Group Sequential Monitoring of Clinical Trials with Multiple Endpoints

Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

Stanford

February 2004

Example 1: A diabetes trial

O'Brien (*Biometrics*, 1984)

The trial was conducted to determine if an experimental therapy resulted in better nerve function, as measured by 34 electromyographic (EMG) variables.

6 subjects randomised to standard therapy,

5 subjects randomised to experimental therapy.

Changes in EMG measurements were recorded after 8 weeks.

Aim: To test

H_0: No treatment difference vs

H_A: Improvements under experimental therapy

— in some or all responses.

Example 2: A crossover trial for treatment of chronic respiratory disease

Pocock, Geller & Tsiatis (*Biometrics*, 1987)

17 patients with asthma or chronic obstructive airways disease were randomised to

	Active drug	then	*Placebo*
or	*Placebo*	then	*Active drug*

Measurements were

1. Peak expiratory flow rate
2. Forced expiratory volume
3. Forced vital capacity

taken at the end of both treatment periods.

Aim: To test

H_0: No treatment difference vs

H_A: Improvements under Active Drug

— for each measure.

Methods of interim monitoring in studies with multiple endpoints

1. Bonferroni adjustment
2. Group sequential χ^2 tests
3. Monitoring a linear combination of response variables
4. Marginal criteria, e.g., monitoring efficacy and safety.

Reference: Jennison & Turnbull, *Group Sequential Methods with Applications to Clinical Trials*, Ch. 15.

1. Bonferroni adjustment

Suppose a set of p endpoints has mean vector

$\boldsymbol{\mu}_A$ for treatment A,

$\boldsymbol{\mu}_B$ for treatment B.

In order to test H_0: $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B$ with type I error rate α:

Create a sequential test with type I error probability α/p for each component.

Stop and reject H_0 if *any* test rejects its null hypothesis.

Then,

$$Pr\{\text{Reject } H_0 \mid \boldsymbol{\mu}_A = \boldsymbol{\mu}_B\}$$

$$\leq \sum_{j=1}^{p} Pr\{\text{Reject } H_{0j} \mid \boldsymbol{\mu}_{A,j} = \boldsymbol{\mu}_{B,j}\}$$

$$= p \times \frac{\alpha}{p} = \alpha.$$

This may not be efficient against important alternatives, especially if endpoints are correlated.

2. Group sequential χ^2 tests

Suppose at analysis k we have summary statistics

$$\boldsymbol{Y}_k = \begin{pmatrix} Y_{1k} \\ \vdots \\ Y_{pk} \end{pmatrix}$$

where $E(Y_{jk})$ depends on $\boldsymbol{\mu}_{Aj} - \boldsymbol{\mu}_{Bj},\ j = 1, \ldots, p$.

For known $\mathbf{Var}(\boldsymbol{Y}_k)$, form standardised statistics

$$\boldsymbol{Z}_k = \begin{pmatrix} Z_{1k} \\ \vdots \\ Z_{pk} \end{pmatrix}$$

where each $Z_{jk} \sim N(0,\ 1)$ when $\boldsymbol{\mu}_{Aj} = \boldsymbol{\mu}_{Bj}$.

Let $\mathbf{Var}(\boldsymbol{Z}_k) = \boldsymbol{V}_k$, then marginally

$$\boldsymbol{Z}_k^T\, \boldsymbol{V}_k^{-1}\, \boldsymbol{Z}_k \ \sim\ \chi_p^2 \qquad \text{under } H_0.$$

(The analogue when $\mathbf{Var}(\boldsymbol{Y}_k)$ is unknown is Hotelling's T-statistic, which has a marginal F-distribution.)

Group sequential χ^2 tests

Jennison & Turnbull (*Biometrika*, 1991) derive the joint distribution of

$$\{\boldsymbol{Z}_k^T \, \boldsymbol{V}_k^{-1} \, \boldsymbol{Z}_k; \; k = 1, \ldots, K\}.$$

Hence, one can calculate group sequential χ^2 tests with specified type I error rates.

- χ^2 tests are of H_0: $\boldsymbol{\mu}_{Aj} = \boldsymbol{\mu}_{Bj}$ vs the *general alternative* $\boldsymbol{\mu}_{Aj} \neq \boldsymbol{\mu}_{Bj}$.

- They suit, say, a bio-equivalence study with p response measurements.

- They are inappropriate if the goal is to demonstrate that one treatment is superior to another — consider rejecting H_0 with a mixture of positive and negative differences.

3. Tests based on a linear combination of responses

O'Brien (*Biometrics*, 1984),

Tang, Gnecco & Geller (*JASA*, 1989)

Suppose responses are

on treatment A: $\boldsymbol{X}_{Ai} \sim N_p(\boldsymbol{\mu}_A, \boldsymbol{V})$

on treatment B: $\boldsymbol{X}_{Bi} \sim N_p(\boldsymbol{\mu}_B, \boldsymbol{V})$

and suppose *high* values of each variable are desirable.

Aim: To test

H_0: $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B$ vs

H_A: Treatment A better than Treatment B.

Restrict attention to the case

$$\mu_{Aj} - \mu_{Bj} = \lambda \delta_j, \quad j = 1, \ldots, p,$$

for specified $\delta_1, \ldots, \delta_p > 0$.

Then test H_0: $\lambda = 0$ vs H_A: $\lambda > 0$.

Linear combination of responses

Response vectors

$$\boldsymbol{X}_{Ai} \sim N_p(\boldsymbol{\mu}_A, \boldsymbol{V}), \quad \boldsymbol{X}_{Bi} \sim N_p(\boldsymbol{\mu}_B, \boldsymbol{V})$$

and we assume $\boldsymbol{\mu}_A - \boldsymbol{\mu}_B = \lambda \boldsymbol{\delta}$.

With m observations on each treatment,

$$\bar{\boldsymbol{X}}_A \sim N_p(\boldsymbol{\mu}_A, m^{-1}\boldsymbol{V})$$

and

$$\bar{\boldsymbol{X}}_B \sim N_p(\boldsymbol{\mu}_B, m^{-1}\boldsymbol{V}).$$

The Generalised Least Squares estimate of λ is

$$\begin{aligned} \widehat{\lambda} &= \frac{\boldsymbol{\delta}^T \boldsymbol{V}^{-1} (\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)}{\boldsymbol{\delta}^T \boldsymbol{V}^{-1} \boldsymbol{\delta}} \\ &\sim N(\lambda, \frac{2}{m \boldsymbol{\delta}^T \boldsymbol{V}^{-1} \boldsymbol{\delta}}). \end{aligned}$$

Let $\widehat{\lambda}^{(k)}$ denote the estimate of λ at analysis k.

Then $\{\widehat{\lambda}^{(1)}, \ldots, \widehat{\lambda}^{(K)}\}$ has the canonical joint distribution of a sequence of parameter estimates.

Linear combination of responses

The GLS estimate of λ at stage k is

$$\widehat{\lambda}^{(k)} = \frac{\boldsymbol{\delta}^T \boldsymbol{V}^{-1} (\bar{\boldsymbol{X}}_A^{(k)} - \bar{\boldsymbol{X}}_B^{(k)})}{\boldsymbol{\delta}^T \boldsymbol{V}^{-1} \boldsymbol{\delta}}.$$

The sequence $\{\widehat{\lambda}^{(1)}, \ldots, \widehat{\lambda}^{(K)}\}$ satisfies

$$(\widehat{\lambda}^{(1)}, \ldots, \widehat{\lambda}^{(K)}) \sim \text{multivariate normal},$$

$$\widehat{\lambda}^{(k)} \sim N(\lambda,\, 1/\mathcal{I}_k) \quad \text{for each } k,$$

$$Cov(\widehat{\lambda}^{(k_1)},\, \widehat{\lambda}^{(k_2)}) = 1/\mathcal{I}_{k_2} \quad \text{for } k_1 < k_2.$$

Thus, a standard group sequential test for a *univariate* parameter can be employed.

Note from the form of $\widehat{\lambda}^{(k)}$ that the data vector for each subject could have been reduced to the scalar quantity $\boldsymbol{\delta}^T \boldsymbol{V}^{-1} \boldsymbol{X}_i$ at the outset.

In some instances, investigators choose a univariate score for each subject directly.

Example 3:

Women's Health Initiative, Hormone Replacement Trial

Freedman et al (*Cont. Clin. Trials*, 1996)

The overall response was defined as a weighted sum:

	Weight
Incidence of coronary heart disease	0.5
Incidence of hip fracture	0.18
Incidence of breast cancer	0.35
Incidence of endometrial cancer	0.15
Death from other causes	0.1

The weights were assigned using data external to the trial.

Example 4: Total parenteral nutrition (TPN) for patients undergoing gastric cancer surgery

Tang, Gnecco & Geller (*JASA*, 1989)

This study investigated whether peri-operative TPN decreases the rate of complications in nutritionally compromised patients in the week following surgery.

Baseline rates:

25% Major complications,

45% Minor complications

Power 0.8 required to detect a reduction to:

15% Major complications,

30% Minor complications

Treatment was compared to control using a linear combination of major and minor complication rates.

Example 4: Total parenteral nutrition

The authors prove the general result that

> "the multivariate test based on all endpoints is more powerful than the similar univariate test based on a single endpoint".

In the TPN study, maximum sample sizes for several designs are:

	Minor complications only	Major complications only	Both
1-stage design	324	500	236
3-stage design	336	512	246

The advantage of using both endpoints is clear.

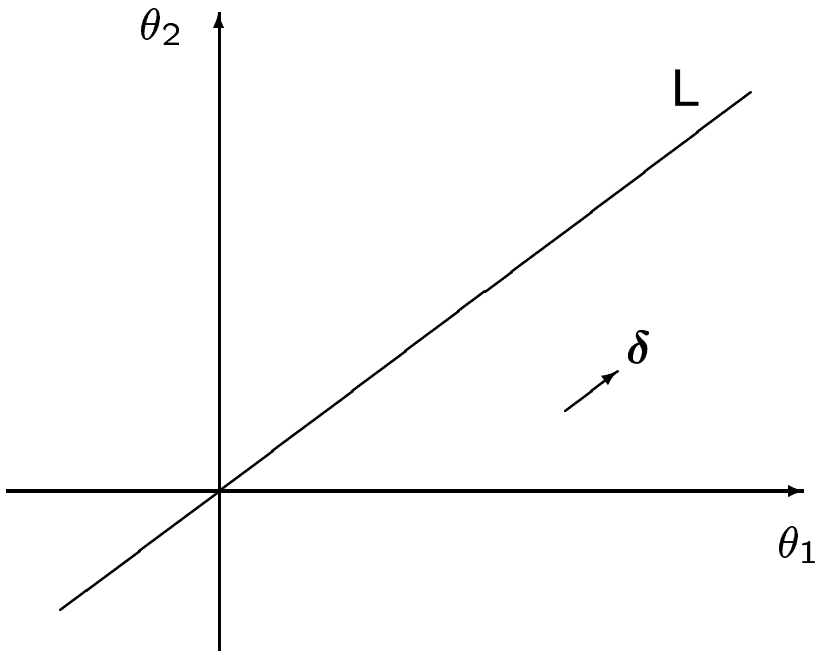
The 3-stage procedure obtains the usual reductions in expected sample sizes for a group sequential test.

When is a linear response combination appropriate?

Let $\boldsymbol{\theta} = \boldsymbol{\mu}_A - \boldsymbol{\mu}_B$, then

θ_1 = Improvement by Treatment A for response 1,

θ_2 = Improvement by Treatment A for response 2.



If we assume $\boldsymbol{\theta} = \lambda\boldsymbol{\delta}$ for given $\boldsymbol{\delta}$, we are assuming $\boldsymbol{\theta}$ must lie on the line L.

But, we must consider what may happen under other values of $\boldsymbol{\theta}$.

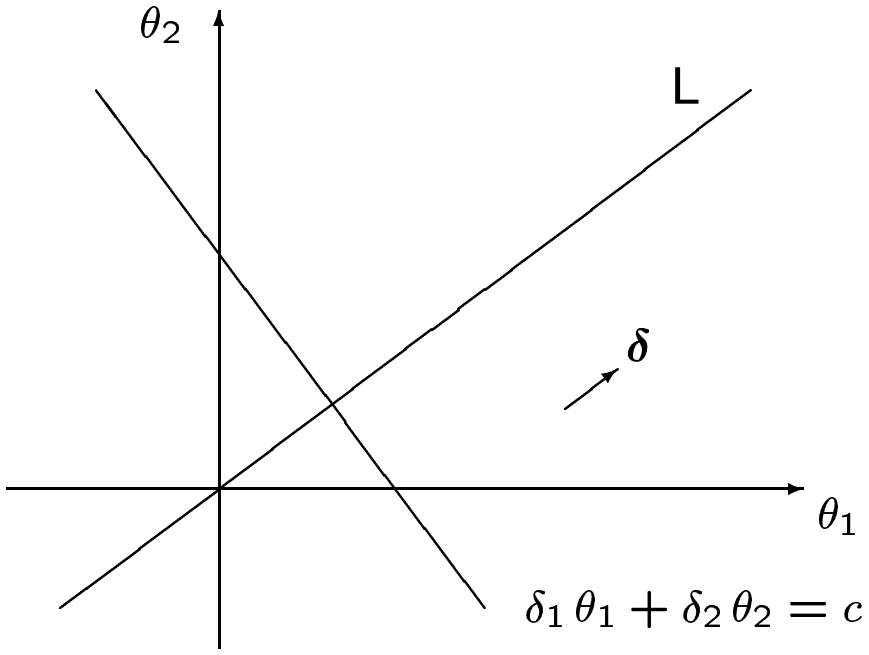
Linear combination of responses

For simplicity, suppose $\boldsymbol{V} = \boldsymbol{I}$.

Assuming $\boldsymbol{\theta} = \lambda\,\boldsymbol{\delta}$, the estimate of λ at stage k is

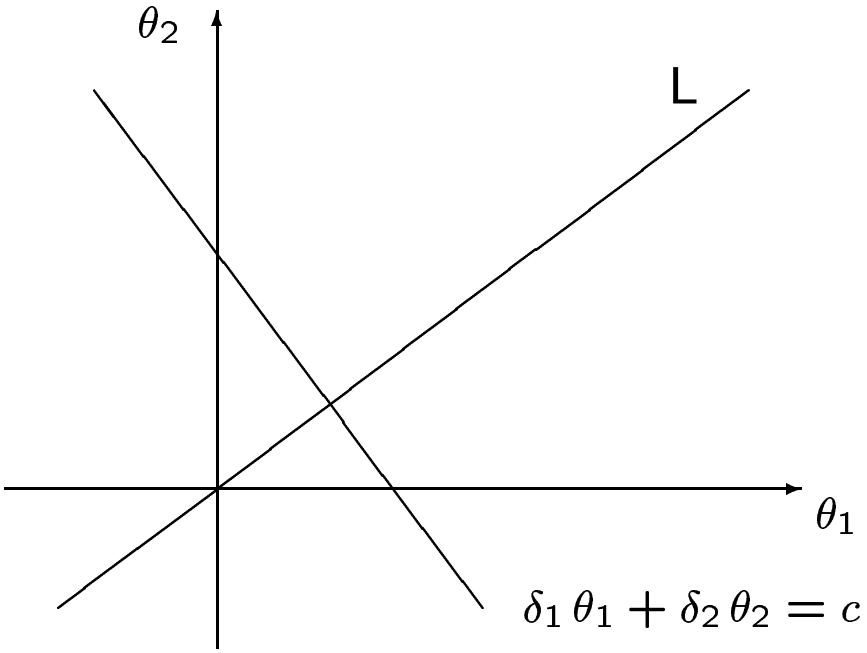
$$\widehat{\lambda}^{(k)} = \frac{\boldsymbol{\delta}^T(\bar{\boldsymbol{X}}_A^{(k)} - \bar{\boldsymbol{X}}_B^{(k)})}{\boldsymbol{\delta}^T\boldsymbol{\delta}}$$

which has mean $\boldsymbol{\delta}^T\boldsymbol{\theta}/(\boldsymbol{\delta}^T\boldsymbol{\delta})$.



The same mean, and the same joint distribution of $(\widehat{\lambda}^{(1)}, \ldots, \widehat{\lambda}^{(K)})$, arises for all values of $\boldsymbol{\theta}$ on a line orthogonal to L.

Linear combination of responses



Using a linear combination of responses “trades” between θ_1 and θ_2.

This *can* be desirable. It may even be reasonable when, say, θ_1 is negative and θ_2 positive.

In other situations, such trading is *not* appropriate — there is not much scope for such trading between *efficacy* and *safety*.

4. Marginal criteria

Studies with *efficacy* and *safety* responses:

Cancer chemotherapy trials

Efficacy: Survival time

Safety: Treatment toxicity

Chronic respiratory disease trial (Example 2)

Efficacy: PEFR, FEV_1, FVC

Safety: Lung mucociliary clearance

A new treatment must usually be shown to be *both* effective and safe.

Reference: Jennison & Turnbull, (*Biometrics*, 1993)

A testing formulation

Reduce measurements for each patient to a pair of responses

Example: A cross-over trial

$X_1 =$ Improvement of condition using active treatment

$X_2 =$ (Severity of side-effects on Placebo) $-$ (Severity of side-effects on Active treatment)

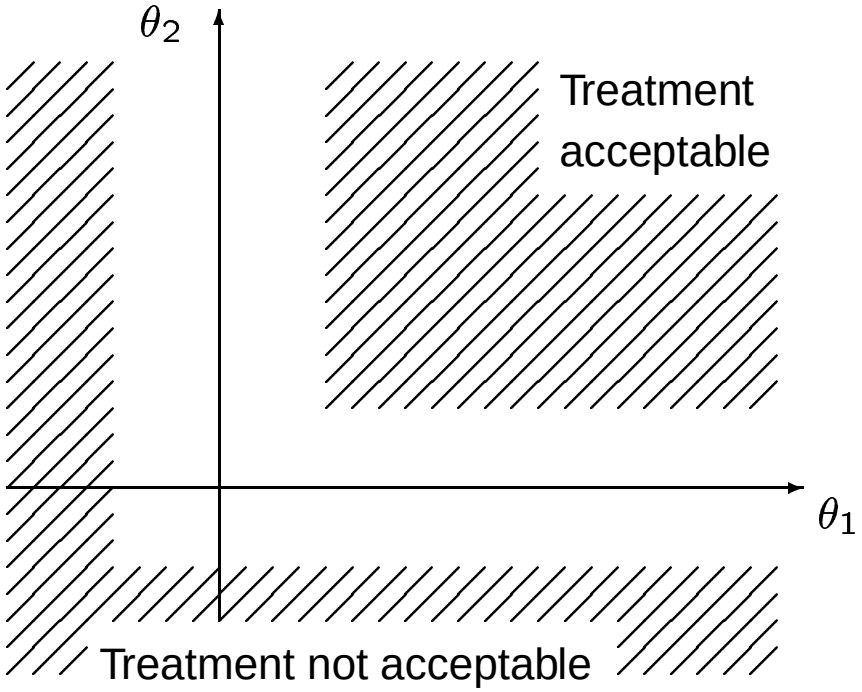
Define responses so a safe and effective treatment yields *high* values of X_1 and X_2.

Letting $\theta_1 = E(X_1)$ and $\theta_2 = E(X_2)$,

$X_1 > 0 \Rightarrow$ Treatment is effective,

$X_2 > 0 \Rightarrow$ Treatment is safe.

Setting type I and type II error rates



Type I error

We do not wish to recommend the new treatment if

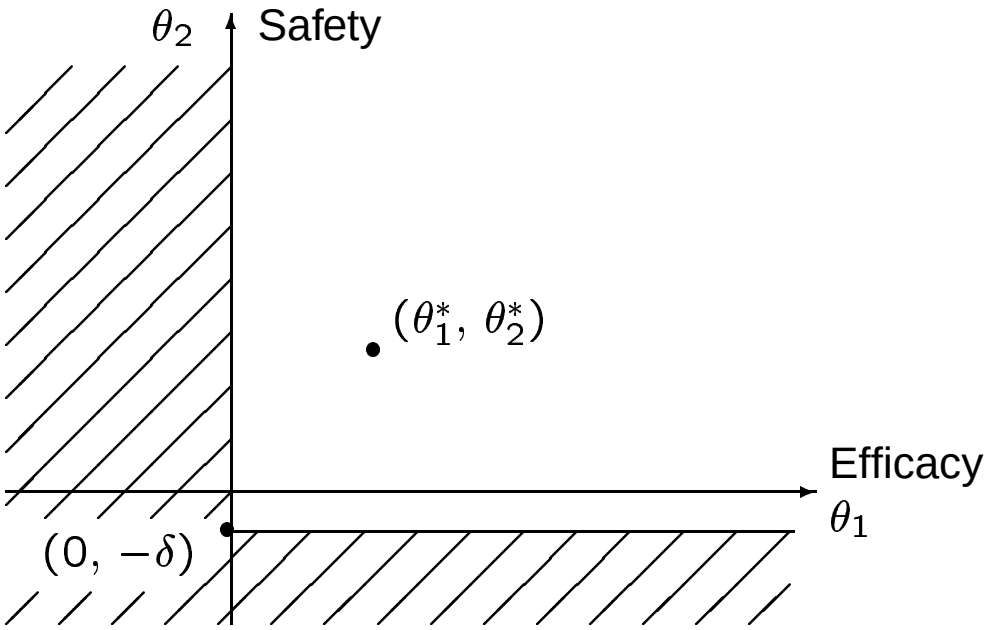
$\theta_1 \leq 0$ treatment is not effective or if

$\theta_2 \leq -\delta$ too many harmful side-effects.

Power

We want to recommend the new treatment if both θ_1 and θ_2 are large.

Type I and type II error rates



Require

$$Pr_{\boldsymbol{\theta}}\{\text{Recommend new treatment}\} \leq \alpha$$

$$\text{if } \theta_1 \leq 0 \text{ or } \theta_2 \leq -\delta, \qquad (1)$$

$$Pr_{\boldsymbol{\theta}}\{\text{Recommend new treatment}\} \geq 1 - \beta$$

$$\text{if } \theta_1 \geq \theta_1^* \text{ and } \theta_2 \geq -\theta_2^*.$$

In (1), highest error rates are at $(0, \infty)$ and $(\infty, -\delta)$.

NB The error rate at $(0, -\delta)$ is not the key concern.

A group sequential bivariate test

Observations

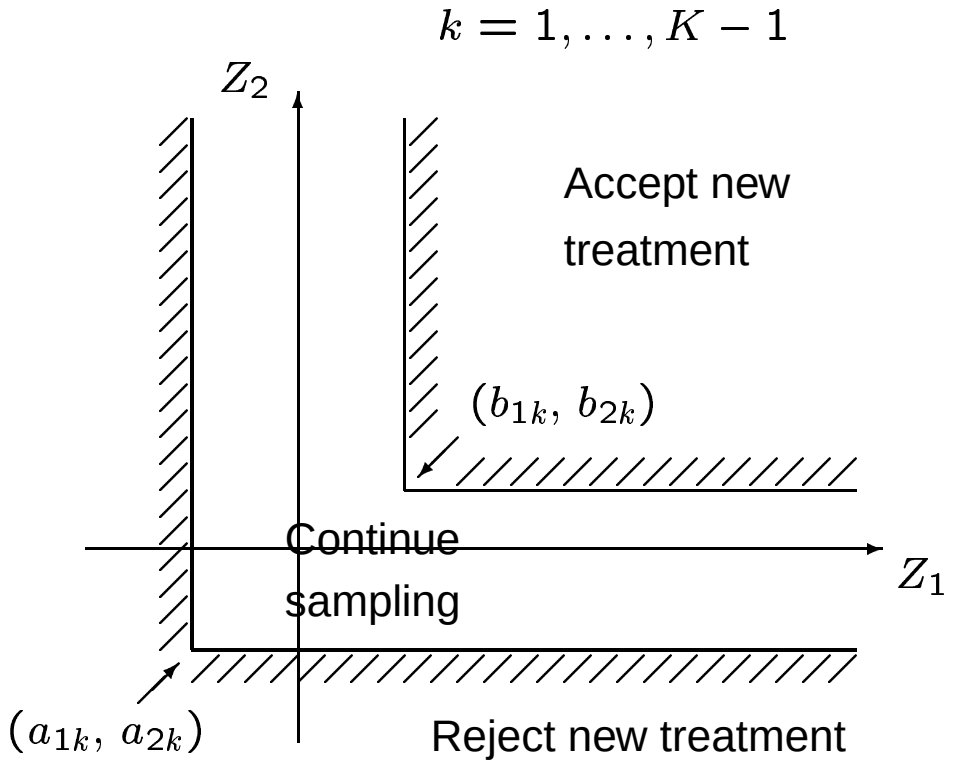
$$(X_1, X_2) \sim N\left((\theta_1, \theta_2), \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

After n observations

$$Z_1 = \bar{X}_1 \sqrt{n}/\sigma \sim N(0, 1) \text{ if } \theta_1 = 0,$$

$$Z_2 = (\bar{X}_2 + \delta) \sqrt{n}/\sigma \sim N(0, 1) \text{ if } \theta_2 = -\delta.$$

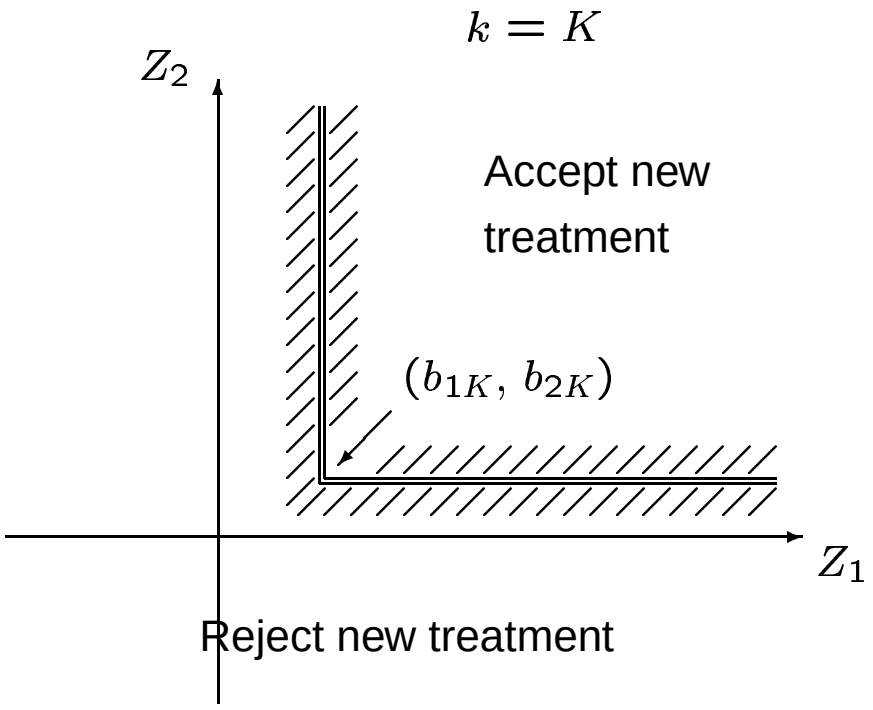
Take K groups of g observations with stopping regions:



A group sequential bivariate test

Up to K groups of g observations, monitored using L-shaped stopping regions.

Final analysis:



The sequence of boundaries must be chosen to give:

Type I error rate α at $\boldsymbol{\theta} = (\infty, -\delta)$ and $\boldsymbol{\theta} = (0, \infty)$,

Power $1 - \beta$ at $\boldsymbol{\theta} = (\theta_1^*, \theta_2^*)$.

Attaining type I error rate α at $\theta = (0, \infty)$

As $\theta_2 \to \infty$, values of Z_{2k} are high. (Extremely safe.)

Thus the test's outcome depends on the direction in which the sequence $\{Z_{11}, \ldots, Z_{1K}\}$ leaves the region $\{(a_{11}, b_{11}), \ldots, (a_{1K}, b_{1K})\}$. (Is the treatment effective?)

Now, $\{Z_{11}, \ldots, Z_{1K}\}$ has the canonical joint distribution

$$(Z_{11}, \ldots, Z_{1K}) \sim \text{multivariate normal},$$

$$Z_{1k} \sim N(\theta_1 \sqrt{\mathcal{I}_k},\, 1) \quad \text{for each } k,$$

$$Cov(Z_{1k_1},\, Z_{1k_2}) = \sqrt{(\mathcal{I}_{k_1} / \mathcal{I}_{k_2})} \quad \text{for } k_1 < k_2,$$

where $\mathcal{I}_k = \sqrt{(kg/\sigma^2)}$.

We can choose any *univariate* group sequential boundary $\{(a_{11}, b_{11}), \ldots, (a_{1K}, b_{1K})\}$ such that

$$Pr_{\theta_1 = 0}\{\text{Exit upper boundary, } Z_{1k} > b_{1k}\} = \alpha.$$

Attaining type I error rate α at $\boldsymbol{\theta} = (\infty, -\delta)$

As $\theta_1 \to \infty$, values of Z_{1k} are high. (Very effective.)

Thus the test's outcome depends on the direction in which the sequence $\{Z_{21}, \ldots, Z_{2K}\}$ leaves the region $\{(a_{21}, b_{21}), \ldots, (a_{2K}, b_{2K})\}$. (Is the treatment safe?)

Now, $\{Z_{21}, \ldots, Z_{2K}\}$ has the canonical joint distribution

$$(Z_{21}, \ldots, Z_{2K}) \sim \text{multivariate normal},$$

$$Z_{2k} \sim N((\theta_2 + \delta)\sqrt{\mathcal{I}_k},\, 1) \quad \text{for each } k,$$

$$Cov(Z_{2k_1},\, Z_{2k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})} \quad \text{for } k_1 < k_2,$$

where $\mathcal{I}_k = \sqrt{(kg/\sigma^2)}$.

We can choose any univariate group sequential boundary $\{(a_{21}, b_{21}), \ldots, (a_{2K}, b_{2K})\}$ such that

$$Pr_{\theta_2 = -\delta}\{\text{Exit upper boundary, } Z_{2k} > b_{2k}\} = \alpha.$$

Attaining power $1 - \beta$ at $\boldsymbol{\theta} = (\theta_1^*, \theta_2^*)$

The mean of (Z_{1k}, Z_{2k}) is augmented by increasing the group size g. We need the value g for which

$$Pr_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}\{\text{Recommend new treatment}\} = 1 - \beta.$$

For general values of $\boldsymbol{\theta}$, acceptance and rejection probabilities depend on the correlation coefficient ρ.

Univariate calculations no longer suffice: group sequential *bivariate* calculations are required.

Given standardised boundary values $a_{1k}, a_{2k}, b_{1k}, b_{2k}$, $k = 1, \ldots, K$, and a value of ρ, we can compute

$$Pr_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}\{\text{Recommend new treatment}\} = 1 - \beta.$$

for any group size g and, hence, search for the group size that meets the power condition.

Example

Parameter values:

$\sigma^2 = 1,\ \rho = 0.2,$

$-\delta = -0.2,\ (\theta_1^*, \theta_2^*) = (0.2, 0),$

$\alpha = 0.05,\ 1 - \beta = 0.8,$

$K = 5$ groups.

Use univariate boundaries from Emerson & Fleming (*Biometrics*, 1989) with parameter $p = 0.5$.

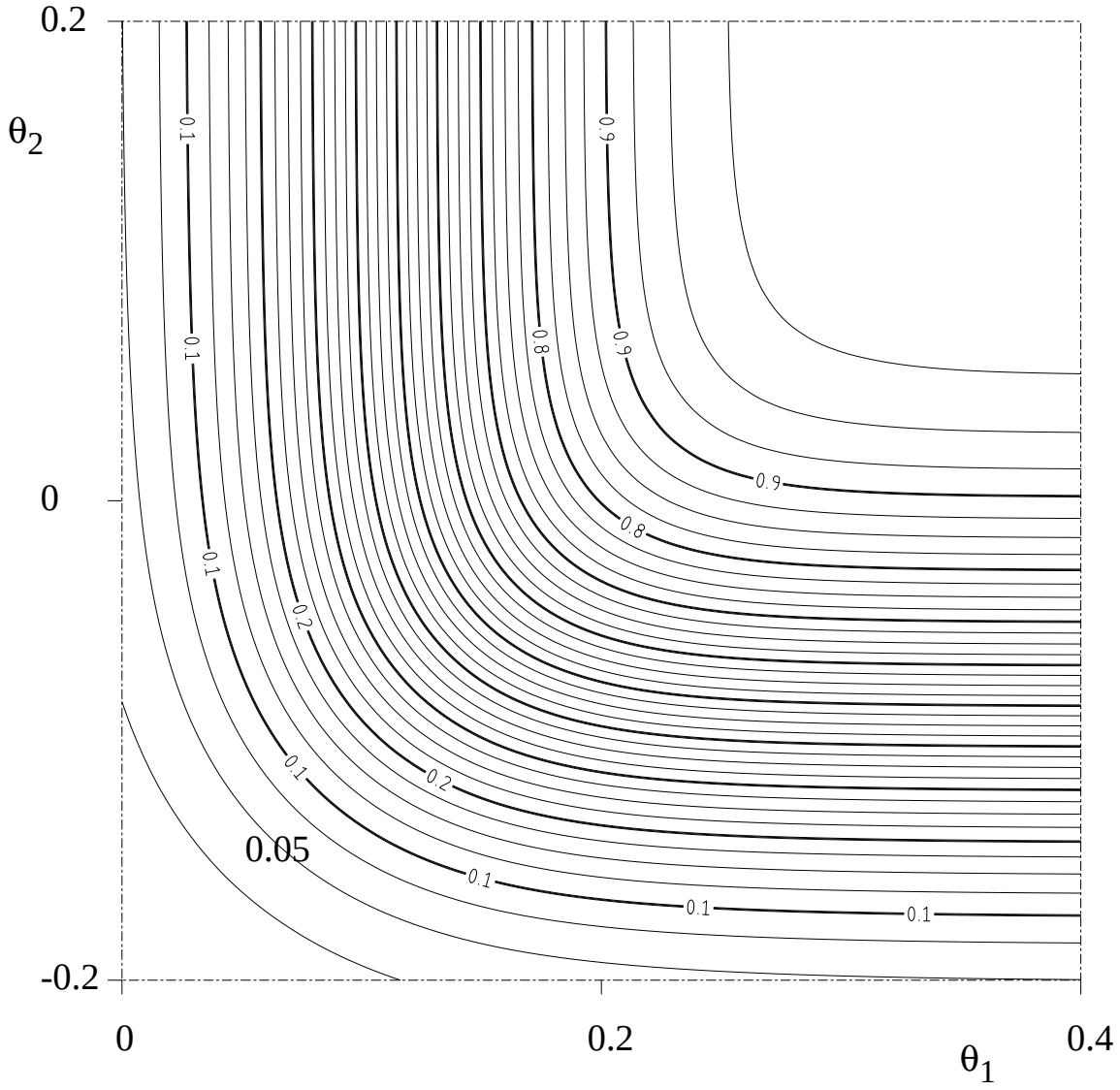
A fixed sample test would require a sample size of 206.

Maximum sample size for the group sequential test is 325.

(For $p = 0,$ maximum sample size would be 225.)

Group sequential bivariate test: Example

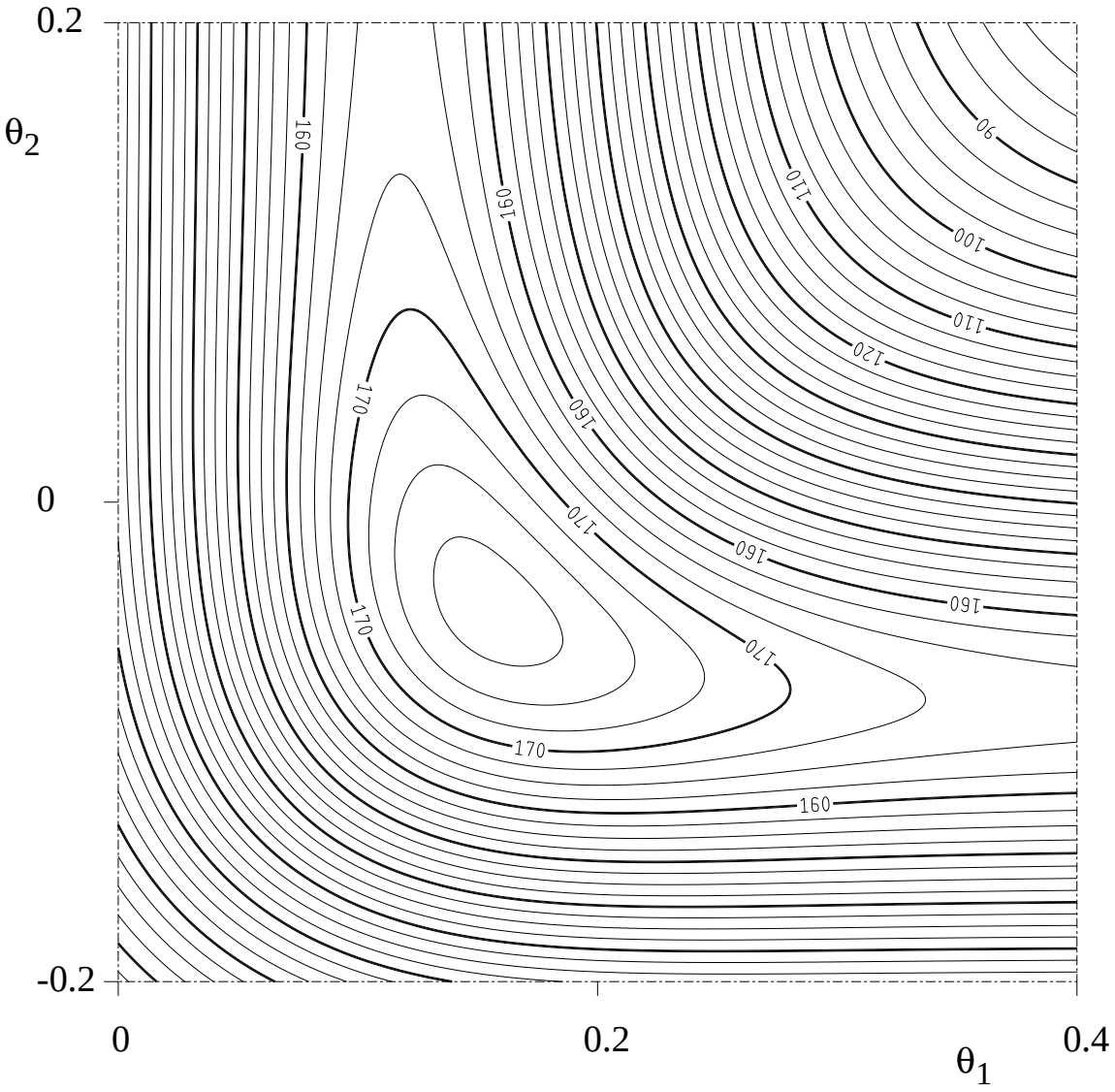
Contour plot of power against $\boldsymbol{\theta}$.



The 0.05 contour has asymptotes at $\theta_1 = 0,\ \theta_2 = -0.2$. This contour passes through $(0.06, -0.14)$; the type I error rate at $(0, -0.2)$ is much lower than 0.05.

Group sequential bivariate test: Example

Contour plot of ASN against $\boldsymbol{\theta}$.



The maximum ASN of just under 180 is well below the fixed sample size of 206.

Conclusions

It is natural to record multiple endpoints

Care must be taken to combine information in an appropriate manner

Once a testing problem is formulated, group sequential designs can be created

Efficiency gains from sequential monitoring are available