. Let X_{ij} be the number of good items in the jth group tested in stage i. If $X_{ij} = m$, the corresponding m good items are put aside to meet the demand. If $X_{ij} < m$, this group will be used as part of the new sample population after the current stage is finished. Let j_i be the number of groups of size m in which the remaining population of stage i is divided (i.e., $j_i = N_i/m$). The random vector $(X_{i1}, \ldots, X_{ij_i})$ completely describes the ith stage, but is, of course, unknown at that time. The quantity observed at the ith stage is

$$(Y_{i1}, \ldots, Y_{ij_i}) = (1_{\{X_{i1}=m\}}, \ldots, 1_{\{X_{ij_i}=m\}}).$$

Let

$$Y_i^{(\ell)} = \sum_{j=1}^{\ell} Y_{ij}, \qquad \ell = 1, \ldots, j_i.$$

Note that Y_{ij} indicates whether the jth group in stage i was clean or not, whereas $mY_i^{(\ell)}$ is the number of good items identified by the first ℓ tests of stage i. The demand requirement is first met or exceeded after

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

$$T_A = \inf\left\{\sum_{i=1}^{k-1} j_i + \ell \,|\, k \in \mathbb{N}, \ell \in \{1,\ldots,j_k\}, m\sum_{i=1}^{k-1} Y_i^{(j_i)} + mY_k^{(\ell)} \geq D\right\}$$

group tests.

It may happen that there are not enough good items in the population (e.g., if $P(G < D) > 0$); in this case, $T_A = \inf\emptyset = \infty$.

Model B: In this model, we define N_i to be the residual population size just before the ith sampling, whereas G_i is the number of good items left among these N_i untested ones. Let X_i be the number of good items in the ith group. We successively observe the indicators $Y_i = 1_{\{X_i = m\}}$. The sequences N_i and G_i follow a simple recursive pattern: $N_1 = N$, $G_1 = G$ and, for $i > 1$,

$$N_i = \begin{cases} N_{i-1} - m & \text{if } Y_{i-1} = 1 \text{ (i.e., } X_{i-1} = m) \\ N_{i-1} & \text{if } Y_{i-1} = 0 \text{ (i.e., } X_{i-1} < m), \end{cases} \tag{2.1}$$

$$G_i = \begin{cases} G_{i-1} - m & \text{if } Y_{i-1} = 1 \\ G_{i-1} & \text{if } Y_{i-1} = 0. \end{cases} \tag{2.2}$$

Clearly, $R_k = \sum_{i=1}^{k} Y_i$ is the number of uncontaminated groups among the first k. Hence, the number of group tests needed to satisfy the demand requirement is

$$T_B = \inf\{k \,|\, mR_k \geq D\}, \tag{2.3}$$

where, again, $\inf\emptyset = \infty$.

The distributions of T_A and T_B are crucial in the analysis and optimization of the two models. We will develop algorithms to compute these distributions in the next section.

Regarding cost and constraints, we use some or all of the following.

Assumptions:

1. *The cost of testing a group of size m is $c(m)$. For simplicity, we assume in most cases that $c(m)$ has a linear and a fixed component (i.e., is given by $c(m) = c_0 + c_1 m$ for two constants $c_0, c_1 \geq 0$, not both being 0). c_0 can be considered as a setup cost for a test, the further cost of which is then assumed to be proportional to the number of items tested. In the numerical analysis, we also take $c(m) = c_0 m^{1/2}$.*
2. *There is an alternative option to buy good items (which do not need testing) from another (secure) source at the price of b per item.*
3. *There is a budget limit C for the entire process. If $c_1 = 0$, this constraint means that the maximum number h of group tests is the integer part of C/c_0. If $c_1 > 0$, the restriction concerns h and the selected group size m; in this case, $h = h(m)$ is the integer part of $C/(c_0 + c_1 m)$. Under Assumption 2, we must, of course, suppose that $C < bN$.*

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

The stopping times for Models A and B are given by

$$T_{A,h} = \min[T_A, h], \qquad T_{B,h} = \min[T_B, h],$$

respectively. At the end of the procedure, the number of good items failing to meet the demand (i.e., still to be purchased at the price of b per item if this option exists) is

$$R_A = m\left(d_1 - \sum_{i=1}^{k-1} Y_i^{(j_i)} - Y_k^{\ell}\right)_+ \quad \text{for Model A,} \tag{2.4}$$

where $\sum_{i=1}^{k-1} j_i + \ell = T_{A,h}$, $0 \le \ell < j_k$, and

$$R_B = m\left(d_1 - \sum_{i=1}^{T_{B,h}} Y_i\right)_+ \quad \text{for Model B.} \tag{2.5}$$

We will consider several *optimization problems*. In the first three, N and the distribution of G are given and m is the only decision variable.

Problem 1: Under cost assumption 1, we want to minimize the expected cost of group testing (i.e., $\mathrm{E}(T_A)$ or $\mathrm{E}(T_B)$, respectively).

Problem 2: If we additionally consider Assumption 3, we want to minimize $\mathrm{E}(T_{A,h})$ or $\mathrm{E}(T_{B,h})$, respectively, subject to meeting the demand requirement with a certain prespecified probability, say $1 - \alpha$ for some small $\alpha > 0$.

Problem 3: Consider Assumptions 1–3. Let

$$C_A(m) = c(m)T_{A,h} + bR_A, \qquad C_B(m) = c(m)T_{B,h} + bR_B.$$

These objective functions combine the cost of group testing and the cost of buying missing good items from a secure source. We want to minimize $\mathrm{E}(C_A(m))$ and $\mathrm{E}(C_B(m))$, respectively.

Problem 4: In addition to Assumptions 1–3, we assume that the population to be tested first has to be purchased at the price of b_0 per item. (Usually, b is much larger than b_0, reflecting a penalty for not being able to satisfy the demand requirement from the tested population.) Now, let N and m be the decision variables. We have to assume some distribution for G, which will, of course, depend on our choice of N. We consider two cases: (a) G is binomially distributed with parameters N and q and (b) a fixed fraction of N. The total costs whose expectation is to be minimized are then

$$C_A(N, m) = b_0 N + (c_1 + c_2 m)T_{A,h} + bR_A \quad \text{for Model A}$$

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

and

$$C_B(N,m) = b_0 N + (c_1 + c_2 m) T_{B,h} + bR_B \quad \text{for Model B.}$$

3. GENERAL ALGORITHMS FOR THE DISTRIBUTIONS OF STOPPING TIMES

Using the following approach, we can construct an algorithm to compute all quantities of importance for the optimization problems in the framework of both models.

Model A: We consider the successive stages recursively. At the beginning of the ith stage, we assume that we know the residual population size N_i and the number of good items still needed, which we denoted by md_i. We first derive the distribution of $(X_{i1},\dots,X_{ij_i})$; recall that $j_i = N_i/m$ is the number of groups of size m in which the remaining population is divided and X_{ih} is the number of good items in the hth group tested at stage i. Let G_i be the number of good items in the remaining population. The distribution of G_i, the number of good items at that stage, is known from the former stages. Then our sampling method yields

$$P(X_{i1} = m_1,\dots,X_{ij_i} = m_{j_i} \mid G_i = g) = \frac{\binom{g}{m_1}\binom{N_i - g}{m - m_1}\binom{g - m_1}{m_2}\binom{N_i - g - m + m_1}{m - m_2}\cdots}{\binom{N_i}{m,\dots,m}}$$

(where the multinomial coefficient contains j_i m's). After deconditioning with respect to G_i, we obtain the joint distribution of $(X_{i1},\dots,X_{ij_i})$, G_i and, thus, also that of $(Y_{i1},\dots,Y_{ij_i}) = (1_{\{X_{i1}=m\}},\dots,1_{\{X_{ij_i}=m\}})$. Note that Y_{il} indicates whether the lth group in stage i was clean or not. Now, we get the distribution of G_{i+1}, the number of good items present at the beginning of the $(i+1)$st stage, and can continue. Clearly,

$$N_{i+1} = N_i - \sum_{l=1}^{j_i} X_{il} 1_{\{X_{il}=m\}} = N_i - m \sum_{l=1}^{j_i} Y_{il}$$

and

$$G_{i+1} = G_i - m \sum_{l=1}^{j_i} Y_{il},$$

so that $P_{G_{i+1}}$, the marginal distribution of G_{i+1}, can be computed from $P_{(G_i, X_{i1},\dots,X_{ij_i})}$, the joint distribution of $(G_i, X_{i1},\dots,X_{ij_i})$.

At the beginning of the first stage, we have $N_1 = N$, $md_1 = D$, and G_1 has a given distribution (e.g., it can be deterministic or binomially distributed with parameters N

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

and q). Thus, we can start the above-described procedure and carry on until we have D or more good items or the maximally permitted number of stages is reached.

The distribution of the stopping time T_A (i.e., the number of performed group tests) can be expressed in terms of the joint distribution of the Y_{il} as follows. We consider the untruncated case (i.e., T_A itself), which is set equal to ∞ if we do not eventually obtain D good items. (The truncation leading to $T_{A,h}$ only requires a minor modification.) Let l_i be the maximum number of groups tested at stage i. Note that $l_1 = N/m$, whereas $l_2, l_3, \ldots$ are random variables: $l_i = N_i/m$. Clearly,

$$P(T_A > k) = P\left(\sum_{l=1}^{k} Y_{1l} < D\right) \quad \text{if } k \le l_1,$$

$$P(T_A > k) = P\left(\sum_{l=1}^{l_1} Y_{1l} + \sum_{l=1}^{k-l_1} Y_{2l} < D\right) \quad \text{if } l_1 < k \le l_1 + l_2,$$

and so on. Our algorithm provides us with the conditional distribution of $\sum_{l=1}^{k} Y_{il}$, given everything that happened before stage i. Actually, at the beginning of stage i, all information about the past that is important is contained in the triple (N_i, d_i, G_i). Therefore, we can recursively determine the distributions of

$$\sum_{l=l_1}^{k} Y_{1l} + \cdots + \sum_{l=1}^{l_{i-1}} Y_{i-1,l} + \sum_{l=1}^{k-l_1-\cdots-l_{i-1}} Y_{il},$$

where

$$\sum_{j=1}^{i-1} l_j < k \le \sum_{j=1}^{i} l_j,$$

so that we also obtain that of T_A.

Model B: The algorithm in this case is much simpler. Let N_i and the distribution of G_i, as defined for this model, be given. Then X_i has a conditional hypergeometric distribution, given G_i, since it is equal to the number of good items in a sample without replacement of size m from a population of G_i good and $N_i - G_i$ defective items. Hence,

$$P(X_i = l \,|\, G_i = g) = \frac{\binom{g}{l}\binom{N_i - g}{m - l}}{\binom{N_i}{m}}. \tag{3.1}$$

Clearly, $N_{i+1} = N_i - m1_{\{X_i = m\}}$ and $G_{i+1} = G_i - m1_{\{X_i = m\}}$. Therefore, at stage $i+1$ we now know N_{i+1} and the distribution of G_{i+1} and can continue. This method yields the joint distribution of $Y_1, \ldots, Y_k$ for every $k \in \mathbb{N}$ and, therefore, also the distribution of

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

T_B (again, the untruncated stopping time), as it is related to the sequence $(Y_i)_{i\in\mathbb{N}}$ by the relation

$$P(T_B > k) = P\left(\sum_{i=1}^{k} Y_i < D\right).$$

4. AN EXAMPLE FOR MODEL B

In this section, we consider Model B in the case when the initial number of good items is some known constant $g < N$. Let

$$\alpha_i = \frac{\binom{g-im}{m}}{\binom{N-im}{m}}, \qquad i = 0,1,\ldots . \tag{4.1}$$

We will give explicit formulas for the distributions of R_k and T_B. In order to get the flavor of the proof of Proposition 1 below, we will first derive the probabilities $P(R_k = l)$ for $l = 0,1,2$ and then present the general expression.

The event $\{R_k = 0\}$ occurs if and only if $(Y_1,\ldots,Y_k) = (0,\ldots,0)$, so that

$$P(R_k = 0) = (1-\alpha_0)^k.$$

The event $\{R_k = 1\}$ occurs if and only if $(Y_1,\ldots,Y_k) = (y_1,\ldots,y_k)$, where $(y_1,\ldots,y_k)$ consists of one 1 and $k-1$ 0's. The probability that $y_{i_1} = 1$ and $y_i = 0$ for $i \in \{1,\ldots,k\}\backslash\{i_1\}$ is $(1-\alpha_0)^{i_1-1}\alpha_0(1-\alpha_1)^{k-i_1}$. Hence, in order to obtain $P(R_k = 1)$, we have to sum over all k possible permutations. This results in

$$\begin{aligned} P(R_k = 1) &= \alpha_0 \sum_{j_1=1}^{k} (1-\alpha_0)^{j_1-1}(1-\alpha_1)^{k-j_1} \\ &= \frac{\alpha_0[(1-\alpha_1)^k - (1-\alpha_0)^k]}{\alpha_0 - \alpha_1}. \end{aligned} \tag{4.2}$$

Similarly, the event $\{R_k = 2\}$ occurs if and only if $(Y_1,\ldots,Y_k) = (y_1,\ldots,y_k)$, where $(y_1,\ldots,y_k)$ consists of two 1's and $k-2$ 0's. For this event with the two 1's at the positions i_1 and i_2, $1 \le i_1 < i_2 \le k$, the probability is $(1-\alpha_0)^{i_1-1}\alpha_0(1-\alpha_1)^{i_2-i_1-1}\alpha_1(1-\alpha_2)^{k-i_2}$. To obtain $P(R_k = 2)$, we need to sum over the $\binom{k}{2}$ possible permutations. This yields

$$P(R_k = 2) = \alpha_0\alpha_1 \sum_{j_1=1}^{k} \sum_{j_2=j_1+1}^{k} (1-\alpha_0)^{j_1-1}(1-\alpha_1)^{j_2-j_1-1}(1-\alpha_2)^{k-j_2}, \tag{4.3}$$

which can be simplified similarly to (4.2).

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

In general, we have the following proposition.

PROPOSITION 1:

$$P(R_k = l) = \begin{cases} (1-\alpha_0)^k & \text{if } l = 0 \\ \left(\prod_{i=0}^{l-1} \alpha_i\right) \sum_{j_1=1}^{k} \sum_{j_2=j_1+1}^{k} \cdots \sum_{j_l=j_{l-1}+1}^{k} (1-\alpha_l)^{k-j_l} \prod_{i=1}^{l} (1-\alpha_{i-1})^{j_i - j_{i-1}-1} \\ \qquad \text{if } l = 1,\ldots,k. \end{cases} \tag{4.4}$$

PROOF: Let $l \in \{1,\ldots,k\}$. The event $\{R_k = l\}$ occurs if and only if $(Y_1,\ldots,Y_k) = (y_1,\ldots,y_k)$ for some $(y_1,\ldots,y_k)$ consisting of l 1's and $k-l$ 0's. The probability of this event with the l 1's placed in positions $j_1 < j_2 < \cdots < j_l$ and 0's in all remaining positions is

$$\left(\prod_{i=1}^{l-1} \alpha_i\right)(1-\alpha_l)^{k-j_l} \prod_{i=1}^{l-1} (1-\alpha_{i-1})^{j_i - j_{i-1}-1}.$$

By summing over all possible permutations, we obtain the desired result. ■

PROPOSITION 2: *Let $d = D/m$. For $k \in \{d, d+1,\ldots\}$, the distribution of the stopping time T_B defined by (2.3) is given by*

$$P(T_B = k) = \alpha_{d-1} P(R_{k-1} = d-1). \tag{4.5}$$

PROOF: For $k \in \{d, d+1,\ldots\}$, we clearly have

$$\{T_B = k) = \{R_{k-1} = d-1\} \cap \{Y_k = 1\}.$$

Hence,

$$P(T_B = k) = P(Y_k = 1 | R_{k-1} = d-1) P(R_{k-1} = d-1),$$

and the result follows. ■

5. PROCESSING TIMES

Processing times in the context of group testing were first considered by Bar-Lev et al. [2] for independent binomial schemes. Let us assume that each group test is time-consuming and that there is a prespecified threshold time c by which the group testing process must be finished. The processing times are supposed to be i.i.d. positive random variables with the common distribution function F. They are inde-

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

pendent of the outcomes of the tests. Again, our two models need to be treated separately.

5.1. Model A

Let us start with the first stage. Let $V_{1j}, j \in \mathbb{N}$, denote the time required to run the jth group test in the first stage. We are interested in

$$T_{1,c} = \inf\left\{k \le N/m \,\middle|\, \sum_{j=1}^{k} V_{1,j} \ge c\right\}$$

(recall the convention $\inf \varnothing = \infty$) and $T_{A,c} = \min[T_A, T_{1,c}]$. Our aim is to derive the distribution of $T_{A,c}$. Note that according to our definition, a group during whose testing the time limit is exceeded would still be accepted if it is tested perfect. An alternative would be to reject such a group; this model can be handled similarly.

We have

$$P(T_{A,c} = k) = P(T_A = k < T_{1,c}) + P(T_A = k = T_{1,c}) + P(T_A > k = T_{1,c}). \quad \textbf{(5.1)}$$

The probabilities on the right-hand side of (5.1) are given by

$$\begin{aligned} P(T_A = k < T_c) &= P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1, \sum_{i=1}^{k} V_{1i} < c\right) \\ &= P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1\right) P\left(\sum_{i=1}^{k} V_{1i} < c\right) \\ &= P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1\right) F^{*k}(c-) \end{aligned}$$

(where F^{*k} denotes the k-fold convolution of F with itself) and by

$$\begin{aligned} P(T_A = k = T_c) &= P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1\right) P\left(\sum_{i=1}^{k-1} V_{1i} < c \le \sum_{i=1}^{k} V_{1i}\right) \\ &= P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1\right)[F^{*k-1}(c-) - F^{*k}(c-)], \\ P(T_A > k = T_c) &= P\left(\sum_{i=1}^{k} Y_{1i} \le d-1\right) P\left(\sum_{i=1}^{k-1} V_{1i} < c \le \sum_{i=1}^{k} V_{1i}\right) \\ &= P\left(\sum_{i=1}^{k} Y_{1i} \le d-1\right)[F^{*k-1}(c-) - F^{*k}(c-)]. \end{aligned}$$

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

In the special case when the Y_{1i} are i.i.d. and $B(1, q^m)$ distributed, we obtain

$$P\left(\sum_{i=1}^{k-1} Y_{1i} = d-1, Y_{1k} = 1\right) = \binom{k-1}{d-1} q^{mk}(1-q^m)^{k-d},$$

$$P\left(\sum_{i=1}^{k} Y_{1i} \le d-1\right) = \sum_{j=0}^{d-1} \binom{k}{j} q^{mj}(1-q^m)^{k-j}.$$

In the second and further stages, one can determine the probabilities $P(T_{A,c} = k)$, $k > j_1 = N/m$, recursively, using (5.1) and the distributions of $(Y_{i1}, \ldots, Y_{ij_i})$, which have been derived in Section 3.

5.2. Model B

Now, let V_j be the processing time of the jth group test. The threshold c is first reached or exceeded at time $V_1 + \cdots + V_{T_c}$, where

$$T_c = \inf\left\{r : \sum_{i=1}^{r} V_i \ge c\right\}. \tag{5.2}$$

We need to know the distribution of

$$T_{B,c} = \min[T_B, T_c]. \tag{5.3}$$

The following result can be found in Bar-Lev et al. [2] in a different context.

PROPOSITION 3: *Let $\{Y_i\}$ be a sequence of Bernoulli random variables (r.v.'s) (not necessarily independent) and $\{V_i\}$ be a sequence of i.i.d. positive r.v.'s with common distribution F, such that the two sequences are independent. Let the stopping times T_B, T_c, and $T_{B,c}$ be defined by (2.3), (5.2), and (5.3), respectively. Then, the distribution of $T_{B,c}$ is given by*

$$P(T_{B,c} = k) = \begin{cases} P\left(\sum_{i=1}^{k-1} Y_i = d-1, Y_k = 1\right) F^{*k}(c-) \\ \quad + P\left(\sum_{i=1}^{k-1} Y_i = d-1, Y_k = 1\right)[F^{*k-1}(c-) - F^{*k}(c-)] \\ \quad + P\left(\sum_{i=1}^{k} Y_i \le d-1\right)[F^{*k-1}(c-) - F^{*k}(c-)] \\ \qquad\qquad \text{if } k = d, d+1, \ldots \\ [F^{*k-1}(c-) - F^{*k}(c-)], \quad \text{if } k = 1, \ldots, d-1, \end{cases} \tag{5.4}$$

with $F^{*0}(b) \equiv 1$.

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

Since the Y_i''s in Model B are dependent Bernoulli r.v.'s, Propositions 2 and 3 yield the following corollary.

COROLLARY: *In Model B, the distribution of $T_{B,c}$ defined by (5.3) is given by*

$$P(T_{B,c}=k)=\begin{cases}\alpha_{d-1}P(R_{k-1}=d-1)F^{*k}(c-)\\ \quad+\alpha_{d-1}P(R_{k-1}=d-1)[F^{*k-1}(c-)-F^{*k}(c-)]\\ \quad+P(R_k\le d-1)[F^{*k-1}(c-)-F^{*k}(c-)] \qquad \text{if } k=d,d+1,\ldots\\ [F^{*k-1}(c-)-F^{*k}(c-)] \quad \text{if } k=1,\ldots,d-1.\end{cases} \tag{5.5}$$

6. NUMERICAL RESULTS

Let us now present a few numerical results. In all of the examples, the running of the algorithms takes much computing time; the results were checked by (also very time-consuming) simulations.

In Tables 1 and 2 we display, for Model A, the values of

$$P^* = P(\text{meeting the demand requirement})$$

TABLE 1. Model A for $G \sim U_d(115,120)$; $N = 120$, $h = 10$, and Different Values of m and D

	m				
D	5	10	20	30	40
10	1.000	1.000	—	—	P^*
	2.224	1.236	—	—	$\mathrm{E}(T_{A,h})$
20	1.000	1.000	1.000	—	—
	4.446	2.472	1.530	—	—
30	0.999	1.000	—	0.9935	—
	6.668	3.708	—	1.986	—
40	0.918	1.000	0.998	—	0.949
	8.790	4.943	3.080	—	2.771
60	0	0.978	0.964	0.901	—
	10	7.395	4.782	4.378	—
80	0	0.655	0.798	—	0.659
	10	9.284	6.567	—	5.991

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

TABLE 2. Model A for $G \sim B(120,117.5/120)$; $N = 120$, $h = 10$, and Different Values of m and D

	m				
D	5	10	20	30	40
10	1.000	1.000	—	—	P^*
	2.222	1.234	—	—	$\mathrm{E}(T_{A,h})$
20	1.000	1.000	0.9996	—	—
	4.444	2.470	1.526	—	—
30	0.998	1.000	—	0.991	—
	6.664	3.703	—	1.949	—
40	0.930	0.999	0.994	—	0.954
	8.793	4.934	3.067	—	2.633
60	0.000	0.973	0.964	0.919	—
	10.00	7.368	4.684	4.189	—
80	0.000	0.708	0.840	—	0.706
	10.00	9.348	6.424	—	5.901

and the expected value of $T_{A,h}$ for $N = 120, h = 20$, and various values of D and m under the initial distributions $U_d(115,120)$, the discrete uniform distribution on $\{115,\ldots,120\}$, and $B(120,117.5/120)$, the binomial distribution with n trials and success probability 117.5/120, respectively. Since we assume that m divides D and N, a few boxes are empty. In both Tables 1 and 2 the values of $\mathrm{E}(T_{A,h})$ are decreasing in m, whereas P^* is unimodal in m. The two initial distributions for G have the same means, but obviously there are relevant differences between the values in Table 1 and Table 2.

In Table 3 we provide the values of P^*, $\mathrm{E}(T_{A,h})$, and $\mathrm{E}(R_A)$ dependent on the group size m for certain fixed values of the other parameters. We present the objective function $\mathrm{E}(C_{A,m})$ relative to its minimum in terms of

$$\mathrm{E}(C_{A,m})^* = \frac{\mathrm{E}(C_{A,m})}{\min\{\mathrm{E}(C_{A,j}), j \in \{3,4,5,6,10,15,20,30\}\}}$$

for several choices of $c(m)$ and b. Thus, the expression (1.00) in the lower part of Table 3 indicates the optimal group size. We see that P^* is first increasing and then decreasing, whereas $\mathrm{E}(T_{A,h})$ and $\mathrm{E}(R_A)$ are first decreasing and then increasing. The cost functions show huge differences for different group sizes. We have also included two nonlinear (square root) examples for $c(m)$.

In Table 4 we give a few values of P^*, $\mathrm{E}(T_{B,h})$ and $\mathrm{E}(R_B)$ for Model B under the same assumptions as in Table 3. Since the objective functions are simple functions of $\mathrm{E}(T_{B,h})$ and $\mathrm{E}(R_B)$, minimization is easy.

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

TABLE 3. Model A; $N = 120$, $D = 60$, $h = 20$, several values of m, initial distribution $G \sim B(120,0.9)$, and different cost functions $C_{A,m} = c(m)T_{A,h} + bR_A$

			m							
	$c(m)$	b	3	4	5	6	10	15	20	30
P^*			0.000	0.264	0.562	0.694	0.510	0.325	0.212	0.105
$E(T_{A,h})$			20.000	19.701	18.689	17.686	16.938	17.747	18.413	19.132
$E(R_A)$			16.266	8.289	4.805	3.603	9.620	19.105	28.375	42.741
$E(C_{A,m})^*$	$2m + 10$	30	1.63	1.21	1.04	(1.00)	1.60	2.58	3.56	5.27
$E(C_{A,m})^*$	$4m + 10$	20	1.17	1.03	(1.00)	1.03	1.58	2.47	3.39	5.09
$E(C_{A,m})^*$	$13\sqrt{m}$	10	1.04	1.01	(1.00)	1.01	1.34	1.83	2.29	3.03
$E(C_{A,m})^*$	$15\sqrt{m}$	7	(1.00)	1.02	1.04	1.07	1.37	1.84	2.26	2.95

TABLE 4. Model B for $G \sim B(120,0.9)$; $N = 120$, $h = 20$ and Different Values of m and D

	m		
D	10	20	30
30	0.908	P^*	0.457
	10.194	$\mathrm{E}(T_{B,h})$	14.714
	1.184	$\mathrm{E}(R_B)$	16.281
40	0.768	0.482	—
	13.537	15.463	—
	3.500	13.825	—
60	0.355	0.184	0.099
	18.143	18.700	19.200
	14.271	30.158	43.298

Finally, Figures 1 and 2 display the exact probability function of $T_{A,h}$ for the parameter values $N = 120$, $D = 60$, $m = 15$, $h = 120$, and different initial distributions G, using a logarithmic scale. They show $p = P(T_{A,h} = k)$. In the case $G \sim U_d(112,120)$, the function $p \mapsto P(T_{A,h} = k)$ can clearly be approximated by a straight line with high accuracy, so that the distribution is "almost" geometric. In the case $G \sim B(120,0.967)$, the main part of the (logarithmic) graph shows a nonlinear decrease.

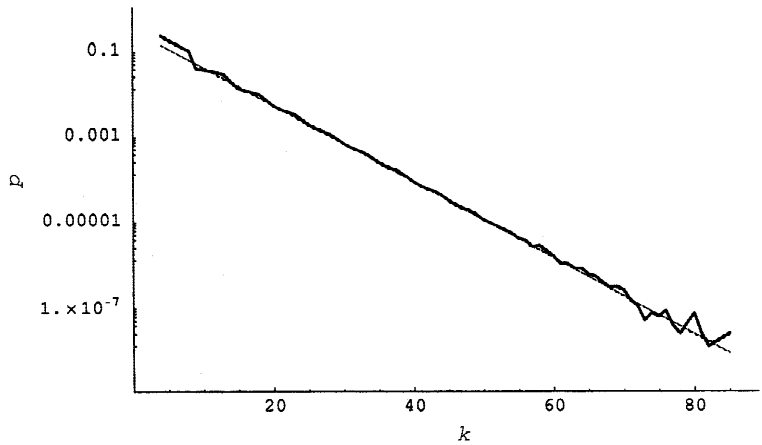


FIGURE 1. $G \sim U_d(112,120)$. The gray line corresponds to the function $0.3e^{-0.205k}$.

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press

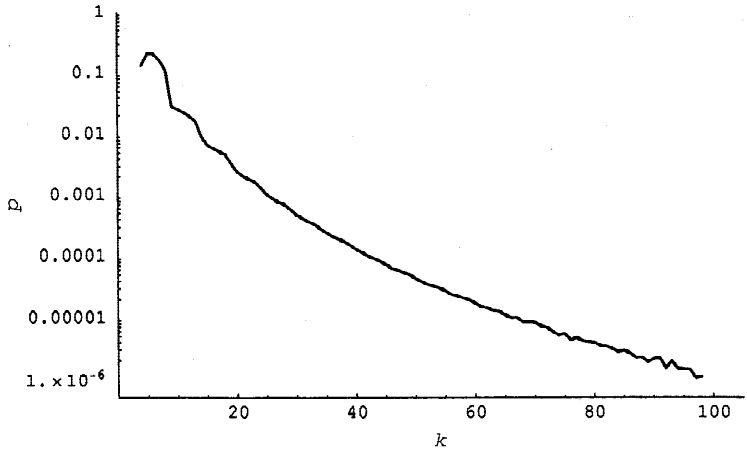


FIGURE 2. $G \sim B(120,0.967)$.

Acknowledgments

The authors would like to thank Andreas Löpker for his assistance in the numerical part of this study.

References

1. Bar-Lev, S.K., Boneh, A., & Perry, D. (1990). Incomplete identification models for group-testable items. *Naval. Research Logistic* 37: 647–659.
2. Bar-Lev, S.K., Stadje, W., & Van der Duyn Schouten, F. (2003). Group testing models with incomplete identification and processing times. *Methodology and Computing in Applied Probability*, forthcoming.
3. Ding-Zhu, D. & Hwang, F.K. (2000). *Combinatorial group testing and its applications*, 2nd ed. Singapore: World Scientific.
4. Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* 14: 436–440.
5. Hammick, P.A. & Gastwirth, J.L. (1994). Group testing for sensitive characteristics: Extensions to higher prevalence. *International Statistical Review* 62: 319–331.
6. Hung, M.C. (2000). Use of binomial group testing in tests of hypotheses for classification or quantitative covariables. *Biometrics* 56: 204–212.
7. Hwang, F.K., Pfeifer, C.G. & Enis, P. (1981). An optimal hierarchical procedure for a modified binomial group-testing problem. *Journal of the American Statistical Association* 76: 947–949.
8. Litvak, E., Tu, X.M., & Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* 89: 424–434.
9. Macula, A.J. (1999). Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. *Annals of Combinatorics* 3: 61–69.
10. Macula, A.J. (1999). Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. *Journal of Combinatorial Optimization* 2: 385–397.
11. Tu, X.M., Litvak, E., & Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* 82: 287–297.
12. Wein, L.M. & Zenios, S.A. (1996). Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research* 44, 543–569.

https://doi.org/10.1017/S0269964803173032 Published online by Cambridge University Press