Chapter 1

PHARMACEUTICAL APPLICATIONS OF THE MULTI-STAGE GROUP TESTING METHOD

Brian Bond and Valeri Fedorov
Smithkline Beecham Pharmaceuticals
Statistical Sciences Department
Brian_C_Bond@sbphrd.com and Valeri_V_Fedorov@sbphrd.com

Matthew Jones and Anatoly Zhigljavsky
Cardiff University
School of Mathematics
JonesMC1@Cardiff.ac.uk and ZhigljavskyAA@Cardiff.ac.uk

Abstract An important problem in pharmaceutical research is whether individual testing of components should be made, or alternatively, the components should be tested in groups. Of more importance is that the cost of the experiment is economically viable, for multi-stage procedures the cost of additional stages must be taken into consideration along with the cost of testing the mixtures of components. Optimum group sizes are calculated for the two-stage, three-stage (both members of Li's family of algorithms) and the row-and-column procedures, along with the minimum number of tests required to determine all of the active components. Finally, comparisons are made between the costs of the one, two, and three-stage procedures using two different cost functions for the cost of testing mixtures of components.

1. INTRODUCTION

1.1 PHARMACEUTICAL BACKGROUND

High-throughput (HTP) drug screening is an important and widely used strategy for identifying novel compounds as leads for future work in the pharmaceutical industry. Leads are later submitted to further biological investigation through functional assays on whole cells or tissue. They are also used in ini-

tiating chemical exploration for better compounds through techniques such as quantitative structure activity relationships (QSAR).

As company collections of compounds grows technologies have been and continue to be developed to efficiently screen them using the minimal amount of resources in smaller time scales whilst maintaining the quality of the resulting data. One method implemented by a number of companies is pooling samples. If a pooled sample produces a positive result then the individual components would be assayed separately to identify which individual compound or compounds 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 and false negative error rates?
- How does this strategy compare to a single stage screen of assaying all individual compounds.

Additionally there are a number of practical constraints:

- Much of the assay work is performed by robots and hence the final strategy needs to be compatible with robotics.
- Random access to samples of single compounds in the compound library is performed manually and adds substantial amount of time to a multi-stage process.

New assay technologies such as miniaturisation, single molecule detection and chip-based assays are fast improving throughput and out-pacing the time and resource savings a two-stage procedure makes. Complete data from all compounds is perceived to be a clear advantage of running the single step process when data-mining of the resulting database is expected to become the norm.

1.2 STATISTICAL BACKGROUND

For simplicity, assume that we are only interested in activity of a large number of different components, potential drug compounds, against certain enzymes relating to certain diseases.

In a particular test, either an individual component or a mixture of them is tested for activity. The result is given in the form of a numerical value. If we were to look at the distribution of the activity of all the components we would find a positively-skewed distribution. We are only interested in the components that are on the far right tail of the activity scale. In a typical experiment we would search for $d = 50$ the most active components among the total number $n = 500\,000$ components.

A cut-off point is specified whereby if a component has an activity reading greater than this point then it is deemed to be a "hit", however if the activity is less the component is judged to be inactive. When a component is labelled as being active when in fact it isn't, it is called a *false positive*. In cases where the component is judged to be inactive when it is active, it is said to be a *false negative*. After all of the active components have been identified they are ranked in ascending order of activity, which determines their importance relative to one another.

The problem with *false positives* could be overcome in an obvious way: following the main tests components are individually retested for their activity. The issue about *false negatives* is only related to those components fringeing around the cut-off point. We assume here that the errors in the tests are reasonably small, therefore the *false negative* problem does not concern the very active components.

There are two slightly different ways to formalise the problem as 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, the problem is mathematically identical to a problem which is called search for defectives (cf. Sobel and Groll (1958)). Alternatively, d could be the upper bound for the number of defectives. Another way, known as *binomial group testing problem*, is to assume that the probability to find an active component by one simple trial is $p = d/n$, the activity of different components are independent.

An important problem in pharmaceutical research is whether individual testing of components should be made or, alternatively, the components are to be tested in groups. (We shall call a group active if it contains at least one active component, we assume that in an experiment we are able to detect the activity of a group without error.)

The main difference between the present study and the papers on group testing mentioned below is the consideration of costs (penalties) for both additional stages along with the number of components in a test group.

Let Λ represent the cost incurred between successive stage and c_s be the cost of testing a mixture of s components, we shall assume that $c_1 = 1$ (that is the cost of individually testing the components is 1). Let λ represent the normalised cost between successive stages i.e. $\lambda = \frac{\Lambda}{n}$.

Two simple cost functions that can be used are as follows:

(i) $c_s = 1 + \kappa s^{\gamma}$ with $0 \leq \kappa < 1$ and $0 \leq \gamma \leq 1$

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

We thus parameterise the costs with additional two or three parameters, namely λ, κ, and perhaps γ. If $\kappa = 0$, the cost function in case (i) is 1 and the cost of the experiment is exactly the number of tests.

2. MINIMISING THE 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 we ignore the costs (that is assume $\lambda = 0$ and $c_s = 1$ for all s) and characterise different methods by the number of tests only.

Let us provide references for some of the most well-known results always assuming (which is in agreement with the practical requirements) 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.

The origin of group-testing is creditable to R. Dorfman (1943) and it is from his work that future studies stemmed. Sobel and Groll (1959) extensively studied the binomial group testing model. In their main procedure components that are proven to be active or inactive are never used in the subsequent tests, aside from such components at every stage the components are required to be separated into at most two sets. One, which is called the *defective set* (contains at least one active component) and the other the *binomial set* (the components act like independent binomial chance variables with probability p of being active). Let EN denote the expected number of group tests remaining to be performed, Sobel and Groll show that

$$EN \cong -n \log_2 \left(\frac{1}{(1-p)} \right) + np \log_2 \left[\log_2 \frac{1}{(1-p)} \right]^{-1}, \quad n \to \infty . \quad (1.1)$$

Li's s-Stage algorithm (1962) was set to minimise the worst case number of tests using combinatorial group testing to detect the d active components. At the first stage the n components are divided into g_1 groups of k_1 (some possibly $k_1 - 1$) components. All groups are then tested and those that are inactive are removed. In general at stage i, $2 \leq i \leq s$, items from the active groups of stage i -1 are pooled and arbitrarily divided into g_i groups of k_i (some possibly $k_i - 1$) components, and a test is performed on each group. k_s is set to be 1; thus every component is identified at stage s. Let N denote the number of tests required to determine the active components, it has been found that

$$N \leq e\, d(\log n - \log d), \quad \text{where } \; e = 2.7182818... \quad (1.2)$$

Hwang's (1972) generalised binary splitting algorithm is an extension of the binary splitting procedure, which involves the partitioning of n components into two disjoint groups such that neither has size exceeding $2^{\lceil \log_2 n \rceil - 1}$ ($\lceil x \rceil$ denotes the smallest integer value larger than or equal to x), then test one

group and the outcome will indicate either the test group or the other is active. Hwang suggested a way to co-ordinate the d (number of active components) applications of binary splitting such that the total number of tests can be reduced. The algorithm was as follows:

1. If $n \leq 2d-2$ then perform individual testing. If $n \geq 2d-1$, define $\alpha = \lfloor \log_2((n-d+1)/d) \rfloor$ ($\lfloor x \rfloor$ denotes the smallest integer value smaller than or equal to x).
2. Test a group of size 2^α, if inactive, the group is identified as good, set $n = n - 2^\alpha$ and return to the step 1. If the outcome is active use binary splitting to identify an active component and an unspecified number say x of inactive components. Let $n = n - 1 - x$ and $d = d - 1$. Return to step 1.

If n is large enough then the number of tests for this algorithm satisfies

$$N \leq d(\log_2 n - \log_2 d + 3) \tag{1.3}$$

General formulas for the expected number of tests to determine active components in multi-stage procedures are also discussed in Patel (1962).

Alternative literature on the hypergeometric group testing problems 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 Renyi (1965) and it has been successfully continued by many authors, some examples are listed in Du and Hwang (1993). For a fixed number of active components d and $n \to \infty$, the best known upper bound has been derived in Dyachkov, Rykov and Rashad (1989), see also Du and Hwang (1993), p.68: $N \leq dA_d(1+o(1)) \log_2 n$ where

$$\frac{1}{A_d} = \max_{0\leq q\leq 1} \max_{0\leq Q\leq 1} \Big\{ -(1-Q)\log_2(1-q^d) + dQ \log_2 \frac{q}{Q} + d(1-Q)\log_2 \frac{1-q}{1-Q} \Big\} \tag{1.4}$$

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

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

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 Zhigljavsky (1998): $N \leq N^* = N^*(n,d,s)$ where N^* is the minimum over $k = 1, 2, \ldots$ such that

$$\frac{1}{2}\sum_{i=0}^{d-1} \binom{n}{i\ \ d-i\ \ d-i\ \ n-2d+i} \left(1 - 2\cdot \frac{\binom{n-t}{s} - \binom{n-2d+i}{s}}{\binom{n}{s}}\right)^k < 1 \tag{1.5}$$

where $\binom{n}{a\ b\ c\ d} = n!/(a!b!c!d!)$ is the multinomial co-efficient. When $n \to \infty$, the results in Zhigljavsky (1998) imply that $N(n,d,s) = \lceil N^{(\mathrm{as})}(n,d,s) + o(1) \rceil$ where

$$N^{(\mathrm{as})}(n,d,s) = \frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1 - 2\frac{s}{n}(1-\frac{s}{n})^d)}\,. \qquad (1.6)$$

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

Optimisation of the right-hand side in (1.6) with respect to s, the size of the test groups, gives $s_{opt} = s(n) = n/(d+1)$ and

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

The approximations (upper bounds) for the length of different group testing strategies are compared in Table 1 for $n = 500\,000$ and $d = 10, 50,$ and 100 (The corresponding values of p are $0.00002,\ 0.0001,$ and 0.0002.)

As we see from the Table 1, the multi-stage strategies are about d times better than the best one-stage procedures. But the situation totally changes when the cost λ for additional stages are taken into account. As we show below, see Section 6. and Fig. 3(b), 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. The same holds for the other sequential algorithms.

The formula (1.6) giving the upper bound for the length of optimal one-stage procedure can easily be extended to calculate the costs: (1.6) implies that there exist one-stage procedures with the normalised cost

$$C^{(\mathrm{as})}(n,d,s) = \frac{c_s}{n} \cdot \frac{(d+1)\log n - \log(d-1)! - \log 2}{-\log(1 - 2\frac{s}{n}(1-\frac{s}{n})^d)}\,. \qquad (1.7)$$

For the cost function $c_s = 1 + \kappa \log s$, optimisation of the right-hand side of (1.7) 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 the 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 Dorfman (1943), a short description could also be found in Feller (1960), Chapter 9, Excercise 26), it is a particular case of the Li's family of algorithms and is described as follows. The motherplate consists of m columns and k rows, which gives in total km cells, with each cell containing a different component (assume for

Table 1: Approximations for expected number of tests in various procedures for $n = 500\,000$.

Procedure	d=10 (p = 0.00002)	d=50 (p= 0.0001)	d=100 (p=0.0002)
Sobel and Groll procedure(1.1)	137	566	1 032
Li's *s*-Stage algorithm (1.2)	128	544	1 006
Hwang algorithm (1.3)	187	815	1 197
One-stage Algorithm(1.6)	1 801	35 702	131 402

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.

Let us follow the binomial group testing model and assume that prior to the experiment the probability that a component is active is p, in practise p is very small with a typical value being $p = 0.0001$ (this would correspond to $d = 50$ and $n = 500\,000$). We also assume that the activity of every component is independent of every other component.
We have:
Pr(a component is inactive) = $1 - p$;
Pr(a group of components is inactive) =$(1 - p)^m$.
Thus,
Pr(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. The normalised cost is then

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

In practise p is small and therefore by Taylor's expansion $P_{m,p} \sim mp$ $(p \to 0)$, and we have:

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

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

This implies in particular that when p is small and $n \to \infty$, the expected cost tends to infinity but the variance of the cost stays 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 $\kappa = 0$,

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

Figure 1(a) shows the mean number of tests required to detect the active components for the optimum two-stage procedure, Figure 1(b) 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 that of the two-stage procedure, that is creating a mixture of the m components in each row from the motherboard, components from the active mixture 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 then tested individually for activity.

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

$$C(n,m,l,p,\Lambda) = \frac{n}{m}c_m + \frac{k'm}{l}c_l + lk'' + 2\Lambda \tag{1.8}$$

where $k' \sim Bin(k, P_{m,p})$ and $k'' \sim Bin(k', P_{l,p'})$.

The first term in (1.8) 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 we test these components 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: $k'' \sim Bin(k', P_l, p')$ where p' is the posterior probability of an individual component being active. Since p is small, the probability of two or more active components in a group of m items in the first stage is negligible, so we may assume that $p' = \frac{1}{m}$.

Thus, when p is small, the expected normalised cost is

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

The optimum values of m and l for minimising the number of tests to find active components could be found by means of numerical optimisation.

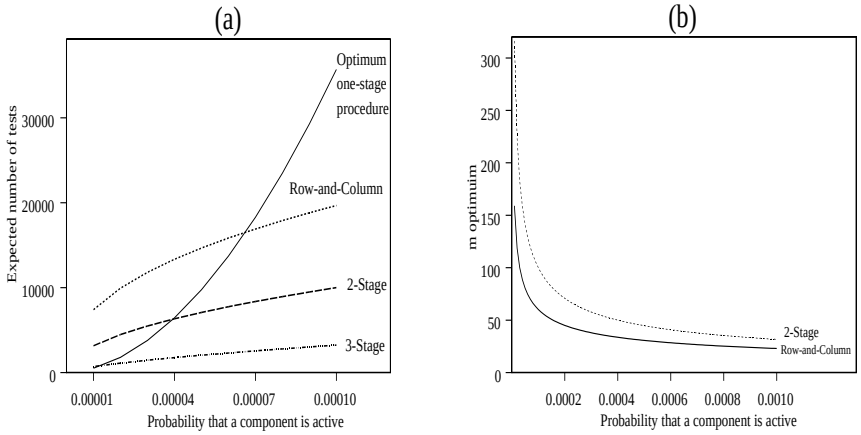

Figure 1.1 (a) shows the mean number of tests as a function of p for the one-stage, two-stage, three-stage (section 4.) and row-and-column (section 5.) procedures with optimum parameters; (b) shows the optimum values of m minimising the mean number of tests for the 2-stage and the row-and-column procedures

Figure 1(a) shows the mean number of tests required to detect the active components for the optimum three-stage procedure, Figure 2(a) shows the optimum values of m and l required to minimise the number of tests.

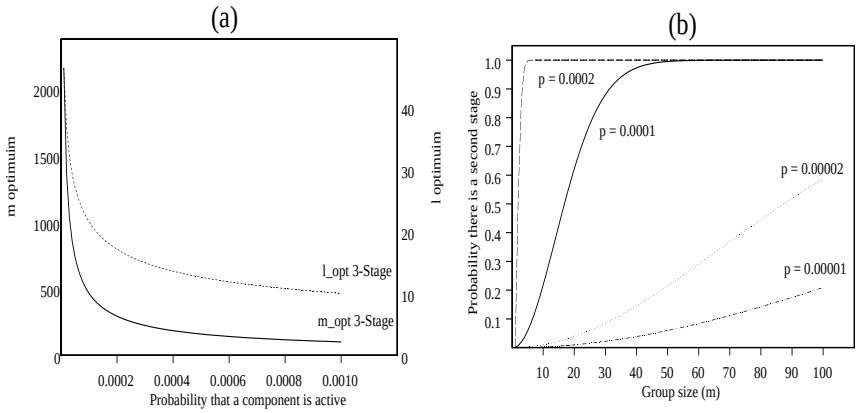

Figure 1.2 (a) shows the optimum values of m and l minimising the mean number of tests for the 3-stage procedure; (b) shows the probability that there is a second stage in the row-and-column procedure, as a function of m

5. ROW-AND-COLUMN PROCEDURE

For the row-and-column procedure the motherplate consists of $m > 1$ columns and $k > 1$ rows giving in total $k \cdot m$ cells, with 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, at the first stage we thus make $\frac{n}{k} + \frac{n}{m}$ tests in total. The number of active components in each motherplate is ξ, which is a random variable with a binomial distribution, $\xi \sim Bin(km, p)$, thus:

$$p_s = \Pr(\xi = s) = \left(\begin{array}{c} km \\ s \end{array} \right) p^s (1-p)^{km-s}, \quad s = 0, \ldots, km$$

$$E(\xi) = kmp, \quad Var(\xi) = kmp(1-p).$$

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) for activity. 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 motherboard 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 (if the active components are in the same row or column then the number of tests is smaller, the reason for this being that the number of intersection points to test for the active components will be less). Also, when there are $\xi \geq k$ active components then at most mk (the full motherplate) tests will be required. This implies that 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:

$$\tilde{C}(n, m, k, p) \leq \frac{n}{mk}(p_0 \cdot 0 + p_1 \cdot 0 + p_2 \cdot 2^2 + \ldots + p_m \cdot m^2 + (p_{m+1} + \ldots + p_{km})) \cdot km) \qquad (1.9)$$

If the value of p is small the following estimator may be used to estimate the expected number of tests at the second stage:

$$E[\tilde{N}(n, m, k, p)] = \frac{n}{mk}\left(4 \left(\begin{array}{c} km2 \end{array} \right) p^2 (1-p)^{km-2} + (1 - p_0 - p_1 - p_2)km \right)$$

since $(p_3 3^2 + \ldots + p_m m^2 + (p_{m+1} + \ldots + p_{km})km \leq (1 - p_0 - p_1 - p_2)km$.

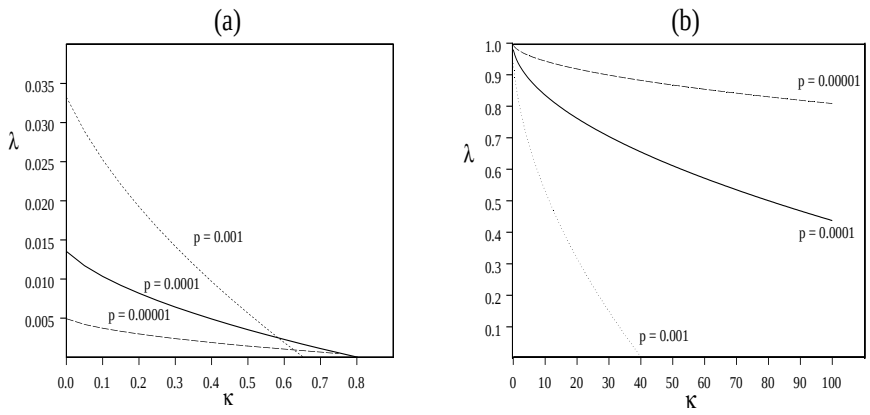

Figure 1.3 (a) shows the values of λ as a function of κ such that the 2-stage procedure has the same cost as the 3-stage procedure with $c_s = 1 + \kappa s$; (b) shows the values of λ as a function of κ such that 1-Stage procedure has the same cost as 2-stage procedure with $c_s = 1 + \kappa \log s$

Therefore, the expected normalised cost of the row-and-column procedure may be estimated as follows:

$$E[\tilde{N}(n,m,k,p)] \leq c_m \frac{1}{m} + c_k \frac{1}{k} + \frac{1}{mk}(2km(km-1)p^2(1-p)^{km-2} + (1-p_0-p_1-p_2)km) + \lambda \quad (1.10)$$

It is easy to estimate the probability that there is a second stage for the row-and-column procedure. Indeed, we do not need the second stage if there is never more than one active component in the motherplate. We have $r = n/(mk)$ motherplates and the probability of having two or more active components on each motherplate is

$$Q = 1 - p_0 - p_1 = 1 - (1-p)^{km} - kmp(1-p)^{km-1}.$$

Therefore, the probability that there is a second stage is $1-(1-Q)^r$. These probabilities, for the optimum case $k = m$, are plotted in Figure 2(b). 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.

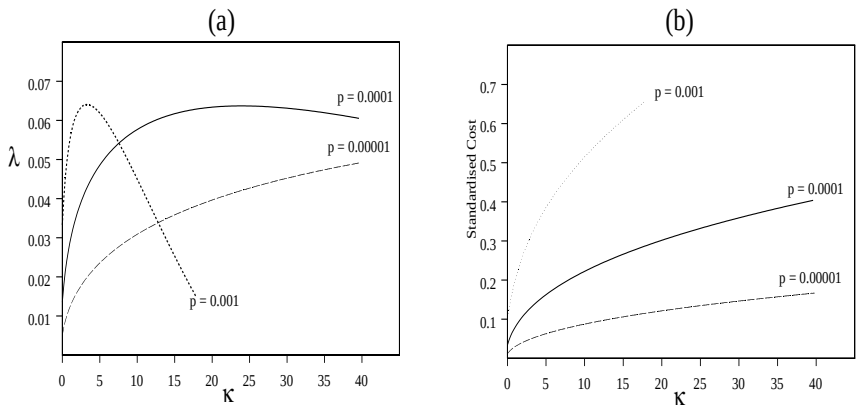

Figure 1.4 (a) shows the values of λ as a function of κ such that 2-Stage procedure has the same cost as 3-stage procedure with $c_s = 1 + \kappa \log s$; (b) shows the standardised cost as a function of κ such that 2-Stage procedure has the same cost as 3-stage procedure with $c_s = 1 + \kappa \log s$

6. CONCLUSIONS

It was found that if the number of active components, d, is reasonably small (say, $d \leq 10$) then the optimum one-stage procedures could be much more cost-effective than the best multi-stage procedures.

By increasing the number of components in the mixtures in the two-stage procedure we can significantly reduce the number of tests required to detect active components (the number of tests could be reduced by a factor of approximately 4 in the standard case, that is p=0.0001).

The reduction in the number of tests can be made even bigger if we apply a multi-stage procedure. However, if we take into account the costs associated with the number of components in a mixture and especially the penalties for extra stages, the three- and more stage procedures could be much less effective. The two-stage procedure is often a good compromise.

The row-and-column method is typically worse that the two-stage standard with related parameters. However, the number of components to be tested at the second stage for the row-and-column procedure is much smaller than for the two- and three-stage procedures. (With some probability, see Section 5., the row-and-column procedure is even a non-adaptive, one-stage procedure.)

References

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

Du D. Z. and Hwang F. K. (1993) *Combinatorial Group Testing*, World Scientific, Singapore.

Dyachkov A.G., Rykov V.V and Rashad A.M. (1989) *Superimposed distance codes* Problems Control Inform. Thy., **18**, 237–250.

Feller W. (1960) *An introduction to probability theory and its Applications*, Wiley Publications, Volume I, Second Edition, 225.

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

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

Patel M. (1962) *Group screening with more than two stages*, Technometrics **4**, 209–217.

Renyi A. (1965) *On theory of random search*, Bull. Amer. Math. Soc., **71**, 809–828.

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

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