Group Testing Procedures with Quantitative Features and Incomplete Identification

Shaul K. Bar-Lev[*], Onno Boxma[†],
Andreas Löpker[‡], Wolfgang Stadje[§] and Frank A. Van der Duyn Schouten[¶]

Abstract

We present a group testing model for items characterized by marker random variables. An item is defined to be good (bad) if its marker is below (above) a given threshold. The items can be tested in groups; the goal is to obtain a prespecified number of good items by testing them in optimally sized groups. Besides this group size, the controller has to select a threshold value for the group marker sums, and the target number of groups which by the tests are classified to consist only of good items. These decision variables have to be chosen so as to minimize a cost function, which is a linear combination of the expected number of group tests and an expected penalty for missing the desired number of good items, subject to constraints on the probabilities of misclassifications. We study two models of this kind: the first one is based on an infinite population size, while the second one is a two-stage model for a finite number of available items. All cost functionals are derived in closed form and bounds and approximations are also given. In several examples the dependence of the cost function on the decision variables is studied.

[*] Department of Statistics, University of Haifa, Haifa 31905, Israel (barlev@stat.haifa.ac.il)

[†] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[‡] Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (lopker@eurandom.tue.nl)

[§] Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany (wolfgang@mathematik.uos.de)

[¶] Center for Economic Research, Tilburg University, 5000 LE Tilburg, The Netherlands (f.a.vdrduynschouten@uvt.nl)

1 Introduction

Group testing, i.e., the use of procedures based on pooled samples, is often a cost-efficient technique, provided the screening can be designed so as to provide test results with sufficiently high accuracy, sensitivity and specificity. The objective is to classify the items of some finite population according to certain categories, one of which may be called 'good' or 'clean' and one or more others are 'defective' or 'contaminated'. The basic idea of group testing is to conduct the tests using pooled samples. While good groups are considered to consist only of clean samples, those classified differently either have to be subject to further screening or have to be scrapped. Employing suitably designed procedures of this kind leads to a significant reduction of the number of required tests and thus of screening cost, under controlled probabilities of misclassifications.

In [5] it was proposed to classify group testing models according to the following five dichotomies: (i) probabilistic versus combinatorial; (ii) complete versus incomplete identification; (iii) reliable versus unreliable testing; (iv) binomial versus multinomial; (v) time constraints versus arbitrary processing times. As these features can be combined freely, this leads to 32 possible types of basic group testing models. Let us in particular discuss (ii).

The objective of complete identification is a correct classification of the whole population into good or defective items via repeated group testing; the main goal is to find optimal pooling policies in order to minimize the expected number of required group tests. However, for reasonably large population sizes no optimal policies have been found; only suboptimal policies have been suggested. For a thorough survey of group testing models with complete identification the reader is referred to the monograph [8] (and the references cited therein).

Incomplete identification means that the population is not necessarily exhaustively examined until all defective items are identified. Often the testing process serves the goal of meeting some prespecified demand requirement for good items so that testing is terminated once this objective has been reached. Accordingly, groups which have been declared to be clean are aggregated for meeting the demand requirement, while contaminated groups are set aside (but perhaps recorded for other possible uses). These models lead to optimal stopping rules and optimization problems under constraints; see [3, 4, 5, 6] where some of the combinations of the features mentioned above are dealt with.

The question of how to proceed with groups that are found contaminated depends on various aspects. In many medical applications retesting of all items in contaminated groups is called for because the aim is to establish a diagnosis for all patients involved. However, in many industrial applications as well as in blood screening in blood banks, the further processing of contaminated groups heavily depends on various retesting costs. There may also be a residual economic value, however reduced, to items belonging to contaminated groups. Accordingly, group testing procedures for incomplete identification are called for when the objective is purely economic (profit-raising or cost-decreasing); then they have to reflect the underlying profit and cost functionals.

In this paper we add one more dichotomy to the ones listed above, namely: (vi) quantitative versus qualitative. Many tests (in medical as well as in industrial applications) provide not only a qualitative result (i.e., whether a sample is contaminated or not) but also give a quantitative value, for example the continuous measurement of some marker. An item is classified as high (positive) or low (negative) risk according to whether the corresponding marker value is greater or less than a certain threshold (cut-off value). Associated with a threshold is then the probability of a true positive (i.e., the sensitivity) and the probability of a true negative (i.e., the specificity). The effectiveness of continuous diagnostic markers in distinguishing between low and high risk populations is well studied in the biostatistical and medical literature (e.g. [9, 10]).

To think of a concrete example, consider samples of well-water which are collected in small, sterile bottles and taken to a laboratory to be tested for bacterial contamination. Small amounts of the water are pooled and then cultivated in a special dish; after a predetermined cultivation period the number of bacteria colonies is counted. If this number exceeds a prespecified acceptance level, the pooled water sample is denoted as 'contaminated'.

In our model we make the simplifying assumption that any single item can be classified as being good or deficient with complete certainty by measuring its marker, so that an item is good if and only if its marker does not exceed the given threshold. However, for pooled samples a new threshold has to be determined, depending on the group size, such that the probabilities of misclassifications are sufficiently small.

In Section 2 we present the above model in detail, describe its relevant features and assumptions, formulate the objective functions together with the associated constraints, and derive explicit analytic formulas as well as bounds and approximations for the functionals. The analysis leads to an optimization problem under constraints. In Section 3 we consider the analogous model with a finite population size. We propose a two-stage policy in which the groups accepted in the first stage are supplemented by the 'best' groups among those that have not been selected before. Again the objective function can be determined in closed form. Section 4 is devoted to several examples; we use simulation to study the dependence of the objective function on the decision variables involved. Further possible extensions are discussed in Section 5.

2 Grouping in a quantitative model

We first describe the model in detail. We formulate an expected cost minimization problem subject to probabilistic constraints. Three decision variables (group size, threshold value for group tests and the parameter of the natural family of stopping rules) have to be determined so as to minimize the expected cost.

2.1 Model description

We are given a virtually infinite population whose members (called items) can be classified into two categories: good or defective. Each item is assumed to contain a random number of certain particles (e.g., antibodies) and there exists an accepted threshold t such that an item is classified as 'good' if the number of particles it contains does not exceed t, and 'defective' otherwise. We also assume that items are independent of each other. Let X be a generic random variable (called marker) which denotes the number of particles in an item; its distribution is assumed to be known (as is the case in most biostatistical and medical studies). Let $p = \mathbb{P}(X > t)$ be the known proportion of defective items and set $q = 1 - p$.

We assume that a prespecified demand for d good items has to be satisfied. The aggregation of good items is conducted successively via grouping of the population in groups of size m, our first decision variable. We only consider group sizes that divide d. A group which is found good is kept and recorded for meeting the demand requirement while a contaminated group is put aside but may be recorded for other possible uses.

We denote by X_i the marker random variable counting the number of particles in item i. By the meaning of the threshold t, an individual item i is defined to be 'good' if $X_i \leq t$. Of course, if $\sum_{i=1}^{m} X_i \leq mt$, there is no guarantee that $X_i \leq t$ for each of the items $i = 1, \ldots, m$. It seems much more reasonable to take a smaller threshold $s = t(m)$ depending on m, where $t(m) < mt$ for $m \geq 2$. The choice of $t(m)$ is not simple and one can use various criteria. For example, one may employ the Youden index (cf. [11]) which maximizes (sensitivity+specificity-1) over all threshold values. Our objective is to choose the decision variable s in a cost-minimizing way subject to certain constraints. When group testing m items, there are two possibilities of undesired classifications:

- At least one of the items in the considered group has a marker greater than t but the group is declared to be good; this has conditional probability

$$p_1(m,s) = \mathbb{P}\left(\max_{i=1,\dots,m} X_i > t \mid \sum_{i=1}^{m} X_i \leq s\right). \tag{2.1}$$

- None of the items has a marker greater than t but the group is declared to be contaminated; this has conditional probability

$$p_2(m,s) = \mathbb{P}\left(\max_{i=1,\dots,m} X_i \leq t \mid \sum_{i=1}^{m} X_i > s\right). \tag{2.2}$$

We want these probabilities to be small: $p_1(m,s) \leq \varepsilon_1$ and $p_2(m,s) \leq \varepsilon_2$ for certain prespecified $\varepsilon_i \in (0,1)$.

2.2 The stopping time and the cost functionals

Now assume that independent groups of size m are tested successively and s is the selected threshold value. Define

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

Then $Y_i \sim B(1,\rho)$, where

$$\rho = \rho(m,s) = \mathbb{P}\left(\sum_{i=1}^{m} X_i \leq s\right). \tag{2.3}$$

In order to compute ρ we let A_j denote the event that exactly j of the m items are good, $j = 0,\dots,m$. Then $\mathbb{P}(A_j) = \binom{m}{j} q^j p^{m-j}$ and, by symmetry,

$$\mathbb{P}\left(\sum_{i=1}^{m} X_i \leq s \mid A_j\right) = \mathbb{P}\left(\sum_{i=1}^{m} X_i \leq s \mid X_1\dots,X_j \leq t \text{ and } X_{j+1},\dots,X_m > t\right), \tag{2.4}$$

so that we get

$$\begin{aligned}\rho &= \sum_{j=0}^{m} \mathbb{P}\left(\sum_{i=1}^{m} X_i \leq s \mid A_j\right)\binom{m}{j} q^j p^{m-j} \\ &= \sum_{j=0}^{m} \binom{m}{j} \mathbb{P}\left(\sum_{i=1}^{m} X_i \leq s,\ X_1\dots,X_j \leq t \text{ and } X_{j+1},\dots,X_m > t\right).\end{aligned} \tag{2.5}$$

We want to obtain d good items, so that the number of group tests we have to conduct is at least $\inf\{n \mid \sum_{j=1}^{n} Y_j = d/m\}$. We propose to consider the stopping rules

$$T(m,s,c) = \inf\{n \mid \sum_{j=1}^{n} Y_j = c\}, \quad c = d/m, d/m+1, \ldots$$

Our model thus contains three decision variables m, s, c, where

- m is the group size, a divisor of d;

- s is the threshold value for the sum of the markers in each tested group;

- c is the number of groups classified as good after which testing is stopped.

Costs are incurred due to the conducted number of group tests and a penalty in the case that the goal of obtaining d good items is missed. Let $Z(m,s,c)$ be the (random!) number of good items among the ones classified as good and let $a > 0$ be the penalty per missing item. Then the cost function is composed of the following ingredients:

- $\mathbb{E}(T(m,s,c))$, the expected number of group tests;

- $\mathbb{E}(a(d - Z(m,s,c))^+)$, the expected total penalty.

Thus, we deal with the following *optimization problem*:

$$\text{Minimize} \ \ \mathbb{E}[T(m,s,c)] + a\mathbb{E}[(d - Z(m,s,c))^+] \tag{2.6}$$

$$\text{subject to} \ \ p_1(m,s) \le \varepsilon_1, \ \ p_2(m,s) \le \varepsilon_2. \tag{2.7}$$

The distribution of $T(m,s,c)$ is negative binomial and the associated parameter ρ has been computed in (2.5) so this distribution and its expected value are available in closed form:

$$\mathbb{P}(T(m,s,c) = c+k) = \binom{c+k-1}{k}(1-\rho)^k \rho^c, \quad k = 0,1,2,\ldots \tag{2.8}$$

$$\mathbb{E}[T(m,s,c)] = \frac{c}{\rho}. \tag{2.9}$$

To compute $\mathbb{E}((d - Z(m,s,c))^+)$, let W_i be the number of good items in the ith group if it has been classified as good; otherwise set $W_i = 0$. Then

$$\mathbb{P}(Z(m,s) = l) = \sum_{k=0}^{\infty} \mathbb{P}(T(m,s,c) = c+k) \times \mathbb{P}\left(\sum_{i=1}^{c+k} W_i = l \mid T(m,s,c) = c+k\right). \tag{2.10}$$

Let $\mu_{m,s} = \mathbb{P}_{W_1 | \sum_{i=1}^m X_i \leq s}$ be the conditional distribution of W_1, given that the first group has been accepted. We have

$$\mu_{m,s}(j) = \binom{m}{j} \mathbb{P}\left(\max_{1\leq i\leq j} X_i \leq t, \min_{j<i\leq m} X_i > t \mid \sum_{i=1}^m X_i \leq s \right). \tag{2.11}$$

The condition $T(m,s,c) = c+k$ means that there are exactly c groups among the first $c+k$ ones that are classified as good. Therefore, $\mathbb{P}(\sum_{i=1}^{c+k} W_i = l \mid T(m,s,c)) = c+k)$ is equal to $\mu_{m,s}^{*c}(l)$, where $\mu_{m,s}^{*c}$ denotes the $c$fold convolution of $\mu_{m,s}$ with itself. Note that this probability is independent of k. It thus follows from (2.10) that

$$\mathbb{P}(Z(m,s) = l) = \mu_{m,s}^{*c}(l). \tag{2.12}$$

Eq. (2.12) yields the second cost functional:

$$\mathbb{E}[(d - Z(m,s,c))^+] = \sum_{l=0}^{d-1} (d-l)\mu_{m,s}^{*c}(l). \tag{2.13}$$

The convolution probabilities in (2.13) have to be computed from (2.11).

The constraint probabilities $p_1(m,s)$ and $p_2(m,s)$ are defined in (2.1)–(2.2) in terms of the underlying distribution of the X_i and thus also known.

2.3 Integral expressions

Analytic formulas are available for all quantities in the optimization problem. Let F be the distribution function of X and let

$$I_{m,j}(s,t) = \int \dots \int_{0\leq x_1,\dots,x_j \leq t,\ t<x_{j+1},\dots,x_m\leq s,\ x_1+\dots+x_m\leq s} dF(x_1)\dots dF(x_m),$$
$$j = 0,\dots,m.$$

Note that in our model t is fixed and that

$$I_{m,0}(s,0) = F^{*m}(s).$$

Then we have

$$p_1(m,s) = 1 - \frac{I_{m,m}(s,t)}{I_{m,0}(s,0)}, \tag{2.14}$$

$$p_2(m,s) = \frac{F(t)^m - I_{m,m}(s,t)}{1 - I_{m,0}(s,0)}, \tag{2.15}$$

$$\rho = \sum_{j=0}^{m} \binom{m}{j} I_{m,j}(s,t), \tag{2.16}$$

$$\mu_{m,s}^{*c}(l) = \frac{1}{I_{m,0}(s,0)^c} \sum_{0 \le j_1,\ldots,j_c \le l,\ j_1+\ldots+j_c=l} \binom{m}{j_1} \cdots \binom{m}{j_c} \times I_{m,j_1}(s,t) \ldots I_{m,j_c}(s,t), \tag{2.17}$$

$$\mathbb{E}[(d - Z(m,s,c))^+] = \sum_{l=0}^{d-1} (d-l) \frac{1}{I_{m,0}(s,0)^c} \sum_{0 \le j_1,\ldots,j_c \le l, j_1+\ldots+j_c=l} \binom{m}{j_1} \cdots \binom{m}{j_c} \times I_{m,j_1}(s,t) \ldots I_{m,j_c}(s,t). \tag{2.18}$$

Since $\mathbb{E}(T(m,s,c)) = c/\rho$, all functionals in our optimization problem can be written in terms of the integrals $I_{m,j}(s,t)$ by means of (2.14)-(2.18).

2.4 Bounds and approximations

We now derive bounds and approximations for some of the quantities required in the optimization problem. We first establish bounds for the probabilities p_1 and p_2 of (2.1) and (2.2). Define the two functions

$$\begin{aligned} f(x_1, x_2, \ldots, x_m) &= \mathbb{1}\{\sum_{i=1}^{m} x_i \le s\} \\ g(x_1, x_2, \ldots, x_m) &= \mathbb{1}\{\min_{i=1,\ldots,m} x_i \le t\}. \end{aligned}$$

Since f and g are non-increasing, it follows that the random variables $f(X_1, \ldots, X_m)$ and $g(X_1, \ldots, X_m)$ are positively correlated (see for instance [12]) so that

$$\mathbb{E}[f(X_1, \ldots, X_m) g(X_1, \ldots, X_m)] \ \ge \ \mathbb{E}[f(X_1, \ldots, X_m)] \mathbb{E}[g(X_1, \ldots, X_m)].$$

Consequently, after division by $\mathbb{E}[f(X_1, \ldots, X_m)] = \mathbb{P}(\sum_{i=1}^{m} x_i \le s)$,

$$\mathbb{P}(\max_{i=1,\ldots,m} X_i \le t | \sum_{i=1}^{m} X_i \le s) \ge \mathbb{P}(\max_{i=1,\ldots,m} X_i \le t).$$

Similarly, since $1-f$ and $1-g$ are non-decreasing,

$$\mathbb{P}(\max_{i=1,\dots,m} X_i > t | \sum_{i=1}^{m} X_i > s) \geq \mathbb{P}(\max_{i=1,\dots,m} X_i > t).$$

Thus we have proved the two inequalities

$$p_1 \leq 1 - F(t)^m, \quad p_2 \leq F(t)^m. \tag{2.19}$$

It follows for example that with $F(t) = 0.6$ and $m \geq 9$, we already achieve $p_2 \leq 0.01$, so that the constraint $p_2 \leq \varepsilon_2$ is fulfilled in most cases discussed here.

Next we derive an approximation for the expected number $\mathbb{E}[T(m,s,c)]$ of group tests. Recall that

$$\mathbb{E}[T(m,s,c)] = \frac{c}{\mathbb{P}(\sum_{i=1}^{m} X_i \leq s)}.$$

Hence, assuming that m is large enough to imply that $(\sigma\sqrt{m})^{-1}\sum_{i=1}^{m}(X_i - \mu)$ is approximately normally distributed, we arrive at

$$\mathbb{E}[T(m,s,c)] \approx c \Big/ \Phi\left(\frac{s-\mu m}{\sigma\sqrt{m}}\right),$$

where Φ is the standard normal distribution function. To find a bound for $\Phi(x)$, let $a \in (0,2)$ and define for $x < a$ the function

$$G_a(x) = \Phi(x) - \sqrt{\frac{\pi}{2}} \cdot \frac{\Phi'(x)}{a-x}.$$

Then

$$G_a'(x) = \frac{1}{2} e^{-x^2/2} \big(\sqrt{\frac{2}{\pi}} - \frac{1 - x(a-x)}{(a-x)^2} \big).$$

Note that $1 - x(a-x) > 0$ since $a \in (0,2)$. Hence $G_a'(x) \to -\infty$ as $x \uparrow a$ and $G_a'(x) \uparrow 0$ as $x \to -\infty$. The equation $G_a'(x) = 0$ has exactly one solution if

$$\big(\sqrt{\frac{2}{\pi}} - 1\big) y^2 + ya - 1 = 0,$$

has exactly one solution, where $y = a - x$. This is the case if $a = \kappa = 2\sqrt{1 - \sqrt{2/\pi}} \approx 0.899$, and thus $G_\kappa'(x)$ stays non-positive for all $x < \kappa$. Hence $G_\kappa(x)$ is non-increasing and since $\lim_{x \to -\infty} G_\kappa(x) = 0$, it follows that $G_\kappa(x) < 0$ for all $x < \kappa$, which yields

$$\Phi(x) < \sqrt{\frac{\pi}{2}} \cdot \frac{\Phi'(x)}{\kappa - x} = \frac{e^{-x^2/2}}{2(\kappa - x)}. \tag{2.20}$$

We note that for values $x \in (-7/2, 0)$ this bound is better than the classical estimate $\Phi(x) < -\Phi'(x)/x$ for $x < 0$.

It follows that

$$\begin{aligned} \mathbb{E}[T(m,s,c)] &\approx c\Big/\Phi\left(\frac{s-\mu m}{\sigma\sqrt{m}}\right) \\ &> 2c\left(\kappa - \frac{s-\mu m}{\sigma\sqrt{m}}\right)\exp\left\{\frac{1}{2}\left(\frac{s-\mu m}{\sigma\sqrt{m}}\right)^2\right\}. \end{aligned} \tag{2.21}$$

For the penalty term $\mathbb{E}[(d - Z(m,s,c))^+]$ we argue as follows. Let $\widetilde{Z}(m,c)$ denote the number of good items found in c group tests (regardless of their classification). Clearly $\widetilde{Z}(m,c) = \sum_{i=1}^{c}\widetilde{W}_i$, where $\widetilde{W}_i$ denotes the number of good items in the ith group (cf. (2.12)). By a correlation argument as above,

$$\mathbb{P}(W_i \leq k \mid \sum_{i=1}^{m} X_i < s) = \mathbb{P}(\widetilde{W}_i \leq k \mid \sum_{i=1}^{m} X_i < s) \leq \mathbb{P}(\widetilde{W}_i \leq k).$$

It follows that

$$\mathbb{E}[(d - Z(m,s,c))^+] \quad \leq \quad \mathbb{E}[(d - \widetilde{Z}(m,c))^+].$$

Since $\widetilde{W}_i$ has a binomial distribution with parameters m and $F(t) = \mathbb{P}(X_1 \leq t)$, $\widetilde{Z}(m,c)$ has a binomial distribution with parameters mc and $F(t)$. Writing $\tilde{\mu} = cmF(t)$ for its mean and $\tilde{\sigma} = \sqrt{mcF(t)(1-F(t))}$ for its standard deviation, we obtain the approximation

$$\begin{aligned} \mathbb{E}[(d - \widetilde{Z}(m,c))^+] &= \sum_{k=0}^{d}\binom{mc}{k}(d-k)F(t)^k(1-F(t))^{mc-k} \\ &\approx \int_{-\infty}^{d}(d-y)\ \mathrm{d}\Phi(\frac{y-\tilde{\mu}}{\tilde{\sigma}}) \\ &= (d-\tilde{\mu})\cdot\Phi(\frac{d-\tilde{\mu}}{\tilde{\sigma}}) + \frac{1}{\sqrt{2\pi}}\tilde{\sigma}e^{-\frac{(\tilde{\mu}-d)^2}{2\tilde{\sigma}^2}}. \end{aligned}$$

For $d \gg \tilde{\mu}$ this yields the intuitive approximation

$$\mathbb{E}[(d - \widetilde{Z}(m,c))^+] \approx (d-\tilde{\mu})^+ = (d - cmF(t))^+. \tag{2.22}$$

3 A policy in the case of finite population size

The model presented above assumes an infinite population of items. Under this assumption it is possible to achieve or exceed the required number of good items with probability arbitrarily close to 1 by using a stopping rule $T(m, s, c)$ with sufficiently large c. This is not the case if the population size is finite consisting of, say, N items available for grouping and testing. In the following we only consider values of m that are divisors of N. Assume that for a given group size m and threshold s satisfying (2.1)-(2.2) the total number of accepted items, $m \sum_{j=1}^{N/m} Y_j$, after the population has been completely tested in groups has not reached the desired level mc. Then it may be worthwhile to add to these items a few of the groups not accepted so far (because for them the threshold s was surpassed), in order to reach the target value. It is reasonable to take those groups for which (a) the sum of the markers is maximal and (b) the probability of containing a bad item is sufficiently small. This idea leads to the following *two-stage policy.* After fixing the decision variables m and s satisfying (2.1)-(2.2) choose c and use the stopping rule $\min[T(m, s, c), N/m]$. (Note that N/m is the maximum available number of groups of size m.) Next choose a (small) $\delta \geq \varepsilon_1$ as the maximal probability permissible for a group in the second stage to contain a bad item.

If $T(m, s, c) \leq N/m$, the procedure is finished. If $T(m, s, c) > N/m$, consider the K groups not selected so far and denote their marker sums by $S_1, S_2, \ldots, S_K$ (in the order in which the groups were tested). Note that K is a random variable. Now select in addition successively those groups whose marker sums S_i satisfy $f_m(S_i) > \delta$, where $f_m(u)$, $u > 0$, denotes the probability that a group with marker sum u contains a bad item, i.e.,

$$f_m(u) = \mathbb{P}\left(\max_{i=1,\ldots,m} X_i > t \mid \sum_{i=1}^{m} X_i = u \right). \tag{3.1}$$

It is intuitively obvious that the functions f_m are nondecreasing. Indeed, this assertion can be proved by induction on m by conditioning on X_1, yielding

$$f_m(u) = \int_0^u f_{m-1}(u - v) \; \mathbb{P}(X_1 \in dv),$$

so that the monotonicity of f_{m-1} implies that of f_m. Moreover,

$$p_1(m, u) \leq f_m(u),$$

because

$$p_1(m, u) = \int_0^u f_m(v) \; \mathbb{P}\left(\sum_{i=1}^{m} X_i \in dv \mid \sum_{i=1}^{m} X_i \leq u \right) \leq f_m(u).$$

In particular there exists a constant $K \geq s$ such that $f_m(S_i) > \delta$ is equivalent either to $S_i < K$ or to $S_i \leq K$. Note that $K = K(m, \delta)$ is not an independent decision variable but a function of the decision variable m and the prespecified δ.

Summarizing, under the proposed policy one accepts as many groups as possible with marker sum less than s according to the truncated stopping rule $\min[T(m,s,c),N/m]$ and then supplements the set of accepted items by the groups having a marker sum in the interval $[s,K)$. (Alternatively, we could consider the closed interval $[s,K]$.) Note that also under this policy it is possible that there will not be enough selected groups to reach the desired number d of good items. The decision variables have to be chosen so as to keep the error probabilities as small as is specified by the error probability constraints specified by ε_1, ε_2 and δ.

The corresponding objective function (2.8) can be given in closed, albeit intricate form. By (2.10), the expected number of group tests is

$$\begin{aligned}\mathbb{E}[\min[T(m,s,c),N/m]] = &\sum_{k=0}^{(N/m)-c-1}\binom{c+k-1}{k}(c+k)(1-\rho)^k\rho^c\\ &+\frac{N}{m}\sum_{k=(N/m)-c}^{\infty}\binom{c+k-1}{k}(1-\rho)^k\rho^c, \quad (3.2)\end{aligned}$$

where ρ is given by (2.3). To determine the total expected penalty, we have to compute $\mathbb{E}[(d-Z(m,s,c))^+]$, where we again denote by $Z(m,s,c)$ the number of good items among the accepted ones. If n groups are accepted in the first stage, let $U_1,\dots,U_n$ be the successive numbers of good items in these groups, let $V_1,\dots,V_{(N/m)-n}$ be the numbers of good items in the groups not accepted in stage one, and let $S_1,\dots,S_{(N/m)-n}$ be their marker sums. Using conditioning and similar arguments as in Section 2.2 we have

$$\begin{aligned}&\mathbb{P}(Z(m,s)=l)\\ &=\sum_{k=0}^{(N/m)-1}\mathbb{P}(T(m,s,c)=c+k)\mu_{m,s}^{*c}(l)\\ &\quad+\sum_{n=0}^{c-1}\mathbb{P}\Big(\sum_{j=1}^{N/m}Y_j=n\Big)\mathbb{P}\Big(U_1+\ldots+U_n\\ &\qquad\qquad+V_1 1_{\{S_1<K\}}+\cdots+V_{(N/m)-n}1_{\{S_{(N/m)-n}<K\}}=l\mid\sum_{j=1}^{N/m}Y_j=n\Big).\end{aligned} \tag{3.3}$$

Conditional on $\sum_{j=1}^{N/m}Y_j=n$, the random variables

$$U_1,\dots,U_n,V_1 1_{\{S_1<K\}},\dots,V_{(N/m)-n}1_{\{S_{(N/m)-n}<K\}}$$

are independent, $U_1, \dots, U_n$ have the common distribution

$$\mu_{m,s} = \mathbb{P}_{W_1 | \sum_{i=1}^m X_i \le s},$$

which is given by (2.13), and $V_1 1_{\{S_1 < K\}}, \dots, V_{(N/m)-n} 1_{\{S_{(N/m)-n} < K\}}$ all have the distribution $\nu_{m,s,K}$ given by

$$\begin{aligned}\nu_{m,s,K}(j) &= \mathbb{P}\Big(V_1 1_{\{S_1<K\}} = j \mid \sum_{j=1}^{N/m} Y_j = n\Big) \\ &= \binom{m}{j} \mathbb{P}\left(\max_{1\le i\le j} X_i \le t, \min_{j<i\le m} X_i > t, \sum_{i=1}^m X_i < K \mid \sum_{i=1}^m X_i \ge s\right), \\ &\qquad\qquad j = 0, \dots, m.\end{aligned}$$

It now follows from (3.3) that

$$\begin{aligned}\mathbb{P}(Z(m,s) = l) = \mathbb{P}\big(T(m,s,c) &\le c + (N/m) - 1\big)\mu_{m,s}^{*c}(l) \\ &+ \sum_{n=0}^{c-1} \binom{N/m}{n} \rho^n (1-\rho)^{(N/m)-n} (\mu_{m,s}^{*n} * \nu_{m,s}^{*[(N/m)-n]})(l).\end{aligned} \tag{3.4}$$

(3.4) and (3.2) provide explicit formulas for the two terms of the objective function (2.6).

4 Numerical analysis and simulation

The representations (2.14)-(2.18) show that in order to determine the objective function and the constraint probabilities p_1 and p_2, one needs to calculate the m-dimensional integrals

$$I_{m,j}(s,t) = \int \dots \int_{0 \le x_1, \dots, x_j \le t, \; t < x_{j+1}, \dots, x_m \le s, \; x_1 + \dots + x_m \le s} dF(x_1) \dots dF(x_m).$$

Moreover, for each triple m, s, c a large sum of products of these integrals $I_{m,j}$ has to be computed to arrive at the expected number of group tests $\mathbb{E}[T(m,s,c)]$ and the expected penalty $\mathbb{E}[(d - Z(m,s,c))^+]$, which give the objective function

$$\Omega(m,s,c) = \mathbb{E}[T(m,s,c)] + a\mathbb{E}[(d - Z(m,s,c))^+].$$

Solving the optimization problem thus requires a considerable numerical effort. Therefore it is advisable to simulate the model with a sufficiently large number of samples, rather than implementing the exact formulas (2.14)-(2.18). In what follows we present results of Monte Carlo simulations of the group test model studied in Section 2.

We assume that $d = 1000$ items are demanded and that the marker variables X_i have a lognormal distribution with mean 100 and standard deviation 30, i.e.,

$$\mathbb{P}(X_i \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{1}{u} \exp\left\{\frac{-(\log(u) - \mu)^2}{2\sigma^2}\right\} du,$$

where $\mu = 4.562$ and $\sigma = 0.293$ are the mean and standard deviation of the associated normal random variable $\log(X)$. Without going into details we mention that the classical Box-Muller algorithm (see [7]) to generate normal variates and a subsequent exponentiation is well-suited for our purposes, and no extra effort has been made to shorten the duration of the simulations.

For the data presented here $10,000$ sequences of group tests were carried out for each choice of the decision variables. We take $t = 103.178$, so that the probability of having more than t particles in one item is given by $1 - F(t) = 0.4$, which is not an unrealistic assumption for the intended applications.

4.1 Dependence on the threshold value s

For the data shown in Figure 1 we chose $c = 60$, $m = 20$ and let the threshold value s vary from 1800 to $mt = 2063$. The two solid graphs show the expected number of group tests, $\mathbb{E}[T(m,s,c)]$, in black and the approximation given by (2.21) in grey, for different values of s. The dashed curves in Figure 1 show the expected penalty $\mathbb{E}[(d - Z(m,s,c))^+]$ (black) and its approximation $(d - cmF(t))^+$ (grey), as given in (2.22). The dotted grey line in Figure 1 shows the corresponding values of the constraint probability $p_1 = \mathbb{P}(\max_{i=1,\ldots,m} X_i > t \mid \sum_{i=1}^m X_i \leq s)$. The probability p_2 was indistinguishable from 0 throughout this simulation (from (2.19) we have $p_2 \leq F(t)^m = 3.6 \cdot 10^{-5}$).

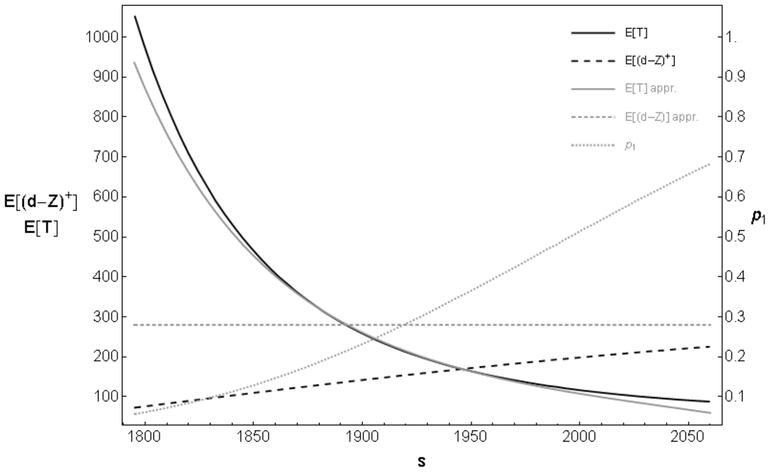

Fig. 1: $\mathbb{E}[T(m,s,c)]$, $\mathbb{E}[(d-Z(m,s,c))^+]$ *and their approximate values and* p_1.

According to Figure 1,

- $\mathbb{E}[T(m,s,c)]$ decreases with s;
- $\mathbb{E}[(d-Z(m,s,c))^+]$ increases with s;
- the approximation (2.21) for $\mathbb{E}[T(m,s,c)]$ is surprisingly close, even for smaller values of s;
- the approximation for $\mathbb{E}[(d-Z(m,s,c))^+]$ is not too close, but it still provides a good upper bound for the penalty;
- $s \mapsto p_1(m,s)$ is increasing.

In Figure 2 the objective function $\Omega(m,s,c) = \mathbb{E}[T(m,s,c)] + a\mathbb{E}[(d - Z(m,s,c))^+]$ is displayed for different values of a. It is seen to have a proper minimum, which is actually achieved for some $s \leq mt$.

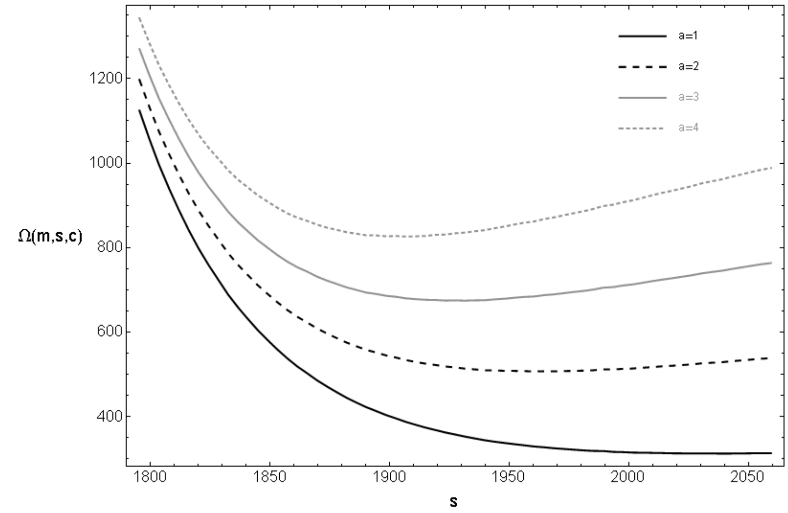

Fig. 2: *The objective function for different values of a.*

4.1.1 Dependence on c

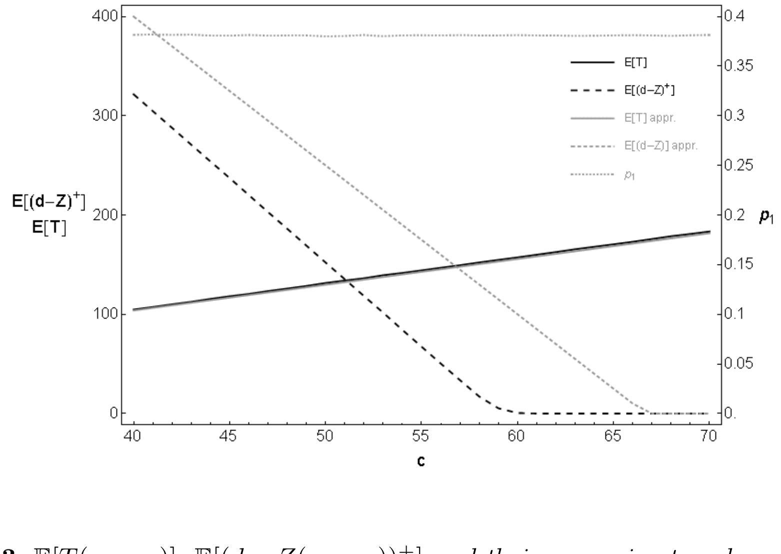

Fig. 3: $\mathbb{E}[T(m,s,c)]$, $\mathbb{E}[(d-Z(m,s,c))^+]$ *and their approximate values and* p_1.

The next plots shows the same quantities as Figure 1, namely $\mathbb{E}[T(m,s,c)]$ (solid) and $\mathbb{E}[(d-Z(m,s,c))^+]$ (dashed), but now for varying c (with $s =$

2450 and $m = 25$).

Figure 3 suggests that

- $c \mapsto \mathbb{E}[T(m,s,c)]$ is increasing;
- $c \mapsto \mathbb{E}[(d - Z(m,s,c))^+]$ is decreasing until it hits zero, and is zero thereafter;
- The approximation for the expected number of group tests is very close again;
- The bound for the penalty $\mathbb{E}[(d-Z(m,s,c))^+]$ is not sharp but roughly shows the almost linear dependence on c.

Note that p_1 is independent of c by definition. The objective function for different values of a is drawn in Figure 4 and shows a minimum close to $c = 60$.

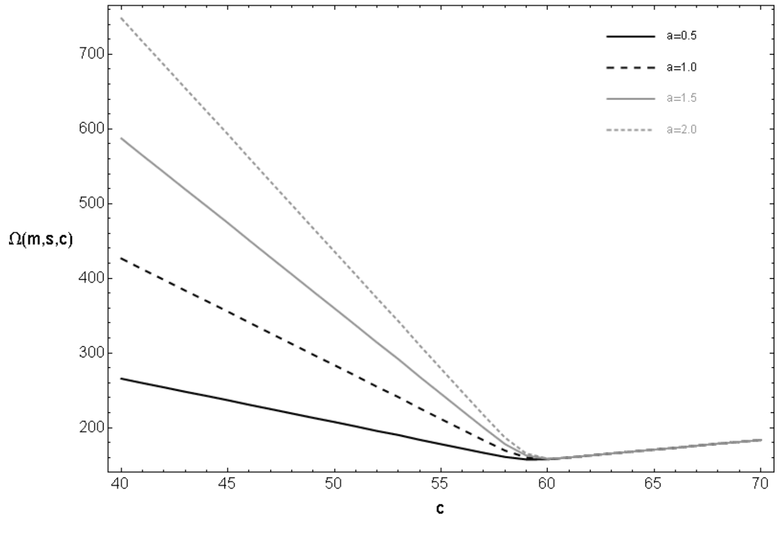

Fig. 4: *The objective function for different values of a.*

4.1.2 Dependence on the group size m

It turns out that for values of m larger than about 110% of s/t the term $\mathbb{E}[T(m,s,c)]$ becomes very large. Since always $m \geq s/t$, there are only a few values of m left that produce reasonable results, too few, in fact, to draw significant diagrams. We therefore introduce the variable

$$\xi = \xi(s,m,t) = \frac{s}{mt} \in (0,1]$$

and study the behavior of the objective function for fixed $\xi = 0.95$, $c = 60$ and varying m (so that s varies implicitly). The resulting Figure 5 looks similar to the previously discussed Figure 3 and suggests that

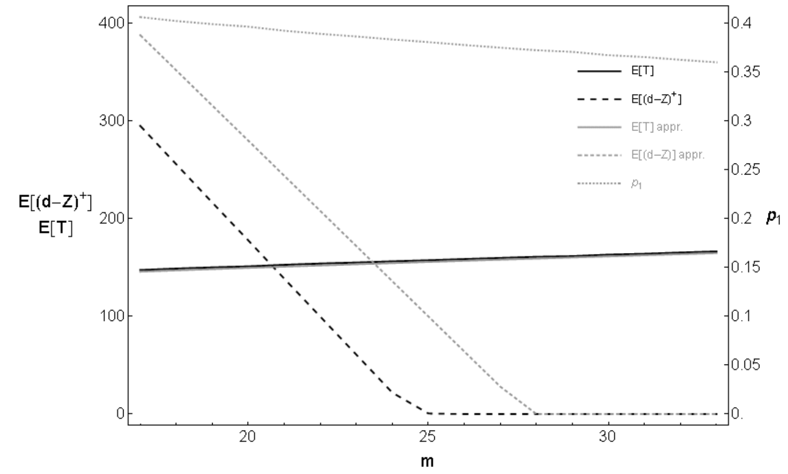

Fig. 5: $\mathbb{E}[T(m,s,c)]$, $\mathbb{E}[(d - Z(m,s,c))^+]$ *and their approximate values and* p_1.

- $m \mapsto \mathbb{E}[T(m,s,c)]$ is slightly increasing;
- $m \mapsto \mathbb{E}[(d - Z(m,s,c))^+]$ is decreasing until it hits zero, and is zero thereafter;
- The approximation for $\mathbb{E}[T(m,s,c)]$ is almost exact;
- The bound for $\mathbb{E}[(d - Z(m,s,c))^+]$ shows roughly the linear decrease of the penalty;
- $m \mapsto p_1(m,s)$ is decreasing.

Again, different values for a are chosen in Figure 6, all leading to a minimum of the objective function near $m = 25$.

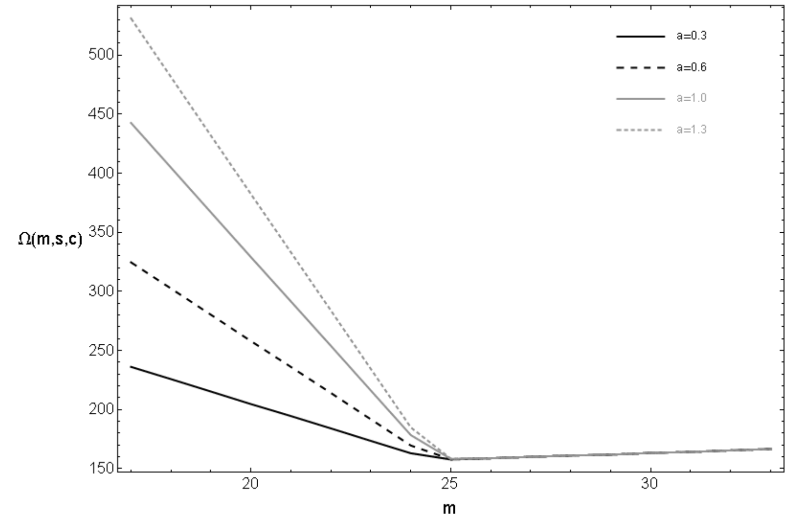

Fig. 6: *The objective function for different values of a.*

4.2 Three-dimensional plots

To gain more insight into the joint influence of the decision variables m, s and c on the objective function, we now fix one of the variables and let the other two vary. The simulation output is presented in three-dimensional diagrams. As before we assume that the demand is given by $d = 1000$ and the probability that the marker is larger than t is $1 - F(t) = 0.4$. Moreover we set $a = 2$. Recall that $\xi = s/(mt)$.

4.2.1 Dependence on m, ξ

First we fix $c = 60$ and consider the objective function for varying m and ξ. Figure 7 displays the $\Omega(m, s, c)$ surface. Darker areas belong to smaller values of p_1. For example, the darkest area corresponds to values of (m, ξ, Ω) meeting the constraint $p_1 \leq 0.1$. (The jagged shape of the equiprobability lines $p_1 = const.$ in the figure is due to the grid size in the simulation.)

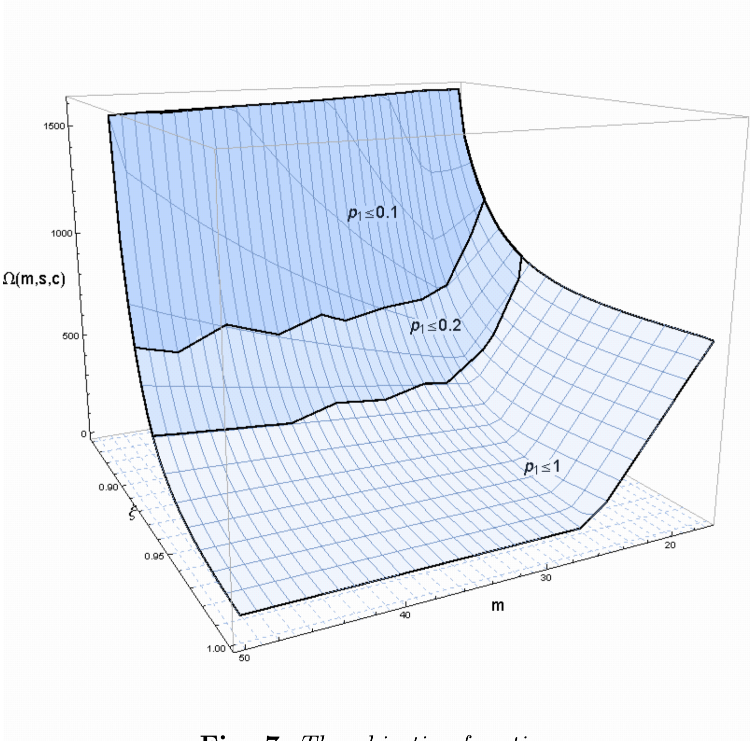

Fig. 7: *The objective function.*

The plot reflects the behavior already discussed for the Figures 1 and 5:

- $\xi \mapsto \Omega(m, s, c)$ is decreasing (it is however increasing for different choices of $a > 2$);
- $m \mapsto \Omega(m, s, c)$ has a minimal value located near $m \approx 25$ (almost independent of ξ);
- Given the constraint $p_1 \leq \varepsilon_1$, the global minimum of the objective function is attained on the curve $p_1 = \varepsilon_1$ in the (m, ξ)-plane, parallel to the c-axis;
- Given that $p_1 = \varepsilon_1$, the objective function is decreasing in m, but Ω is almost constant if m is large (Figure 8 below shows this behavior for $p_1 \approx 0.2$ and varying m, ξ).

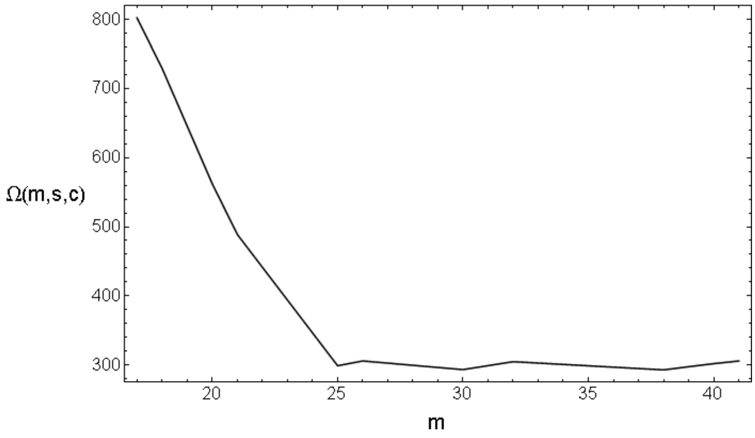

Fig. 8: $\Omega(s, m, c)$ *for values of* m *and* ξ *fulfilling* $p_1(m, s) = 0.2$.

4.2.2 Dependence on m, c

Next we fix $\xi = 0.93$ and plot the objective function for varying c and m in Figure 9. Note that $mc \geq d = 1000$. Figure 9 shows the following:

- $m \mapsto \Omega(m, s, c)$ and $c \mapsto \Omega(m, s, c)$ have minima, located near $mc \approx 1500$ (compare with the dashed line, marking points (m, c, Ω) with $mc = 1500$ and $\Omega = 730$). The minimal value is decreasing for m increasing or c decreasing (see Figure 10).
- For $mc > 1500$ (right of the dashed curve) the Ω-surface is close to a plane, here $\mathbb{E}[(d - Z(m, s, c))^+]$ dominates $\mathbb{E}[T(m, s, c)]$.
- For $mc < 1500$ (left of the dashed line) $\Omega(m, s, c)$ is increasing very fast, due to a domination of the $\mathbb{E}[T(m, s, c)]$ term.
- Given the constraint $p_1 \leq \varepsilon_1$, the minimal value of the objective function is attained on the curve $p_1 = \varepsilon_1$ in the (m, ξ)-plane.

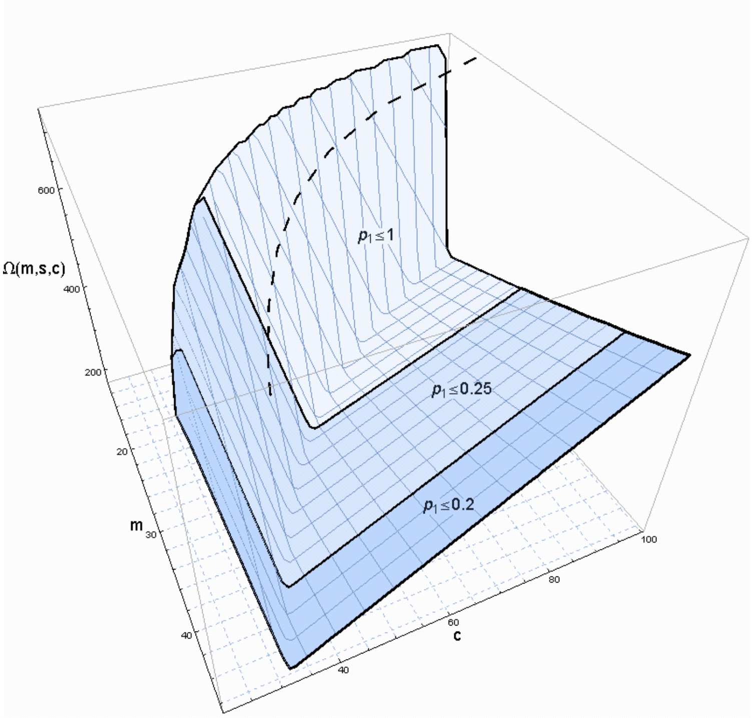

Fig. 9: *The objective function for different values of c and m; the dashed line indicates (m, c, Ω) with $mc = 1500$, $\Omega = 730$.*

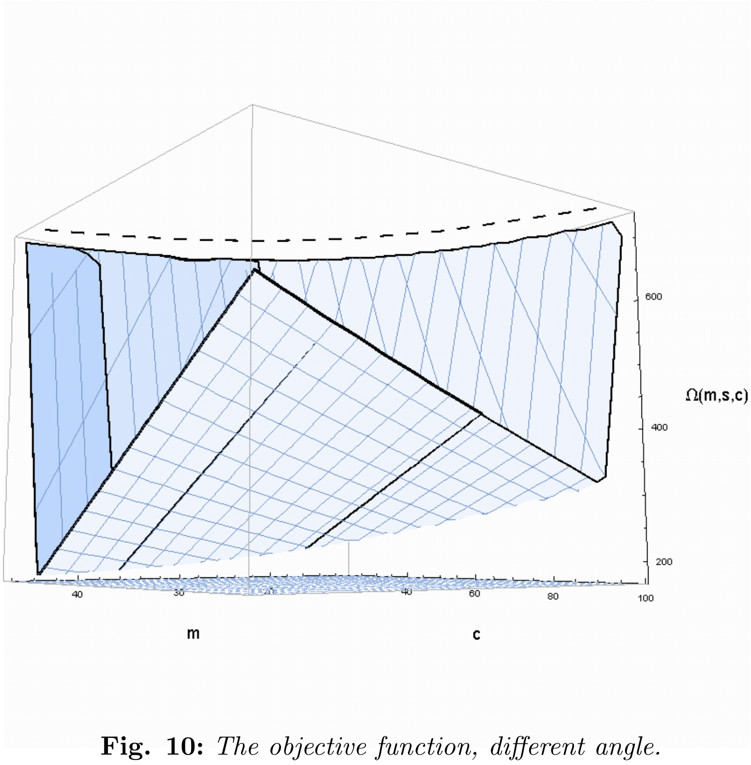

Fig. 10: *The objective function, different angle.*

4.2.3 Dependence on c, ξ

Figure 11 shows $\Omega(m, s, c)$ for fixed $m = 25$ and varying c and ξ. We find that

- $\xi \mapsto \Omega(m, s, c)$ is decreasing;
- $c \mapsto \Omega(m, s, c)$ has a minimum near $c \approx 60$.

Note that p_1 does not depend on c.

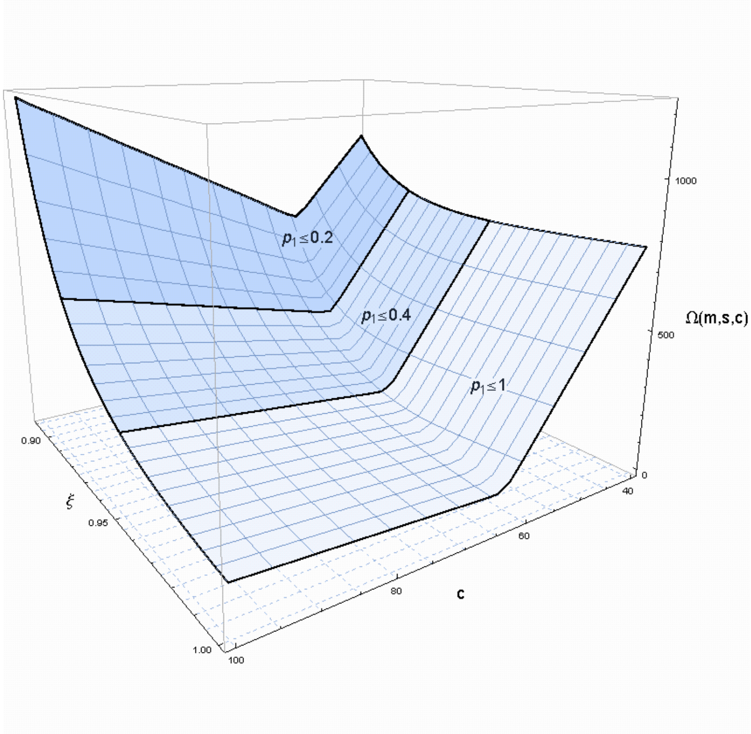

Fig. 11: *The objective function for different values of c and ξ.*

4.3 Dependence on the probability $F(t)$

A final simulation was carried out to reveal the dependence of the previous results on the choice of the probability $1 - F(t) = \mathbb{P}(X > t)$. For Figure 12 we have chosen $c = 60$ and $\xi = 0.93$ and let m and t vary.

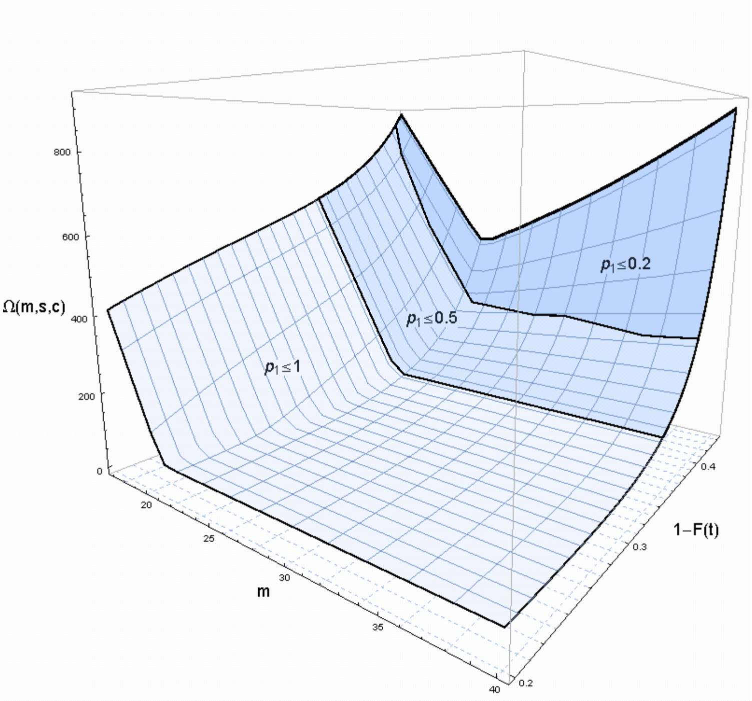

Fig. 12: *Objective function for different values of m and $1 - F(t)$.*

- In general Ω increases with $1 - F(t)$, the increase being steep for large values of $1 - F(t)$;
- The influence of the penalty term in $\Omega(m, s, c)$ becomes less important for higher values of $1 - F(t)$.

5 Possible extensions

The models studied in this paper can be extended in several directions. Let us briefly mention two possibilities.

1. Inconclusive testing. There are situations in which marker values in a certain intermediate interval are considered to be 'inconclusive'. An item is declared positive if its marker is above some threshold t_2, negative if it is less than some $t_1 < t_2$, and inconclusive if it lies in the interval $[t_1, t_2]$. For group tests of m items we may consider thresholds $t_1(m)$ and $t_2(m)$ so that a group of size m passes or fails the inspection if its marker sum is below $t_1(m)$ or above $t_2(m)$, respectively, and declared to be inconclusive otherwise. One may then again consider the problem of aggregrating sufficiently many items to meet a certain prespecified demand. In the minimization problem for the cost function one now has to take into account new constraints, for example those on the probability of a misclassification of inconclusive groups. If the demand requirement is not met by group testing all available items, one may think of getting back to the groups that have been classsified as inconclusive. For example, one could put them into one pool and start testing new groups of size m from this pool, use the same procedure with the same two thresholds $t_1(m)$ and $t_2(m)$ to classify groups according to the three categories (clean, inconclusive or contaminated) and stop testing if the residal demand is met or no groups are left in the pool.

2. Unreliable results. In this paper it was assumed that each test is fully reliable, i.e., the marker values (or their sums) are measured with total precision. In practice all kinds of measurement errors can and will occur and should be incorporated in the stochastic modeling. In such more realistic models one has to take into account the probabilities of misclassifications due to error variables perturbing the measurements of the markers or their sums. Such an approach will lead to various new constraints in the cost optimization problem to ensure the quality of the selected items and avoid misclassifications of good ones.

References

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

[2] Bar-Lev, S.K., Parlar, M., Perry, D., Stadje, W. and van der Duyn Schouten, F.A. (2007). Applications of bulk queues to group testing models with incomplete identification. *European Journal of Operational Research* 183, 226-237.

[3] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2003) Hypergeometric group testing models with incomplete information. *Probability in the Engineering and Informational Sciences* 17, 335-350.

[4] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2004) Optimal group testing with processing times and incomplete identification. *Methodology and Computing in Applied Probability* 6, 55-72.

[5] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2005) Multinomial group testing models with incomplete identification. *Journal of Statistical Planning and Inference* 135, 384-401.

[6] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2006) Group testing procedures with incomplete identification and unreliable testing results. *Applied Stochastic Models in Business and Industry* 22, 281-296.

[7] Box, G.E.P. and Muller, M.E. (1958) A note on the generation of random normal deviates. *Ann. Math. Stat.* 29, 610-611.

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

[9] Faraggi, D., Reiser, B. and Schisterman, E.F. (2003). ROC curve analysis for biomarkers on pooled assessments. *Statistics in Medicine* 22, 2515-2527.

[10] Fluss, R, Faraggi, D. and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal* 47, 458-472.

[11] Shapiro, D.E. (1999) The interpretation of diagnostic tests. *Statistical Methods in Medical Research* 8, 113-134.

[12] Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration.*. New York, NY: Springer.