Journal of Applied Mathematics and Stochastic Analysis, **16**:4 (2003), 361-374.

OPTIMIZATION IN HIV SCREENING PROBLEMS

LEV ABOLNIKOV
Loyola Marymount University, Department of Mathematics
Los Angeles, CA 91316 USA

and

ALEXANDER DUKHOVNY
San Francisco State University, Department of Mathematics
San Francisco, CA 94132 USA

(Received August, 2002; Revised March, 2003)

In this article, the authors use both deterministic and stochastic approaches to the analysis of some optimization problems that arise in the group ("pool") HIV screening practice. Two kinds of testing policies are considered. For the first kind, group-individual testing, the optimal size of the group that should be selected for testing is found. For more general group-subgroup testing procedure the authors develop a numerical algorithm for finding the sequence of successively selected subgroups that minimizes the total cost of testing. Assuming that both arriving and testing processes have a random nature, the authors suggest a stochastic model in which the optimal size of the group in the group-individual testing procedure is found by using methods of queueing theory.
Key words: HIV Screening, Group Testing, Service Optimization.

AMS (MOS) subject classification: 60K25, 90B22

1 Introduction

The rapidly spreading AIDS epidemic and the limited resources of diagnostic facilities in the third world countries make researchers and public health professionals look for new ways to increase the effectiveness of HIV screening procedures. Pooled (or group) testing method is one of the possible ways in this direction. The idea of this method is simple. Instead of traditional testing of a number of m, $m > 1$, blood samples individually one-by one, this procedure suggests to pool all (or a part) of them and then test the pool using a single test. If it indicates the presence of HIV positive blood samples in the pool then either each of the samples is tested individually (we will call it "group-individual testing") or the pool is divided into subgroups for further testing ("group-subgroup testing"). On the other hand, if the pool turns to be HIV negative then this single test would replace m individual tests which may result in great savings of time and money without losing the accuracy of the testing procedure. For example, as mentioned in [10], "the field studies show that the pooling methods (in HIV screening) may be as sensitive and specific as individual testing, and can result in cost savings from 5% to 80%."

However, the pooling methods pose a number of important questions which need to be answered. Here are some of them.

1. **Is a group testing procedure (group-individual or group-subgroup) always better than the individual testing?**

In this article, we will show that the answer is: "No, not always", and we will find the conditions under which the group testing is better (in a certain sense) than the individual one. It is clear, however, that if the probability q ("seroprevalence") that an individual blood sample is HIV positive is small, then, to minimize the total number of tests necessary to identify all HIV positive individuals, it makes sense to take more blood samples for the pool test because there is a good chance that the individual testing would not be required. On the other hand, the pool size should not be too large because no matter how small q is, a large pool, most likely, would contain an HIV positive blood sample and then the pool test would be a waste.[1].

If, however, q is large, then it seems better to test blood samples in small groups or, possibly, to cancel the pool testing altogether, leaving individual testing only. Thus, one can expect that for some values of q a group-individual testing is more advantageous, while for other values of q it is better to use the traditional individual testing procedure.

2. **Which of the two group testing procedures (group-individual or group-subgroup) is more effective (that is, requires less tests to identify all HIV positive blood samples in a pool of the same size)?**

As in the previous case, the answer depends on the value of q. Clearly, if q is small, it seems better to use some kind of group-subgroup testing, while if q is large a group-individual testing may be more effective.

3. **What is the optimal size of the group taken for testing? What are possible criteria of optimality?**

They may be, for example, the average number of tests per item in a group or the total cost of testing per blood sample. In more general cases, it may be also important to take into account the average number of items waiting for testing, or the average time an untested item spends in the system, and so on.

4. **What should the testing strategy be if at the beginning of a testing procedure there are less than the optimal number of blood samples available in the waiting room? Should the group test start immediately or is it better to wait until the optimal number of blood samples become available?**

A rigorous approach to the solutions of all of these problems requires a considerable mathematical effort. R. Dorfman [6] was, presumably, the first who used a simple statistical approach to show how group testing can be employed to efficiently eliminate all defective items from a large population. This approach found an immediate application in syphilis screening practice for WW II draftees where it resulted in considerable savings (see [10]). Recent publications show that the pooling methods are extremely important for HIV screening practice and a large literature now exists on this topic (see, e.g., [4, 7, 8, 10]).

[1]There is another reason to restrict the size of the pool: the dilution effect ([10]). It means that the pool test may be unable to detect the presence of an HIV positive blood sample if the pool is too large.

2 Static and Dynamic Optimizations

In this article, we consider two kinds of optimization problems which arise in HIV screening practice. In the first one, we suppose that there always exists an unlimited number of individuals ("customers") to be tested ("served"), so that a testing facility ("the system") is never idle. Also, we suppose that the amount of time necessary to test a single blood sample ("service time") is constant and the same for each customer (the pool testing time may be different). Optimization problems considered under these conditions will be called static.

The second kind of optimization problems take into account a dynamic nature of the arrival and testing processes. In this case, we suppose that customers arrive to the system in groups of different and random size at random moments of time and the service time is random and different for each customer. These assumptions, of course, make optimization problems more difficult, but they also make them much closer to real life situations. The problems of this kind are called in this article dynamic optimization problems. (Observe that the first three problems mentioned in the previous section can be treated both as static or dynamic, while the last problem is obviously dynamic). It should be noted that in the existing literature all optimization problems related to HIV screening practice were considered only in their static version. For example, a widely referred formula obtained by Dorfman in [6] for the optimal size of the group in the group-individual testing was derived under the condition of the constant presence in the system of a large number of customers needed to be tested and can not be used for any dynamic optimization problem. However, a lot of important problems related to static and especially dynamic optimizations in the HIV screening practice remain unsolved. In this article, some of these problems are solved; some other are formulated and possible methods of solving them are discussed. In Section 3, we consider a general static optimization problem related to HIV screening practice and then, in Section 4, using methods of queueing theory we will solve and discuss some dynamic optimization problems.

3 Finding the Optimal Group size (Static Approach)

3.1 One-step Optimal Subgroup Selection from a Large Population

Consider a population of N individuals which are supposed to be tested with the help of a group test $T(m)$, $m = 1, \ldots, n$, where m is the group size of the test and n is the capacity of the testing facilities (i.e., the maximum number of individuals which can be tested simultaneously), and suppose that $N \gg n$. A group testing process of the whole population is, in fact, a sequence of tests $\{T(m_j)\}$, $j = 1, 2, \ldots,$ where a group of m_j individuals is selected from the general population or from another subgroup which turned out to be HIV positive. Once the first group is selected from the general population, the problem arises of finding an optimal "service schedule" that would minimize the total "service cost" for the selected group. However, what should be the size m of a subgroup selected for the first test? Under all conditions being equal, it is natural to define a group test $T(m_1)$ more effective than a group test $T(m_2)$ if a single use of the first test leads to more tested individuals (on the average) than a single use of the second test. Therefore, the effectiveness of any two group tests can be

compared by the average number of individuals which are determined HIV negative as a result of using a single group test. According to this optimization criterion, the optimal size m_{opt} of the group taken for testing can be found as a maximum of the function $Neg(m) = mp^m$, $p = 1 - q$, $1 < m < n$. Clearly, $Neg(m)$ increases for $m < -1/lnp$ and decreases for $m > -1/lnp$. Let η be an integer such that $-1/lnp \in [\eta, \eta + 1)$. Now, we have

$$m_{opt} = \begin{cases} min(n, \eta) & \text{if } Neg(\eta) \geq Neg(\eta + 1) \\ min(n, \eta + 1) & \text{otherwise.} \end{cases} \tag{3.1}$$

For example, if $q = 0.054$, $M_{opt} = 34$. We will call the test $T(m_{opt})$ where m_{opt} is determined from (3.1), the most effective single group test of a large size population. If the first group tested by the most effective single group test turned to be HIV positive, subgroups of this group can be selected according to optimal testing policies determined in Sections 3.2 and 3.3.

Another optimization criterion of a group testing procedure may be the total number of tests necessary to identify all HIV positive blood samples in a pool of a certain size (or per individual blood sample). If the group testing time depends on the size of a group, then this criterion can be replaced by the total required time (or the total required cost) necessary to identify all HIV positive samples in the pool of a certain size. Some of these criteria are used in this article.

3.2 Group-Individual Testing

In this section, we consider only group-individual testing procedures. First, we determine under which conditions a group-individual inspection is "better" than a traditional individual testing procedure. The average number of tests necessary to identify all HIV positive samples in a group of size m will be used as a criterion of the comparison.

Let N_m be the number of acts of individual inspection which are necessary to perform after the group stage in order to identify all HIV positive samples (if any) in a group of size m. Then, obviously, $N_m = m$ with probability $1 - mp^{m-1}q - p^m + (m-1)p^{m-1}q = 1 - p^{m-1}$ (the group contains more than one HIV positive sample, or it contains only one but it was not tested last), $N_m = m - 1$ with probability $p^{m-1}q$ (the group contains only one HIV positive sample, and it turned to be last), and $N_m = 0$ with probability p^m (the group does not contain any HIV positive samples). Comparing the overall number of service acts

$$1 + E(N_m) = 1 + (m-1)p^{m-1}q + m(1 - p^{m-1})$$

for the group-individual and m service acts for the individual testing procedures, we find that a group-individual service is better than the individual inspections if

$$mp^m > 1 - p^{m-1}q,\ m = 2, 3, 4, \ldots \tag{3.2}$$

Since for $m \gg 1$ and (or) small q the second term in the right-hand side of (3.2) is negligibly small, we can approximate (3.2) by a more convenient relation $mp^m > 1$ and conclude that under these conditions, a group-individual inspection is better when either $mp^m > 1$, or, if $q < q_0 \approx 1 - e^{-1/e} \approx 0.308$, when m is such that $m^{-1/m} < 1 - q$. For other values of q and m, the general condition (3.2) should be used.

For example, if $q = 0.054$, a group-individual procedure is more advantageous than the individual service only if $m < 80$. For $m > 80$ a group-individual testing becomes

worse than the individual service. (It is not surprising: if the pool is too large the first part of the group-individual testing may be a waste - see "Introduction".)

When the average testing times b (for the group stage) and g (for an individual test) differ, the group-individual and individual service disciplines can be compared by the expected time $E(G_m)$ necessary to service a group of m customers. Similar to the above calculations, we obtain from (3.2) that for the group-individual one-by-one testing, $E(G_m) = b + gE(N_m) = b + (m-1)gp^{m-1}q + mg(1-p^{m-1})$. Comparing that with the average time mg for the individual service, we denote $\delta = b/g$ and obtain the following modification of (3.2):

$$mp^m > \delta - p^{m-1}q. \tag{3.3}$$

(Observe that b and g may differ even when the same server is used for the group and individual stages, as the group stage may require setup time.). To find out for which values of δ and q there exist such $m \geq 2$ that (3.3) is satisfied, we investigate whether the function $h(x) = \delta - qp^{x-1} - xp^x$ can assume negative values on the interval $[2, \infty)$. By a routine analysis one can see that at $x_0 = -(p + qlnp)/plnp$ function $h(x)$ attains its absolute minimum value $h(x_0) = \delta - p^{-q/p}/(e\, lnp)$.

Under a natural assumption that $q < 0.5$, one can see that $x_0 > 2$ (in fact, calculations by Mathematica show that to be true for $q < 0.553567$), so condition (3.3) can be satisfied by some $m \geq 2$ if and only if $h(x_0) < 0$.

When q is very small, $x_0 \simeq -1/lnp$ and $h(x_0) \simeq \delta - 1/(e\, lnp)$. Therefore, for small q condition (3.3) can be satisfied by some $m \geq 2$ if and only if $1 + \delta e\, lnp > 0$, which can also be written as $q < 1 - e^{-1/(\delta e)}$. More generally, two testing modes can be compared by the expected cost of testing a group of m customers. In this case, an inequality similar to (3.3) appears, where δ is the ratio of the average costs of the group test and an individual inspection.

If the expected service times of the group and individual stages are equal, it may be reasonable to selects a group size m that minimizes the average number of service acts per customer $[1 + N_m]/m$. In a more general situation, when those expected times are not equal, the following optimization problem can be considered:

Minimize the average service time per customer in a full group of m customers

$$\frac{E(G_m)}{m} = \frac{b - gp^{m-1}q - gmp^m}{m} + g \tag{3.4}$$

subject to the restriction that the one-by-one group-individual service is better than the individual service procedure in either of the two available modes, that is, that $E(G_m) < min\{mg, mb\}$. This is equivalent to the following problem:

Minimize

$$f(m) = \frac{\delta - p^{m-1}q - mp^m}{m}, \quad m \geq 2, \tag{3.5}$$

subject to

$$h(m) = \delta - p^{m-1}q - mp^m < min\{0,\, m(1-\delta)\}. \tag{3.6}$$

As $m \to \infty$, $h(m) \to \delta > 0$, so the set of possible values of m specified by (3.6) has to be finite. Once the upper bound U of this set is estimated, the minimum of $f(m)$ can be found by simply evaluating $f(m)$ for $m = 2, .., U$. To estimate U, let $x_1 = -1 - 2/lnp$. A routine analysis shows that at $x = x_1$ the function $(x+1)p^{x/2}$ assumes

its absolute maximum value M on $[2,\infty)$, $M=-2/(e\sqrt{p}\,lnp)$. Let $x_2=2\,log_p\,(\delta/M)=2\,log_p(-\delta e\sqrt{p}\;ln\sqrt{p})$ Clearly, for any x, $x>x_2$, $p^{x/2}<\delta/M$. Now, let

$$U=\max\{x_1,x_2\}=\max\{-1-2/lnp,\;2\,log_p(-\delta e\sqrt{p}\;ln\sqrt{p})\,\}. \tag{3.7}$$

Under a natural assumption that $q<p$, for any m,m$>U$, $\delta-p^{m-1}q-mp^m>\delta-(m+1)p^m>\delta-[(m+1)p^{m/2}/M](Mp^{m/2})>0$, so values of the group size m that satisfy (3.6) have to belong to $[2,U]$. For special cases when $\delta=1$, it can be shown that for $q=0.05$, 0.03 and 0.01 the optimal (that is, minimizing the average number of service acts per customer in group-individual testing) sizes of the group taken for service are, respectively, $m=5$, 6, and 10. (These results show that the employment of large size groups ($m\approx 80$) in HIV screening practice mentioned in [10] is presumably unlikely to be optimal!)

3.3 Group-Subgroup Testing

In this section, we consider a general model related to the group-subgroup testing procedure. The main two differences between this model and the previous one are the following:

(i) Unlike the model with group-individual testing, we suppose that if the group test indicates the presence of HIV positive customers in a group of the size m_0, the server uses a certain decision rule to select a subgroup of the size $m_1=f(m_0)$, $m_1<m_0$, for further testing. We call it a subgroup of the "first generation" and its complement subgroup of the size m_0-m_1 a complement subgroup of the first generation. We suggest the following algorithm for determining all HIV positive samples in the group. If the subgroup of the first generation turns to be also HIV positive, the server selects a sub-subgroup of the "second generation" of the size $m_2=f(m_1)$, $m_2<m_1$, out of those m_1, and so on, until the test for the first time shows that a subgroup of the k-th generation contains no HIV positive blood samples or it consists of a single HIV positive blood sample. After that, the server immediately starts testing the complement group of the k-th generation, using the same testing policy, then the complement subgroup of the $(k-1)$-st generation, and so on, till the complement group of the first generation. This process continues until all HIV positive blood samples are identified.

(ii) The distribution of the testing time may depend on the size of a subgroup.

(iii) We also suppose that each blood sample has a "value" (or a "length") which is a discrete random variable distributed identically for each sample. (As a special case, we can assume for simplicity that the value is an integer, the same for each sample). Every time a sample participates in a testing procedure, its value is reduced by a certain discrete amount. A sample having zero value can not participate in any further testing. (This assumption reflects a real situation in blood screening practice where a certain amount of sera of each participating blood sample is lost in every test and, therefore, the number of subgroup tests must be limited as described above).

A decision rule is supposed to be found which minimizes the average number of tests per sample or the total cost of the testing procedure per sample subject to certain

restrictions which include the number of samples untested as a result of losing their values (see (iii)) during the testing procedure.

In this general form, the problem seems to be rather difficult but, because of its obvious importance for HIV screening practice, several attempts have been made to solve it in some particular cases (using, for example, heuristic algorithms [10]). However, without taking into account conditions (ii) and (iii) of the aforementioned model and without an estimation of the closeness of the obtained solution to the optimal one, such attempts could be apparently considered only as first steps towards the ultimate solution.

We will consider two different approaches to the problem of finding the optimal testing strategy for group-subgroup testing procedures. The first one is based on the idea which was described in the beginning of this section.

Once a group of m is selected from the waiting line, the problem arises of finding an "optimal" service schedule that would minimize the total "service cost" for the selected group. (The "service cost" may be the total expected number of tests, or the total expected time of tests, or, indeed, the actual expected cost of the procedure.) In particular, what should be the size n of a subgroup selected for the first test?

Let A be the event that a group of m items tested positive and let B_n be the event that its subgroup of n will test negative

Lemma 3.1: $P(B_n/A) = \frac{p^n - p^m}{1-p^m}$.

Proof: Clearly, $P(A) = 1 - p^m$, $P(B_n) = p^n$, $P(\overline{B_n}) = 1 - p^n$. Since $\overline{B_n} \subset A$, $P(\overline{B_n}/A) = P(\overline{B_n})/P(A) = (1 - p^n)/(1 - p^m)$.

Let us denote the minimal cost of complete service of a group of $m > 0$ by $\Psi(m)$, given that the group has not yet been tested, or by $\Phi(m)$, given that the group as a whole has tested positive. It is clear that $\Phi(1) = 0$, $\Psi(1) = C(1)$, where $C(n)$ is a cost of one test of a group of n items.

Theorem 3.1: *For $m > 1$, the following recursive relations hold (with $\Psi(0) = 0$):*

$$\Phi(m) = \min_{1 \le n < m} \{C(n) + \frac{p^n - p^m}{1 - p^m}\, \Phi(m-n) + \frac{1-p^n}{1-p^m}\, [\Phi(n) + \Psi(m-n)]\}, \tag{3.8}$$

$$\Psi(m) = \min_{1 \le n \le m} [C(n) + \Psi(m-n) + (1-p^n)\Phi(n)]. \tag{3.9}$$

Proof: First, suppose that the group of m has not been tested. Testing a subgroup of n samples, we incur a cost of $C(n)$. Regardless of the outcome of this test, nothing will be known about the complement subgroup of $m - n$ samples, so the minimal cost of identifying all of them is $\Psi(m - n)$. With the probability of p^n, the subgroup of size n will test negative and will not require any additional cost. With the probability of $1 - p^n$, the subgroup of size n will test positive and the minimal cost of identifying all of its samples will be $\Phi(n)$. Then (3.9) follows.

Suppose now that the group of size m has tested positive. Given that, by Lemma 3.1, the conditional probability that a subgroup of size n will not contain positive items is $\frac{p^n - p^m}{1-p^m}$. In this case, the subgroup will not require any additional cost, the complement subgroup of $m-n$ items will be known to be positive, and the minimal cost of identifying all its samples will be $\Phi(m - n)$. If, on the other hand, the subgroup of size n tested positive (the conditional probability of which, by Lemma 1, is $\frac{1-p^n}{1-p^m}$), the minimal cost of identifying all samples of this subgroup will be $\Phi(n)$, no information will be obtained about the complement subgroup of $m - n$ samples and the minimal cost of identifying

all its samples will be $\Psi(m-n)$. In addition, regardless of the outcome, testing a subgroup of n samples we incur a cost of $C(n)$. Hence (3.8).

Relations (3.8) and (3.9) can be used to find $\Phi(m)$ and $\Psi(m)$ recursively. (Note that $\Phi(m)$ should be found before $\Psi(m)$, since $\Phi(m)$ appears in (3.8).) In the process, we answer the question about the optimal size n of the subgroup to be tested first. If the group of m has not been tested, n is the one for which the minimum is attained in (3.9). If, on the other hand , the group of m has tested positive, n should be the one for which the minimum is attained in (3.8).

In many cases $C(n)$ can be considered a constant. Mathematically, all such cases are equivalent to the case where $C(n) \equiv 1$. Relations (3.8) and (3.9) can be now rewritten in a simpler way:

$$f(m) = 1 - p^m + \min_{1 \le n < m} [p^n f(m-n) - p^n + f(n) + (1-p^n)\Psi(m-n)], \tag{3.10}$$

$$\Psi(m) = \min_{1 \le n \le m} [\Psi(m-n) + f(n)] \tag{3.11}$$

where $f(n) = 1 + (1-p^n)\Phi(n)$. As defined by (3.10) and (3.11), functions $\Psi(m)$ and $f(m)$ have now the following probabilistic meaning: $\Psi(m)$ is the minimal expected number of tests necessary to completely identify all items in a group of m; $f(m)$ is the minimal expected number of tests necessary to completely identify all items in a group of m under the decision that the first test should involve the whole group. The initial values of these functions are $\Psi(1) = f(1) = 1$. Although relations (3.10) and (3.11) can not be solved explicitly, we created a computer program that solves them recursively.

4 Dynamic Optimization: Queueing Approach

As we mentioned above, a dynamic optimization approach takes into consideration a random pattern of the arrival and testing processes and, in many situations, enables us to get more realistic and more comprehensive recommendations to improve the effectiveness of a testing facility. Proceeding on the nature of the problem, one can conclude that methods of queueing theory are the most appropriate mathematical tool for its analysis. In the next section, using some results of [1, 2, 3], we will describe and study the problem in terms of queueing theory.

4.1 Queueing System with a Group-Individual Service

We consider a single server queueing system under the following assumptions about the arrival and service procedures:

1. Customers (blood samples) arrive at the system (testing facility) in i.i.d. groups of a random size α at random moments of time forming a Poisson process with parameter λ; we denote $P(\alpha = k) = a_k$, $E(z^\alpha) = a(z)$, and assume that the expected group size $a = a'(1) < \infty$;

2. Each customer has probability q of being "defective" (HIV positive) and the goal of the service (testing) procedure is to identify all defective customers;

3. The server has a finite "capacity" m (i.e. the maximum number of customers it can serve simultaneously). Three models ("disciplines") of the server behavior after completing a service act will be considered in this article.

a) *Continuously operating server* discipline. According to this discipline, the server, upon a service completion, immediately starts a new service act even if there are no new customers in the queue at that moment. (As, for example, in the case when the server appears at a service station only to pick up customers and leave.)

b) *Waiting server* discipline. In this case, if after a service completion the server finds no customers in the queue, he waits for the next arrival and then starts a new service act.

c) *Quorum* discipline. The server starts a new service act if and only if , after completing the previous service act, he finds at least r, $0 \leq r \leq m$, customers in the queue. Otherwise, the server waits until a necessary number of customers is accumulated in the queue. The number r is called a quorum threshold. (It should be noted that formally the first two disciplines are particular cases of the quorum service discipline for $r = 0$ and $r = 1$. However, since these disciplines are frequently encountered in applications and following the tendency developed in the literature on bulk queueing systems (see, for example, [5]) we will consider them independently from the general case.

In all three cases, a group of $min\{i, m\}$ customers is taken for service where i is the number of customers in the queue when the service begins (in case of a quorum discipline, $i \geq r$).

4 Arriving customers are served according to the group-individual procedure described in the previous section. The group service time does not depend on the size of the group taken for service and is a random variable with the distribution function $B(t)$, $b = \int_0^\infty t dB(t) < \infty$.

If the group does not contain any defective customers, the service act is completed and the customers leave the system. Otherwise, the customers in the group are served individually until all defective customers are identified. We suppose that the results of both group and individual tests are reliable with probability 1. It follows, in particular, that if the group test showed the presence of defective customers in a group of size m, and if the first $m - 1$ of them turned to be not defective, then the last customer in the group is concluded to be defective without additional testing. The individual service time of each customer is distributed according to the distribution function $G(t)$, $g = \int_0^\infty t dG(t) < \infty$. Observe, however, that the distribution $G_i(t)$ of the total service time needed to serve i, $i = 2, 3, ...,$ customers individually one-by-one by one server (given that the group contains defective customers) is not the $i-$fold self-convolution $[G(t)]^{*i}$ of the distribution function $G(t)$ as it seems it should be but

$$G_i(t) = \frac{p^{i-1} - p^i}{1 - p^i} [G(t)]^{*(i-1)} + \frac{1 - p^{i-1}}{1 - p^i} [G(t)]^{*i}, \ \ p = 1 - q. \qquad (4.1)$$

To see this, denote by E_i the event that a group of size i taken for service contains defective customers and by F_{i-1} the event that the first $i - 1$ customers taken from this group are non-defective. The conditional probability that a customer randomly selected from the group for the individual service to be defective is $q[P(E_i)]^{-1} = (1 - p)(1 - p^i)^{-1}$ and, therefore, $P(F_{i-1}|E_i) =$

$P(F_{i-1} \cap E_i)[P(E_i)]^{-1} = p^{i-1}(1-p)(1-p^i)^{-1}$ from which (4.1) follows immediately since it is not necessary to inspect the last customer if the first $i-1$ of them turned to be non-defective. (This fact is especially significant if i is small.)

In some applications, instead of a one-by-one individual service, a parallel service discipline may be used, when the individual inspections are performed independently and simultaneously on all customers of the group by separate identical servers. In this case, the distribution $G_i(t)$ of the total time necessary to serve a group of i, $i \leq m$, customers individually given that the group contains defective customers can be obtained similarly:

$$G_i(t) = \frac{p^{i-1} - p^i}{1-p^i} [G(t)]^{(i-1)} + \frac{1-p^{i-1}}{1-p^i} [G(t)]^i. \tag{4.2}$$

5. When the individual service procedure is over and all defective customers are identified, the service act is completed and all customers in the group leave the system at once. No new customers can be taken for service from the queue until both parts of the service procedure are completed.

This completes the description of the queueing system. In the next section, we introduce an embedded Markov chain which describes the behavior of the queueing process at successive moments of total group-individual service completions. We also determine necessary and sufficient conditions for the ergodicity of this Markov chain and its ergodic distribution for each mentioned above type of service discipline. A detailed analysis is given of a system with the quorum threshold equal to the server capacity.

4.2 Markov Chain $\{Q_n\}$. Ergodicity Condition and Ergodic Distribution

Let $Q(t)$ denote the number of customers in the system at time t and $Q_n = Q(t_n + 0)$, where t_n, $n = 1, 2, 3, ...$ are successive moments of service completions. It is easy to see that under conditions 1-5, $\{Q_n\}$ forms a homogeneous Markov chain. Specifying transition probabilities of this chain, one can see that its transition matrix $A = \{a_{ij}\}$, $i, j = 0, 1, 2, ...$ has a special "m - quazi-triangular" structure ([1]) and, therefore, belongs to the class of so-called Δ_m-matrices. (Markov chains with transition Δ_m- and $\Delta_m^{'}-$matrices were first introduced and studied in [1] and later generalized in [2, 3].)

It proved to be advantageous to analyze the chain $\{Q_n\}$ by using generating functions $A_i(z) = \sum_{j=0}^{\infty} a_{ij}\, z^j$, $i = 0, 1, 2, ...$ of the entries of the i-th row of the matrix A. All these functions are completely specified by the matrix A and can be found in an explicit form in the following way. Let $K(z)$ and $L_i(z)$ be the generating functions of the number of customers arriving to the system during the group service stage and the subsequent individual service stage for a group of i customers, respectively. Then $K(z) = B^*(\lambda - \lambda a(z))$, where $B^*(.)$ is the Laplace-Stiltjes transform of B(t). The formula for $L_i(z)$ depends on whether the individual service is one-by-one or parallel. In case of the one-by-one service, proceeding similar to the derivation of (4.1), we obtain:

$$L_i(z) = \frac{p^{i-1} - p^i}{1-p^i} [L(z)]^{(i-1)} + \frac{1-p^{i-1}}{1-p^i} [L(z)]^i, \quad i \geq 1, \tag{4.3}$$

where $L(z) = G^*(\lambda - \lambda a(z))$.

For the parallel service

$$L_i(z) = \int_0^\infty exp\{t[\lambda a(z) - \lambda]\} d\{\frac{p^{i-1} - p^i}{1 - p^i} [G(t)]^{(i-1)} + \frac{1 - p^{i-1}}{1 - p^i} [G(t)]^i\}, \ i \geq 1. \quad (4.4)$$

For both service disciplines, we also define $L_0(z) = 1$. Denote $K_i(z) = K(z)[p^i + (1 - p^i)L_i(z)]$. For all above mentioned disciplines of service,

$$A_i(z) = z^{i-m} K_m(z), \ i \geq m, \quad (4.5)$$

regardless of whether the individual service discipline is one-by-one or parallel. For $i < m$, the formulas for $A_i(z)$ follow from the discipline descriptions given above.

To find $A_i(z)$, $i < m$, for the quorum service discipline, we introduce an auxiliary cumulative random process $S_1 = i + X_1$, $S_2 = S_1 + X_2$, $S_3 = S_2 + X_3, ...,$ where i is the initial number of customers in the queue, $i = 0, 1, 2, ..., r - 1$; r is the minimal number of customers necessary for starting the service (threshold), and X_k, $k = 1, 2, 3, ...,$ are i.i.d. random variables distributed as α. Let $\{c_{ij}(r), \ j \geq r\}$, be the distribution of the value of the first weak excess of the process S_k, over level r. It follows from the definition of the quorum discipline that

$$A_i(z) = \sum_{j=r}^{m-1} c_{ij}(r) K_j(z) + \sum_{j=m}^{\infty} c_{ij}(r) z^{j-m} K_m(z), \ i < r, \quad (4.6)$$

$$A_i(z) = K_i(z), \ r \leq i < m. \quad (4.7)$$

For the discipline a), when $r = 0$, equation (4.6) does not specify any $A_i(z)$. For the discipline b), when $r = 1$, it is obvious that $c_{0j} = a_j$, $j = 1, 2, ...$. The case with the discipline c) will be considered separately.

Since $\{Q_n\}$ is a Markov chain with transition Δ_m - matrix, the general approach to the analysis of the ergodicity of such Markov chains developed in [1, 2, 3] can be applied in this case. Proceeding similar to [3], we can obtain the following result which contains necessary and sufficient conditions for the ergodicity of the chain $\{Q_n\}$.

Theorem 4.1: *Markov chain $\{Q_n\}$ is ergodic if and only if*

$$K'_m(1) - m = K'(1) + (1 - p^m)L'_m(1) - m < 0. \quad (4.8)$$

If the individual service is performed one-by-one, then, on the strength of (4.3)*, we can rewrite* (4.8) *in terms of the parameters of the system as*

$$\kappa - p^{m-1} q\rho + (1 - p^m)m\rho - m < 0, \quad (4.9)$$

where $\kappa = K'(1) = \lambda ab$, $\rho = L'(1) = \lambda ag$.

If, on the other hand, the individual service is parallel, then, using (4.4)*, we rewrite* (4.8) *as*

$$\lambda a\{b + \int_0^\infty t \, d\{(p^{m-1} - p^m)[G(t)]^{m-1} + (1 - p^{m-1})[G(t)]^m\} - m < 0.$$

Denote by p_i, $i = 0, 1, 2, ...,$ the steady-state probabilities of the chain $\{Q_n\}$ and by $P(z) = \sum_{i=0}^\infty p_i z^i$ −the generating function of these probabilities.

Theorem 4.2: *If the ergodicity condition* (4.8) *holds true, then*

$$P(z) = \frac{\sum_{i=0}^{m-1} p_i[z^m A_i(z) - z^i K_m(z)]}{z^m - K_m(z)}. \tag{4.10}$$

This result follows from [8] where a more general Markov chain (with so-called $\Delta_{m,n}$ transition matrix) is analyzed.

As in [3], the presence of the unknown probabilities p_i, $i = 0, 1, ..., m-1$, on the right side of the relation (4.10) prevents us from determining the generation function $P(z)$. One way to find them is based on so-called matrix-analytic methods (see Neuts [9]). This technique has been shown to be a reliable solution method provided that entries a_{ij} of the transition matrix A can be calculated with good precision for sufficiently large i, j. Another method to find the unknown probabilities is based on the following results.

Lemma 4.1: *If the ergodicity condition* (4.8) *is true, then the equation*

$$z^m - K_m(z) = 0 \tag{4.11}$$

has exactly m roots in the unit disk $|z| \le 1$ (counting with multiplicities). One of the roots is $z = 1$; (4.11) may have $k > 1$ roots on the unit circle $|z| = 1$ if and only if m is divisible by k and $K_m(z)$ is a generating function of a lattice distribution with period k. (All such roots have to be $k-$th roots of unity.)

Lemma 4.2: *There is a unique set of p_i, $i = 0, 1, ..., m-1$, such that $P(z)$ defined by* (4.10) *is analytic at the roots of* (4.11) *in $|z| \le 1$ and $P(1) = 1$.*

Lemmas 4.1 and 4.2 lead to a system of equations for p_i, $i = 0, 1, ..., m-1$,whose coefficients depend on the values and (in the case of multiple roots) derivatives of functions $A_i(z)$ at the roots of (4.11) in $|z| \le 1$. It can be proved [2] that if the ergodicity condition is satisfied, this system has a unique solution and, therefore, all unknown probabilities p_i, $i = 0, 1, ..., m-1$, and, by the same token, $P(z)$ can be found.

Special Case: a System with Quorum Threshold Equal to the Server's Capacity.

An important special case arises naturally when level r is chosen to be equal to m. It can be shown that that under this condition $P(z)$ can be found in an explicit form. Using (4.6) and (4.7) under the assumptions of this case, formula (4.10) can be rewritten as

$$P(z) = \frac{K_m(z)\sum_{i=0}^{m-1} p_i[\sum_{j=m}^{\infty} c_{ij}(m) z^j - z^i]}{z^m - K_m(z)}. \tag{4.12}$$

Exploiting a special relation between the quantities c_{ij}(m) and the distribution of arrival group sizes, one can show that the analytic properties of the right-hand side of (4.12) lead to the following theorem (the proof of which is omitted here).

Theorem 4.2: *If the ergodicity condition* (4.8) *holds true, the generating function $P(z)$ for a system with quorum threshold $r = m$ is given by*

$$P(z) = C\frac{[a(z) - 1]K_m(z)\prod_{j=1}^{m-1}(z - \zeta_J)}{[z^m - K_m(z)]}, \tag{4.13}$$

where ζ_j, $j = 1, 2, ..., m-1$, are the roots of (4.11) *in $|z| \le 1$ other than* 1 *and the normalizing constant C is*

$$C = \frac{m - K'_m(1)}{a'(1)\prod_{j=1}^{m-1}(1 - \zeta_J)}. \tag{4.14}$$

4.3 Minimization of the Queue Length and other Optimization Problems

Now we can come back to our main dynamic optimization problem: finding the optimal size of the group taken for testing. Unlike static optimization problems of Sections 2 and 3, in this case it is more appropriate to take as a criterion of optimality the average number of customers waiting for testing (or the total combined cost of the testing procedure related to this value). This optimum is supposed to be found subject to the following restrictions. First of all, it is necessary to take into account restrictions resulting from the condition (4.8) for the ergodicity of the queueing process. In addition, since any real-life problem may have its own specific requirements and restrictions, some further adjustments may be needed. For example, in HIV screening practice the size of the pool of blood samples taken for testing can not be too large because of the above mentioned "dilution effect". After these adjustments the main optimization problem can be formulated. Namely, given q and the other parameters of the system, a value of m is to be found such that the average number $P'(1)$ of customers in the queue is minimum. Although some computational work has to be done for this purpose (finding the roots of $z^m - K(z)$, finding the first weak excess probabilities, setting up and solving the system of equations for p_i, $i = 0, 1, \ldots, m-1$), relations (4.6), (4.7) and (4.10) obtained in this Section, in principle, make it possible to solve this problem.

The results obtained above also enable us to solve other important optimization problems which could not be solved by static optimization methods. One of them is mentioned in Section 1 (problem 4). Now, using formulas (4.6), (4.7) and (4.10) we can determine a better testing policy if at the beginning of the group test the number of customers in the queue is less than optimal. Clearly, to do this it is enough to compare the average queue length $P'(1)$ for both "waiting server" and "quorum" service disciplines under all other parameters of the system and the arrival process being equal.

References

[1] Abolnikov, L.M., Investigation of a class of discrete Markov processes, *Izvestiya Academii Nauk SSSR, Tekhnicheskaya Kibernetika* **2** (1977), 69–82 (English Translation: *Engineering Cybernetics* **15**:2 (1977), 51–63).

[2] Abolnikov, L.M. and Dukhovny, A.M., Necessary and sufficient conditions of ergodicity for the Markov chains with $\Delta_{m,n}(\Delta'_{m,n})$ transition matrix, *J. Appl. Math. Simulation* **1**:1 (1987), 13–24.

[3] Abolnikov, L.M. and Dukhovny, A.M., Markov Chains with transition delta-matrix: Ergodicity conditions, invariant probability measures and applications, *J. of Appl. Math. and Stoch. Anal.* **5**:1 (1992), 83–98.

[4] Behets F. et al, Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models, *AIDS* **4** (1990), 737–741.

[5] Chaudhry, M.L. and Templeton, J.D.C., *A First Course in Bulk Queues*, Wiley, New York 1983.

[6] Dorfman, R., The detection of defective members of large population, *Annals of Math. and Stats.* **44** (1943), 436–441.

[7] Emmanuel, J.C. et al, Pooling sera for human immunodeficiency virus (HIV) testing: An economical method for use in developing countries, *J. of Clinical Pathology* **41** (1988), 582–585.

[8] Litvak, E., Tu, X.M. and Pagano, M., Screening for the presence of a disease by pooling sera samples: Simplified procedures, *J. of Amer. Stat. Assoc.* **89** (1994), 424–434.

[9] Neuts, M.F., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York 1989.

[10] Wein, L.M. and Zenios, S.A., Pooled testing for HIV screening: Capturing the dilution effect, *Operations Research* **44**:4 (1996), 543–569.