SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853-3801

TECHNICAL REPORT NO. 960

February 1991

GROUP SEQUENTIAL TESTS FOR BIVARIATE RESPONSE: INTERIM ANALYSES OF CLINICAL TRIALS WITH BOTH EFFICACY AND SAFETY ENDPOINTS[1]

by

Christopher Jennison[2] and Bruce W. Turnbull[3]

[1]This research was supported by Grant RO1 GM28364 from the U.S. National Institutes of Health and by a U.K. SERC Visiting Fellowship GR/F 72864.

[2]School of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.
Tel: 0225-826468 Fax: 0225-826492 Email: cj@uk.ac.bath.maths

[3]Tel: 607-255-9131 Fax: 607-255-9129 Email: bruce@orie.cornell.edu

Summary

We describe group sequential tests for a bivariate response which treat the two response components separately, rather than through a single summary statistic. Such methods are necessary if the two responses concern different aspects of a treatment; for example, it may be desirable to show that a new treatment is both as effective and as safe as the current standard. We present a formulation of the bivariate testing problem, describe tests which satisfy Type I error conditions, and show how to find the sample size guaranteeing a specified power. We describe how properties of group sequential tests for bivariate normal observations can be computed by numerical integration.

1. Introduction

This paper was motivated by a design problem concerning a proposed clinical trial of an analgesic drug used to relieve arthritic pain. The primary endpoint was an efficacy measure of the amount of pain relief experienced by the patient but a secondary endpoint concerned a possible effect on the arthritic condition of the joint. The latter might be considered a "safety" variable. It was thought that the two outcome measures might be related, if for no other reason than that the drug's success in relieving pain may lead the patient to be less careful in protecting the joint.

Interim analyses in many clinical trials now usually include examination of several patient outcome measures. These endpoints may include several efficacy variables and/or several safety (e.g., toxicity) variables. Lan and Friedman (1986) cite the WHO clofibrate study where incidence of coronary heart disease was the primary endpoint but interim analyses were also performed on total mortality experience. Whitehead (1986) describes an example where outcome measures were mortality and the incidence of acute graft versus host disease in a bone marrow transplantation trial; he gives another example involving both birthweight and length of infants in a trial of treatments during pregnancy. Cox (1989) and Tang, Gnecco and Geller (1989a) describe trials in which there were both major and minor endpoints to be considered. Other examples of multivariate endpoints have been given by O'Brien (1984) and by Pocock, Geller and Tsiatis (1987).

For ease of exposition, we shall consider only the bivariate case, where each patient's data consist of the pair (X_1, X_2). Our methods can be extended to the general multivariate case; however, the formulation is somewhat more complex and the computations more tedious. We shall sometimes refer to X_1 as the "efficacy" variable and X_2 as the "safety" variable although they might instead be primary and secondary outcome measures of either type. For a prototype case we consider the one sample problem where the pairs (X_1, X_2) from different patients have independent bivariate normal distributions each with common mean $\mu = (\mu_1, \mu_2)$ and correlation ρ. We shall take the variances to be known and by appropriate rescaling we can then

assume $\mathrm{Var}X_1 = \mathrm{Var}X_2 = \sigma^2 = 1$, without loss of generality. Also without loss of generality, we can assume that high values of μ_1 and μ_2 are considered desirable. Thus, for example, if X_2 reflects the degree of toxicity, then a high value of μ_2 implies low toxicity.

Group sequential tests for the one-sample bivariate normal, known variance problem can be adapted easily to a wide variety of other situations. Sequences of test statistics with the same joint distributions could arise, for example, in a placebo-controlled comparative trial where test statistics are based on differences between the sample means in the two treatment groups. The same basic methods can also be applied when the variance is unknown, as long as the first group size is large enough to provide a good estimate of σ^2. Sequences of test statistics with approximately the same joint distribution as a sequence of sums of independent normal variates arise in trials with other types of response, for example, survival data (where the sequence of logrank statistics is approximately jointly normal), binary data, stratified 2×2 tables, or problems with covariates; for details see Whitehead (1983, Chap. 3), Pocock *et al.* (1987), and Jennison and Turnbull (1989).

Most methods for handling multiple endpoints have involved reduction to a univariate or ''global'' statistic: to a χ^2, F or Hotelling's T^2 statistic; to a likelihood ratio statistic (Kudo, 1963 and Perlman, 1969); to an approximate likelihood ratio statistic (Tang, Gnecco and Geller 1989b); to a linear combination or generalized least square statistic (O'Brien 1984); or to a Bonferroni adjusted P-value which considers the most extreme univariate P-value (Geller and Pocock, 1987 and Pocock *et al.* 1987). O'Brien (1984) also considers a nonparametric global statistic. These statistics are then used to test a null hypothesis $\mu_1 = \mu_2 = 0$, say, and the corresponding tests can have good or poor power properties depending on the class of alternative hypotheses considered. In the sequential setting, group sequential designs based on χ^2 and F-statistics have been considered by Jennison and Turnbull (1991a) and a group sequential design based on O'Brien's generalized least square statistic has been proposed by Tang *et al.* (1989a).

However, it will often be inappropriate to combine efficacy and safety variables into a single global statistic. Different response variables may appear in quite different ways in the formulation of a testing problem; Pocock *et al.* (1987) remark that "there are often disparate features of patient response unsuitable for combining" and they go on to describe an instance of myocardial infarctions and non-cardiovascular deaths in a coronary heart disease trial. Similarly, it may be inappropriate to combine different responses into a single measure to define a rule for early stopping. In this paper we shall concentrate on the case of a bivariate response and we shall introduce group sequential tests which involve the full bivariate nature of the response.

As mentioned by Goldman (1987), in pharmaceutical trials, proposals to the U. S. F. D. A. are expected to address side effect issues, even though efficacy is of primary interest in Phase II and III trials. For ethical reasons there will be interim monitoring of the safety variables; early stopping will be considered if there appears to be an unacceptable level of harmful side-effects. The role of the efficacy variables in early stopping depends on the circumstances of the trial. In the case of a comparative trial for the treatment of a life-threatening disease, ethical considerations may demand that the trial be stopped early if one treatment proves superior, as happened, for example, in the AZT trial for AIDS (Fischl *et al.* 1987, Barnes 1986). Otherwise, for example in some pharmaceutical industry applications, a trial may be allowed to continue to the planned termination even if efficacy results appear positive early on, in order to obtain the strongest possible evidence on safety and subgroups for submission with the New Drug Application; see Enas *et al.* (1989). Early stopping for economic reasons may be considered if results appear negative, so called "abandoning of lost causes", as discussed by Gould (1983).

In this paper we shall consider both fixed sample and group sequential designs for a study with bivariate response. We formulate the basic one-sided testing problem for a bivariate response in Section 2 and describe a fixed sample solution in Section 3. In Section 4, we propose a general form of group sequential test; details for specific cases are presented in Section 5. Finally, in Section 6, we discuss variations on the

basic method, including adaptation to unpredictable group sizes and connections with other problems, in particular the comparison of two treatments with a control when response is univariate.

2. The One-Sided Hypothesis and Two-Decision Problem

When deciding on actions to take regarding acceptance or rejection of a new treatment, it is useful to set up regions of indifference; see, for example, Meier (1975), Freedman and Spiegelhalter (1983) and Armitage (1987). With regard to variable X_i, $i=1,2$, we shall assume that there are constants $\varepsilon_i < \delta_i$ such that the new treatment is preferred if $\mu_i > \delta_i$ and is unacceptable if $\mu_i \leq \varepsilon_i$, but the region $\varepsilon_i < \mu_i \leq \delta_i$ is considered a region of "equivalence" or "indifference". The methods of Freedman and Spiegelhalter (1983) could be used to elicit these constants in practice. When the two variables are considered jointly, the parameter space for μ is divided into nine regions as shown in Figure 1.

Figures 1 and 2 about here

However, in the decision problem, one of only two, not nine, decisions must be made, namely acceptance or rejection of the treatment. Acceptance might mean, for example, proceeding to submit a New Drug Application to a regulatory body. The idea of considering just two possible actions may be an oversimplification but it is a convenient one and will often be appropriate. The nine regions in Figure 1 can be collapsed into two regions in several different ways; four of the most reasonable are displayed in Figure 2. Situation (a) occurs if the new treatment is acceptable as long as it is not unacceptable on either individual variable, whereas, in situation (b), the treatment must be preferred on both variables in order to be acceptable. Situation (c) might apply when X_1 is an efficacy variable and X_2 a safety variable: for an improvement in efficacy μ_1 must be in the preference zone, but for safety at least comparable with the current standard μ_2 need only be in the equivalence region.

Finally, (d) represents a situation in which the new treatment will be deemed acceptable if it is at least equivalent with respect to both variables and preferred on at least one.

It is convenient if all four situations (a) to (d) can be handled with the same theory. This can be achieved by considering a transformed problem with shifted variables, where

in case (a): $\mathbf{X}$ is replaced by $\mathbf{X} - (\varepsilon_1, \varepsilon_2)$,

in case (b): $\mathbf{X}$ is replaced by $\mathbf{X} - (\delta_1, \delta_2)$,

in case (c): $\mathbf{X}$ is replaced by $\mathbf{X} - (\delta_1, \varepsilon_2)$.

In case (d) there is no simple shift of origin but a compromise is to replace $\mathbf{X}$ by $\mathbf{X} - (\frac{1}{2}(\varepsilon_1+\delta_1), \frac{1}{2}(\varepsilon_2+\delta_2))$. The means μ_1 and μ_2 are transformed correspondingly. From now on, we shall assume we are working with the appropriately translated $\mathbf{X}$ and μ vectors. The advantage of using such a transformation is that there is now a single common region over which to control the maximum Type I error, defined to be

$$\alpha = \max \{\pi(\mu_1, \mu_2);\ \mu_1 \leq 0 \text{ or } \mu_2 \leq 0\} \tag{2.1}$$

where $\pi(\mu_1, \mu_2)$ is the probability of accepting the treatment when $\mu = (\mu_1, \mu_2)$. We shall consider procedures that satisfy a Type I error requirement $\alpha \leq \alpha^*$ where $0 < \alpha^* < 1$ is prespecified, and α is defined by (2.1).

3. Fixed Sample Design

We consider a fixed sample design for the problem of Section 2 in which n bivariate normal observations $\{X_{ij}; i=1, 2, j=1, \ldots, n\}$ are taken, $E(X_{1j}, X_{2j}) = \mu$, $\mathrm{Var}X_{1j} = \mathrm{Var}X_{2j} = \sigma^2 = 1$ and $\mathrm{Corr}(X_{1j}, X_{2j}) = \rho$ for $j = 1, \ldots, n$. We define $\bar{X}_i = n^{-1}(X_{i1} + \ldots + X_{in})$, $i=1, 2$, to be the sample means and compute standardized values $Z_i = \bar{X}_i \sqrt{n}$, $i=1, 2$. We consider the decision rule:

If $\min(Z_1, Z_2) > \Phi^{-1}(1-\alpha^*)$ accept the treatment,

otherwise reject the treatment.

Here α^* is the specified size of the test and Φ is the standard normal *cdf*. By using a coupling argument whereby sample points are mapped from (x_1, x_2) to $(x_1 + v_1, x_2 + v_2)$, it is easy to show that $\pi(\mu_1, \mu_2)$ is monotone increasing in both arguments. Hence, for any value of ρ,

$$\max\{\pi(\mu_1, \mu_2);\ \mu_1 \le 0 \text{ or } \mu_2 \le 0\} \le \max\{\pi(0, \infty), \pi(\infty, 0)\}$$

$$= Pr\{Z_1 > \Phi^{-1}(1-\alpha^*)\} = Pr\{Z_2 > \Phi^{-1}(1-\alpha^*)\} = \alpha^*,$$

i.e., the maximum Type I error equals the specified value α^*. In general the acceptance probability, $Pr\{\min(Z_1, Z_2) > \Phi(1-\alpha^*)\}$, depends on μ as well as the correlation ρ. This can be easily calculated using tables or computer programs for the bivariate normal distribution, thus the operating characteristic of the test with n observations can be constructed and this can be used to choose a suitable sample size. For example, n could be chosen to satisfy a power requirement of the form

$$\pi(\mu_1, \mu_2) \ge 1-\beta^* \quad \text{whenever} \quad \mu_1 \ge \mu_1^* \text{ and } \mu_2 \ge \mu_2^* \tag{2.2}$$

for specified β^* $(0<\beta^*<1)$ and alternative $\boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*)$ where $\mu_1^* > 0$ and $\mu_2^* > 0$.

If the correlation is not known but can be bounded below by some known value ρ_L^* then the Type II error probability requirement can be guaranteed with a sample size found by solving $\pi(\mu_1^*, \mu_2^*) = 1-\beta^*$ with $\rho = \rho_L^*$. This follows from Slepian's inequality (see, for example, Tong, 1980, p. 12), which implies that the acceptance probability is an increasing function of ρ. In particular (2.2) will be satisfied for any ρ if the sample size is calculated for the most extreme case, $\rho = -1$. In this case

$$\pi(\mu_1^*, \mu_2^*) = \{\Phi(\sqrt{n}\,\mu_2^* - c) - \Phi(c - \sqrt{n}\,\mu_1^*)\}^+ \tag{2.3}$$

where $x^+ = \max(x, 0)$ and $c = \Phi^{-1}(1 - \alpha^*)$. Thus, the sample size, n, needed to achieve $\pi(\mu_1^*, \mu_2^*) = 1-\beta^*$ can be found using standard univariate normal tables.

If the correlation is unknown but can be assumed positive, as might be the case if response variables are both measures of efficacy, the power requirement (2.2) can be guaranteed by solving $\pi(\mu_1^*, \mu_2^*) = 1-\beta^*$ with $\rho = 0$. In this case the necessary sample size is found by solving

$$\pi(\mu_1^*, \mu_2^*) = \{1 - \Phi(c - \sqrt{n}\,\mu_1^*)\}\{1 - \Phi(c - \sqrt{n}\,\mu_2^*)\} = 1 - \beta^*$$

and, again, only univariate normal tables are needed.

As an example, suppose $\alpha^* = 0.05$, and Type II error $\beta^* = 0.2$ is to be allowed at $\mu^* = (0.2, 0.2)$. Then $c = \Phi^{-1}(0.95) = 1.645$. Table 1a shows the sample size n needed for various values of ρ. Sample sizes for general $\mu_1^* = \mu_2^* = \mu^*$, say, general σ, and the same α^* and β^* can be obtained by multiplying the entries by the factor $(0.2\sigma/\mu^*)^2$. In practice these values should be rounded up to the next highest integer. It can be seen that, apart from when ρ is very close to 1, the sample sizes are not very sensitive to changes in ρ. Table 1b shows the effect on the power at $\mu^* = (0.2, 0.2)$ as ρ varies for a fixed sample size $n = 206$, i.e., the sample size needed to achieve power 0.8 if $\rho = 0.2$.

Tables 1a and 1b about here

4. Group Sequential Designs

We now generalize the fixed sample test of the previous section to a group sequential design. Suppose the kth interim analysis ($k \geq 1$) takes place after $n(k)$ bivariate observations, $\mathbf{X}_1, \ldots, \mathbf{X}_{n(k)}$ have been sampled. If groups are of equal size, g, then $n(k) = gk$; if $g=1$, we have the fully sequential case. We define the sample mean vector $\overline{\mathbf{X}}_{n(k)} = (\overline{X}_{1,n(k)}, \overline{X}_{2,n(k)}) = \Sigma\, \mathbf{X}_j/n(k)$, where the sum is taken over all $n(k)$ observations in the first k groups. Now define standardized variates $\mathbf{Z}_k = (Z_{1k}, Z_{2k}) = \overline{\mathbf{X}}_{n(k)}\sqrt{n(k)}$. Thus each $\mathbf{Z}_k$ is bivariate normal with mean $\mu\sqrt{n(k)}$, unit variances and correlation ρ. In addition, defining sample sums

$$\mathbf{S}_k = \sum_{j=1}^{n(k)} \mathbf{X}_j = \mathbf{Z}_k\sqrt{n(k)} \qquad (k \geq 1)$$

and $\mathbf{S}_0 = (0, 0)$; the increments $\{\mathbf{S}_k - \mathbf{S}_{k-1}, k \geq 1\}$ are independent bivariate normal with means $m_k \mu$, variances m_k and correlation ρ, where $m_1 = n(1)$ and $m_k = n(k) - n(k-1)$, $k > 1$, is the size of the kth group. Thus

$$\text{Corr}\,(Z_{ik}, Z_{lk'}) = (\rho + (1-\rho)\delta_{il})\,\sqrt{n(k)/n(k')}$$

for $i, l = 1, 2$ and $1 \leq k \leq k'$, where δ_{il} is the Kronecker delta.

We shall construct a group sequential test in which there are to be a maximum of K analyses (also referred to as "stages") where K and, in this development, $n(1), \ldots, n(K)$ are fixed. The decision rule involves specification of continuation, acceptance and rejection regions in $\mathbf{R}^2$ at each of K possible stages. Following on from the fixed sample case discussed in Section 3, we shall restrict our attention to "L-shaped" regions, i.e., sequential decision rules of the form:

If $Z_{1k} > b_{1k}$ and $Z_{2k} > b_{2k}$	stop at stage k and accept the treatment,
if $Z_{1k} < a_{1k}$ or $Z_{2k} < a_{2k}$	stop at stage k and reject the treatment,
otherwise	continue to stage $k+1$.

Here $a_{ik} < b_{ik}$ for $i = 1, 2$ and $1 \leq k \leq K-1$ and $a_{1K} = b_{1K}$ and $a_{2K} = b_{2K}$ in order to ensure termination at stage K. The boundary values $\{a_{1k}, b_{1k}; 1 \leq k \leq K\}$ and $\{a_{2k}, b_{2k}; 1 \leq k \leq K\}$ are to be chosen to satisfy the error probability requirement $\alpha = \alpha^*$ where α, the maximum Type I error, is defined by (2.1).

Operationally, the test can be described as running univariate group sequential tests separately on each variable, stopping the trial to *accept* the treatment only if *both* univariate tests stop to accept at the same stage, but stopping to *reject* as soon as *either* univariate test rejects. At each stage, the plane $\mathbf{R}^2$ is divided into 3 regions,

$$\mathcal{R}_k = \{(Z_1, Z_2) : Z_1 < a_{1k} \text{ or } Z_2 < a_{2k}\},$$

$$\mathcal{Q}_k = \{(Z_1, Z_2) : Z_1 > b_{1k} \text{ and } Z_2 > b_{2k}\},$$

$$\mathcal{C}_k = \mathbf{R}^2 \setminus \mathcal{R}_k \setminus \mathcal{Q}_k.$$

Respectively, these are called the rejection, acceptance and continuation regions at the kth stage. These regions are illustrated in Figure 3.

Figure 3 about here

We define

$$\pi_k(\mu) = Pr_\mu\ (\mathbf{Z}_i \in \mathcal{C}_i, 1 \le i \le k-1 \text{ and } \mathbf{Z}_k \in \mathcal{Q}_k)$$

to be the probability that, for given μ, the treatment is accepted precisely at stage k $(1 \le k \le K)$. The overall probability of acceptance is then

$$\pi\ (\mu) = \sum_{k=1}^{K} \pi_k(\mu)\ .$$

A coupling argument, similar to that in Section 3, shows that $\pi\ (\mu_1\ , \mu_2)$ is a monotone increasing function of both μ_1 and μ_2. Thus, in order to satisfy the Type I error probability requirement that $\pi(\mu) \le \alpha^*$ whenever $\mu_1 \le 0$ or $\mu_2 \le 0$, it is sufficient to ensure that $\pi\ (0, \infty) \le \alpha^*$ and $\pi\ (\infty, 0) \le \alpha^*$. This can be done by choosing constants $\{a_{1k}, b_{1k}\ ; 1 \le k \le K\}$ and $\{a_{2k}, b_{2k}; 1 \le k \le K\}$ from separate univariate group sequential designs that have Type I error α^*, i.e., $\{a_{1k}, b_{1k}\}$ must form the boundary for $\{Z_{1k}\}$ of a size α^* one-sided test of $H_0 : \mu_1 = 0$, *vs* $H_1 : \mu_1 > 0$, *etc.* Proposals for such one-sided testing procedures have been discussed by several authors, including DeMets and Ware (1980, 1982), Whitehead and Stratton (1983), Jennison (1987) and Emerson and Fleming (1989). Some specific boundaries are discussed in the next section.

It is important to note that only the properties of univariate group sequential tests are needed to construct a bivariate procedure of this type satisfying the Type I error requirement that α, as defined by (2.1), is at most α^*. In particular, knowledge of ρ is not needed. If each univariate test has size α^* and power γ at the alternative $\mu = \mu'$ then, not only do $\pi\ (0, \infty) = \pi\ (\infty, 0) = \alpha^*$, but also $\pi\ (\mu', \infty) = \pi\ (\infty, \mu') = \gamma$. Additionally, the expected sample sizes or "average sample numbers" (ASNs) in the limiting cases where μ_1 or μ_2 are infinite, are simply those of the univariate

procedures. However, for general values of (μ_1, μ_2) the operating characteristic and ASN must be computed from bivariate calculations, even if $\rho = 0$. In particular, bivariate calculations are needed to evaluate the acceptance probability at μ^* in order to check requirement (2.2).

Since the joint distribution of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_k)$ is known for any posited values of μ_1, μ_2 and ρ, the operating characteristic and stopping time distribution can be computed for any sequence of cumulative group sizes $\{n(k); k=1, \ldots, K\}$ and for any specified boundary. This is accomplished by the two-dimensional analogue of the recursion formulae given by, for example, Armitage, McPherson and Rowe (1969) or DeMets and Ware (1980). We present details of the numerical integration of up to $2K$-dimensional integrals in the Appendix.

Suppose equally sized groups of g observations are to be used. Since the boundary values $\{a_{ik}, b_{ik}\}$ are defined on the scale of the standardized statistics, $\{\mathbf{Z}_k\}$, it follows that values of $\pi(\mu)$ at $\mu = (0, 0)$, $(0, \infty)$ and $(\infty, 0)$ do not vary with g if $\{a_{ik}, b_{ik}\}$ are held fixed. Thus, having chosen $\{a_{ik}, b_{ik}\}$ to meet the Type I error requirement $\alpha = \alpha^*$, the group size, g, can be selected to give $\pi(\mu_1^*, \mu_2^*) = 1-\beta^*$ and it follows from the monotonicity of $\pi(\mu)$ that the Type II error probability requirement (2.2) will then be satisfied. The required value of g does depend on ρ. However, we shall see in Section 5 that this dependence is slight and a reasonable initial estimate of ρ may suffice for design purposes. In view of the bivariate nature of μ, it may well be helpful to examine a test's operating characteristic, $\pi(\mu)$, over a range of values of μ. A contour plot of π, an example of which appears in Section 5, provides a very helpful tool to this end.

5. Choice of Boundary Values

We now discuss specific choices of the constants $\{a_{ik}, b_{ik}\ ; i=1, 2,\ k=1, \ldots, K\}$. We restrict ourselves to the case of equally spaced analyses, where $n(k)=gk$ for common group size g. Recall that the Type I error probability requirement $\alpha \leq \alpha^*$, where α is given by (2.1), can be met by considering properties of univariate tests.

For each $i=1,2$, we require that the one-sided group sequential test with boundary $\{a_{ik}, b_{ik}; k=1,\ldots,K\}$ for the sequence of standardized statistics $\{Z_{ik}; k=1,\ldots,K\}$ should have probability α^* of exiting by the upper boundary when $\mu_i=0$. The group size, g, can then be chosen to satisfy the Type II error requirement (2.2).

DeMets and Ware (1980, Sec. 3.2) suggested construction of a one-sided univariate test by adapting the two-sided test of Pocock (1977); starting from a two-sided test of $H_0: \mu_i=0$ with Type I error $2\alpha^*$, the lower boundary and the terminal decision to accept H_0 are combined into a single decision "accept H_0", H_0 is rejected only if the upper boundary is exceeded. This procedure gives a boundary of the form $b_{i1}=\ldots=b_{iK}=a=a_{iK}$ and $a_{i1}=\ldots=a_{i,K-1}=-a$. We shall refer to this as Procedure A. The parameter a is chosen to satisfy the α^* condition and is equal to the constant z for $\alpha=2\alpha^*$ of Pocock (1977); values of a for $\alpha^*=0.005$, 0.025 and 0.05 and $1\leq K\leq 10$ are tabulated by Jennison and Turnbull (1989, Table 1).

Tables 2a and 2b about here

We consider the example of Section 3 with $\alpha^*=0.05$ and Type II error $\beta^*=0.2$ to be allowed at $\mu^*=(0.2,0.2)$. However, suppose now observations are to be taken in up to $K=5$ groups of equal size g. From Table 1 of Jennison and Turnbull (1989), we find $a=2.122$, which specifies the test for given g. Table 2a gives the maximum sample size, gK, needed for selected values of ρ. The group size is obtained by dividing this entry by K (=5) and rounding up to the next highest integer. The table also gives the expected sample sizes when $\mu=(0,0)$, $(0.2,0.2)$, $(0,\infty)$ and $(0.2,\infty)$. By symmetry, the ASNs at $\mu=(\infty,0)$ and $(\infty,0.2)$ are the same as at $(0,\infty)$ and $(0.2,\infty)$, respectively. All sample size entries in the table should be multiplied by the factor $(0.2\sigma/\mu^*)^2$ for the problem of general σ, $\alpha^*=0.05$ and $\beta^*=0.2$ at $\mu^*=(\mu^*,\mu^*)$. Comparison with Table 1a shows that reductions in ASN over the fixed sample size test are achieved at $\mu=(0.2,0.2)$ and $(0.2,\infty)$; the larger ASNs at $\mu=(0,0)$ and $(0,\infty)$ are a consequence of the high expected sample size of the

underlying DeMets and Ware (1980) univariate test at $\mu = 0$. Again, it can be seen that the required sample sizes are not very sensitive to changes in ρ except for values of ρ very close to 1; since the maximum Type I error does not depend on ρ at all, a reasonable estimate of ρ will suffice for design purposes. The effect of misspecifying ρ on the power of a test is illustrated in Table 2b which shows power at $\mu^* = (0.2, 0.2)$ for various ρ when $g = 52$ ($gK = 260$), the group size that would be chosen to achieve power close to 0.8 if ρ were estimated to be 0.2.

The two-sided univariate test of O'Brien and Fleming (1979) can be adapted to give a one-sided test in a similar fashion. In the resulting bivariate test we set $b_{ik} = a\sqrt{K/k} = -a_{ik}$ ($1 \le k \le K-1$, $i=1, 2$), and $b_{iK} = a_{iK} = a$. Again a must be chosen to satisfy the α^* requirement; it is the same constant as is needed for a two-sided O'Brien and Fleming test with probability α^* of exiting by the upper boundary when $\mu_i=0$ and values for $\alpha^* = 0.005$, 0.025 and 0.05, and $1 \le K \le 10$ are given in Jennison and Turnbull (1989, Table 1). The O'Brien-Fleming design has the features that very strong evidence is needed to stop a trial at the earliest stages and that the final critical values are close to those used in a fixed sample test. As for Procedure A, the univariate test has a high expected sample size at $\mu = 0$ and this leads to high ASNs for the bivariate test at, for example, $\mu = (0, 0)$ and $(0, \infty)$. To avoid this problem, we must start with a more satisfactory one-sided test.

Emerson and Fleming (1989) have proposed a family of one-sided symmetric group sequential boundaries for the univariate problem. This family contains tests with boundaries of different shapes, indexed by a parameter $p \ge 0$. For given values of p, α, K and g the boundaries for $\{Z_{ik}\}$ of a univariate one-sided test of $H_0 : \mu_i = 0$ *vs* $H_1 : \mu_i > 0$, with Type I error α, are given by

$$b_{ik} = k^{p-\frac{1}{2}}\, c$$

$$a_{ik} = \delta\sqrt{gk} - b_{ik} \qquad (5.1)$$

where $\delta = 2\,c\,K^{p-1}/\sqrt{g}$, ensuring that $a_{iK} = b_{iK}$, $i=1, 2$, and $c = c_{K,p}^{(\alpha)}$ is a constant, independent of g, that can be obtained from Table 1 of Emerson and Fleming (1989).

(Note that Emerson and Fleming express the boundary values in terms of the sample sums $\sqrt{gk}\ Z_{ik}$ and the above values need to be multiplied by the factor $\sqrt{gk}$ to obtain their expressions for $\{a_k, b_k\}$). When plotted on this scale against stage number, k, the boundaries are symmetric about a straight line of slope $\delta/2$ and their shape varies from triangular for $p = 0$ to pear-shaped for $p = 0.5$.

We construct bivariate tests by defining $\{a_{ik}, b_{ik}\,; k = 1, \ldots, K\}$ by (5.1) with $\alpha = \alpha^*$ for i=1, 2. This ensures a maximum Type I error probability of α^* at $(\mu_1, \mu_2) = (0, \infty)$ or $(\infty, 0)$, as required. Incidentally, it also follows that the acceptance probability or operating characteristic function is $1-\alpha^*$ at $(\mu_1, \mu_2) = (\delta, \infty)$ or (∞, δ). A search technique can now be employed to find the group size g that satisfies the Type II error constraint given by (2.2).

Emerson and Fleming (1989) discuss the choice of the parameter p. They explain (p. 908) that a low value of p implies more conservative testing at the earlier analyses. A design with $p = 0.5$ might be considered a one-sided analogue of the Pocock (1977) test, while $p = 0$ is analogous to the procedure of O'Brien and Fleming (1979). If expected sample size is used as a criterion, Table 2 of Emerson and Fleming (1989) suggests that, of these two, the design based on $p = 0.5$ is to be preferred. We shall refer to the bivariate group sequential based on these Emerson and Fleming boundaries with $p = 0.5$ as Procedure B.

Tables 3a and 3b about here

We consider the same example as that considered previously for the fixed sample procedure and for Procedure A. Here $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\mu^* = (0.2, 0.2)$ and up to K=5 groups of equal size g are available. From Table 1 of Emerson and Fleming (1989) we find $c = 2.065$ for $\alpha = 0.05$, $p = 0.5$ and 5 groups of observations. Table 3a gives the maximum sample size, gK, needed for selected values of ρ. The group size g is obtained by dividing this entry by K (=5) and rounding up. From g and c, the boundary values can be calculated using (5.1). Table 3a also gives the ASNs at

$\mu = (0, 0)$, $(0.2, 0.2)$, $(0, \infty)$ and $(0.2, \infty)$. Again all entries should be multiplied by $(0.2\sigma/\mu^*)^2$ for general σ and $\mu^* = (\mu^*, \mu^*)$ in place of $(0.2, 0.2)$. Table 3b shows the effect on the power at $\mu^* = (0.2, 0.2)$ for this procedure as ρ varies for a fixed group size $g = 65$, $(gK = 325)$, chosen to achieve power close to 0.8 at $\rho = 0.2$.

Comparing the results of Tables 1a, 2a and 3a, we see that Procedure B's increased opportunity for early stopping to reject leads to a much lower expected sample size, well below the fixed sample size, when the treatment is poor. The one drawback of Procedure B is its high maximum sample size; this problem could be alleviated by basing the bivariate test on an Emerson and Fleming (1989) test with smaller parameter p or one of the efficient univariate tests with low maximum sample size described by Eales and Jennison (1991). Since the maximum Type I error probability in (2.1) occurs at $\mu = (0, \infty)$ or $(\infty, 0)$, it may be desirable, instead, to restrict attention to Type I errors at more plausible values of μ. A contour plot of Type I error provides a convenient way of displaying the operating characteristic of a bivariate group sequential test. Figure 4a shows the operating characteristic of Procedure B with $g = 65$; the ASN of this test is shown in Figure 4b. These contour plots were produced using the CONICON 3 (Sibson, 1987) contour drawing package, which implements the method of Sibson and Thomson (1981).

Figures 4a and 4b about here

6. Concluding remarks

In some situations, it may not be desirable to stop early for acceptance (see Gould 1983 and Ho 1986, Sec. 2.4). This may be the case if both outcome variables are safety measures. We can then set $b_{ik} = +\infty$ $(i=1, 2;\ k=1, \ldots, K-1)$ in the procedures of Section 4. If we leave the other boundary values $a_{i1}, \ldots, a_{iK}$ and b_{iK} unchanged, the procedure will be conservative in that the Type I error will be lower, but of course the Type II error at μ^* will increase.

An important practical problem is how to cope with unequal and unpredictable group sizes. Our Type I error condition will be satisfied as long as the univariate tests underlying a bivariate test maintain their Type I errors. This can be achieved using an "error spending function" approach; see Jennison (1987) and Eales and Jennison (1991) for a description of the Lan and DeMets (1983) error spending function method applied to one-sided tests. One situation where unequal group sizes will arise is when group sizes are chosen adaptively to achieve a maximum sample size appropriate to estimates of ρ based on the accumulating observations.

A natural next step is to develop procedures for the two-sided problem, testing between the hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ or $H_0 : \mu \in \mathbf{N}$ and $H_1 : \mu \notin \mathbf{N}$ where $\mathbf{N}$ is a neighbourhood of μ_0. The acceptance region $\mathcal{A}_k$ at stage k $(1 \leq k \leq K)$ would consist of a bounded neighbourhood of μ_0, surrounded by a continuation region $\mathcal{C}_k$ (with $\mathcal{C}_K$ empty), the rejection region being $\mathcal{R}_k = \mathbf{R}^2 \setminus \mathcal{A}_k \setminus \mathcal{C}_k$. These regions could be constructed from univariate two-sided group sequential tests in a parallel fashion to the development in Section 3 for one-sided tests. The operating characteristic and expected sample size function can be computed numerically following the method described in the Appendix. With the extra boundaries, there is more freedom and more choices must be made. Similar methods could be applied in bioequivalence problems, where the roles of H_0 and H_1 are reversed.

Two-dimensional repeated confidence sets, generalising repeated confidence intervals (Jennison and Turnbull 1989, 1991b), can be obtained by inverting group sequential bivariate tests. For the one-sided tests of Sections 4 and 5, these sets would be semi-infinite. Such repeated confidence sets could be used to monitor a group sequential study in a flexible manner, rather than with a rigid stopping rule. However, it would be important to tailor their shape so that they provided useful guidance for resolving the underlying decision problem.

It is interesting to note that the problem of comparing two competing treatments with a control treatment can be formulated in a way that leads to a structure similar to the one we have already discussed. Suppose at analysis k, $1 \leq k \leq K$, the statistic Z_{1k}

represents the observed difference between treatment A and the control, and Z_{2k} the difference between treatment B and the control. If responses are normally distributed, the sequence of bivariate statistics, $\{Z_{1k}, Z_{2k}; k=1, \ldots, K\}$, will have the form described previously, the pair Z_{1k} and Z_{2k} being correlated due to both statistics' dependence on the control observations. The choice of stopping rule must depend on the formulation of the testing or selection problem but the same numerical methods developed for multiple endpoints can be used to design and evaluate the resulting sequential procedures.

Another situation where the numerical methods of this paper may be helpful is in studying confidence intervals for the mean of a distribution, observations from which are correlated with the variable used to define a univariate group sequential test. Suppose observations (X_1, X_2) are bivariate normal with mean (μ_1, μ_2) and non-zero correlation ρ and data are collected using a sequential stopping rule defined in terms of the X_1s. Methods for constructing a confidence interval for μ_2 in this situation have been proposed by Whitehead (1986); one application cited is a comparison of two treatments with respect to the lengths of new-born babies in a sequential trial when the stopping rule was based on birthweights. Our methods, applied to appropriately defined bivariate stopping and decision rules, would allow numerical computation of the coverage probabilities of confidence intervals for μ_2 under specific values of μ_1.

ACKNOWLEDGEMENTS

This research was supported by Grant R01 GM28364 from the U.S. National Institutes of Health and by a U.K. SERC Visiting fellowship GR/F 72864.

APPENDIX

Numerical computation for bivariate group sequential tests

The operating characteristic and expected sample size of a bivariate group sequential test can be calculated from the terms Pr_μ (stop to accept at stage k) and Pr_μ (stop to reject at stage k) for $k=1, \ldots, K$. For the acceptance probabilities, we write

$$Pr_\mu \text{ (stop to accept at stage } k) = Pr_\mu (\mathbf{Z}_i \in \mathcal{C}_i, i=1, \ldots, k-1 \text{ and } \mathbf{Z}_k \in \mathcal{Q}_k)$$

$$= \int_{R_{11}} f_{11}(z_{11}) \int_{R_{21}(z_{11})} f_{21}(z_{21}, z_{11}) \cdots \int_{R_{1j}} f_{1j}(z_{1j}, z_{1,j-1}) \int_{R_{2j}(z_{1j})} f_{2j}(z_{2j}, z_{2,j-1}, z_{1j}, z_{1,j-1})$$

$$\cdots \int_{R_{1k}} f_{1k}(z_{1k}, z_{1,k-1}) \int_{R_{2k}(z_{1k})} f_{2k}(z_{2k}, z_{2,k-1}, z_{1k}, z_{1,k-1})\, dz_{2k}\, dz_{1k} \cdots dz_{21}\, dz_{11}.$$

The ranges of integration are, for $j=1, \ldots, k-1$: $R_{1j} = (a_{1j}, \infty)$, $R_{2j}(z_{1j}) = (a_{2j}, \infty)$ if $z_{1j} < b_{1j}$ and $R_{2j}(z_{1j}) = (a_{2j}, b_{2j})$ if $z_{1j} \geq b_{1j}$; and for $j=k$: $R_{1k} = (b_{1k}, \infty)$ and $R_{2k}(z_{1k}) = (b_{2k}, \infty)$. The functions f_{ij} are conditional probability densities: $f_{11}(z_{11})$ is the marginal density of Z_{11}, $f_{21}(z_{21}, z_{11})$ is the conditional density of Z_{21} given $Z_{11} = z_{11}$; for $j > 1$, $f_{1j}(z_{1j}, z_{1,j-1})$ is the density of Z_{1j} conditional on $Z_{1,j-1} = z_{1,j-1}$ and $f_{2j}(z_{2j}, z_{2,j-1}, z_{1j}, z_{1,j-1})$ is the density of Z_{2j} conditional on $Z_{2,j-1} = z_{2,j-1}$, $Z_{1j} = z_{1j}$ and $Z_{1,j-1} = z_{1,j-1}$. For equally sized groups of g observations, the conditional densities are obtained from

$$Z_{11} \qquad \sim N(\sqrt{g}\,\mu_1, 1)$$

$$Z_{21} \mid Z_{11}=z_{11} \qquad \sim N(\sqrt{g}\,\mu_2+\rho(z_{11}-\sqrt{g}\,\mu_1), 1-\rho^2)$$

$$Z_{1j} \mid Z_{1,j-1}=z_{1,j-1} \qquad \sim N(\sqrt{(g/j)}\,\mu_1+\sqrt{\{(j-1)/j\}}\, z_{1,j-1}, 1/j)$$

and

$$Z_{2j} \mid Z_{2,j-1}=z_{2,j-1}, Z_{1j}=z_{1j}, Z_{1,j-1}=z_{1,j-1}$$

$$\sim N(\sqrt{(g/j)}\,\mu_2+\sqrt{\{(j-1)/j\}}\, z_{2,j-1}+\rho[z_{1j}-\sqrt{\{(j-1)/j\}}\, z_{1,j-1}-\sqrt{(g/j)}\,\mu_1], (1-\rho^2)/j).$$

The probabilities of rejection can be expressed as multiple integrals in a similar way.

We have evaluated these multiple integrals using Simpson's rule to replace the integrals by sums and then summing over z_{11} and z_{21}, z_{12} and z_{22}, *etc.* in order. If Simpson's rule is applied each time with a grid of n points, the total number of arithmetic operations is of order Kn^4 and the numerical error decreases as n^{-4}. Modern computers are capable of carrying out these calculations rapidly; exact computations for sequential t-tests described by Jennison and Turnbull (1991a), which also involved keeping track of two statistics from stage to stage, required a similar amount of computation.

In principle, the same computational approach may be used with more than two variables but the computational burden increases rapidly. For a d-variate response the number of arithmetic operations required is of order Kn^{2d} but numerical error is still of order n^{-4}. In particular, for $d > 4$ a better rate of convergence *per numerical operation* would be obtained using Monte Carlo simulation.

REFERENCES

Armitage, P. (1987). Some aspects of Phase-III trials. *Paper presented at ISI Satellite Meeting on Biometry: Clinical Trials and Related Topics, 21 Sept. 1987, Osaka, Japan.*

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, **A**, **132**, 235-244.

Barnes, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science*, **234**, (Oct 3), 15-16.

Cox, D. R. (1989). Discussion of paper by Jennison and Turnbull, *Journal of the Royal Statistical Society*, **B**, **51**, 338.

DeMets, D. L. and Ware, J. H. (1980). Group sequential methods for clinical trials with one-sided hypothesis. *Biometrika*, **67**, 651-660.

DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**, 661-663.

Eales, J. D. and Jennison, C. (1991). An improved method for deriving optimal one-sided group sequential tests. *Submitted for publication.*

Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905-923.

Enas, G. G., Dornseif, B. E., Sampson, C. B., Rockhold, F. W. and Wuu, J. (1989). Monitoring versus interim analysis of clinical trials: a perspective from the pharmaceutical industry. *Controlled Clinical Trials*, **10**, 57-70.

Fischl *et al.* (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine*, **317**, No. 4, 185-191.

Freedman, L. S. and Spiegelhalter, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician*, **32**, 153-160.

Geller, N. L. and Pocock, S. J. (1987). Interim analyses in randomized trials: ramifications and guidelines for practitioners. *Biometrics*, **43**, 213-223.

Goldman, A. I. (1987). Issues in designing sequential stopping rules for monitoring side effects in clinical trials. *Controlled Clinical Trials*, **8**, 327-337.

Gould, A. L. (1983). Abandoning lost causes (Early termination of unproductive clinical trials). *Proceedings of Biopharmaceutical Section of American Statistical Association, Annual Meetings, Toronto, Ontario*, 31-34.

Ho, C-H. (1986). One-sided sequential stopping boundaries for clinical Series trials: classical and Bayesian approaches. *Ph.D. dissertation, University of Minnesota.*

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika*, **74**, 155-165.

Jennison, C. and Turnbull, B. W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society*, **B**, **51**, 305-361.

Jennison, C. and Turnbull, B. W. (1991a). Exact calculations for the sequential t, χ^2 and F tests. *Biometrika*, **78**, to appear.

Jennison, C. and Turnbull, B. W. (1991b). Sequential equivalence testing and repeated confidence intervals with applications to normal and binary responses. *Submitted for publication.*

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403-418.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659-663.

Lan, K. K. G. and Friedman, L. (1986). Monitoring boundaries for adverse effects in long-term clinical trials. *Controlled Clinical Trials*, **7**, 1-7.

Meier, P. (1975). Statistics and medical experimentation. *Biometrics*, **31**, 511- 529.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079-1087.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549-556.

Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics*, **40**, 549-567.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191-199.

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487-498.

Sibson, R. (1987) *CONICON 3 Handbook.* University of Bath.

Sibson, R. and Thomson, G. D. (1981). A seamed quadratic element for contouring. *Computing Journal*, **8**, 820-832.

Tang, D-I., Gnecco, C. and Geller, N. L. (1989a). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*, **84**, 776-779.

Tang, D-I., Gnecco, C. and Geller, N. L. (1989b). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, **76**, 577-583.

Tong, Y. L. (1980). *Probability Inequalities in Multivariate Distributions.* New York: Academic Press.

Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials.* Chichester: Ellis Horwood.

Whitehead, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics*, **42**, 461-471.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, **39**, 227-236.

Table 1a. Sample sizes, n, required for fixed sample procedures with $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\boldsymbol{\mu}^* = (0.2, 0.2)$ and selected values of ρ.

ρ	-1	-0.5	-0.2	0	0.2	0.5	0.8	1
n	214.1	213.8	212.0	209.6	206.0	197.9	183.8	154.6

Table 1b. Power, $1-\beta$, at $\boldsymbol{\mu} = \boldsymbol{\mu}^* = (0.2, 0.2)$ for procedure with $\alpha^* = 0.05$ and $n = 206$, for selected values of ρ.

ρ	-1	-0.5	-0.2	0	0.2	0.5	0.8	1
$1-\beta$	0.780	0.781	0.786	0.792	0.800	0.817	0.843	0.890

Table 2a. Maximum sample sizes and ASNs at several values of μ for Procedure A with $K = 5$ stages, $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\mu^* = (0.2, 0.2)$ and selected values of ρ.

ρ	*Maximum sample size, gK*	*ASN at* $(\mu_1, \mu_2) =$			
		$(0, 0)$	$(0.2, 0.2)$	$(0, \infty)$	$(0.2, \infty)$
-1	270.7	256.9	218.5	256.9	145.0
-0.5	271.0	257.3	204.5	257.2	145.0
-0.2	268.7	255.1	195.7	255.0	144.4
0	265.3	251.9	189.1	251.8	143.6
0.2	260.4	247.3	181.7	247.2	142.3
0.5	249.4	236.9	168.7	236.8	139.3
0.8	230.9	219.2	150.9	219.2	133.9
1	193.9	184.0	121.6	184.0	121.6

Table 2b. Power, $1 - \beta$, at $\mu = \mu^* = (0.2, 0.2)$ for Procedure A with K=5 stages, $\alpha^* = 0.05$ and $gK = 260$ for selected values of ρ.

ρ	-1	-0.5	-0.2	0	0.2	0.5	0.8	1
$1-\beta$	0.776	0.775	0.782	0.789	0.799	0.819	0.848	0.897

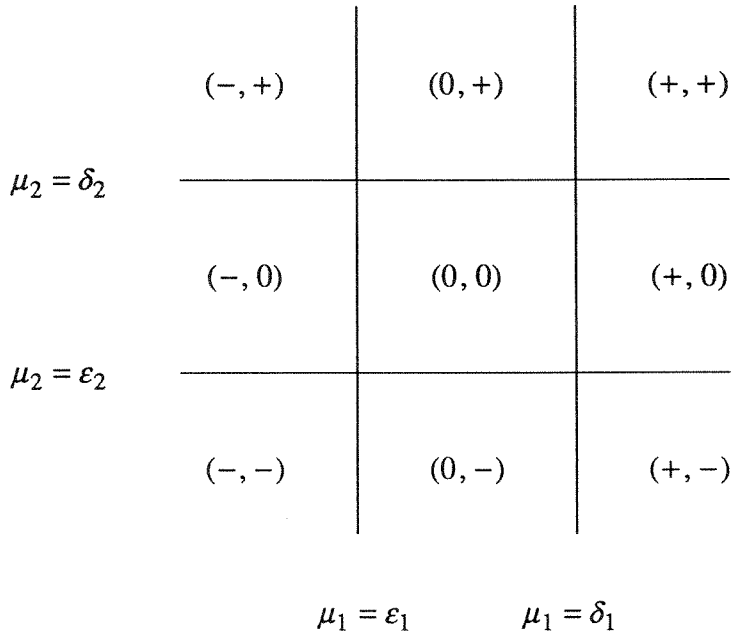
Table 3a. Maximum sample sizes and ASNs at several values of μ for Procedure B with $K = 5$ stages, $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\mu^* = (0.2, 0.2)$ and selected values of ρ.

ρ	*Maximum sample size, gK*	*ASN at* $(\mu_1, \mu_2) =$			
		$(0, 0)$	$(0.2, 0.2)$	$(0, \infty)$	$(0.2, \infty)$
-1	336.5	78.9	198.3	127.1	140.6
-0.5	336.4	86.8	188.3	127.1	140.5
-0.2	333.5	91.4	179.6	126.0	139.8
0	329.8	94.0	173.2	124.6	138.8
0.2	324.2	96.2	166.3	122.5	137.4
0.5	311.2	98.3	154.2	117.6	133.8
0.8	288.0	97.8	137.9	108.8	127.1
1	238.6	90.1	111.1	90.1	111.1

Table 3b. Power, $1 - \beta$, at $\mu = \mu^* = (0.2, 0.2)$ for Procedure B with K=5 stages, $\alpha^* = 0.05$ and $gK = 325$ for selected values of ρ.

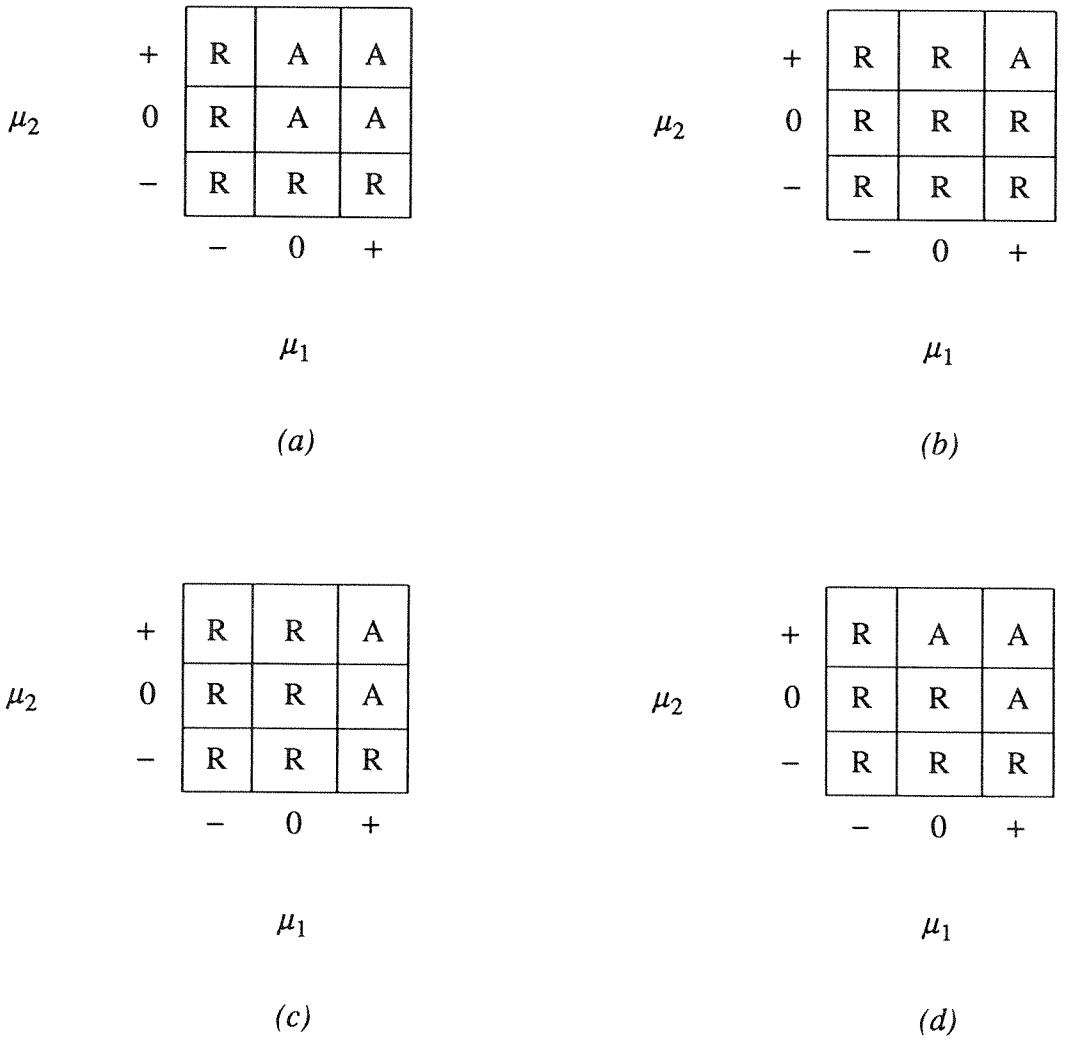
ρ	-1	-0.5	-0.2	0	0.2	0.5	0.8	1
$1-\beta$	0.782	0.782	0.787	0.793	0.801	0.818	0.845	0.894

Figure 1. Preference regions in terms of $\mu = (\mu_1, \mu_2)$.



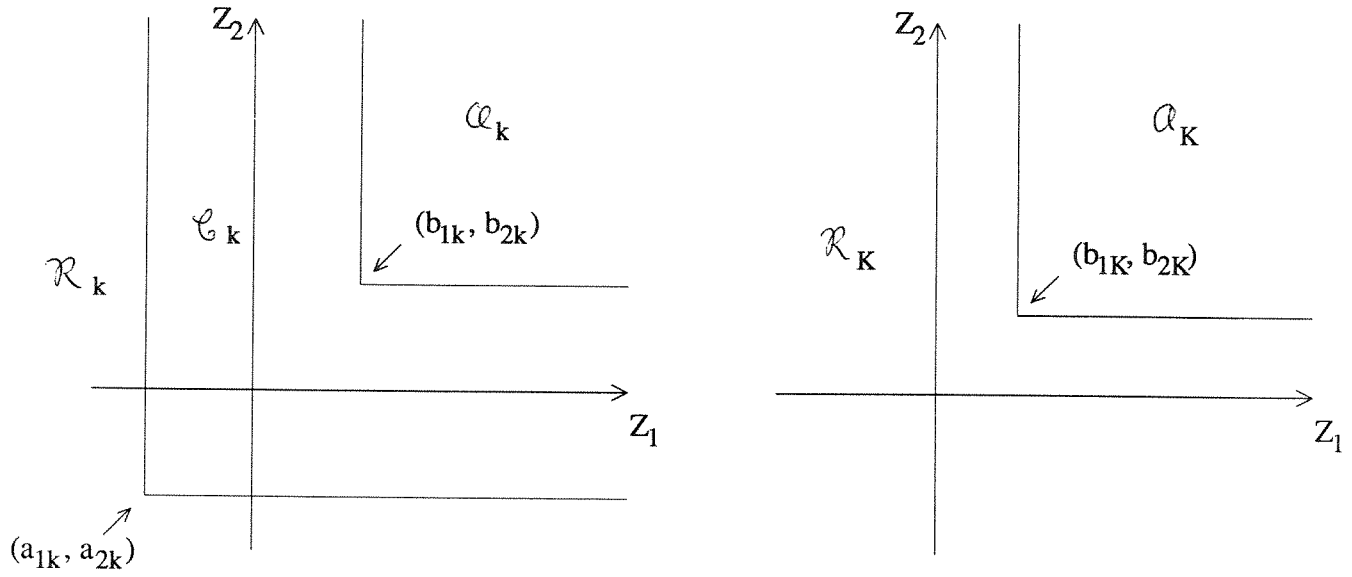
The symbol +, 0 or − in the first position in each pair indicates that the new treatment is preferred, considered equivalent or unacceptable, respectively, with regard to the first variable. The second entry in each pair is similarly defined for the second variable.

Figure 2. Possible acceptance and rejection regions for $\boldsymbol{\mu} = (\mu_1, \mu_2)$.



Key. The action appropriate for specified values of $\boldsymbol{\mu}$ is R: Reject treatment, A: Accept treatment.

Figure 3. Decision regions at stage k



$k = 1, \ldots, K-1$ $\qquad\qquad$ $k = K$

The test stops to accept at stage k if $\mathbf{Z} = (Z_1, Z_2) \in \mathcal{A}_k$, stops to reject if $\mathbf{Z} \in \mathcal{R}_k$, and continues if $\mathbf{Z} \in \mathcal{C}_k$.

Figure 4a. Contour plot of the operating characteristic of Procedure B with $K = 5$, $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\boldsymbol{\mu}^* = (0.2, 0.2)$ and $\rho = 0.2$.

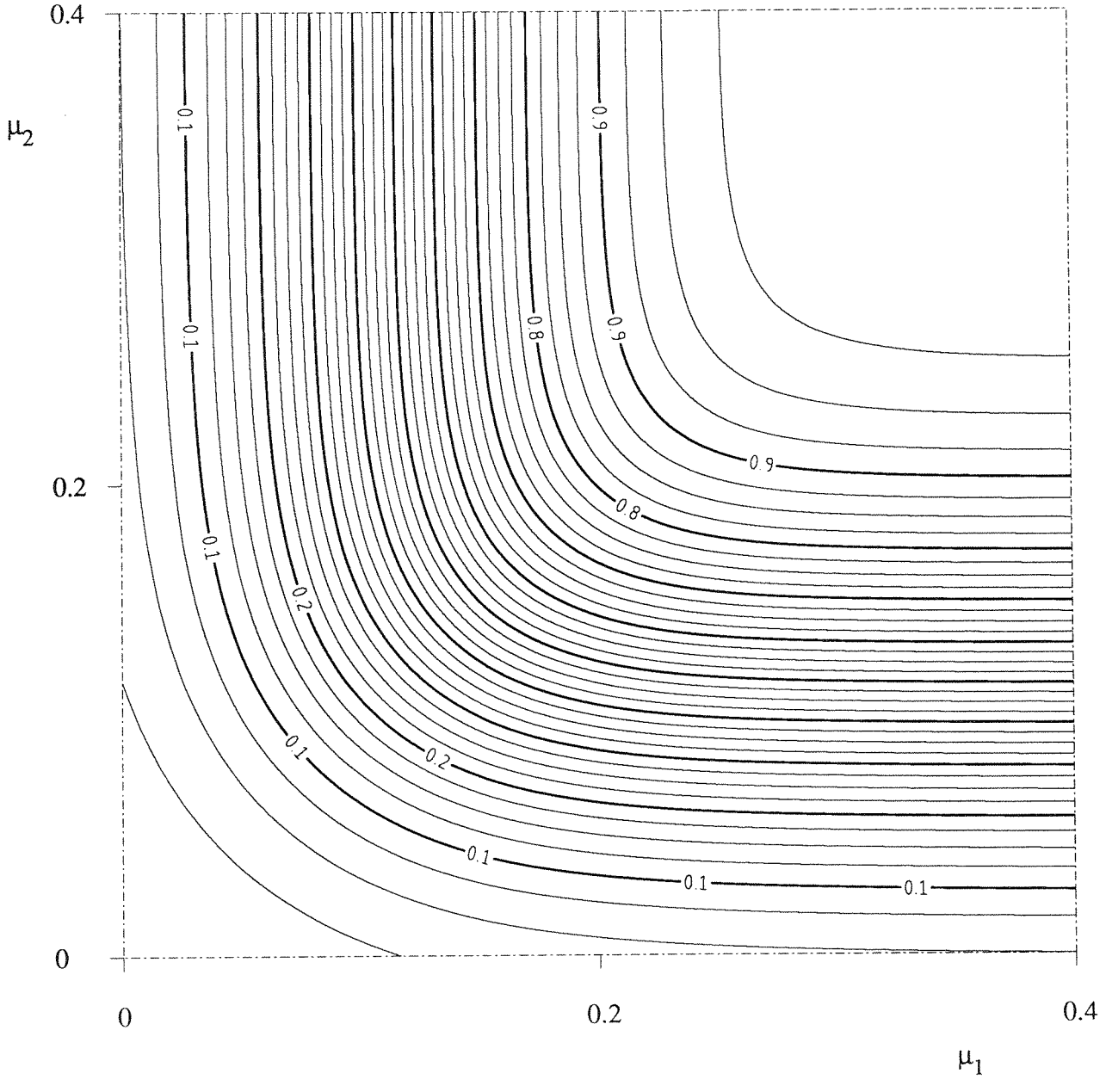


Figure 4b. Contour plot of the expected sample size of Procedure B with $K = 5$, $\alpha^* = 0.05$, $\beta^* = 0.2$ at $\mu^* = (0.2, 0.2)$ and $\rho = 0.2$.