COMPARISON OF COSTS FOR MULTI-STAGE GROUP TESTING METHODS IN THE PHARMACEUTICAL INDUSTRY

C. Matthew Jones and Anatoly A. Zhigljavsky
School of Mathematics
Cardiff University
CF24 4YH, UK.
JonesMC1@Cardiff.ac.uk and ZhigljavskyAA@Cardiff.ac.uk

Key Words: group testing; cost function; two-stage procedure; three-stage procedure; row-and-column procedure.

ABSTRACT

The issue of whether to test components individually, or alternatively in groups, in order to detect certain chemical properties remains an important issue in the pharmaceutical industry. Economic viability is of paramount importance since, for multi-stage procedures, the cost of additional stages must be taken into consideration, along with the cost of testing mixtures of components. Optimum groups sizes are calculated for the two-stage, three-stage (both members of Li's family of algorithms) and the row-and-column procedures. The γ-two-stage design is investigated, which involves using a γ-separating design at the first stage, followed (if necessary) by a strongly separating design at the second stage. Finally, comparisons are made between the costs of single and multi-stage procedures, for both optimum and standard groups sizes, through the use of two different cost functions.

1. INTRODUCTION

As company chemical libraries continue to expand, strategies to efficiently screen compounds remains an important issue. The principle aims are to use the minimal amount of resources in the smallest time scales, whilst maintaining the quality of the resultant data. A method currently in use is the pooling of samples, which entails the analysis of mixtures of compounds. If a pooled sample produces a positive result, the individual components are

assayed separately to identify which component or components were active.

A number of statistical issues arise from this strategy: what is the optimal number of compounds to pool, how will the strategy affect false positive (a component that is labelled as active when it is inactive) and false negative (a component that is labelled as inactive when it is active) error rates, and, how does this strategy compare economically to a single stage screen of assaying all individual compounds.

Additionally there are practical constraints: much of the assay work is performed by robots, thus the final strategy needs to be compatible with an automated procedure, also random access to samples of single compounds in the compound library is performed manually, this adds a substantial amount of time to multi-stage processes.

In a particular test, either an individual component or a mixture of components are tested for activity. The result is given in the form of a numerical value. The distribution of the activity measurements over all components is positively-skewed, and it is only those components in the extreme right-hand tail of the activity scale that are of interest. In a typical experiment between 10 and 100 (d) of the most active components amongst 10^6 (n) components would be searched for. A cut-off point is specified such that a component having an activity reading exceeding this point is deemed to be a hit. If the activity is less than this arbitrary point, then the component is judged to be inactive. To determine the importance of active components relative to one another, they are arranged in ascending order of activity.

The problem of false positives can be overcome in an obvious way: after the active components are identified, they are individually re-tested for their activity. The issue of false negatives is only relevant to those components fringeing around the cut-off point. It is assumed that the errors in the tests are small, therefore it is unlikely that the very active components are classified as false negatives.

There are two slightly different ways of formalising the problem in the form of a rigorous group testing problem. One way is to consider it as the so-called *hypergeometric group testing problem* where the number d of the active components is fixed (d can also be the upper bound for the number of active components). Alternatively, the problem can be considered as a

binomial group testing problem, where it is assumed that the probability of finding an active component by one simple trial is $p = d/n$, and that the activity of different components are independent (cf. [1]).

The principal difference between the present study and previous papers on group testing is the consideration both of costs (penalties) for additional stages, and for the number of components in a test group. These cost functions have been introduced in [2], where some preliminary results have also been reported. The cost associated with the additional stages is due to alterations that must be made to the machinery used in pooling the samples.

Let Λ represent the cost incurred between successive stages and let c_s be the cost of testing a mixture of s components. It is assumed that $c_1 = 1$ (that is the cost of individually testing the components is 1). Let λ represent the normalised cost between successive stages ($\lambda = \frac{\Lambda}{n}$), in pharmaceutical applications typical values are $0.1 \leq \lambda \leq 0.3$.

Two simple cost functions that can be used are:

(i) $c_s = 1 + \kappa(s-1)^\zeta$ with $0 \leq \kappa < 1$ and $0 \leq \zeta \leq 1$

(ii) $c_s = 1 + \kappa \log s$ with $\kappa \geq 0$.

The costs are thus parameterised with additional two or three parameters, namely λ, κ, and perhaps ζ. If $\kappa = 1$ and $\zeta = 1$, the cost function in case (i) is $c_s = s$ (if $\Lambda = 0$), and the cost of the experiment is exactly the number of tests performed. The cost function (ii) seems to be the more natural in the pharmaceutical applications that we have in mind. This is due to the common belief that the number of substances required to test s components is proportional to $\log s$, where s is large. The cost function (i) is very flexible and describes an intermediate case between cost function (ii) and $c_s = s$ (i.e. the number of test performed).

When the cost incurred between stages (Λ) is taken into account, the standard sequential procedures become inefficient (even with respect to individual testing). Thus, we only consider these procedures when the cost is equal to the number of tests.

2. MINIMISING NUMBER OF TESTS

Different group testing strategies, as well as upper and lower bounds for the length of optimal algorithms, have been extensively studied for both formulations of the problem. In

this section costs shall be ignored (that is assume $\lambda = 0$ and $c_s = 1$ for any s) and different methods are characterised by the number of tests only.

References will be provided for some of the most well-known results, with the assumption that the total number of components n is large, the number of active components d is relatively small, and (in the binomial group testing model) the probability that a random component is active, p, is small. These assumptions conform with practical requirements.

The origin of group-testing is credited to [3] from whose work future studies stemmed. [1] extensively studied the binomial group testing model. For their main procedure, it was found that the expected number of group tests required to detect the d active components is as follows:

$$E(N) \cong -n \log_2 \left(\frac{1}{(1-p)} \right) + np \log_2 \left[\log_2 \frac{1}{(1-p)} \right]^{-1}, \quad n \to \infty \,. \tag{1}$$

Li's s-stage algorithm [4] was set to minimise the worst case number of tests using combinatorial group testing to detect the d active components. Let N denote the number of tests required to determine the active components, it has been found that:

$$N \leq e\, d(\log n - \log d), \quad \text{where} \;\; e = 2.7182818... \tag{2}$$

Hwang's generalised binary splitting algorithm [5] is an extension of the binary splitting procedure, Hwang suggested a way to include d (number of active components), into applications of binary splitting such that the total number of tests can be reduced. If n is sufficiently large the number of tests for this algorithm satisfies:

$$N \leq d(\log_2 n - \log_2 d + 3) \tag{3}$$

more precise bounds are given in [6] and it is these values that appear in Table I. General formulas for the expected number of tests to determine active components in multi-stage procedures are also discussed in [7].

An alternative literature on the hypergeometric group testing problem deals with the probabilistic technique of derivation of the existence theorems for the one-stage designs. The pioneering work in this area was done by [8]. Many authors have continued this work,

some examples are listed in [6]. For a fixed number of active components d and $n \to \infty$, the best known upper bound has been derived in [9], see also [6], Theorem 7.2.15:

$$N \le dA_d(1+o(1))\log_2 n$$

where:

$$\frac{1}{A_d} = \max_{0\le q\le 1}\max_{0\le Q\le 1}\left\{-(1-Q)\log_2(1-q^d) + dQ\log_2\frac{q}{Q} + d(1-Q)\log_2\frac{1-q}{1-Q}\right\}$$

and $A_d = \frac{2}{d}\log_2 e(1+o(1))$ as $d\to\infty$. Asymptotically, when both n and d are large,

$$N \le N_*(n,d) \sim \frac{e}{2}\,d^2\log n\,,\quad n\to\infty,\ d=d(n)\to\infty,\ d(n)/n\to 0.$$

In the case where d is fixed and the number of components in every test group, say s, is also fixed, the upper bound for the length of optimum one-stage design is derived in [10]: $N \le N^* = N^*(n,d,s)$ where N^* is the minimum over $k = 1,2,\ldots$ such that

$$\frac{1}{2}\sum_{i=0}^{d-1}\binom{n}{i\ \ d-i\ \ d-i\ \ n-2d+i}\left(1-2\cdot\frac{\binom{n-t}{s}-\binom{n-2d+i}{s}}{\binom{n}{s}}\right)^k < 1$$

where $\binom{n}{a\ b\ c\ d} = n!/(a!b!c!d!)$ is the multinomial co-efficient. When $n\to\infty$, the results in [10] imply that $N(n,d,s) = \lceil N^{(\mathrm{as})}(n,d,s)+o(1)\rceil$ where

$$N^{(\mathrm{as})}(n,d,s) = \frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1-2\frac{s}{n}(1-\frac{s}{n})^d)}\,. \tag{4}$$

Analogous results hold when d is the upper bound for the number of active components.

Optimisation of the right-hand side in (4) with respect to s, the size of the test groups, gives $s_{opt} = s(n) = n/(d+1)$ and

$$N^{(\mathrm{as})}(n,d) \sim \frac{e}{2}\,d^2\log n\,.$$

The approximations (upper bounds) for the lengths of different group testing strategies are compared in Table I for $n = 10^6$ and $d = 10, 50$, and 100 (with the corresponding values of p equal to 0.00001, 0.00005, and 0.0001).

Table I: Approximations for expected number of tests in various procedures for $n = 1000\,000$.

Procedure	p=0.00001 ($d = 10$)	p=0.00005 ($d = 50$)	p=0.0001 ($d = 100$)
Sobel and Groll alg. (1)	176	761	1 421
Li's s-Stage alg. (2)	313	1 347	2 504
Hwang alg. (3)	184	810	1 521
One-stage alg. (4)	1 904	38 110	140 930

As we see from Table I, both the Sobell and Groll and Hwang's algorithm are about d times better than the best one-stage procedures. But the situation alters when the cost λ for additional stages is taken into account. As we show later, see Figures 4, 5 and 6, for reasonable values of λ, multi-stage strategies, from the family of the Li's algorithms, with three or more stages become less efficient than the one-stage and two-stage strategies. This also applies to other sequential algorithms.

The formula (4) giving the upper bound for the length of optimal one-stage procedure can easily be extended to calculate the normalised cost:

$$\tilde{C}^{(\mathrm{as})}(n, d, s) = \frac{c_s}{n} \cdot \frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1 - 2\frac{s}{n}(1-\frac{s}{n})^d)}\,. \tag{5}$$

For the cost function $c_s = 1 + \kappa \log s$, optimisation of the right-hand side of (5) with respect to s again gives the asymptotically optimum rate $s = n/(d+1)$ for s. In the case of the cost function $c_s = 1 + \kappa s$, $0 < \kappa < 1$, the individual testing procedure ($s = 1$) is asymptotically optimum.

3. TWO-STAGE PROCEDURE

A typical procedure used in the pharmaceutical industry to detect active components is essentially the classical Dorfman's procedure (see [3], a short description can also be found in [11]), which is a particular case of Li's family of algorithms.

The motherplate consists of m columns and k rows, giving in total km cells, with each cell containing a different component (assume for simplicity $n = km$). At the first stage, a mixed

sample of the m components in each row is taken and deposited into the daughterplate, the mixtures are then tested for activity. At the second stage, if the mixture is active then it is deemed to be a hit, the m components that make the hit are then tested individually to test their activity.

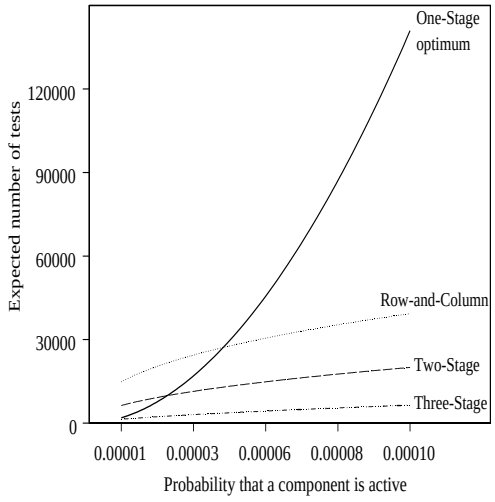

Figure 1: Mean number of tests as a function of p for the one-stage, two-stage, three-stage (Section 4) and row-and-column (Section 5) procedures with optimum parameters.

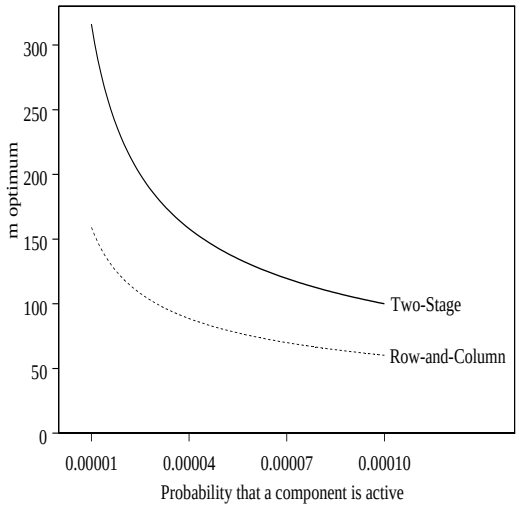

Figure 2: Optimum values of m minimising the mean number of tests for the two-stage and the row-and-column procedures.

The binomial group testing model will be followed, and it is assumed that prior to the experiment the probability that a component is active is p. In practice, p is very small with a typical value being $p = 0.00001$ (this would correspond to $d = 10$ and $n = 10^6$). The activity of every component is assumed to be independent of every other component. A group is deemed active if it contains at least one active component, it is assumed that in an experiment we are able to detect the activity of a group without error. Then we have:

$$\Pr \text{ (a component is inactive)} = 1 - p,$$

$$\Pr \text{ (a group of components is inactive)} = (1 - p)^m,$$

$$\Pr(\text{a group of components is active}) = 1 - (1 - p)^m = P_{m,p}.$$

Hence, the first stage can be modelled by a sequence of $k = n/m$ Bernoulli trials with the probability of success being $P_{m,p}$. The number of successes (that is the number of active groups) is k', which is a random variable with a binomial distribution $k' \sim Bin(k, P_{m,p})$, thus:

$$Pr(k' = i) = \begin{pmatrix} k \\ i \end{pmatrix} {P_{m,p}}^i (1 - P_{m,p})^{k-i}; \;\; E(k') = kP_{m,p}; \;\; Var(k') = kP_{m,p}(1 - P_{m,p})$$

Hence, the standardised cost of determining the number of active components is calculated as follows:

$$\widetilde{C}(m, p, \lambda) = \frac{1}{m} c_m + \frac{k'm}{n} + \lambda$$

By Taylor's expansion $P_{m,p} \sim mp$ $(p \to 0)$, which gives:

$$E[\tilde{C}(m, p, \lambda)] \sim \frac{1}{m} c_m + pm + \lambda \,;$$

$$Var[\tilde{C}(m, p, \lambda)] \sim \frac{Var(k')m^2}{n^2} = \frac{mP_{m,p}(1 - P_{m,p})}{n} \approx \frac{m^2 p}{n}$$

This implies that when p is small and $n \to \infty$, the expected cost tends to infinity but the variance of the cost remains bounded. The optimum value of m for minimising the total cost may be found by numerical optimisation. For the case where the cost function is $c_s = 1 + \kappa s$, including the case $\kappa = 0$,

$$\frac{d\tilde{C}(m, p, \lambda)}{dm} = -\frac{1}{m^2} + p = 0; \;\; \text{thus} \;\; m^2 = \frac{1}{p} \;\; \text{which gives} \;\; m_{opt} = \sqrt{\frac{1}{p}}.$$

Figure 1 shows the mean number of tests required to detect the active components for the optimum two-stage procedure, and Figure 2 shows the optimum value of m required to minimise the number of tests.

4. THREE-STAGE PROCEDURE

The three-stage procedure (again a particular case of the Li's family of algorithms) has the same first stage as the two-stage procedure. However, components from the active mixtures are then analysed in groups of size l rather than individually to detect activity. On the third stage, the groups that were active on the second stage are tested individually for activity.

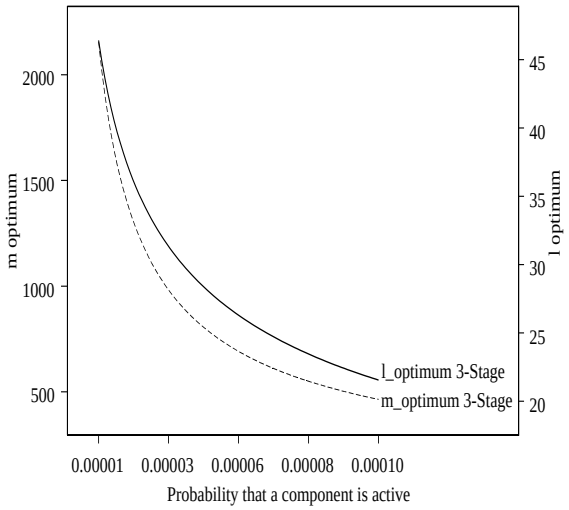

Figure 3: Optimum values of m and l for minimising the mean number of tests in three-stage procedure.

The binomial group testing model is again adopted. The cost of determining the number of active components for the three-stage procedure can be calculated as follows:

$$C(n, m, l, p, \Lambda) = \frac{n}{m} c_m + \frac{k'm}{l} c_l + lk'' + 2\Lambda \tag{6}$$

where $k' \sim Bin(k, P_{m,p})$ and $k'' \sim Bin(k', P_{l,p'})$. The first term in (6) counts the number of tests in the first stage which is $k = \frac{n}{m}$. At the second stage we have $n' = k'm$ components and these components are tested in groups of l items. This gives $n'/l = k'm/l$ tests at the third stage. As a result of the second stage, we have got k'' active groups each of size l, where analogously to the above: $k'' \sim Bin(k', P_l, p')$, with p' being the posterior probability of an individual component being active. We may assume that $p' = \frac{1}{m}$ (as p is small).

Thus, the expected normalised cost is

$$E[\widetilde{C}(m, l, p, \lambda)] \sim \frac{1}{m} c_m + \frac{pm}{l} c_l + lp(1 - (1 - \frac{1}{l})^m) + 2\lambda, \quad p \to 0.$$

Optimum values for m and l which minimise the number of tests to find active components can be found by means of numerical optimisation.

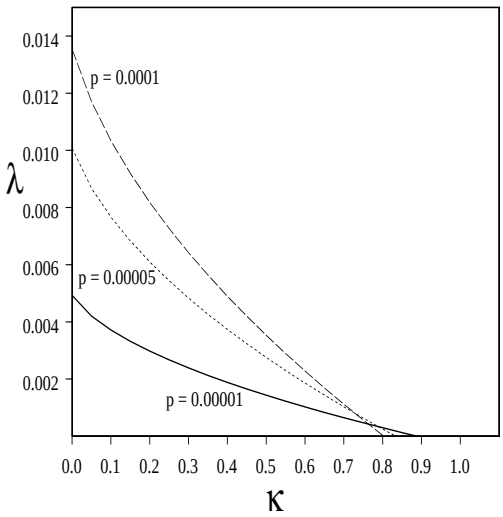

Figure 4: Values of λ as a function of κ such that the two-stage procedure has the same cost as the three-stage procedure with $c_s = 1 + \kappa s$.

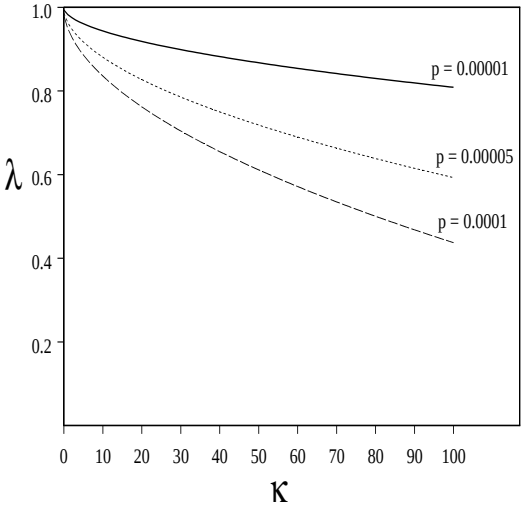

Figure 5: Values of λ as a function of κ such that one-stage procedure has the same cost as two-stage procedure with $c_s = 1 + \kappa \log s$.

Figure 1 shows the mean number of tests required to detect the active components for the optimum three-stage procedure, Figure 3 shows the optimum values of m and l required to minimise the number of tests.

5. ROW-AND-COLUMN PROCEDURE

A procedure that is popular in the pharmaceutical industry, is the row-and-column procedure. In this procedure the motherplate consists of $m > 1$ columns and $k > 1$ rows giving in total km cells, each cell containing a different component. Without loss of generality we assume that $m \leq k$. The number of motherplates to be tested is $r = \frac{n}{km}$, for simplicity we assume that r is an integer (in a typical experiment r is large).

At the first stage, a mixed sample of the m components in each row is taken, along with mixed samples of the k components in each column, these are then deposited in the daughterplate. The mixtures are tested for activity. We thus make $\frac{n}{k} + \frac{n}{m}$ tests in total at the first stage.

The number of active components in each motherplate is ξ, which is a random variable

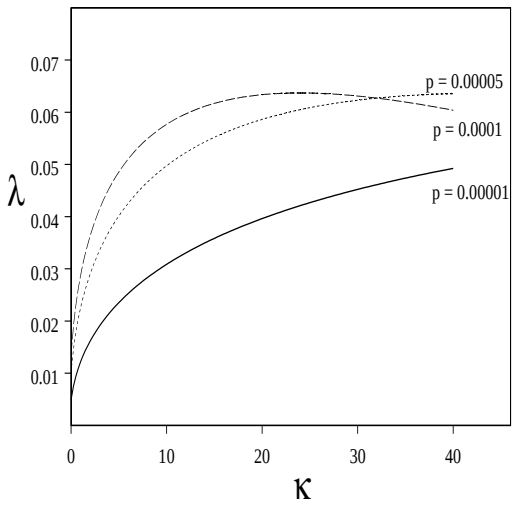

Figure 6: Values of λ as a function of κ such that two-stage procedure has the same cost as three-stage procedure with $c_s = 1 + \kappa \log s$.

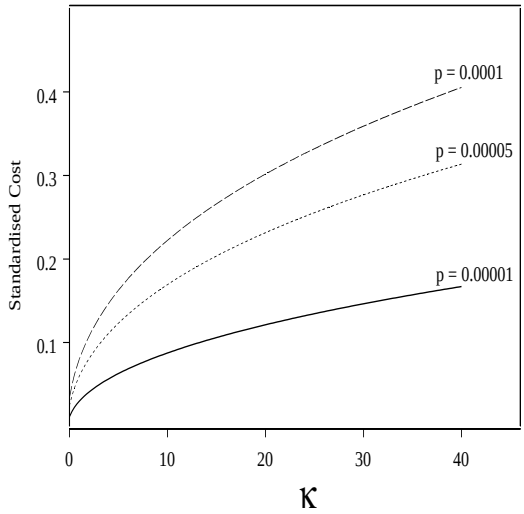

Figure 7: Standardised cost as a function of κ such that two-stage procedure has the same cost as three-stage procedure with $c_s = 1 + \kappa \log s$.

with a binomial distribution, $\xi \sim Bin(km, p)$, thus:$E(\xi) = kmp,\ Var(\xi) = kmp(1-p)$.

At the second stage we test the components that could be active. (These components are located at the intersections of the active rows and columns). If there is either zero or one active component in the motherplate then no further tests are required at the second stage.

However, if the number of active components in the motherplate ξ is larger than 1, then we must at most test all the intersections of the active rows and columns to detect the active components. If the active components are in different columns and rows, this will require at most ξ^2 further tests, for the case where the active components are in the same row or column the number of tests is smaller, due to the number of intersection points to test for the active components being less.

When there are $\xi \geq k$ active components then at most mk (the full motherplate) tests will be required. This implies the upper bound for the expected number of tests required to determine the number of active components at the second stage may be estimated as follows:

$$E[N(n,m,k,p)] \leq \frac{n}{mk}(p_2 \cdot 2^2 + ... + p_m \cdot m^2 + (p_{m+1} + ... + p_{km}) \cdot km)$$

where p_x is the probability that there are x active components in the motherplate (the terms related to the cases of zero and one active components in the motherplate do not require additional testing at the second stage).

Since $(p_3 3^2 + ... + p_m m^2 + (p_{m+1} + ... + p_{km}))km \leq (1-p_0-p_1-p_2)km$, we get:

$$E[N(n,m,k,p)] \leq \frac{n}{mk}\left(4\begin{pmatrix} km \\ 2 \end{pmatrix} p^2(1-p)^{km-2} + (1-p_0-p_1-p_2)km\right) \tag{7}$$

This inequality is sharp when p is small (which is typical in practice). Therefore, the right hand side of the inequality (7) can be used as an estimate of the average number of tests required to determine the active components at the second stage.

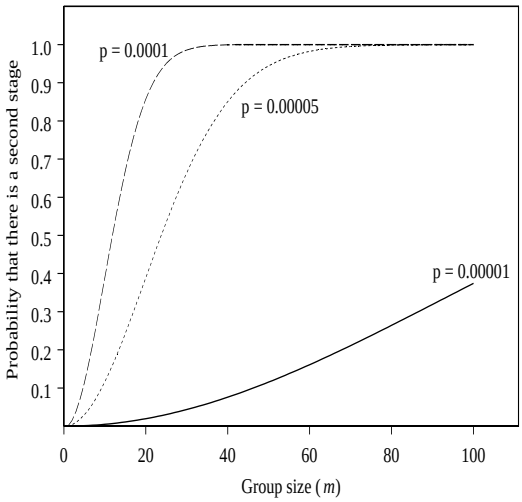

Figure 8: Probability that there is a second stage in the row-and-column procedure, as a function of m.

The expected normalised cost of the row-and-column procedure can be estimated as follows:

$$E[\tilde{C}(n,m,k,p)] \leq c_m\frac{1}{m} + c_k\frac{1}{k} + \frac{1}{mk}\left(2km(km-1)p^2(1-p)^{km-2} + (1-p_0-p_1-p_2)km\right) + \lambda Q$$

where Q is the probability that there is a second stage. The second stage is only required when the number of active components on a motherplate exceeds one. As $r = n/mk$ repre-

sents the number of motherplates and the probability of having two or more active components on each motherplate is $1 - p_0 - p_1 = 1 - (1-p)^{km} - kmp(1-p)^{km-1}$. This gives:

$$Q = 1 - ((1-p)^{km} + kmp(1-p)^{km-1})^r.$$

These probabilities, for the optimum case $k = m$, are plotted in Figure 8. We see from this plot that for practical values of k and m the probability that the row-and-column is actually a one-stage procedure is large for small values of p. The reason why we assume that $k = m$ is that the expected number of tests is always smaller when $m = k$ than $m < k$, if say mk = constant.

6. γ-TWO-STAGE PROCEDURE

Through the use of probability it is possible to derive upper bounds for the length of optimal non-sequential designs; such bounds are called existence theorems.

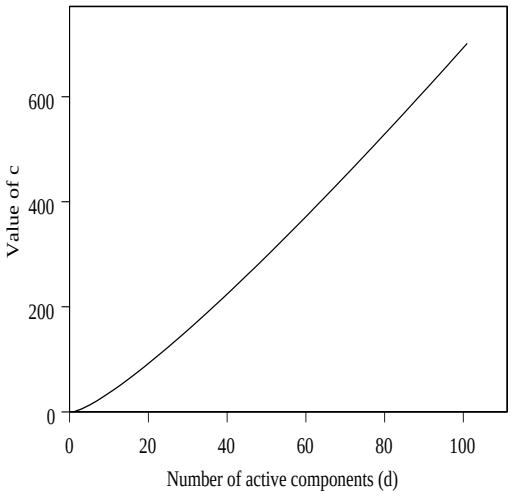

Figure 9: Values of c (see Equation (9)) against different values of d.

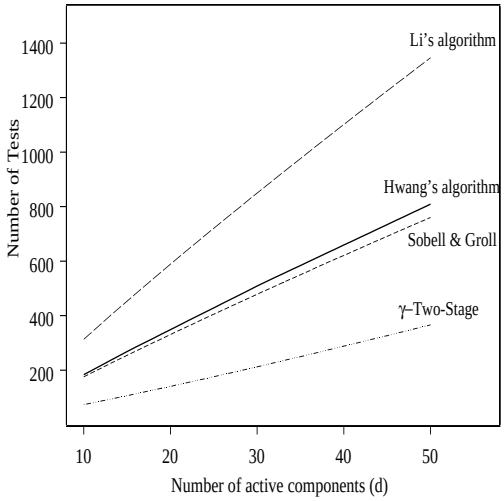

Figure 10: Number of tests required to detect different number of active components.

Let the target T be an unknown group of active components, and $\mathcal{T}$ be a set of all possible collections of groups containing d (or less) active components. Let X be a group of test components in a particular test, and $\mathcal{X}$ be a set of all admissible test groups. For

a pair of targets $T, T' \in \mathcal{T}$, $X \in \mathcal{X}$ separates T and T' if $f(X,T) \neq f(X,T')$. A design $\mathcal{D}_N = \{X_1, ..., X_N\}$ separates T in $\mathcal{T}$ if for any $T' \in \mathcal{T}$, such that $T' \neq T$, $\exists\ X \in \mathcal{D}_N$ which separates the pair (T, T'). A design $\mathcal{D}_N$ is a one-stage design if all T in $\mathcal{T}$ are separated.

A design $\mathcal{D}_N$ is $\gamma - \textit{separating}$ if :

$$\frac{|\{T \in \mathcal{T}\} : \text{design } \mathcal{D}_N \text{ separates } T \text{ in } \mathcal{T}|}{|\mathcal{T}|} \geq 1 - \gamma$$

where γ is constant, $0 \leq \gamma \leq 1$, and $|A|$ denotes the number of elements in a set A. Designs with $\gamma = 0$ are called strongly separating.

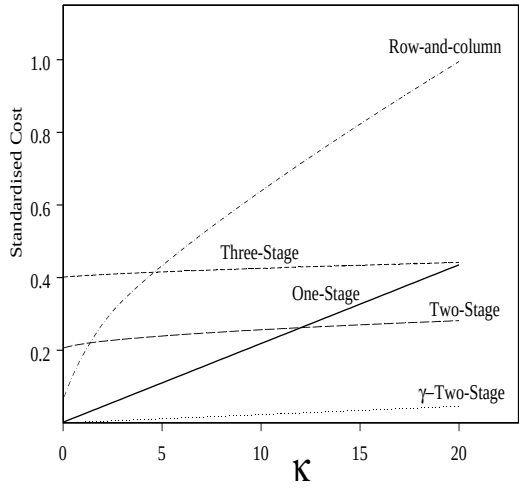

Figure 11: Optimum standardised cost for one-stage, two-stage, row-and-column and γ-two-stage procedures for $c_s = 1 + \kappa \log s$, $\lambda = 0.2$ and $p = 0.00001$.

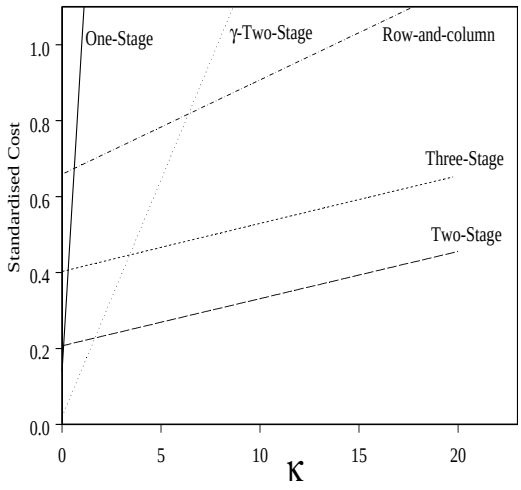

Figure 12: Standardised cost for one-stage, two-stage, row-and-column and γ-two-stage procedures for $c_s = 1 + \kappa \log s$, $\lambda = 0.2$, $s = 500$ and $p = 0.00001$.

The γ-two-stage procedure is investigated in this section. At the first stage a γ-separating design is applied, this uniquely determines the unknown collection of active components with probability 1-γ (assuming that every collection of active components is equally possible). The second stage is only needed if the collection of active components is not uniquely determined at the first stage; that happens with probability γ. At this stage we ignore previous results and apply the strongly separating design.

According to [12] the number of tests with group size s required to detect active components in a γ-separating design asymptotically (as $n \to \infty$) is:

$$N_\gamma(n,d) = \frac{n}{2s}\log_2 n - \log_2\gamma + c \tag{8}$$

where $c = c(d)$ is the solution of

$$\sum_{p=1}^{d-1} 2^{-c(d-p)/d}\frac{d!}{(p!(d-p)!)^2} = 1. \tag{9}$$

In the asymptotic case c is disregarded, values of c for typical values of d may be seen in Figure 9.

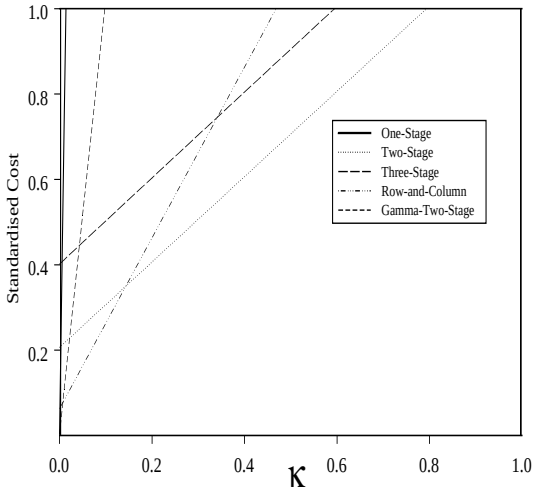

Figure 13: Optimum standardised cost for one-stage, two-stage, row-and-column and γ-two-stage procedures for $c_s = 1 + \kappa s$, $\lambda = 0.2$ and $p = 0.00001$.

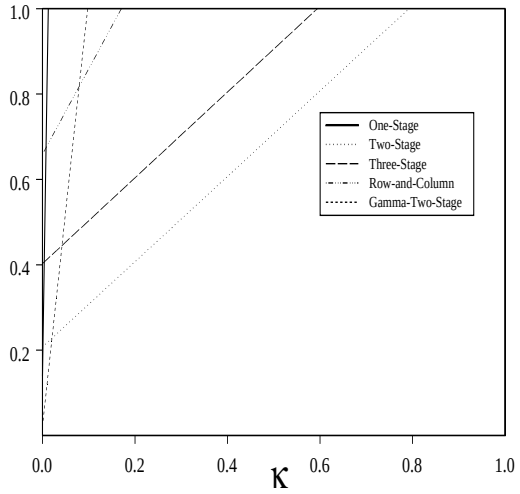

Figure 14: Standardised cost for one-stage, two-stage, row-and-column and γ-two-stage procedures for $c_s = 1 + \kappa s$, $\lambda = 0.2$, $s = 500$ and $p = 0.00001$.

Through using expressions (8) and (4), the number of tests performed using a γ-two-stage design to detect all active components may be derived as follows:

$$N_\gamma(n,d) = \frac{n}{2s}\ \log_2 n - \log_2\gamma + c + \gamma\left(\frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1 - 2\frac{s}{n}(1-\frac{s}{n})^d)}\right) \tag{10}$$

with the cost to detect all active components being:

$$C_\gamma(n,d) = \left(\frac{n}{2s}\,\log_2 n - \log_2\gamma + c\right)c_s + \gamma\left(\frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1-2\frac{s}{n}(1-\frac{s}{n})^d)}\right)c_s + \gamma\Lambda$$

The number of tests to determine the active components for Sobell & Groll procedure (1), Li's s-stage algorithm (2), Hwang's algorithm (3) and optimal γ-two-stage-separating procedure (10), against the number of active components can be seen in Figure 10. Optimum values for γ were found to be negligible, this implies that the γ-separating procedure almost becomes a one-stage procedure when minimising the number of tests. Figures 11 and 13 compare the standardized costs of all procedures with optimum parameters, for the cost functions $c_s = 1 + \kappa s$ and $c_s = 1 + \kappa \log s$, respectively. Figure 12 and 14 compare the standardized costs for all procedures, again with both cost functions, but with the group sizes fixed at 500 (a value that is often applied in practice). For small s the γ-two-stage-separating procedure is inefficient.

7. SUMMARY

For small values of d (say $d \leq 20$) the optimum one-stage procedure is as efficient as the multi-stage procedures in determining active components.

Multi-stage procedures reduce then number of tests required to detect active components considerably. However, when the additional costs associated with testing pooled samples, and in particular the cost of additional stages are taken into account, the three- and more stage procedures are inefficient. Often, the two-stage procedure can be considered to be a good compromise.

The row-and-column procedure proves to be worse than a two-stage procedure with related parameters. However, if d is small and λ is large, then this procedure can prove to be very cost-effective. Moreover, the number of components to be tested at the second stage for the row-and-column procedure is far smaller than that of the two- and three-stage procedures. With some probability the row-and-column procedure is even a one-stage procedure.

For $d \leq 100$ it was found that the optimum γ-two-stage design is more efficient than Li's s-stage algorithm, Sobell and Groll's procedure and Hwang's generalised binary splitting algorithm in determining the active components. When penalty costs for additional stages are

taken into consideration the γ-two-stage procedure proves to be by far the most economical of the four procedures.

The γ-two-stage design proves more cost effective than both the optimum one-stage and multi-stage procedures for the logarithmic cost function, $c_s = 1 + \kappa \log s$.

BIBLIOGRAPHY

[1] Sobel, M.; Groll, P.A. Group testing to eliminate efficiently all defectives in a binomial sample. Bell System Tech. J., **1959**,*38*, 1179–1252.

[2] Bond, B.; Fedorov, V.; Jones, C.M.; Zhigljavsky, A. A. Pharmaceutical applications of a multi-stage group testing method. Optimum Design 2000 (eds. Atkinson A.C., Bogacka B., Zhigljavsky A.), Kluwer Academic Publishers, **2001**, 155–166.

[3] Dorfman, R. The detection of defective numbers of large population. Ann. Math. Statist., **1943**,*14*, 436–440.

[4] Li, C.H. A sequential method for screening experimental variables. J. Amer. Statist. Assoc., **1962**, *57*, 455–477.

[5] Hwang, F.K. A method for detecting all defective members in a population by group testing. J. Amer. Statist. Assoc., **1972**,*67*, 605–608.

[6] Du, D. Z.; Hwang, F. K. Combinatorial Group Testing (second edition). World Scientific, Singapore, **2000**.

[7] Patel, M. Group screening with more than two stages. Technometrics, **1962**, *4*, 209–217.

[8] Renyi, A. On theory of random search. Bull. Amer. Math. Soc., **1965**,*71*, 809–828.

[9] Dyachkov, A.G.; Rykov, V.V; Rashad, A.M. Superimposed distance codes. Problems Control Inform. Th., **1989**,*18*, 237–250.

[10] Zhigljavsky, A.A. Asymptotic upper bounds for the optimal design length in factor screening experiments. MODA 5 - Advances in Model-Oriented Data Analysis and Experimental Design (eds. Atkinson A.C., Pronzato L. and Wynn H.P.), **1998**, 85–93.

[11] Feller, W. An introduction to probability theory and its Applications. Wiley Publications, **1960**, V. I (Second Edition, Chapter 9, Excercise 26), 225.

[12] Zhigljavsky, A.A. Probabilistic existence theorems in group testing, J. Statist. Planning and Inference (to appear). **2001**.