No. 2002-75

GROUP TESTING MODELS WITH PROCESSING TIMES AND INCOMPLETE IDENTIFICATION

By Shaul K. Bar-Lev, Wolfgang Stadje, Frank A. Van der Duyn Schouten

August 2002

ISSN 0924-7815

Discussion paper

GROUP TESTING MODELS WITH PROCESSING TIMES AND INCOMPLETE IDENTIFICATION

Shaul K. Bar-Lev[*], Wolfgang Stadje[†]
and Frank A. Van der Duyn Schouten[‡]

Abstract

We consider the group testing problem for a finite population of possibly defective items with the objective of sampling a prespecified demanded number of nondefective items at minimum cost. Group testing means that items can be pooled and tested together; if the group comes out clean, all items in it are nondefective, while a "contaminated" group is scrapped. Every test takes a random amount of time and a given deadline has to be met. If the prescribed number of nondefective items is not reached, the demand has to be satisfied at a higher (penalty) cost. We derive explicit formulas for the distributions underlying the cost functionals of this model. It is shown in numerical examples that these results can be used to determine the optimal group size.

1 Introduction

Since the pioneering work of Dorfman (1943), various group testing procedures have been introduced and discussed in the literature. Their general

[*]Department of Statistics, University of Haifa, Haifa 31905, Israel

[†]Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany

[‡]Center for Economic Research, Tilburg University, 5000 LE Tilburg, The Netherlands

purpose is to reduce the number of tests needed to decide for each item in a given "contaminated" population whether it is good or defective. The basic idea is to pool samples from the given population, screen them together and observe, for any such sample 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 when the fraction of defective items is rather small.

Group testing procedures have been applied in various areas, for example for analysing blood or urine samples to detect syphilis, HIV or other diseases as well as for DNA screening, but also in quality control for industrial production systems. One of the key references is the monograph by Ding-Zhu and Hwang (2000) 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 (1994), Litvak, Tu and Pagano (1994), Tu, Litvak and Pagano (1995), Wein and Zenios (1996), and Hung and Swallow (2000) who used binomial grouping in hypotheses testing for the classification of quantitative covariables. Combinatorial questions in the context of DNA library screening were recently studied by Macula (1999a,1999b).

Many of the group testing models deal with the problem of a complete identification of all items in a population, requiring a correct classification of each item as good or defective. The main goal in early group testing models has been to find optimal group pooling policies (e.g., to find the optimal group size at any stage of the testing process) in order to minimize the expected number of tests required for a complete identification (cf. Hwang, Pfeifer and Enis (1981)). However, no such optimal policies have been found for reasonably large population sizes and only suboptimal policies have been suggested.

Moreover, the majority of models assumes that each group test provides a correct answer: a group labeled clean or contaminated really is of this type. In recent papers on the subject (see e.g. Litvak, Tu and Pagano (1994) on HIV screening) it was pointed out that the phenomena of "false negative" and "false positive" outcomes prevails in some situations, especially in clinical trials. Groups can be declared clean although they contain contaminated material, while groups labeled "defective" can be free of contaminations.

Group testing ideas are also often useful in industrial contexts such as production systems and inventory models. Consider for example quality control of batteries or electronic circuits. The group tests in such situations can be conducted by connecting electronic items in series. Here complete identification of the defective items is not the only interesting objective. In practice, one often has to meet a given demand requirement of good items and wants to minimize expenditures. A sequential quality control problem of this kind, i.e. with group testing and incomplete identification, was studied by Bar-Lev, Boneh and Perry (1990). In this real-world example the production department of some electronic company needed one million chips and had two options for purchasing them: paying \$2.5 per chip with a 100% quality guarantee or paying \$0.5 per chip with a 99% probability for each of them to be of acceptable quality. The chips of the second category were group testable in two phases. In the first phase a set of chips was exposed to heating in a helium environment. If a chip in this set was defective, then some helium would penetrate the chip. In the second phase, the same set was exposed to a helium sensor recording a helium leak if and only if at least one chip is defective. Since the chips of the second alternative were much cheaper than the first one, it had been chosen. The problem was how many chips to acquire and what group size to choose for testing the chips in order to fulfill the demand requirement of one million good chips at minimum cost.

Incomplete identification group testing processes can be considered to belong to the theory of optimization and optimal stopping under probabilistic constraints. The aim of this paper is to extend the approach of Bar-Lev, Boneh and Perry (1990), introducing several new features. We consider two incomplete identification group testing models with processing times. The two models (denoted by I and II) assume stochastic independence between the items constituting the population. Model I assumes that tests of groups of m items are conducted sequentially by one machine, while in Model II $h \geq 1$ of such machines are available to work in parallel. Although Model I is the special case of the second model with $h = 1$, we have preferred to consider it first separately for the ease of exposition.

It is assumed that N items constitute the contaminated population, that the probability of an item being good is q (so that the expected proportion of good items is also q) and that the demand requirement is d. The time for running a group test is a random variable. Moreover, there is a prespecified threshold time b by which the testing must be finished or stopped. The process ends when either the demand requirement is met or the total time for running the group tests reaches b or no more groups are left. The conse-

quences of running out of time before the required number of good items has been found can be of different types. It may occur that missing items can always be purchased at higher cost. However, it is also possible that the decision whether to use a contaminated population or a clean population cannot be revoked, implying that an insufficient number of good items cannot be supplemented afterwards. In this case a penalty clause for the supplier will apply against which he can take an insurance.

We assume that the underlying distributions as well as q and b are known. A detailed description of the models is presented in Section 2. In Sections 2.1 and 2.2 we define the stopping times for the two models and derive suitable formulas for their distributions. In Section 3 we introduce appropriate objective functions and deterministic as well as probabilistic constraints depending on the various model parameters. The resulting optimization problems turn out to be analytically intractable. Our results can, however, be used for a numerical analysis. This is illustrated in Section 4, which shows the dependence of the optimal group size and the objective functions on the model parameters in several examples.

2 The Models

The following assumptions are common to both models.

(i) The contaminated population consists of N items which are testable in groups of any size m. (In practice, there may be a constraint $m \leq m_0$ for the group size.)

(ii) For every group test there are two possible outcomes: "clean", implying that all group items are good, or "contaminated", implying that at least one item in the group tested has to be defective. Under this assumption, outcomes like "false negative" or "false positive" for tested groups are excluded.

(iii) Every item is good with probability q independently of the others. The expected proportion q of good items in the population is assumed to be known in advance and will generally be close to 1.

(iv) The demand requirement is for d good items.

(v) The cost for testing a group of size m is $c(m)$ and thus depends on the group size. For simplicity we assume that $c(m)$ has a fixed and a linearly

increasing part, i.e. $c(m) = c_1 + c_2 m$ for some known cost parameters c_1 and c_2.

(vi) The time for running a group test is a random variable. For simplicity we assume that its distribution does not depend on the group size. The running times for different tests are i.i.d.

(vii) There is a fixed (deterministic) deadline b by which the testing process must be finished. b can be a business constraint or a deterioration time after which the items are no longer considered to be usable.

(ix) Groups which are found clean are kept and recorded for meeting the demand requirement. Contaminated groups are set aside but recorded for, perhaps, other possible uses.

(x) The group testing process ends as soon as one of the following events occurs: Either the demand requirement is met, or the aggregated time for running the group tests exceeds the predetermined deterioration survival time threshold b, or no more items are available for further group testing.

(xi) In order to avoid computational and analytic complexity, we only consider group sizes m that divide both d and N. This assumption avoids some computational and analytic complexity and causes only a negligible loss of generality in practical situations. If $\lceil x \rceil, x \in \mathbb{R}$, denotes the ceiling function, then clearly $d_e = \lceil \frac{d}{m} \rceil$ is the total number of good groups of size m needed to satisfy the demand requirement, and $l = \lceil \frac{N}{m} \rceil$ is the maximum number of possible group tests.

(xii) The cost of purchasing (or producing) an item belonging to N is b_1. We assume there exists an alternative by which it is possible to buy good items at the price b_2 per item, where $b_2 > b_1$. We make such an alternative available in order to fulfill the demand requirement in case the group testing process does not end with enough good items. Finally, there might be a budget limit C for the entire testing process.

2.1 The single-machine case: stopping times and analysis

In this model independent groups of size m are tested sequentially by one machine. If a group is found clean, it is kept and aggregated to meet the demand requirement, otherwise it is discarded. The basic random variables

are

$$Y_i = \begin{cases} 1, & \text{if the } i\text{th group of size } m \text{ is found clean} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

By our assumptions, $Y_i \sim B(1, q^m)$, for any $i \in \mathbb{N}$, so that $\sum_{i=1}^{r} Y_i \sim B(r, q^m)$ for $r \in \mathbb{N}$.

Let $V_i, i \in \mathbb{N}$, be the time required to run the group test at stage i. The V_i's are independent with a common distribution function F and are independent of the Y_i's. The stopping times associated with this model are:

$$T_{GI} = \inf \left\{ r : \sum_{i=1}^{r} Y_i \geqslant d_e \right\}, \tag{2.2}$$

$$T_{bI} = \inf \left\{ r : \sum_{i=1}^{r} V_i > b \right\}, \tag{2.3}$$

$$T_{bGI} = \min \left\{ T_{GI}, T_{bI} \right\}, \tag{2.4}$$

and

$$T_I = \min \left\{ T_{bGI}, l \right\} = \min \left\{ T_{GI}, T_{bI}, l \right\}. \tag{2.5}$$

T_I is the total number of group tests conducted during the process, which stops if the number of detected good items is d, the time constraint is reached or the whole population has been tested, whichever comes first. Note that we assume that the test running at the time limit can be completed and its outcome used. Clearly, the support of T_{GI} is $S_{T_{GI}} = \{d_e, d_e + 1, ...\}$, whereas that of T_{bI} is $S_{T_{bI}} = \mathbb{N}$. The number of good items identified is

$$D_I = m \sum_{i=1}^{T_I} Y_i = m \sum_{i=1}^{l} Y_i I(T_I = l) + dI(T_I = T_{GI}) + m \sum_{i=1}^{T_{bI}} Y_i I(T_I = T_{bI}). \tag{2.6}$$

At the end of the process, the remaining number of clean groups of size m required to fulfill the demand requirement is

$$R_I = d_e - \sum_{i=1}^{T_I} Y_i = \left(d_e - \sum_{i=1}^{l} Y_i \right) I(T_I = l) + \left(d_e - \sum_{i=1}^{T_{bI}} Y_i \right) I(T_I = T_{bI}). \tag{2.7}$$

The exact distributions of the stopping times T_{bGI} and T_I are needed in the optimization problems. They are derived with the help of the following

Proposition 1 *Let $\{Y_i\}$ be a sequence of $\{0,1\}$-valued r.v.'s (not necessarily independent) and $\{V_i\}$ be a sequence of i.i.d. nonnegative r.v.'s with common distribution function F, such that the two sequences are independent of each other. Let F^{*k} denote the kfold convolution of F with itself and let the stopping times T_{GI},T_{bI}, and T_{bGI} be as defined by (2.2), (2.3) and (2.4), respectively. Then, for any $k \in \mathbb{N}$, the probability function of T_{bGI} is given by*

$$P(T_{bGI}=k)=\begin{cases} P(\sum_{i=1}^{k-1} Y_i = d_e-1, Y_k=1)F^{*k}(b)+ \\ P(\sum_{i=1}^{k-1} Y_i = d_e-1, Y_k=1)\left[F^{*k-1}(b)-F^{*k}(b)\right]+ \\ P(\sum_{i=1}^{k} Y_i \le d_e-1)\left[F^{*k-1}(b)-F^{*k}(b)\right], \text{ if } k=d_e,d_e+1,... \\ \\ \left[F^{*k-1}(b)-F^{*k}(b)\right], \qquad \text{if } k=1,...,d_e-1, \end{cases} \tag{2.8}$$

with $F^{*0}(b) \equiv 1$, and that of the stopping time T_I is

$$P(T_I=k)=\begin{cases} P(T_{bGI}=k), & k=1,...,l-1 \\ \sum_{i=l}^{\infty} P(T_{bGI}=i) & k=l. \end{cases} \tag{2.9}$$

Proof. If $k \in \{1,...,d_e-1\}$, then, since $T_{GI} \ge d_e$, the event $(T_{bGI}=k)$ can occur only if the event $(\sum_{i=1}^{k-1} V_i \le b < \sum_{i=1}^{k} V_i)$ occurs, where $\sum_{i=1}^{k-1} V_i$ is defined to be 0 if $k=1$. This proves the second part of equation (2.8). For the first part of (2.8), let $k \in \{d_e, d_e+1, ...\}$, then the event $(T_{bGI}=k)$ can be represented as the union of three disjoint sets:

$$(T_{bGI}=k)=(T_{GI}=k<T_{bI}) \cup (T_{GI}=k=T_{bI}) \cup (T_{GI}>k=T_{bI}). \tag{2.10}$$

By using the stochastic independence between the two sequences $\{Y_i\}$ and $\{V_i\}$, we obtain expressions for the probabilities of the three events on the right-hand side of (2.10). These are given, respectively, by (2.11), (2.12) and

(2.13):

$$\begin{aligned} P(T_{GI} = k < T_{bI}) &= P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1, \sum_{i=1}^{k} V_i \leq b) \\ &= P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1) P(\sum_{i=1}^{k} V_i \leq b) \\ &= P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1) F^{*k}(b), \end{aligned} \tag{2.11}$$

$$\begin{aligned} P(T_{GI} = k = T_{bI}) &= P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1) P(\sum_{i=1}^{k-1} V_i \leq b < \sum_{i=1}^{k} V_i) \\ &= P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1) \left[F^{*k-1}(b) - F^{*k}(b) \right], \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} P(T_{GI} > k = T_{bI}) &= P(\sum_{i=1}^{k} Y_i \leq d_e - 1) P(\sum_{i=1}^{k-1} V_i \leq b < \sum_{i=1}^{k} V_i) \\ &= P(\sum_{i=1}^{k} Y_i \leq d_e - 1) \left[F^{*k-1}(b) - F^{*k}(b) \right], \end{aligned} \tag{2.13}$$

which is the desired result (2.8). Equation (2.9) is obvious. □

Corollary 1 *In Model I the probability functions of T_{bGI} and T_I are given by*

$$P(T_{bGI} = k) = \begin{cases} \binom{k-1}{d_e - 1} q^{md_e} (1 - q^m)^{k - d_e} F^{*k}(b) + \\ \binom{k-1}{d_e - 1} q^{md_e} (1 - q^m)^{k - d_e} \left[F^{*k-1}(b) - F^{*k}(b) \right] + \\ \sum_{j=0}^{d_e - 1} \binom{k}{j} q^{mj} (1 - q^m)^{k-j} \left[F^{*k-1}(b) - F^{*k}(b) \right], \\ \qquad\qquad for \;\; k = d_e, d_e + 1, ... \\ \left[F^{*k-1}(b) - F^{*k}(b) \right], \qquad for \;\; k = 1, ..., d_e - 1, \end{cases} \tag{2.14}$$

and

$$P(T_I = k) = \begin{cases} P(T_{bGI} = k), & k = 1, ..., l-1 \\ \sum_{i=l}^{\infty} P(T_{bGI} = i) & k = l \quad . \end{cases} \tag{2.15}$$

Proof. The proof is straightforward since the Y_i's are i.i.d. with $Y_i \sim B(1, q^m)$ so that

$$P(\sum_{i=1}^{k-1} Y_i = d_e - 1, Y_k = 1) = \binom{k-1}{d_e - 1} q^{md_e} (1 - q^m)^{k-d_e} \tag{2.16}$$

and

$$P(\sum_{i=1}^{k} Y_i \leq d_e - 1) = \sum_{j=0}^{d_e - 1} \binom{k}{j} q^{mj} (1 - q^m)^{k-j}. \quad \square \tag{2.17}$$

2.2 The multiple-machine case: stopping times and analysis

Model II generalizes Model I with one major difference: $h \geq 1$ machines are available for simultaneous group testings at any stage. For simplicity we assume that each of the h machines tests groups of size m. We also assume that all h machines are used at the same time and work simultaneously with full capacity. A new stage in the group testing process starts only after all machines have completed their job in the previous stage. All other assumptions are the same as in Model I. The basic counting variables are now

$$Y_{ij} = \begin{cases} 1, & \text{if in stage } i \text{ machine } j \text{ finds its group clean} \\ 0, & \text{otherwise} \end{cases} \tag{2.18}$$

where $i \in \mathbb{N}$ and $j \in \{1, ..., h\}$. Let

$$Y_{i\cdot} = \sum_{j=1}^{h} Y_{ij}, \tag{2.19}$$

denote the total number of clean groups of size m found at stage i by the h machines. Since the Y_{ij}'s are independent and $Y_{ij} \sim B(1, q^m)$, we have

$Y_{i\cdot} \sim B(h, q^m)$ and $\sum_{i=1}^{r} Y_{i\cdot} \sim B(rh, q^m)$, for $r \in \mathbb{N}$. Let V_{ij} be the time for running group test i on machine j, where $i \in \mathbb{N}$ and $j \in \{1, ..., h\}$. As in Model, the V_{ij}'s are assumed to be independent of the Y_{ij}'s and i.i.d. with common distribution F. Then $V_{(i)} = \max_{j\in\{1,...,h\}} \{V_{ij}\}$, $i \in \mathbb{N}$, is the running time of stage i. The distribution function of $V_{(i)}$ is given by $G(\cdot) = F^h(\cdot)$. The basic stopping times are now defined as follows, analogously to (2.2), (2.3) and (2.4) in Model I:

$$T_{GII} = \inf\left\{r : \sum_{i=1}^{r} Y_{i\cdot} \geqslant d_e\right\}, T_{bII} = \inf\left\{r : \sum_{i=1}^{r} V_{(i)} > b\right\}, \tag{2.20}$$

and

$$T_{bGII} = \min\{T_{GII}, T_{bII}\}, \ T_{II} = \min\{T_{bGII}, l\} = \min\{T_{GII}, T_{bII}, l\}. \tag{2.21}$$

Clearly, Proposition 1 and its corollary are also applicable for Model II with appropriate changes in (2.8) in which $Y_{i\cdot}$, $V_{(i)}$, and G have to replace Y_i, V_i, and F, respectively. This leads to the following proposition.

Proposition 2 *In model II the probability functions of T_{bGII} and that of T_{II} are given by*

$$P(T_{bGII} = k) = \begin{cases} G^{*k}(b)\sum_{j=0}^{h-1}\binom{h(k-1)}{d_e-h+j} q^{m(d_e-h+j)}(1-q^m)^{h(k-1)-(d_e-h+j)} \\ \quad \times\left(\sum_{r=h-j}^{h}\binom{h}{r} q^{rm(}(1-q^m)^{h-r}\right) \\ \quad + \sum_{j=0}^{h-1}\binom{h(k-1)}{d_e-h+j} q^{m(d_e-h+j)}(1-q^m)^{h(k-1)-(d_e-h+j)} \\ \quad \times\left(\sum_{r=h-j}^{h}\binom{h}{r} q^{rm(}(1-q^m)^{h-r}\right)\left[G^{*(k-1)}(b) - G^{*k}(b)\right] \\ \quad + \sum_{j=0}^{d_e-1}\binom{hk}{j} q^{mj}(1-q^m)^{hk-j}\left[G^{*(k-1)}(b) - G^{*k}(b)\right], \\ \qquad\qquad\qquad\qquad for \ \ k = \lceil d_e/h\rceil, \lceil d_e/h\rceil + 1, ... \\ \left[G^{*(k-1)}(b) - G^{*k}(b)\right], \qquad for \ \ k = 1, ..., \lceil d_e/h\rceil - 1 \end{cases} \tag{2.22}$$

and

$$P(T_{II} = k) = \begin{cases} P(T_{bGII} = k), & k = 1, ..., l-1 \\ \sum_{i=l}^{\infty} P(T_{bGII} = i) & k = l \end{cases}. \tag{2.23}$$

Proof. (2.23) is obvious. In a manner analogous to that used in Proposition 1, the last part (line) of equation (2.22) follows. Indeed if $k \in \{1, ..., \lceil d_e/h \rceil - 1\}$, then the event $\{T_{GII}\}$ cannot occur since the support of T_{GII} is $S_{T_{GII}} = \{\lceil d_e/h \rceil, \lceil d_e/h \rceil + 1, ...\}$. Hence for $k \in \{1, ..., \lceil d_e/h \rceil - 1\}$, the event $(T_{bGI} = k)$ can occur only if the event $\left\{ \sum_{i=1}^{k-1} V_{(i)} \leq b < \sum_{i=1}^{k} V_{(i)} \right\}$ occurs, where $\sum_{i=1}^{k-1} V_{(i)}$ is void if $k = 1$. Since $P(V_{(i)} \leq v) = G(v) = F^h(v)$, and the $V_{(i)}$'s are independent, we obtain the last part of equation (2.22).

Now let $k \in \{d_e, d_e + 1, ...\}$, then relation (2.10) holds with T_{GII},T_{bII} and T_{bGII} replacing T_{GI},T_{bI} and T_{bGI}, respectively. Accordingly,

$$(T_{bGII} = k) = (T_{GII} = k < T_{bII}) \cup (T_{GII} = k = T_{bII}) \cup (T_{GII} > k = T_{bII}). \tag{2.24}$$

Now the respective first term on the right hand side of (2.24) can be written as:

$$\begin{aligned} P(T_{GII} = k < T_{bII}) = \; & P\Bigg(\bigcup_{j=0}^{h-1} \left[\{ \sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j \} \cap \{ h - j \leq Y_{k\cdot} \leq h \} \right] \\ & \cap \{ \sum_{i=1}^{k} V_{(i)} \leq b \} \Bigg) \\ = & \sum_{j=0}^{h-1} P(\sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j) P(h - j \leq Y_{k\cdot} \leq h) \\ & \times P(\sum_{i=1}^{k} V_{(i)} \leq b) \\ = & \sum_{j=0}^{h-1} \binom{h(k-1)}{d_e - h + j} q^{m(d_e - h + j)} (1 - q^m)^{h(k-1) - (d_e - h + j)} \\ & \times \left(\sum_{r=h-j}^{h} \binom{h}{r} q^{rm} (1 - q^m)^{h-r} \right) G^{*k}(b), \end{aligned} \tag{2.25}$$

where in (2.25) we have used the following facts: The $Y_{i\cdot}$'s are mutually independent and independent of the $V_{(i)}$'s, $Y_{i\cdot} \sim B(h, q^m)$, $\sum_{i=1}^{k-1} Y_{i\cdot} \sim B(h(k-1), q^m)$, and $V_{(i)} \sim G = F^h$. Similarly, the probability of the second term on

the right hand side of (2.24) is

$$
\begin{aligned}
& P(T_{GII} = k = T_{bII}) \qquad (2.26)\\
&= P\Bigg(\Bigg[\bigcup_{j=0}^{h-1} \{ \sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j \} \cap \{ h - j \le Y_{k\cdot} \le h \} \Bigg] \\
&\qquad \cap \{ \sum_{i=1}^{k-1} V_{(i)} \le b \le \sum_{i=1}^{k} V_{(i)} \} \Bigg) \\
&= \sum_{j=0}^{h-1} P(\sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j) P(h - j \le Y_{k\cdot} \le h) \\
&\quad \times P(\sum_{i=1}^{k-1} V_{(i)} \le b \le \sum_{i=1}^{k} V_{(i)})) \\
&= \sum_{j=0}^{h-1} \binom{h(k-1)}{d_e - h + j} q^{m(d_e - h + j)} (1 - q^m)^{h(k-1) - (d_e - h + j)} \\
&\quad \times \left(\sum_{r=h-j}^{h} \binom{h}{r} q^{rm} (1 - q^m)^{h-r} \right) \left[G^{*(k-1)}(b) - G^{*k}(b) \right] .
\end{aligned}
$$

Finally, the probability of the last term on the right hand side of (2.24) has the form

$$
\begin{aligned}
P(T_{GII} > k = T_{bII}) &= P(\sum_{i=1}^{k} Y_{i\cdot} \le d_e - 1) P(\sum_{i=1}^{k-1} V_{(i)} \le b \le \sum_{i=1}^{k} V_{(i)}) \qquad (2.27)\\
&= \sum_{j=0}^{d_e - 1} \binom{hk}{j} q^{mj} (1 - q^m)^{hk - j} \left[G^{*(k-1)}(b) - G^{*k}(b) \right] , \qquad (2.28)
\end{aligned}
$$

which completes the proof. □

The stopping time T_{II} counts the number of group tests conducted during the testing process. In a manner similar to defining D_I and R_I for Model I (see (2.6) and (2.7)), we define for Model II, D_{II} and R_{II} as the respective number of good items identified during the testing process and the remaining number of clean groups of size m required to fulfill the demand requirement. These are clearly given by

$$
R_{II} = d_e - \sum_{i=1}^{T_{II}} Y_{i\cdot} \quad . \qquad (2.29)
$$

3 Objective functions and constraints

For simplicity we formulate several optimization problems only for Model II since Model I is just the special case $h = 1$. In Assumption (v) of Section 1 we have assumed that the cost for testing groups of size m by one machine is given by $c_1 + c_2 m$. Let P be the price of running one machine. The cost for purchasing an item from the contaminated population is b_1. Hence the total cost of the group testing process is

$$C_G = b_1 lm + (c_1 + c_2 m) h T_{II} + hP + b_2 R_{II}, \tag{3.1}$$

which is composed of the cost of (i) purchasing $N = lm$ units from the contaminated population, (ii) running the group testing process, (iii) using h machines, and (iv) purchasing missing items at higher acquisition cost. The expected value of C_G depends on the decision variables m, l and h; therefore we denote it by

$$g(m,l,h) = E(C_G) = b_1 lm + (c_1 + c_2 m) h E(T_{II}) + hP + b_2 E(R_{II}). \tag{3.2}$$

We denote by $e(m,l,h)$ the expected value of T_{II}:

$$e(m,l,h) = E(T_{II}) = \sum_{k=1}^{l-1} k P(T_{bGII} = k) + l \left(1 - \sum_{k=1}^{l-1} P(T_{bGII} = k) \right). \tag{3.3}$$

An expression for $P(T_{bGII} = k)$ is given by (2.22).

We define the stopping time

$$T_{lbII} = \min \{ T_{bII}, l \}, \tag{3.4}$$

whose distribution is clearly given by

$$\begin{aligned} & P\left(T_{lbII} = k\right) \\ = & \begin{cases} P(T_{bII} = k), & k = 1, ..., l-1 \\ 1 - \sum_{i=1}^{l-1} P(T_{bII} = i), & k = l \end{cases} \\ = & \begin{cases} \left[G^{*(k-1)}(b) - G^{*k}(b) \right], & k = 1, ..., l-1 \\ 1 - \sum_{i=1}^{l-1} \left[G^{*(i-1)}(b) - G^{*i}(b) \right], & k = l. \end{cases} \end{aligned} \tag{3.5}$$

The first constraint that we impose is

$$p_1 = P\left(T_{GII} \leq T_{lbII}\right) \geq 1 - \alpha \tag{3.6}$$

for some preassigned value $\alpha \in (0,1)$, where in practice α is of course small. This constraint assures that, with reliability of at least $100(1-\alpha)\%$, the demand requirement is met before the deadline is reached or no items are left for group testing. Since T_{GII} is independent of T_{bII} and therefore independent of T_{lbII}, the probability of the event in (3.6) is given by

$$\begin{aligned}
& P\left(T_{GII} \leq T_{lbII}\right) \\
= & \sum_{k=d_e}^{l} \sum_{r=k}^{l} P(T_{GII} = k) P(T_{lbII} = r) \\
= & \sum_{k=d_e}^{l} \sum_{r=k}^{l} \left(\sum_{j=0}^{h-1} P(\sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j) P(h - j \leq Y_{k\cdot} \leq h) \right) P(T_{lbII} = r) \\
= & \sum_{k=d_e}^{l} \left(\sum_{j=0}^{h-1} P(\sum_{i=1}^{k-1} Y_{i\cdot} = d_e - h + j) P(h - j \leq Y_{k\cdot} \leq h) \right) \\
& \times \left(\sum_{r=k}^{l-1} P(T_{bII} = r) + \left(1 - \sum_{i=1}^{l-1} P(T_{bII} = i) \right) \right) \\
= & \sum_{k=d_e}^{l} \left(\sum_{j=0}^{h-1} \binom{h(k-1)}{d_e - h + j} q^{m(d_e - h + j)} (1 - q^m)^{h(k-1) - (d_e - h + j)} \right. \\
& \times \left. \left(\sum_{r=h-j}^{h} \binom{h}{r} q^{rm} (1 - q^m)^{h-r} \right) \right) \\
& \times \left[\sum_{r=k}^{l-1} \left(G^{*(r-1)}(b) - G^{*r}(b) \right) + \left(1 - \sum_{i=1}^{l-1} \left(G^{*(i-1)}(b) - G^{*i}(b) \right) \right) \right], \qquad (3.7)
\end{aligned}$$

where in (3.7) we have used (2.26).

As a second constraint we consider

$$p_2 = P\left(b_1 lm + (c_1 + c_2) T_{II} \leq C\right) \geq 1 - \beta, \tag{3.8}$$

for some preassigned $\beta \in (0,1)$. This constraint implies that, with reliability of at least $100(1-\beta)\%$, the total cost incurred by the group testing process does not exceed the preassigned budget limit C.

We now formulate two optimization problems. The decision variables are the group size m, the number l of possible groups of size m (or equivalently,

the number of groups of size m that should be purchased), and the number of machines h that should be used simultaneously at any stage of the group testing process. The parameters involved in these optimization problems are $c_1, c_2, b_1, b_2, C, b, \alpha, \beta, m_0, H$, and q.

Problem 1. The objective is to minimize the expected total cost incurred by the group testing process subject to two essential constraints (3.12) and (3.13) below. Accordingly, the optimization problem is given by

$$\min_{m,l,h} g(m,l,h) \tag{3.9}$$

subject to

$$\frac{N}{m} = l > d_e = \frac{d}{m}, \quad l \in \mathbb{N}, d_e \in \mathbb{N}, \tag{3.10}$$

$$1 \leq m \leq m_0, \quad 1 \leq h \leq H, \quad m \in \mathbb{N}, h \in \mathbb{N}, \tag{3.11}$$

$$p_1 = P\left(T_{GII} \leq T_{lbII}\right) \geq 1 - \alpha, \tag{3.12}$$

$$p_2 = P\left(b_1 lm + (c_1 + c_2 m) T_{II} \leq C\right) \geq 1 - \beta. \tag{3.13}$$

Problem 2. The objective is to minimize the expected number of group tests required in the testing process and is subject to the same constraints as in Problem 1:

$$\min_{m,l,h} e(m,l,h) \tag{3.14}$$

subject to constraints (3.10), (3.11), (3.12), and (3.13).

Obviously, our group testing model gives rise to several other meaningful optimization problems. Using the explicit expressions derived above they can be solved numerically. In our final section we discuss some examples.

4 Numerical analysis

In the numerical examples of this section we take the processing times as exponentially distributed random variables with mean 1. This assumption

facilitates (indeed trivializes) the handling of convolutions. Moreover, we only consider the case $H = 1$ of one machine. The population size N is assumed to be fixed so that the only decision variable is the group size m. Of course, various other optimization problems can be treated numerically in the same way, and the sensitivity analysis can be conducted similarly. .

First we let the demand requirement d vary and fix the other parameters as follows:

$N = 720$	(population size)
$q = 0.99$	(probability of an item to be nondefective)
$C = 1500$	(budget limit)
$b = 100$	(deadline)
$c_1 = 5, c_2 = 2$	($c_1 + c_2 m$ being the cost of a group test of size m)
$b_1 = 1$	(price per item in the population)
$b_2 = 11$	(price per good item from a secure source)

In each of the subsequent Tables 1-4 another value of d is chosen. They give, for the possible group sizes m, the corresponding values of

- l and d_e (the maximum and the minimum number of performed tests),
- p_1 and p_2 (the probabilities in the constraints (3.12) and (3.13)),
- and the objective functions g and $E(T_{II})$.

Table 1 is for $d = 120$. It shows that for this demand requirement g is unimodal with the minimum of g attained for the group size $m^* = 15$. The two constraints on p_1 and p_2 are satisfied for very small values of α and β, namely for $\alpha = 10^{-7}$, $\beta = 10^{-7}$. (Note that the probabilities given as 1.000000 are not exactly equal to one, as there is always a positive probability that the demand cannot be met.)

m	ℓ	d_e	p_1	p_2	g	$E(T_{II})$
1	720	120	0.0000000	1.0000000	1427.00	101.00
2	360	60	1.0000000	1.0000000	1270.96	61.22
3	240	40	1.0000000	1.0000000	1173.47	41.22
4	180	30	1.0000000	1.0000000	1126.00	31.23
5	144	24	1.0000000	1.0000000	1098.55	25.24
6	120	20	1.0000000	1.0000000	1081.13	21.24
8	90	15	1.0000000	1.0000000	1061.37	16.26
10	72	12	1.0000000	1.0000000	1051.72	13.27
12	60	10	1.0000000	1.0000000	1047.17	11.28
15	48	8	1.0000000	1.0000000	1045.56	9.30
20	36	6	1.0000000	0.9999967	1050.11	7.34
24	30	5	1.0000000	0.9999137	1057.29	6.36
30	24	4	1.0000000	0.9994565	1071.49	5.41
40	18	3	0.9999985	0.9920668	1101.18	4.48
60	12	2	0.9988474	0.9288587	1176.63	3.65
120	6	1	0.8817248	0.6560884	1441.57	2.95

Table 1: $d = 120$; minimum at $m^* = 15$ yielding $g^* = 1045.56$.

In Table 2 we consider $d = 240$. Again $m^* = 15$ is the optimal group size, as long as one is satisfied with $p_1 \geq 1 - 10^{-7}$ and $p_2 \geq 0.974$. If a stronger constraint on p_2 is required, one has to choose a smaller group size.

m	ℓ	d_e	p_1	p_2	g	$E(T_{II})$
1	720	240	0.0000000	1.0000000	1427.00	101.00
2	360	120	0.0000000	0.0000000	1629.00	101.00
3	240	80	1.0000000	0.0000000	1626.94	82.45
4	180	60	1.0000000	0.0896286	1532.00	62.46
5	144	48	1.0000000	0.8896824	1477.11	50.47
6	120	40	1.0000000	0.9538497	1442.27	42.49
8	90	30	1.0000000	0.9935817	1402.75	32.51
10	72	24	1.0000000	0.9926342	1383.44	26.54
12	60	20	1.0000000	0.9771914	1374.34	22.56
15	48	16	1.0000000	0.9738097	1371.12	18.60
16	45	15	1.0000000	0.9726243	1371.82	17.62
20	36	12	1.0000000	0.9271561	1380.22	14.67
24	30	10	1.0000000	0.8374891	1394.58	12.73
30	24	8	0.9999944	0.8203720	1422.99	10.82
40	18	6	0.9992090	0.6559637	1482.24	8.97

48	15	5	0.9936655	0.4580517	1536.83	8.09
60	12	4	0.9628223	0.4357783	1623.14	7.23
80	9	3	0.8469702	0.2381818	1753.24	6.26
120	6	2	0.5784851	0.2152198	1914.97	4.88

Table 2: $d = 240$; minimum at $m^* = 15$ yielding $g^* = 1371.12$.

In Tables 3 and 4 we take $d = 360$ and $d = 600$ and, for the higher demands, also larger budget limits $C = 1900$ and $C = 2500$, respectively. From now on only those values of m for which $p_1 \geq 0.8$ and $p_2 \geq 0.8$ are displayed. The optimal group sizes for $d = 360$ and $d = 600$ are $m^* = 15$ and $m^* = 12$, respectively.

m	ℓ	d_e	p_1	p_2	g	$E(T_{II})$
5	144	72	1.0000000	0.9121741	1855.66	75.71
6	120	60	1.0000000	0.9932918	1803.40	63.73
8	90	45	1.0000000	0.9989887	1744.12	48.77
9	80	40	1.0000000	0.9988748	1727.10	43.79
10	72	36	1.0000000	0.9987512	1715.15	39.81
12	60	30	1.0000000	0.9959470	1701.51	33.85
15	48	24	1.0000000	0.9877209	1696.68	27.91
18	40	20	0.9999998	0.9680841	1702.61	23.97
20	36	18	0.9999969	0.9651321	1710.33	22.01
24	30	15	0.9998742	0.9205772	1731.85	19.09
30	24	12	0.9969061	0.8362332	1774.12	16.22

Table 3: $d = 360$, $C = 1900$; minimum at $m^* = 15$ yielding $g^* = 1696.68$.

m	ℓ	d_e	p_1	p_2	g	$E(T_{II})$
8	90	75	0.9986095	0.8869988	2426.82	81.28
10	72	60	0.9817695	0.9637853	2377.75	66.31
12	60	50	0.9264780	1.0000000	2351.44	56.26

Table 4: $d = 600$, $C = 2500$; minimum at $m^* = 12$ yielding $g^* = 2351.44$.

The value $m = 15$ actually turned out to be optimal in several variations of N and d with N/d kept fixed. In Table 5 we give the feasible results for $N = 480$ and $d = 240$, i.e. $N/d = 2$, and $C = 2000$. The same robustness could be seen if b was varied.

m	ℓ	d_e	p_1	p_2	g	$E(T_{II})$
3	160	80	1.0000000	1.0000000	1386.94	82.45
4	120	60	1.0000000	1.0000000	1292.00	62.46
5	96	48	1.0000000	1.0000000	1237.11	50.47
6	80	40	1.0000000	1.0000000	1202.27	42.49
8	60	30	1.0000000	1.0000000	1162.75	32.51
10	48	24	1.0000000	1.0000000	1143.44	26.54
12	40	20	1.0000000	1.0000000	1134.34	22.56
15	32	16	0.9999998	1.0000000	1131.12	18.60
16	30	15	0.9999990	1.0000000	1131.82	17.62
20	24	12	0.9999203	1.0000000	1140.22	14.67
24	20	10	0.9989552	1.0000000	1154.49	12.73
30	16	8	0.9901037	1.0000000	1181.87	10.80
40	12	6	0.9358330	1.0000000	1231.40	8.84
48	10	5	0.8612057	1.0000000	1266.67	7.79

Table 5: $N = 480$, $d = 240$, $C = 2000$; minimum at $m^* = 15$ yielding $g^* = 1131.118600$.

On the other hand, the optimal m^* is highly sensitive to changes of q. An example is shown in Table 6, where the optimal values of m and the objective function g are given for several values of q approaching 1. It is seen that m^* increases from 6 to 180.

q	m^*	g^*
0.9500	6	2107.551
0.9990	45	1514.999
0.9995	72	1492.316
0.9999	180	1463.251

Table 6: $N = 720$, $d = 360$, $C = 2200$. Optimal values of m and the corresponding minimal values of the objective function for $q1$ approaching 1.

Acknowledgement: S. Bar-Lev was partially supported by NWO grant no. B61-493. The authors would like to thank Andreas H. Löpker for his help in the numerical analysis.

References

[1] Ding-Zhu, Du and Hwang, F.K.(2000). *Combinatorial Group Testing and Its Applications* (2nd ed), Singapore: World Scientific.

[2] Hung, M.C.(2000). "Use of binomial group testing in tests of hypotheses for classification or quantitative covariables", *Biometrics* 56, 204-212.

[3] Macula, A.J.(1999a). "Probabilistic nonadaptive group testing in the presence of errors and DNA library screening", *Ann. Comb.* 3, 61-69.

[4] Macula, A.J.(1999b). "Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening", *J. Comb. Optim.* 2, 385-397.

[5] Wein, L.M. and Zenios, S.A.(1996). "Pooled testing for HIV screening: capturing the dilution effect", *Oper. Res.* 44, 543-569.

[6] Litvak, E., Tu, X.M. and Pagano,M.(1994). "Screening for the presence of a disease by pooling sera samples", *J. Amer. Statist. Assoc.* 89, 424-434.

[7] Dorfman, R.(1943). "The detection of defective members of large populations", *Ann. Math. Statist.* 14, 436-440.

[8] Hammick, P.A. and Gastwirth, J.L.(1994). "Group testing for sensitive characteristics: extensions to higher prevalence", *Int. Statist. Rev.* 62, 319-331.

[9] Tu, X.M., Litvak, E. and 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.

[10] Bar-Lev, S.K., Boneh, A. and Perry, D.(1990). "Incomplete identification models for group-testable items", *Naval Res. Logist.* 37, 647-659.

[11] Hwang, F.K., Pfeifer, C.G. and Enis, P.(1981). "An optimal Hierarchical procedure for a modified binomial group-testing problem", *J. Amer. Statist. Assoc.* 76, 947-949.