Discussion on

Are Flexible Designs Sound?

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Bruce Turnbull

Department of Statistics,

Cornell University

http://legacy.orie.cornell.edu/∼bruce

Joint Statistical Meetings, Seattle,

August 8, 2006

Features of sound designs

The authors focus on the ***validity*** of inference following flexible designs.

A related question is their ***efficiency:***

- Institutional Review Boards (Ethics Committees) and studies' Monitoring Boards should be concerned that patients are used as effectively as possible.

- Flexibility in design can come at the price of reduced efficiency — a test of the same size and power could be conducted with, on average, fewer subjects.

Example of a flexible, adaptive design

Shen & Fisher (*Biometrics*, 1998) propose a ***variance spending*** test in which Z-statistics from successive groups of observations are combined as

$$Z = w_1 Z_1 + w_2 Z_2 + \ldots + w_m Z_m.$$

Each w_j can depend on responses in groups 1 to $j - 1$.

The final w_m is chosen to that $w_1^2 + \ldots + w_m^2 = 1$.

Then, $Z \sim N(0, 1)$ under H_0: $\theta = 0$.

Rejecting H_0 for $Z > \Phi^{-1}(1 - \alpha)$ ensures type I error probability α.

Variance spending design

Shen & Fisher start from a fixed sample design with n_f observations which gives power 0.9, say, at $\theta = \delta$.

They observe interim estimates of θ.

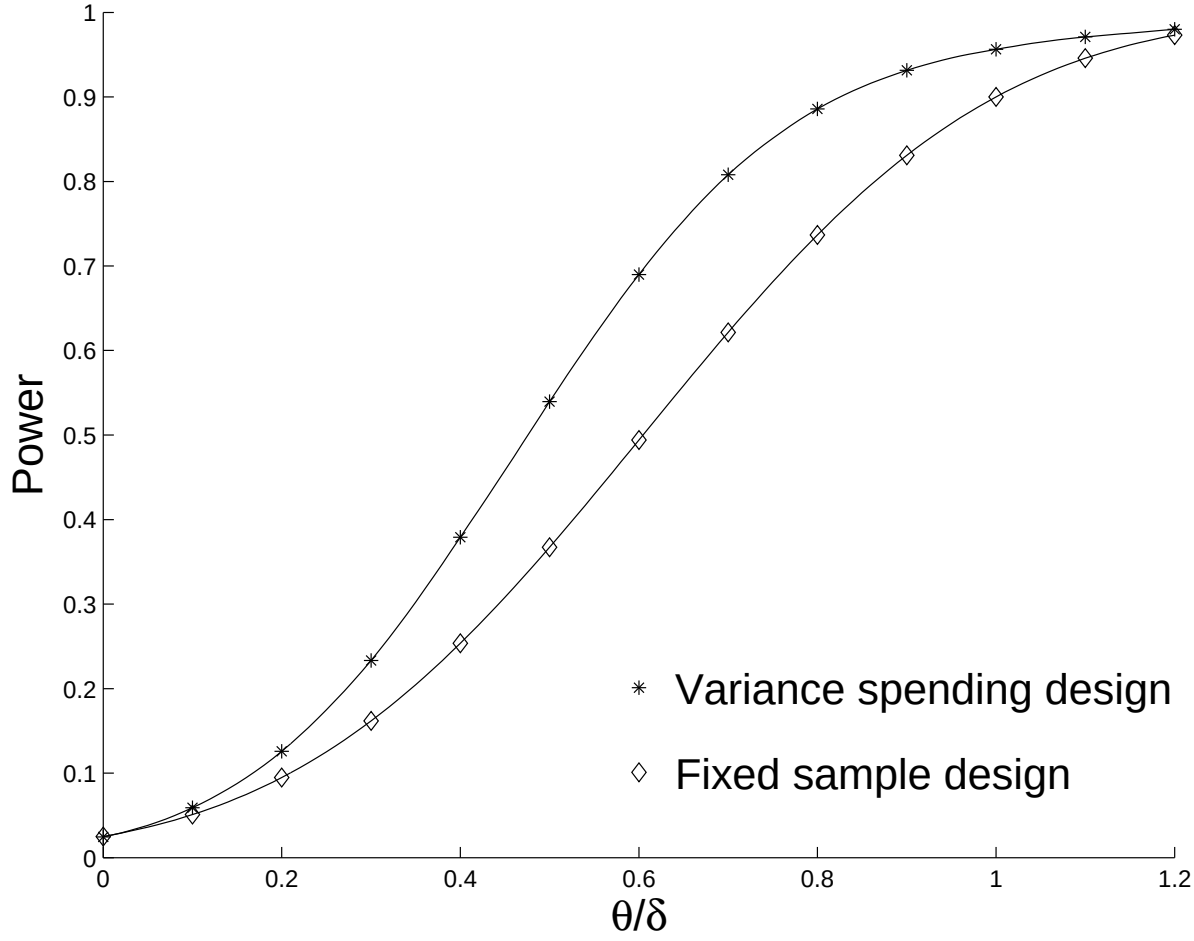
If $\widehat{\theta} < \delta$, sample size is increased beyond n_f in order to try and achieve power 0.9 under the true θ.

Weights w_j are amended accordingly.

We consider an example with up to 10 groups of observations and a maximum sample size of $2\,n_f$.

(Full details are in our written discussion.)

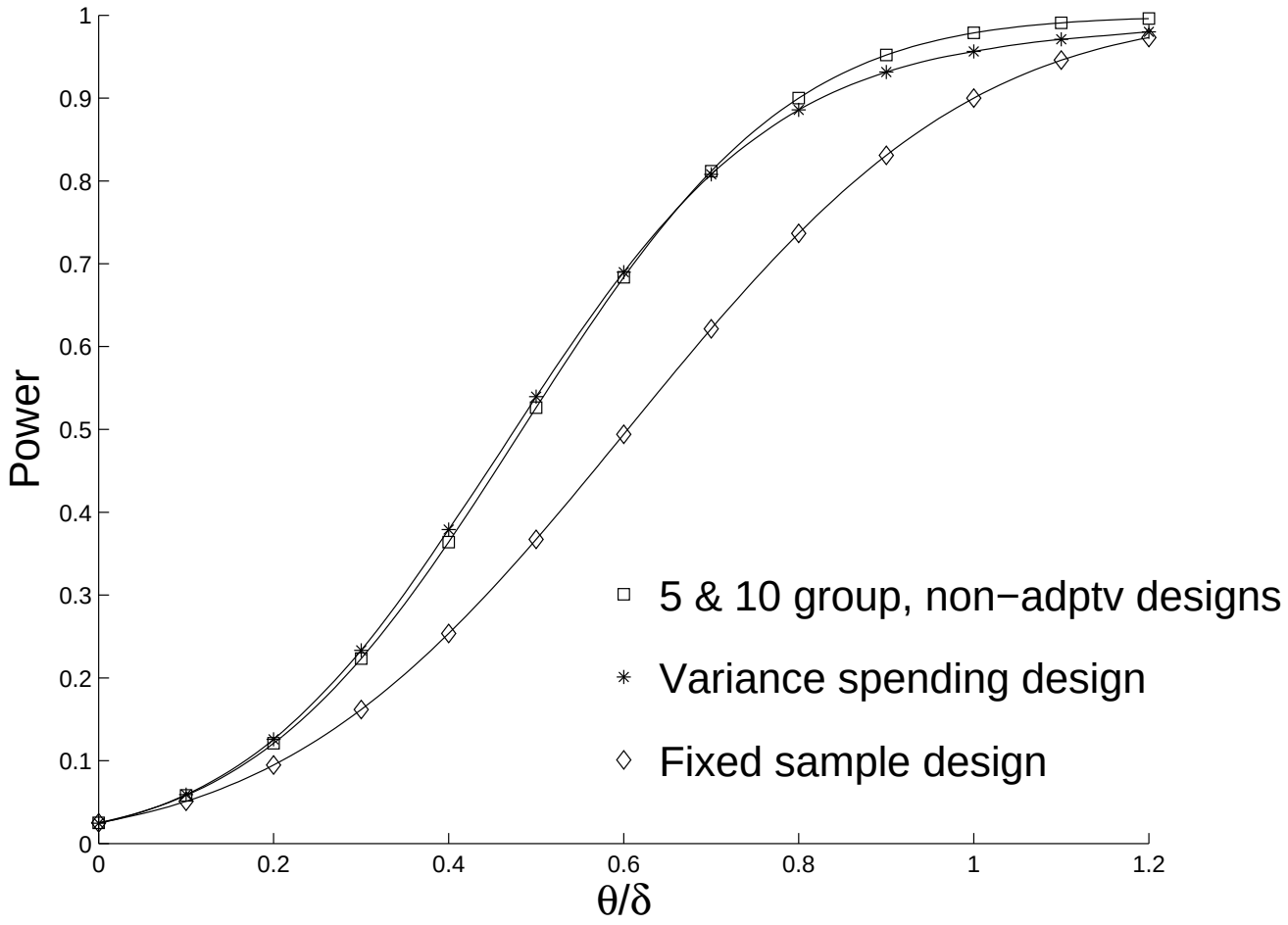
Variance spending design



The adaptation is successful in increasing power beyond that of the fixed sample test.

Note: The variance spending test is pre-specified, this power curve can be computed in advance — and one can consider other procedures achieving similar power.

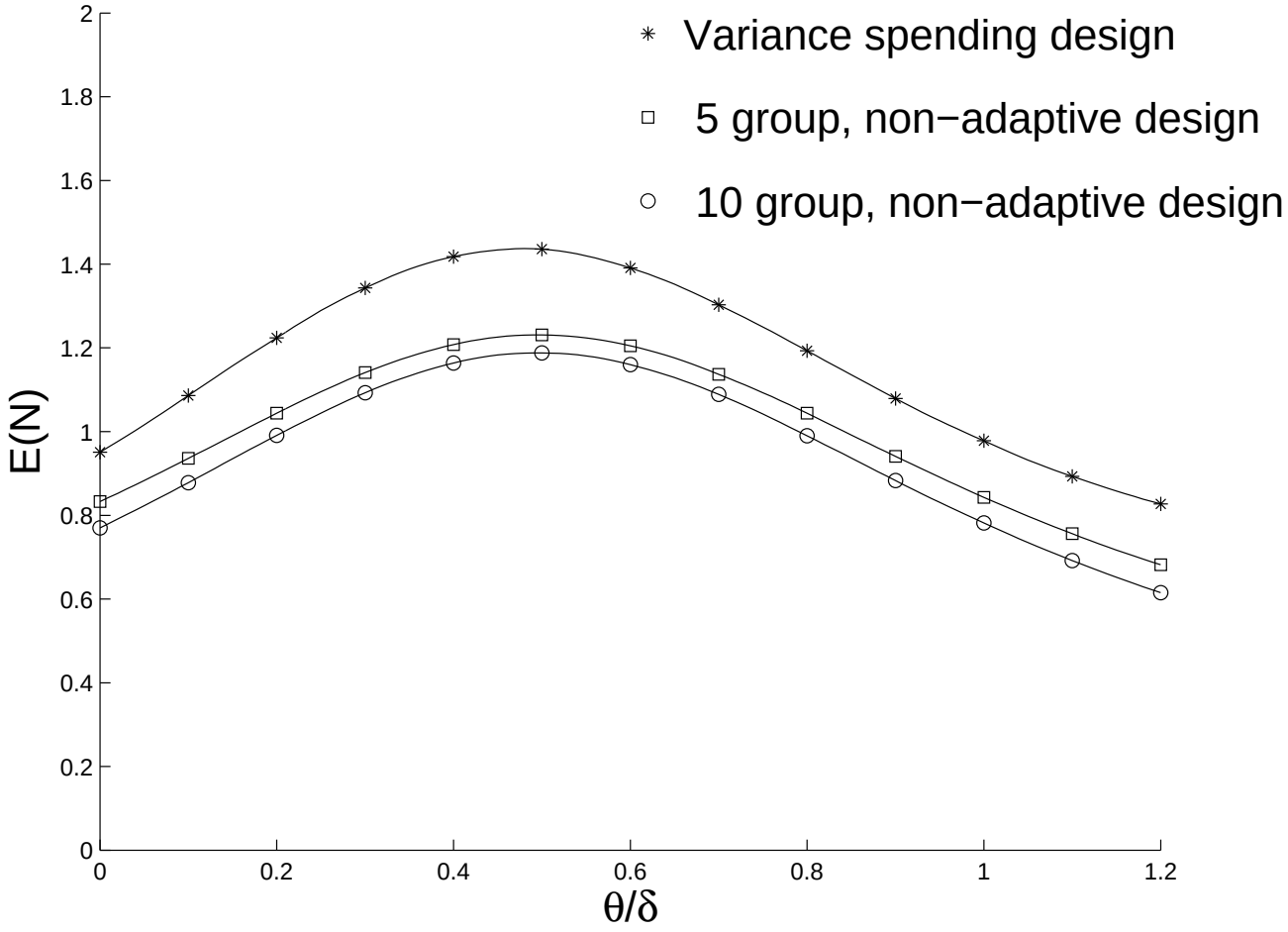
Variance spending design



Non-adaptive, group sequential tests can produce a similar power curve . . .

(Power curves for 5 and 10 group designs are identical.)

Variance spending design



. . . and the group sequential tests have smaller expected sample sizes by 10 to 15%.

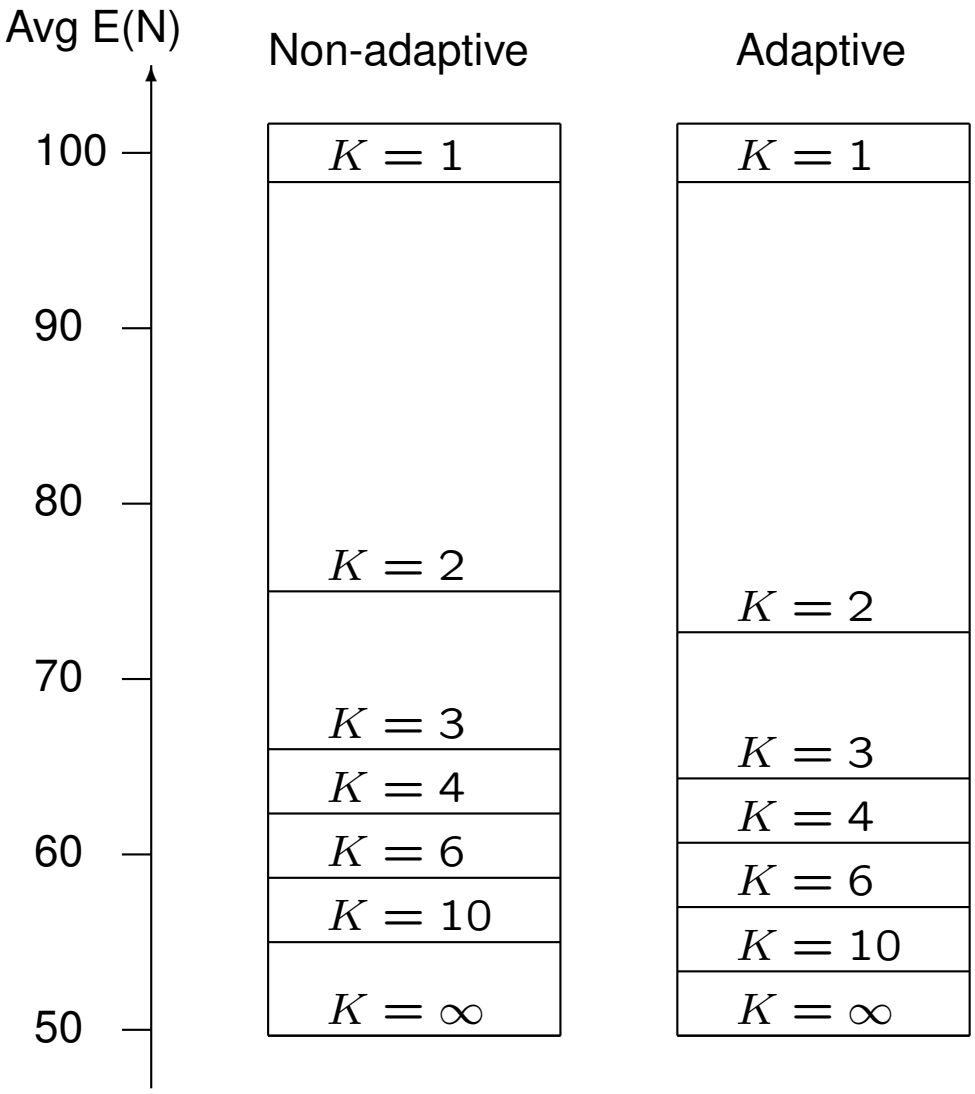
Conclude: The variance spending design is inefficient.

This is quite typical for an adaptive design aiming for conditional power under effect size $\widehat{\theta}$.

Efficiency of adaptive and non-adaptive designs

Reference: Jennison & Turnbull (*Biometrika*, 2006)

Suitably defined adaptive designs can be efficient — and make small improvements on non-adaptive group sequential tests with the same number of groups, K.



Sources of inefficiency in flexible, adaptive designs

1. Use of non-sufficient statistics

Example: The variance spending statistic

$$Z = w_1 Z_1 + w_2 Z_2 + \ldots + w_m Z_m.$$

With "adaptive" weights w_j, this statistic is not a function of the sample sum.

JT (2006) show that all admissible designs (adaptive or non-adaptive) are Bayes procedures.

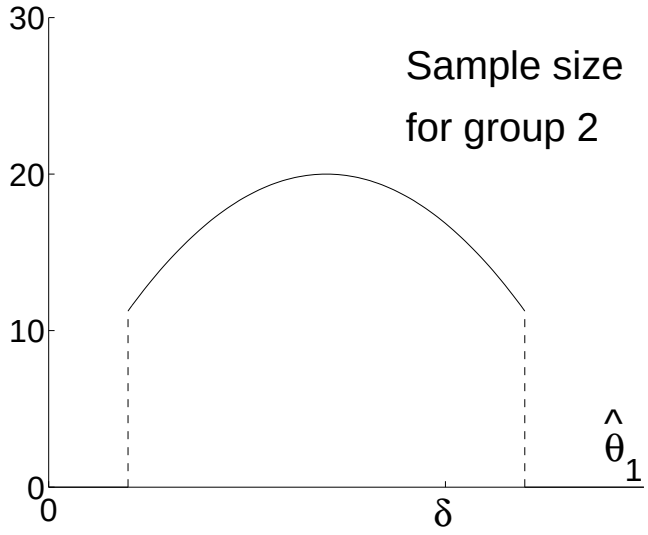
Hence, admissible designs have decision rules and sample size rules based on sufficient statistics.

Designs based on non-sufficient statistics are, thus, inadmissible.

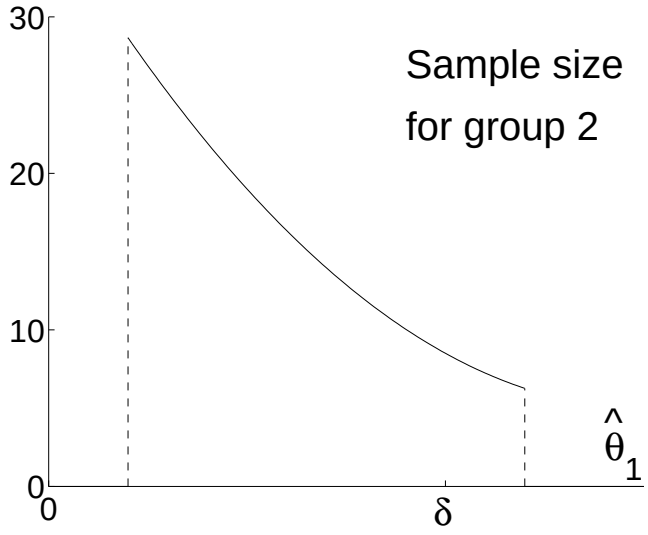
Sources of inefficiency

2. Sub-optimal sample size modification rule

Typical sample size function for an optimal adaptive test



Typical sample size function for a conditional power adaptive design



“Conditional power” sample size modification rules differ qualitatively from those of optimal adaptive designs.

Sources of inefficiency

Burman & Sonneson's Example 4

The authors propose a Likelihood Ratio test for this example.

The test is Bayes, and hence admissible among decision rules ***given*** this sample size modification rule.

However, the ***sample size rule*** itself is inadmissible among all possible adaptive design rules.

We have found two-stage, non-adaptive group sequential designs with similar power and expected sample sizes lower by 10 to 30%.

Conclusion

Burman & Sonneson state:

> "The most fundamental question is . . . not whether flexible designs are efficient but rather what inference following a flexible design is valid."

We would not wish this view to imply that inefficient statistical procedures are "acceptable".

Validity and efficiency are important and inter-related, and ***both*** should be central to the discussion.