Changing study design at an interim analysis:

Is this efficient? Is this necessary?

Chris Jennison,

Department of Mathematical Sciences, University of Bath, UK

and

Bruce Turnbull,

Department of Statistical Science, Cornell University, Ithaca, NY

Belgian Statistical Society

10th Annual Meeting

Kerkrade,

18–19 October 2002

http://www.bath.ac.uk/∼mascj

Plan of talk

1. Motivation for adaptive sample size designs.

2. “Variance spending” and related methods.

3. *Example:* A hypothesis test with a single, final analysis.

4. Formulating the real testing problem.

5. Group sequential tests.

6. Adaptive group sequential tests.

A variety of adaptive and flexible procedures

- Adapting the sample size to estimates of nuisance parameters.
- Adaptive randomisation rules designed to allocate fewer subjects to the inferior treatment.
- Flexibility to change treatment, outcome or response during a study.
- Re-assessing the power requirement in response to interim data.

§1 Motivation: Prototype example

Balanced parallel design

$$X_{Ai} \sim N(\mu_A, \sigma^2), \quad X_{Bi} \sim N(\mu_B, \sigma^2)$$
$$Y_i = X_{Ai} - X_{Bi} \sim N(\theta, 2\sigma^2)$$
$$\theta = \mu_A - \mu_B$$

The MLE of θ is $\widehat{\theta} = \overline{X}_A - \overline{X}_B$.

Without loss of generality, suppose $2\sigma^2 = 1$.

Aim: to Test H_0: $\theta = 0$ versus H_1: $\theta > 0$

with Type I error rate α, e.g. $\alpha = 0.025$.

Fixed sample design

Initially aim for power $1 - \beta$ at target effect size $\theta = \delta$.

Hence set sample size

$$n \;=\; (z_\alpha + z_\beta)^2 \, \frac{2\sigma^2}{\delta^2} \;=\; \left(\frac{z_\alpha + z_\beta}{\delta} \right)^2$$

per treatment arm, where $z_\alpha = \Phi^{-1}(1 - \alpha)$, etc.

(Recall $2\sigma^2 = 1$.)

Data at an intermediate stage

After a fraction r of the sample size (information) is collected,

$$\begin{aligned} \widehat{\theta}_1 &\sim N(\theta, \tfrac{1}{rn}), \\ S_1 &\sim N(\theta rn, rn). \end{aligned}$$

Intermediate results may be examined, even though a formal interim analysis was not planned.

Disappointing results

- Suppose $\widehat{\theta}_1$ is positive but smaller than the hoped for effect size δ.
- It is unlikely that H_0 will be rejected (low conditional power).
- However, the magnitude of $\widehat{\theta}_1$ is clinically meaningful.
- It appears the original target effect size δ was over-optimistic.

Can this trial be “rescued” ?

Revising the sample size

- Let $\xi = \delta/\widehat{\theta}_1$ and suppose $\xi > 1$.
- With hindsight, we wish we had designed the test with power $1 - \beta$ at $\theta = \delta/\xi$ rather than at $\theta = \delta$.
- This would have required the larger sample size $\xi^2 n$ instead of n.
- One might collect extra observations in the remainder of the study to make a total sample size of $\xi^2 n$.

Naive test leads to inflated Type I error

Suppose we behave as if the sample size $\xi^2 n$ was pre-planned and compute

$$Z = \left(\overline{X}_A - \overline{X}_B\right) \sqrt{\xi^2 n}.$$

Since ξ is a function of the first stage data, Z is *not* $N(0,1)$.

The test that rejects when $Z > z_\alpha$ does not have Type I error α.

Type I error rate is inflated

- typically by 30% to 40% (Cui, Hung & Wang, *Bmcs,* 1999)
- can more than double (Proschan, Follmann & Waclawiw, *Bmcs,* 1992).

§2 Variance spending

A fixed sample of n observations can be divided into

stage 1: $S_1 = \sum_{i=1}^{rn}(X_{Ai} - X_{Bi})$,

stage 2: $S_2 = \sum_{i=rn+1}^{n}(X_{Ai} - X_{Bi})$.

Then

$$\begin{aligned} S_1 &\sim N(rn\theta,\, rn), \\ S_2 &\sim N(\{1-r\}n\theta,\, \{1-r\}n), \\ S_1 + S_2 &\sim N(n\theta,\, n) \end{aligned}$$

and

$$Z = \frac{S_1 + S_2}{\sqrt{n}} \sim N(0,1) \quad \text{under } H_0\text{: } \theta = 0.$$

Variance spending — continued

If the stage 2 sample size is modified to $\gamma(1-r)n$ after seeing S_1,

$$S_1 \sim N(rn\theta,\, rn)$$

and, conditionally on S_1,

$$S_2' \sim N(\gamma\{1-r\}n\theta,\, \gamma\{1-r\}n).$$

Under H_0: $\theta = 0$,

$$\gamma^{-1/2}\, S_2' \sim N(0,\, \{1-r\}n)$$

unconditionally. Hence

$$Z = \frac{S_1 + \gamma^{-1/2}\, S_2'}{\sqrt{n}} \sim N(0,1) \quad \text{under } H_0.$$

Lloyd Fisher, *Statistics in Medicine*, 1998

Fisher explains "variance spending" as the construction of a Z statistic from components with pre-specified variances.

Under H_0,

$$W_1 = \frac{S_1}{\sqrt{n}} \sim N(0, r),$$

$$W_2 = \frac{S_2'}{\sqrt{\gamma n}} \sim N(0, 1-r)$$

and

$$Z = W_1 + W_2 \sim N(0,1).$$

Cui, Hung & Wang, *Biometrics*, 1999

Cui et al consider the joint distribution of weighted sample sums.

They show that, under H_0,

$$(S_1,\, S_1 + \gamma^{-1/2}\, S_2')$$

has the same joint distribution as the original

$$(S_1,\, S_1 + S_2).$$

This result generalises to a group sequential setting with K analyses and one or more re-design points.

Conditional Type I error probability

In the original test, the conditional Type I error probability after stage 1 is

$$P_{\theta=0}\{S_1 + S_2 > z_\alpha\sqrt{n} \mid S_1 = s_1\}. \qquad (1)$$

If stage 2 sample size is modified and a rule defined that preserves the conditional error probability (1), overall Type I error rate α is maintained.

- The methods of Fisher and Cui et al do this.
- Jennison & Turnbull (2002) show that any unplanned design modification *must* have this property.
- Müller & Schäfer (2001) and Denne (2001) use this construction in adaptive group sequential designs.

Variance spending — notes

- For $\gamma > 1$, second stage observations are down-weighted. The final statistic Z is not sufficient for θ — so the efficiency of this approach is suspect.
- The distribution of Z under $\theta \neq 0$ is not simple. The inter-relation of stages 1 and 2 needs to be properly treated in power calculations.

We shall assess power and average sample size of this method in an example with a specific rule for the stage 2 sample size.

§3 Example

Original fixed sample design:

To test H_0: $\theta = 0$ with Type I error rate α and power $1 - \beta$ at $\theta = \delta$.
The study needs $n = (z_\alpha + z_\beta)^2/\delta^2$ observations.

After stage 1:

From rn observations, we find $\widehat{\theta}_1 = \delta/\xi$ and decide to aim for power $1 - \beta$ at $\theta = \delta/\xi$.

We modify the second stage sample to $\gamma(1 - r)n$ and follow the variance spending approach, creating

$$Z = (S_1 + \gamma^{-1/2} S_2')/\sqrt{n}.$$

Sample size rule, with truncation

Aim for power $1-\beta$ at $\theta = \delta/\tilde{\xi}$ where

$$\tilde{\xi} = \tilde{\xi}(\widehat{\theta}_1) = \begin{cases} 4 & \text{for} \quad \widehat{\theta}_1 \leq \delta/4, \\ \delta/\widehat{\theta}_1 & \quad\quad\quad \delta/4 < \widehat{\theta}_1 < 2\delta, \\ 0.5 & \quad\quad\quad \widehat{\theta}_1 \geq 2\delta. \end{cases} \tag{2}$$

Note that reduction in sample size is possible for high values of $\widehat{\theta}_1$.

If the interim look is at the halfway point, i.e., $r = 0.5$, the second stage inflation factor is

$$\gamma(\widehat{\theta}_1) \;=\; 4\{\tilde{\xi}(\widehat{\theta}_1) - 0.5\}^2 \;\in\; (0, 49).$$

Example: full details

Initial test:

Type I error rate: $\alpha = 0.025$.

Power: $1 - \beta = 0.9$ at $\theta = \delta$.

Planned sample size: $n = 10.5/\delta^2$ per treatment arm.

Modification:

Intermediate look after $n/2$ observations per treatment arm.

Inflation factor $\gamma(\widehat{\theta}_1) = 4\{\tilde{\xi}(\widehat{\theta}_1) - 0.5\}^2 \in (0, 49)$.

Total sample size is in the range $(0.5n,\ 25n)$.

Also, stop for "futility" at stage 1 and accept H_0 if $\widehat{\theta}_1/\delta < -0.173$,
in which case conditional power under $\theta = \delta/4$ is less than 0.8.

Figure 1. Power functions of Variance Spending test and Fixed Sample test with power 0.9 at $\theta = \delta$.

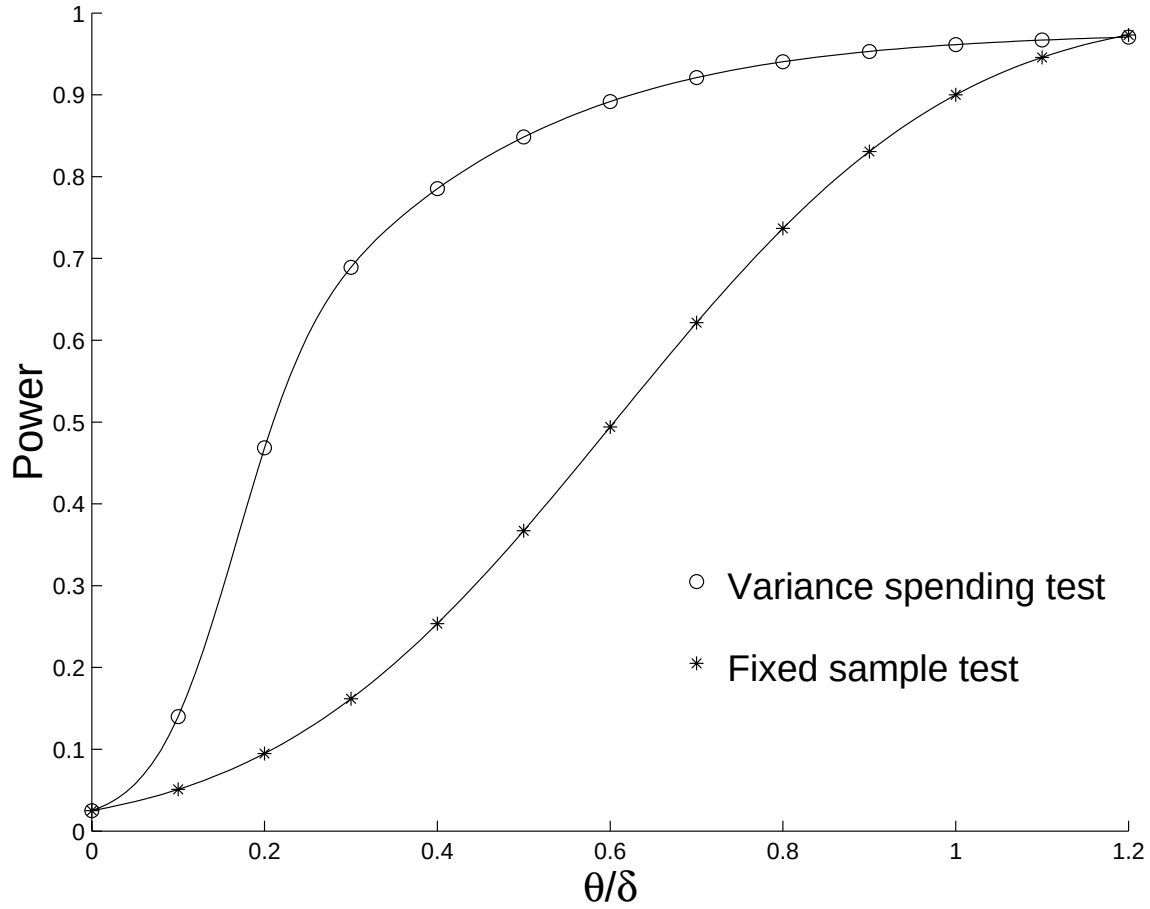


Figure 2. Power functions of Variance Spending test and Fixed Sample test with power 0.9 at $\theta = 0.6\,\delta$.

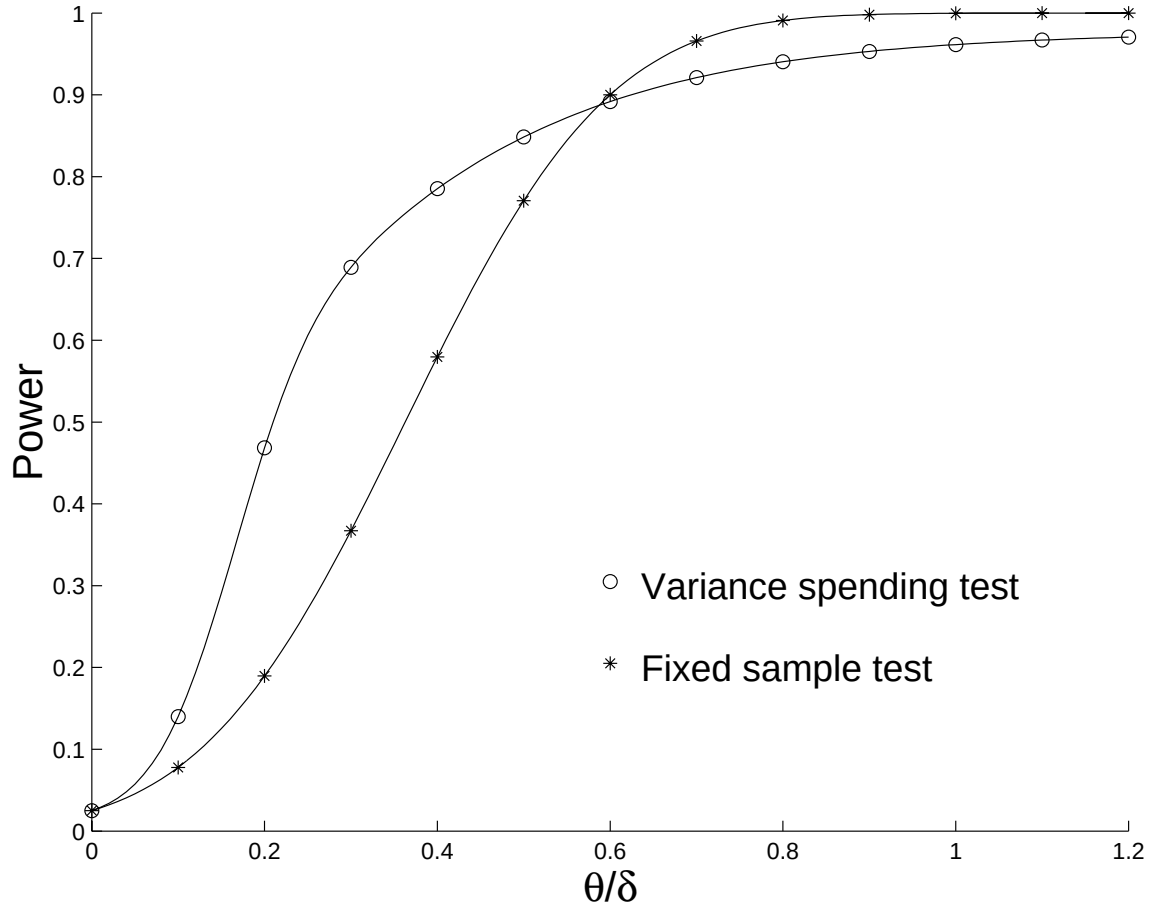
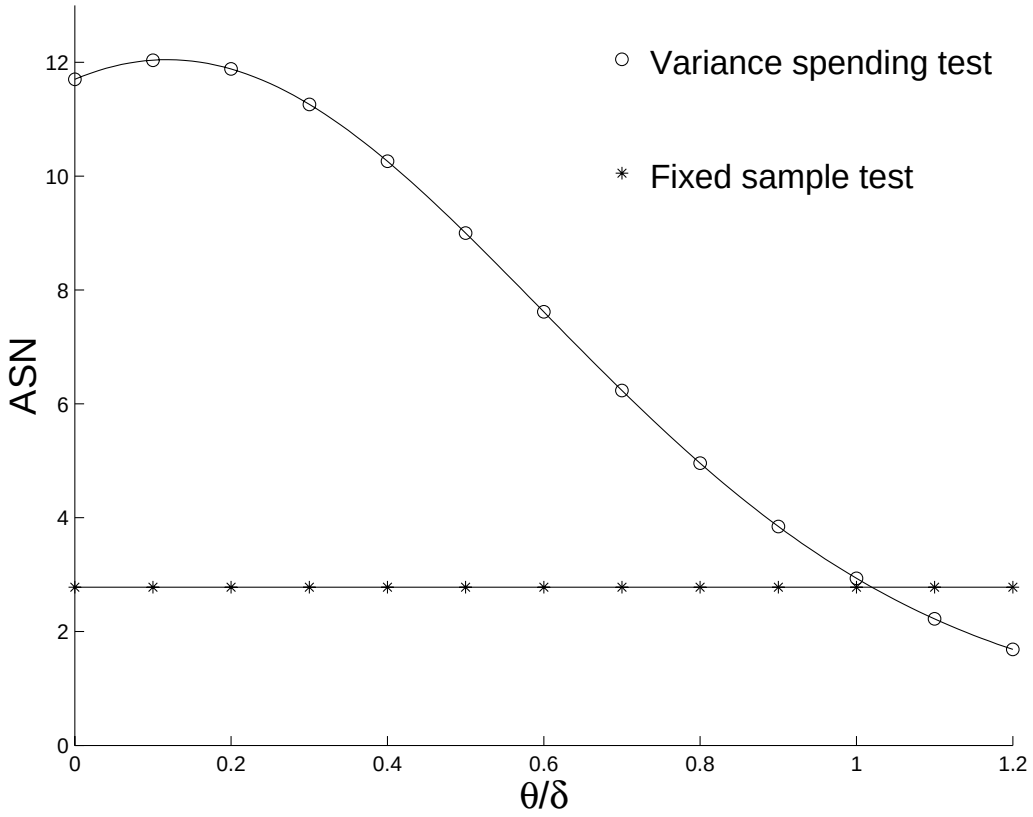


Figure 3. ASN curves of Variance Spending test and Fixed Sample test with power 0.9 at $\theta = 0.6\,\delta$.



ASN (average sample number) expressed as multiples of the fixed sample size, n.

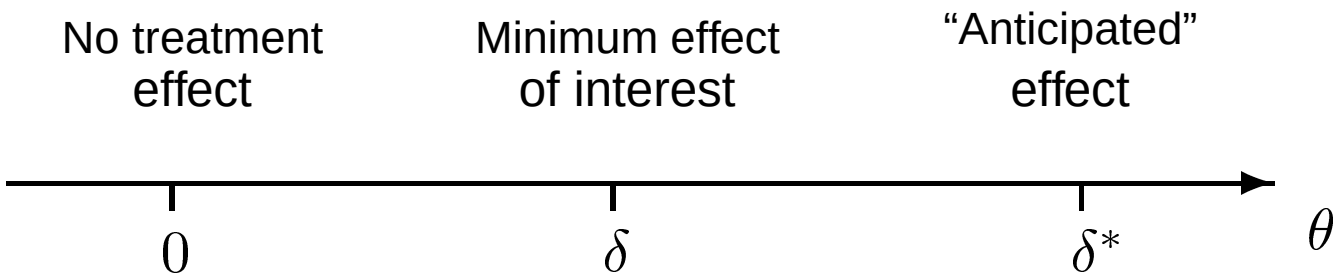
§4 Formulating the testing problem

Test H_0: $\theta = 0$ with:

Type I error rate α,

power $1 - \beta$ at $\theta = \delta$,

low ASN at $\theta = \delta^* \gg \delta$.

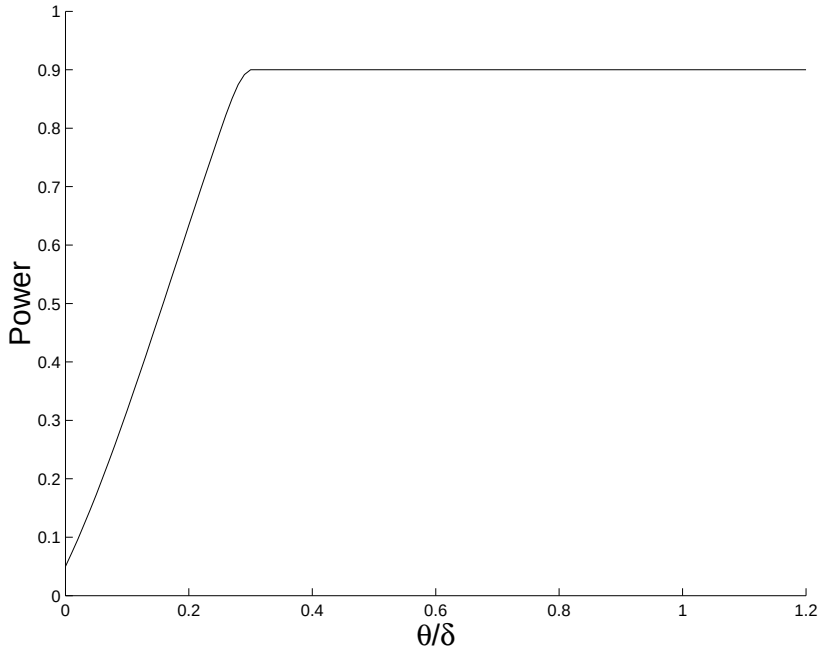


It should not be necessary to see $\widehat{\theta}_1 = \delta$ before realising a treatment effect of this size is (just) worth pursuing.

Setting power: a strange philosophy

Shen and Fisher (1999) refer to setting power $1 - \beta$ at effect size δ where δ is an *estimate* of θ.

This implies a target power function of the following form (!)



§5 Group sequential tests

Analyse data after $n_1, n_2, \ldots, n_K$ observations, with early stopping to reject H_0: $\theta = 0$ or to accept H_0.

Standard group sequential test:

Fix targets for $n_1, \ldots, n_K$ — maybe not equally spaced.

Sequentially planned sequential test:

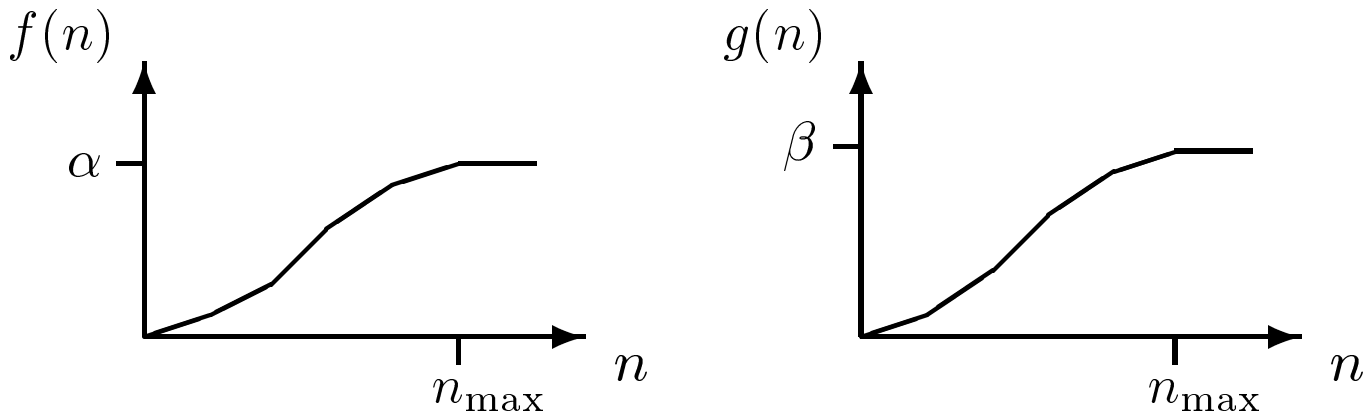
Allow n_k to depend on data at analysis $k-1$ (Schmitz, Springer-Verlag, 1993) — as in adaptive tests.

Efficient tests

Optimal tests or families of efficient tests can be found within these frameworks (Barber & Jennison, *Bmka*, 2002).

Error spending tests

One-sided error spending tests: Functions $f(n)$ and $g(n)$ specify Type I and Type II error to spend when n observations have been observed.



At analysis k with cumulative sample size n_k, set boundaries so that

$$P_{\theta=0}\{\text{Reject } H_0 \text{ by analysis } k\} = f(n_k),$$

$$P_{\theta=\delta}\{\text{Accept } H_0 \text{ by analysis } k\} = g(n_k).$$

Power family of error spending tests

Take

$$f(n) = \begin{cases} \alpha \left(\frac{n}{n_{\max}}\right)^{\rho} & n < n_{\max} \\ \alpha & n \geq n_{\max} \end{cases}$$

$$g(n) = \begin{cases} \beta \left(\frac{n}{n_{\max}}\right)^{\rho} & n < n_{\max} \\ \beta & n \geq n_{\max} \end{cases}$$

Choose $n_{\max}$ so that boundaries meet up at $n = n_{\max}$ for, say, K equally sized groups.

Setting $\rho = 1$ gives a boundary similar to a Pocock test,

$\rho = 3$ approximates an O'Brien & Fleming test.

Figure 4. Five group, one-sided error spending test with $\rho = 1$. Type I error rate is 0.025 and power 0.9 is attained at $\theta = 0.33\,\delta$.

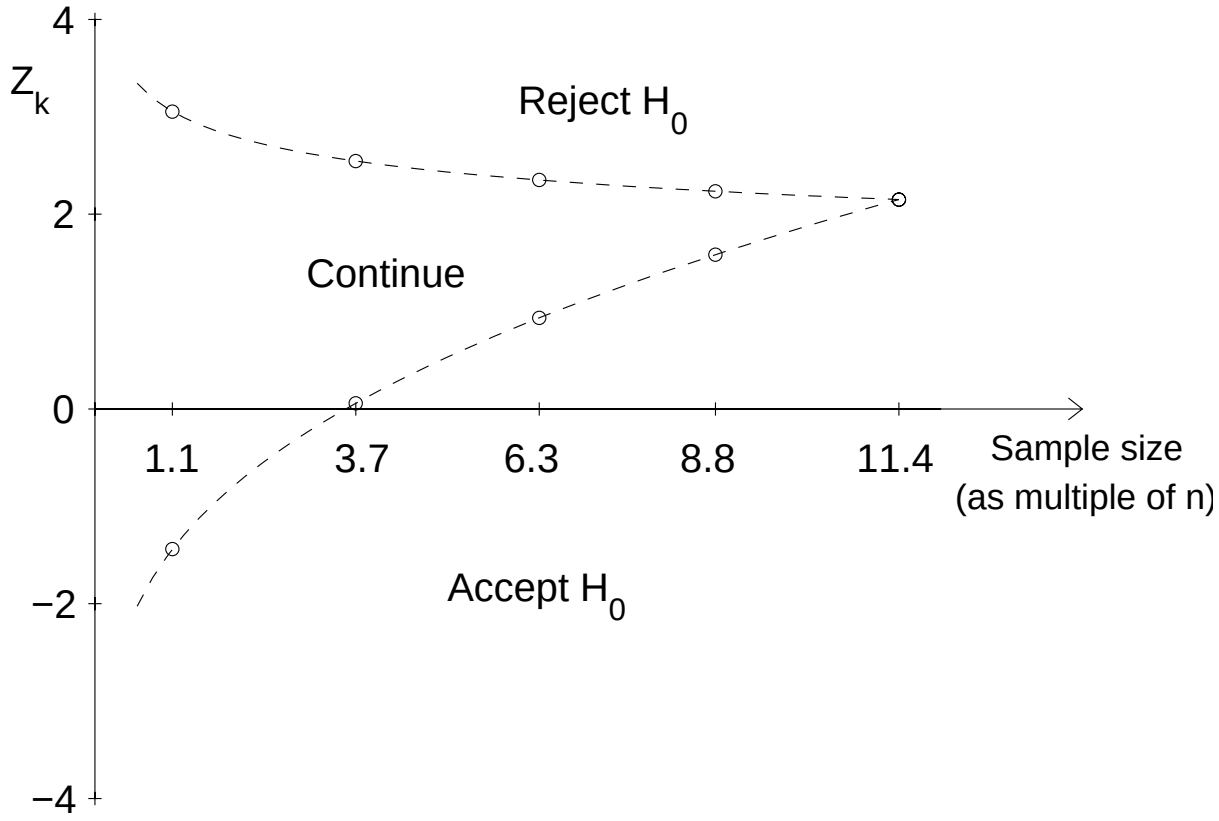


Figure 5. Power functions of Variance Spending test and 5 Group test with power 0.9 at $\theta = 0.33\,\delta$.

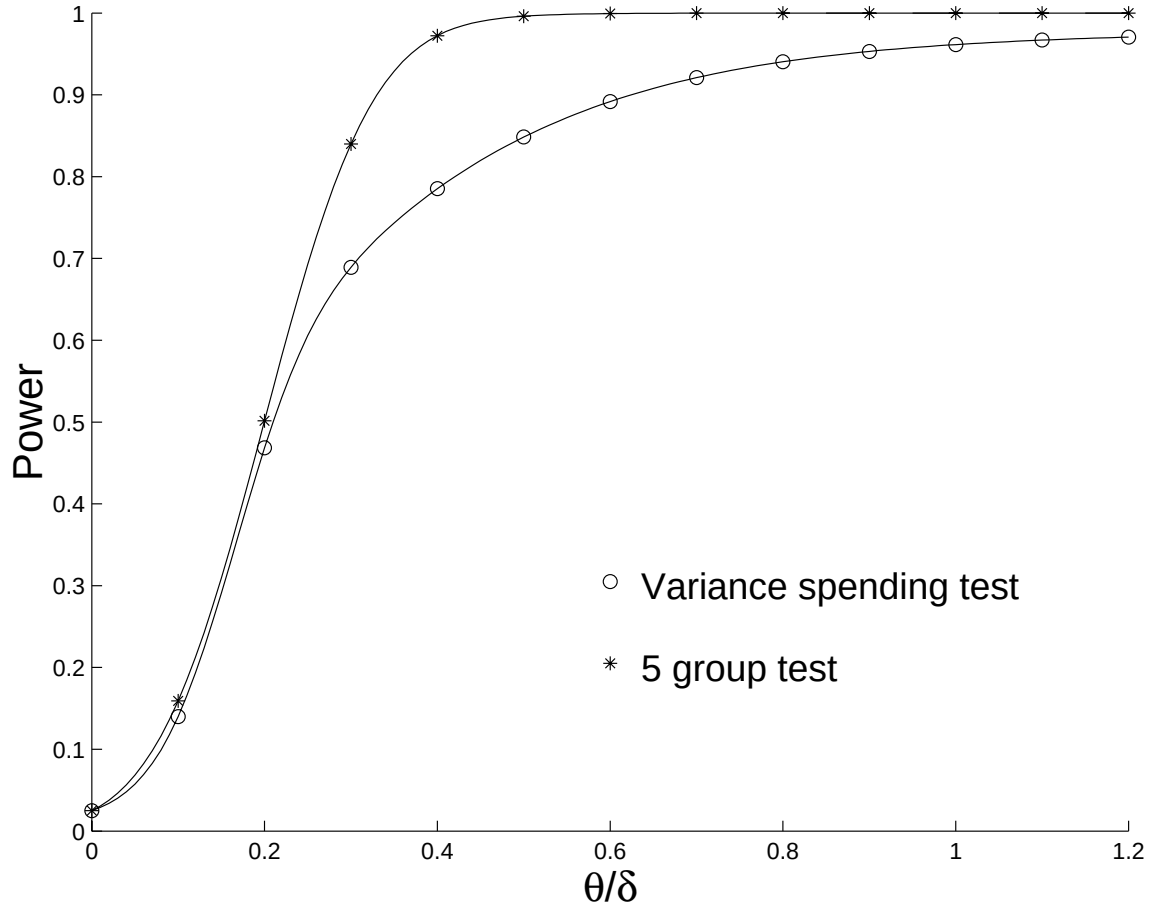
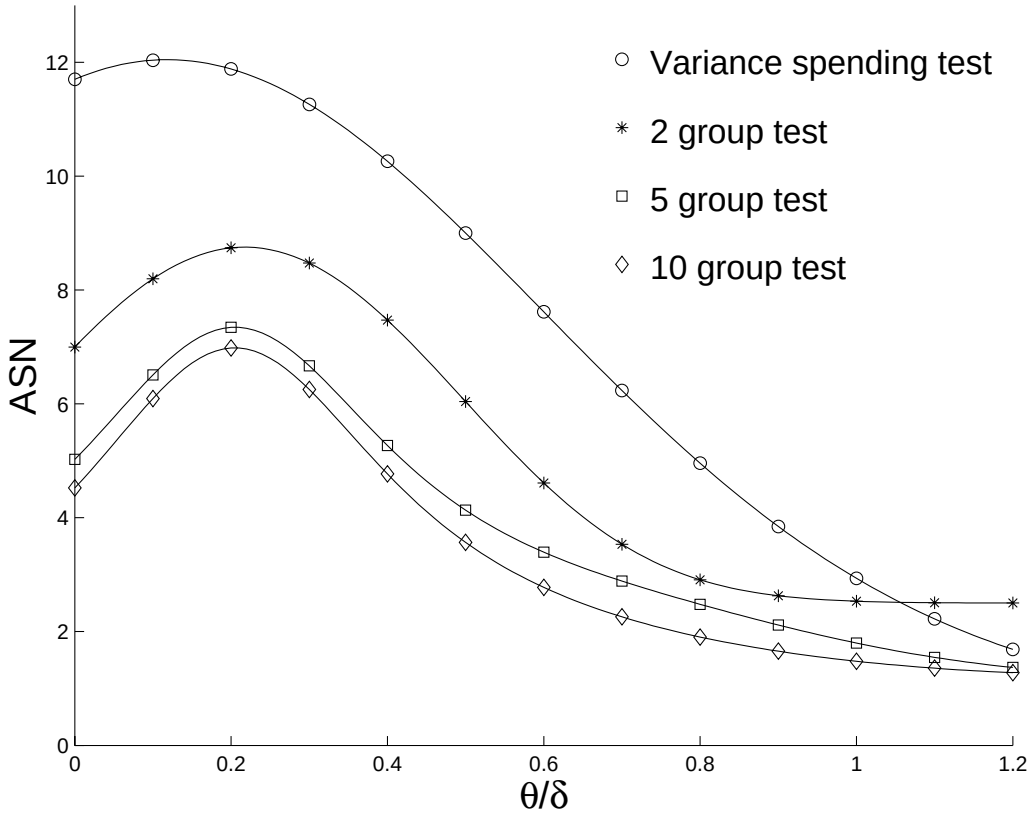


Figure 6. ASN curves of Variance Spending test and 2, 5 and 10 Group tests with power 0.9 at $\theta = 0.33\,\delta$.

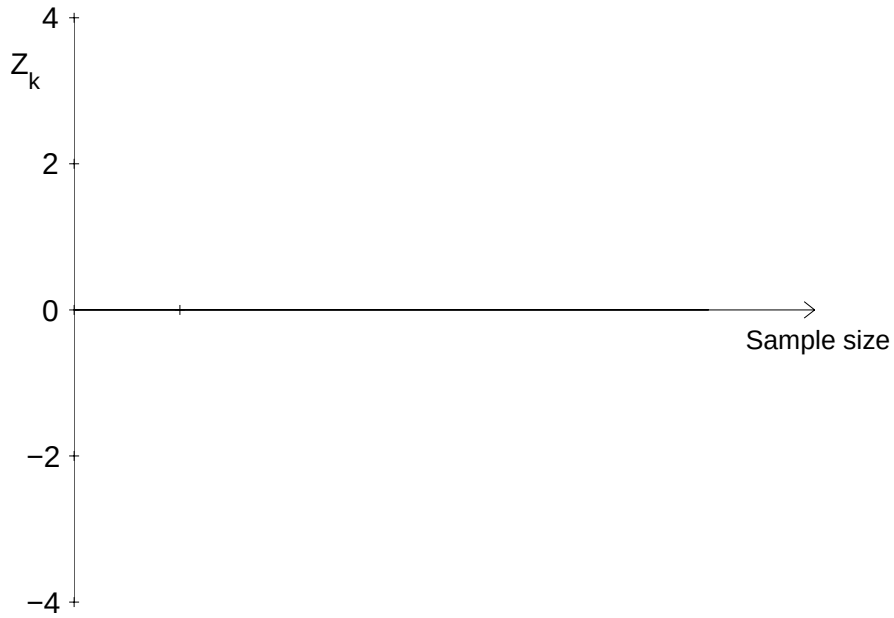


ASN scale is in multiples of the original fixed sample size, n.

§6 Adaptive group sequential tests

Adaptivity provides extra freedom in a group sequential design.

Can this flexibility bring a significant gain in efficiency?



Example

To test H_0: $\theta = 0$ versus H_1: $\theta > 0$

with Type I error rate $\alpha = 0.025$

and power $1 - \beta = 0.8$ at $\theta = \delta$.

Aim for low values of:

$$\frac{1}{3}\{E_{\theta=0}(N) + E_{\theta=\delta}(N) + E_{\theta=2\delta}(N)\}.$$

Constraints:

Maximum sample size $= 1.2 \times$ fixed sample size.

Maximum number of analyses $= K$.

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*
2	70.7
3	59.8
4	55.8
6	52.6
8	51.1
10	50.3

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*	*Optimal adaptive group sequential design*
2	70.7	66.1
3	59.8	57.8
4	55.8	54.0
6	52.6	50.8
8	51.1	49.4
10	50.3	48.6

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*	*Non-adaptive, optimised group sizes*	*Optimal adaptive group sequential design*
2	70.7	66.4	66.1
3	59.8	58.5	57.8
4	55.8	55.1	54.0
6	52.6	52.1	50.8
8	51.1	50.7	49.4
10	50.3	49.8	48.6

Conclusions

- It is possible to rescue a study found, at an interim stage, to be lacking in power — but the flexibility to do this has a price.
- Better practice is to

  think through power requirements fully

  specify θ values at which low sample size is most important

  before embarking on a study.
- Standard types of non-adaptive group sequential tests meet these needs effectively and provide easily interpretable results.
- A little planning can save a lot in sample size and credibility!