Flexible Sample Size: Still no Free Lunch!

Chris Jennison,

Department of Mathematical Sciences, University of Bath, UK

and

Bruce Turnbull,

Department of Statistical Science, Cornell University, Ithaca, NY

Joint Statistical Meetings,
New York, August 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 1:* A hypothesis test with a single, final analysis.

4. Formulating the real testing problem.

5. A catalogue of group sequential tests.

6. *Example 2:* A group sequential test with adaptive re-design.

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 $2\sigma^2 = 1$.)

MLE, Z and score statistics

For this test:

$$\begin{aligned}
\widehat{\theta} &= \overline{X}_A - \overline{X}_B = \overline{Y} && \sim N(\theta,\ n^{-1}) \\
Z &= \widehat{\theta}\sqrt{n} && \sim N(\theta\sqrt{n},\ 1) \\
S &= \widehat{\theta} n = \sum Y_i && \sim N(\theta n,\ n)
\end{aligned}$$

Working with information rather than sample size, we can generalise to

- other designs (e.g. crossover, general linear model)
- other endpoints (e.g. binary data, survival data).

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).

Should we worry about inflation of Type I error?

Pocock:

> “Control of Type I error is a vital aid to prevent a flood of false positives into the medical literature.”

Why not just start over?

Perhaps we should just throw away the data and start again with a new, larger trial.

This is inefficient and wasteful of data.

This procedure would also inflate the Type I error rate. If repeated, it leads to a Type I error rate of almost one! ("sampling to a foregone conclusion", Cornfield, *JASA,* 1966.)

“Flexible/adaptive” procedures

Bauer and Köhne (1994). *Biometrics.*

Proschan and Hunsberger (1995). *Biometrics.*

Wassmer (1998). *Biometrics.*

Lehmacher and Wassmer (1999). *Biometrics.*

Fisher, Lloyd (1998). Self-designing clinical trials. *Statist. in Med.*

Cui, Hung and Wang (1999). *Biometrics.*

Chi and Liu (1999). *J. Biopharm. Statist.*

Müller and Schäfer (2001). *Biometrics.*

Denne (2001). *Statist. in Med.*

Jennison and Turnbull (2002). *Submitted.*

§2 Variance spending

A fixed sample of n observations can be divided into

$$\begin{aligned} \text{stage 1:} \qquad S_1 &= \textstyle\sum_{i=1}^{rn}(X_{Ai} - X_{Bi}), \\ \text{stage 2:} \qquad S_2 &= \textstyle\sum_{i=rn+1}^{n}(X_{Ai} - X_{Bi}). \end{aligned}$$

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

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

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 1

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

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

Choice of γ

Treating γ as fixed (!) we obtain

$$E(Z) \;=\; \{r + \sqrt{\gamma}(1-r)\}\sqrt{n}\,\theta.$$

A test designed for power $1-\beta$ at δ/ξ has sample size $\xi^2 n$ and statistic

$$Z' \;\sim\; N(\xi\sqrt{n}\,\theta,\, 1).$$

Equating $E(Z)$ and $E(Z')$ gives

$$\xi = r + \sqrt{\gamma}(1-r) \quad \text{or} \quad \gamma = \left(\frac{\xi - r}{1-r}\right)^2 \qquad (2)$$

to determine our modified sample size.

Sample size rule, with truncation

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

$$\tilde{\xi} = \tilde{\xi}(\widehat{\theta}_1) = \begin{cases} 4 & \text{for} \quad \widehat{\theta}_1 \leq \delta/4, \\ \delta/\widehat{\theta}_1 & \quad\quad \delta/4 < \widehat{\theta}_1 < 2\delta, \\ 0.5 & \quad\quad \widehat{\theta}_1 \geq 2\delta. \end{cases} \tag{3}$$

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, from (2), is

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

Properties of the test

Power

$$P_\theta\{\text{Reject } H_0\} = P_\theta\{Z > z_\alpha\} = \int P_\theta\{Z > z_\alpha | \widehat{\theta}_1\} f_\theta(\widehat{\theta}_1)\, d\widehat{\theta}_1$$

where $f_\theta(\widehat{\theta}_1)$ is the $N(\theta,\, 1/(rn))$ density of $\widehat{\theta}_1$ and

$$Z \;=\; \{S_1 + \gamma(\widehat{\theta}_1)^{-1/2}\, S_2'\}/\sqrt{n}.$$

Average Sample Number

$$ASN(\theta) = rn + (1-r)n \int \gamma(\widehat{\theta}_1) f_\theta(\widehat{\theta}_1)\, d\widehat{\theta}_1.$$

Example

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 $\theta = \delta$.

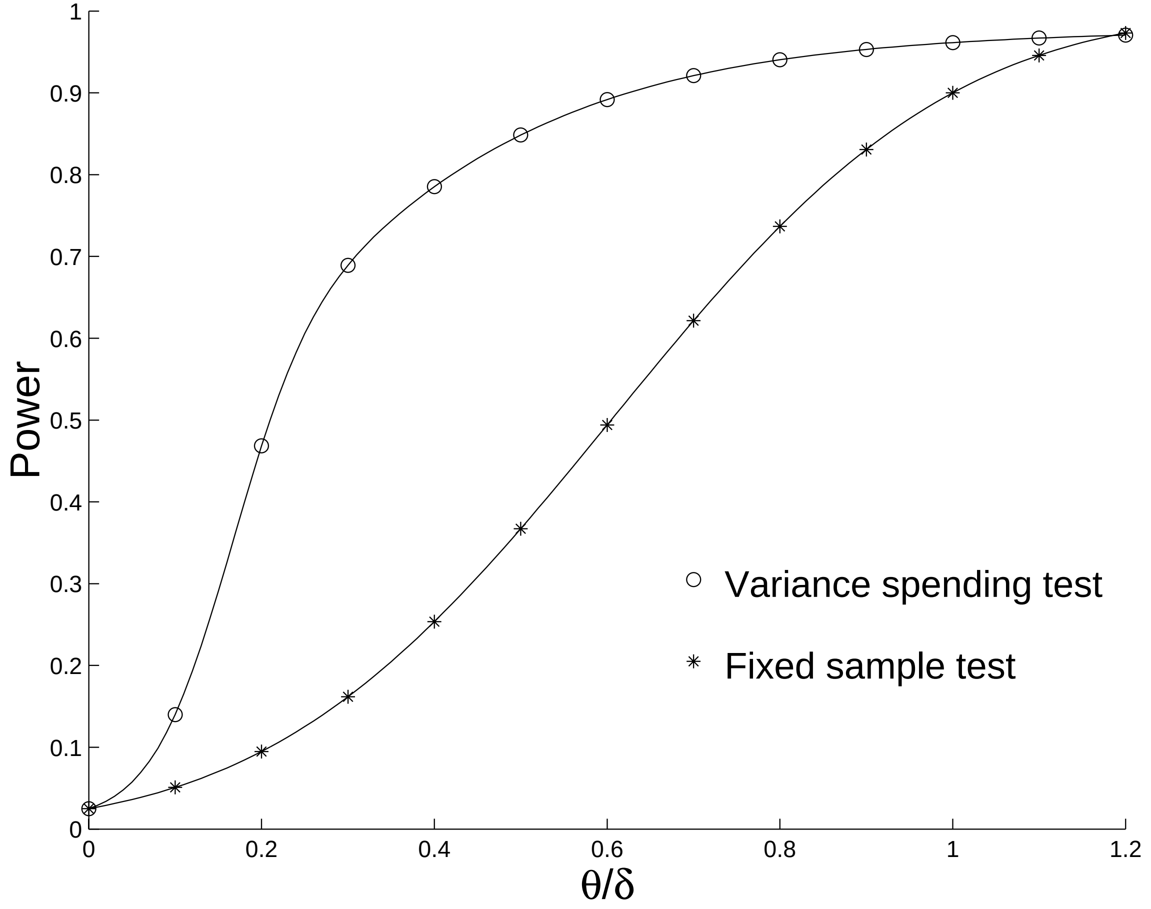


Figure 2. Power functions of Variance Spending test and Fixed Sample test with power 0.9 at $\theta = 0.6\,\delta$.

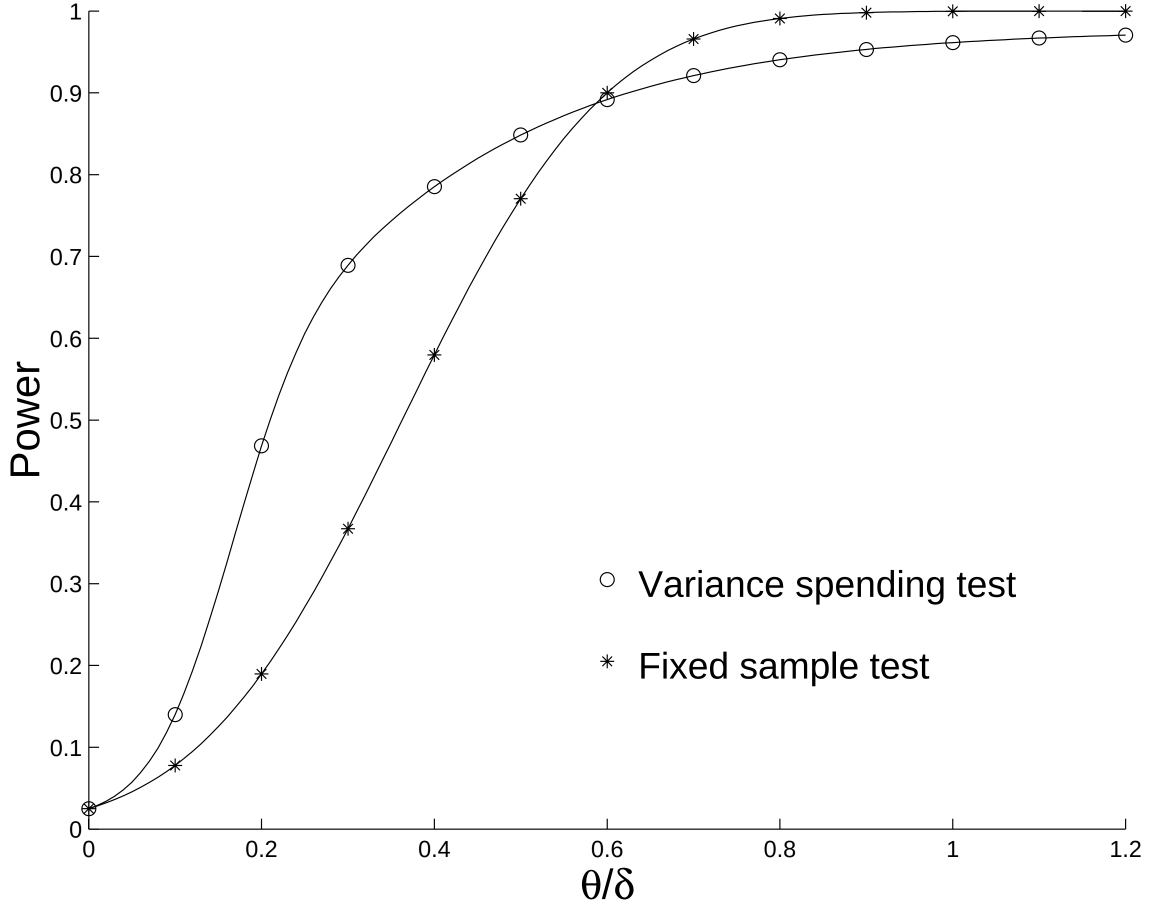
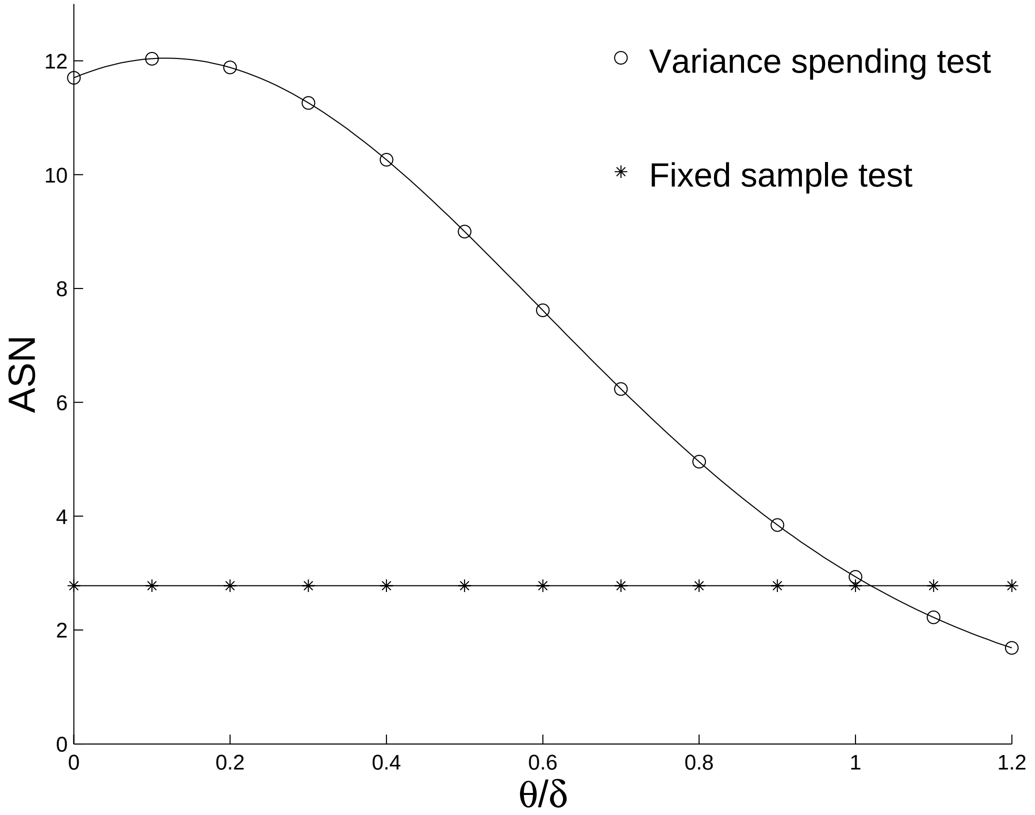


Figure 3. ASN curves of Variance Spending test and Fixed Sample test with power 0.9 at $\theta = 0.6\,\delta$.



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

Inefficiency: Use of a non-sufficient statistic

Total sample size is $N = rn + \gamma(1-r)n$.

Ignoring randomness in γ, the final statistic has distribution

$$\{S_1 + \gamma^{-1/2}\,S_2'\}/\sqrt{n} \;\sim\; N([r + \gamma^{1/2}\{1-r\}]\sqrt{n}\theta, 1)$$

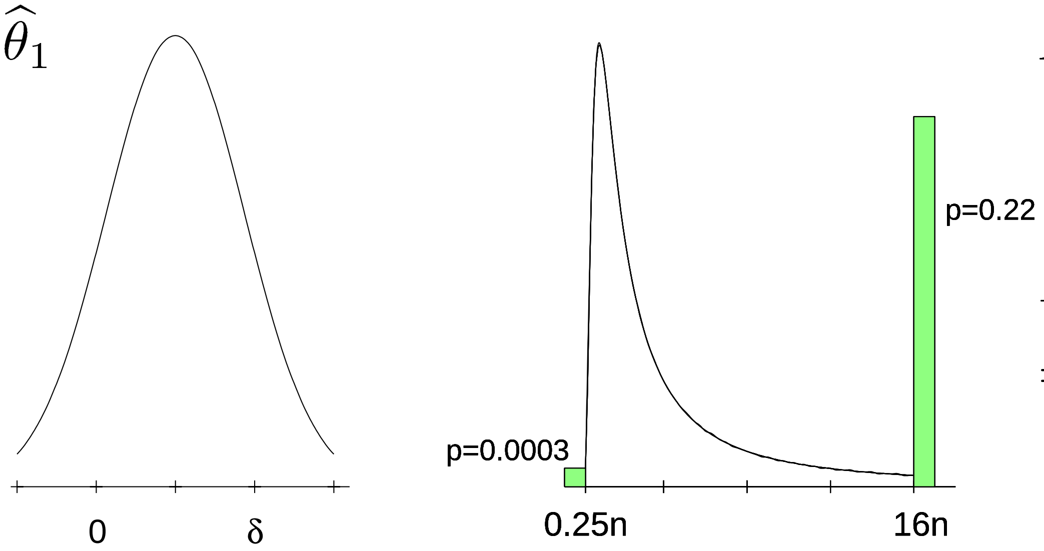
so the effective sample size is $N_{\textit{eff}} = (r + \gamma^{1/2}\{1-r\})^2 n$.

For $r = 1/2$, the “inefficiency” $N/N_{\textit{eff}}$ is:

γ	0	0.5	1	2	4	10	49	∞
Inefficiency	2	1.03	1	1.03	1.11	1.27	1.56	2

Inefficiency: Variable sample size, based on noisy $\widehat{\theta}_1$

For $\theta = 0.5\delta$



$N_{\textit{eff}}$

$E(N_{\textit{eff}})$
$= 6.17n$

A fixed sample test with $6.17n$ observations would have power 0.98.

The variance spending design gives power 0.85 at $\theta = 0.5\delta$.

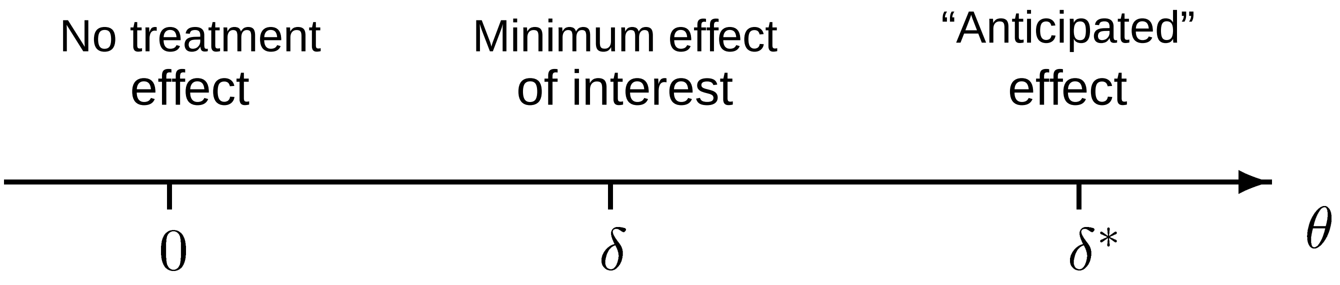
§4 Formulating the testing problem

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.

Group sequential setting

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:

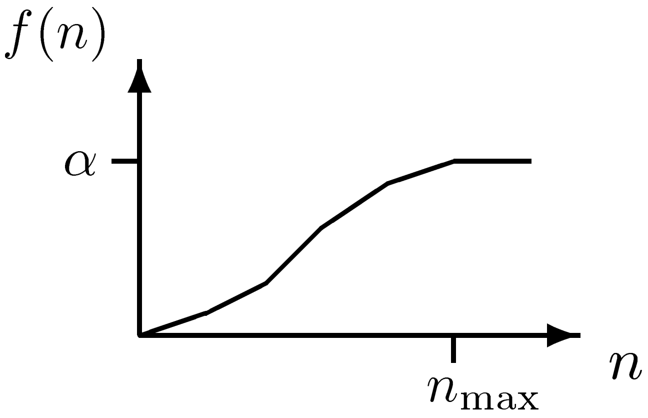
Allow n_k to depend on data at analysis $k-1$ (Schmitz, Springer-Verlag, 1993) — as in adaptive tests.

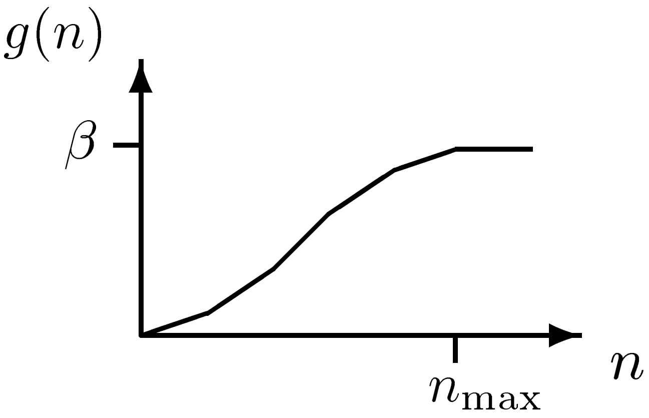
Efficient tests

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

§5 Group sequential 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.

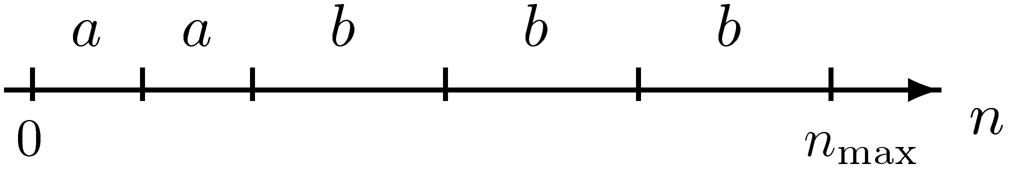
Attaining low ASN under high values of θ

Values $\rho = 1$ or $\rho = 0.75$ spend error at a high rate early on.

Also, a few *very* early analyses are desirable.

1. *Small groups / large groups*

M groups of a observations, followed by $K - M$ groups of size b.



2. *Geometric pattern*

$$n_k = \gamma^{K-k}\, n_{\max} \quad (\gamma < 1)$$

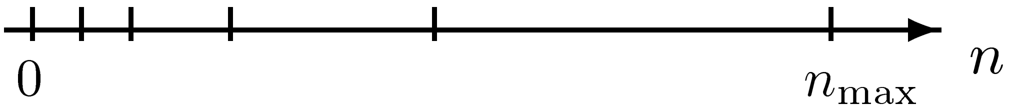


Figure 4. Five group, one-sided error spending test with $\rho = 1$. Type I error rate is 0.025 and power 0.9 is attained at $\theta = 0.33\,\delta$.

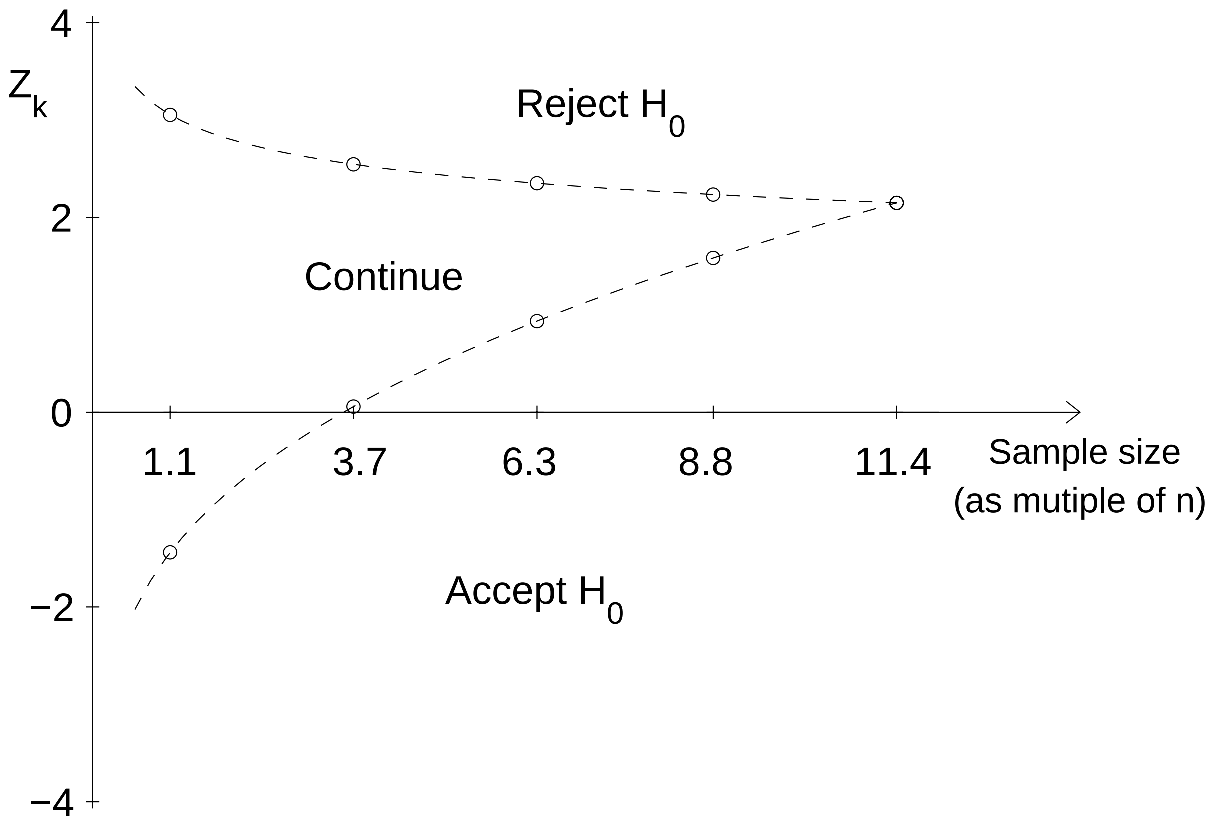


Figure 5. Power functions of Variance Spending test and 5 Group test with power 0.9 at $\theta = 0.33\,\delta$.

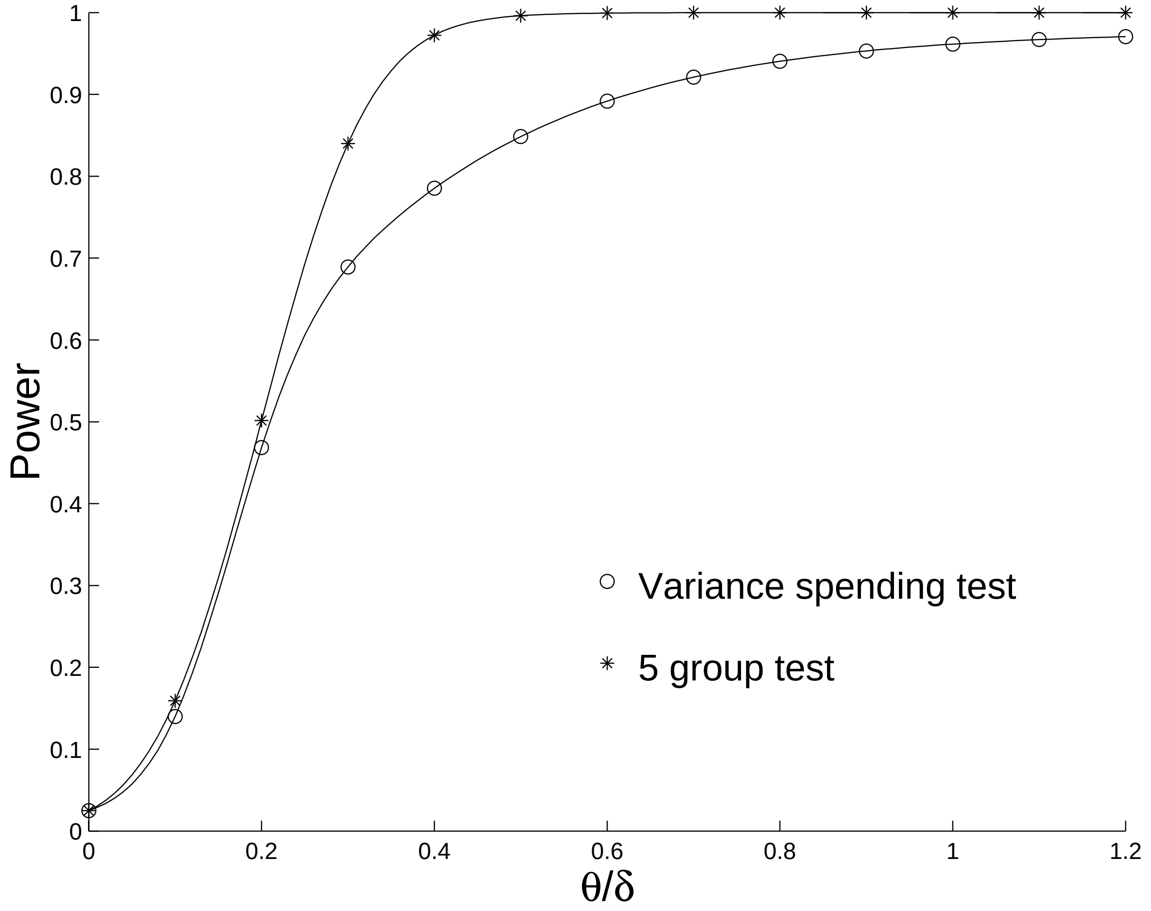
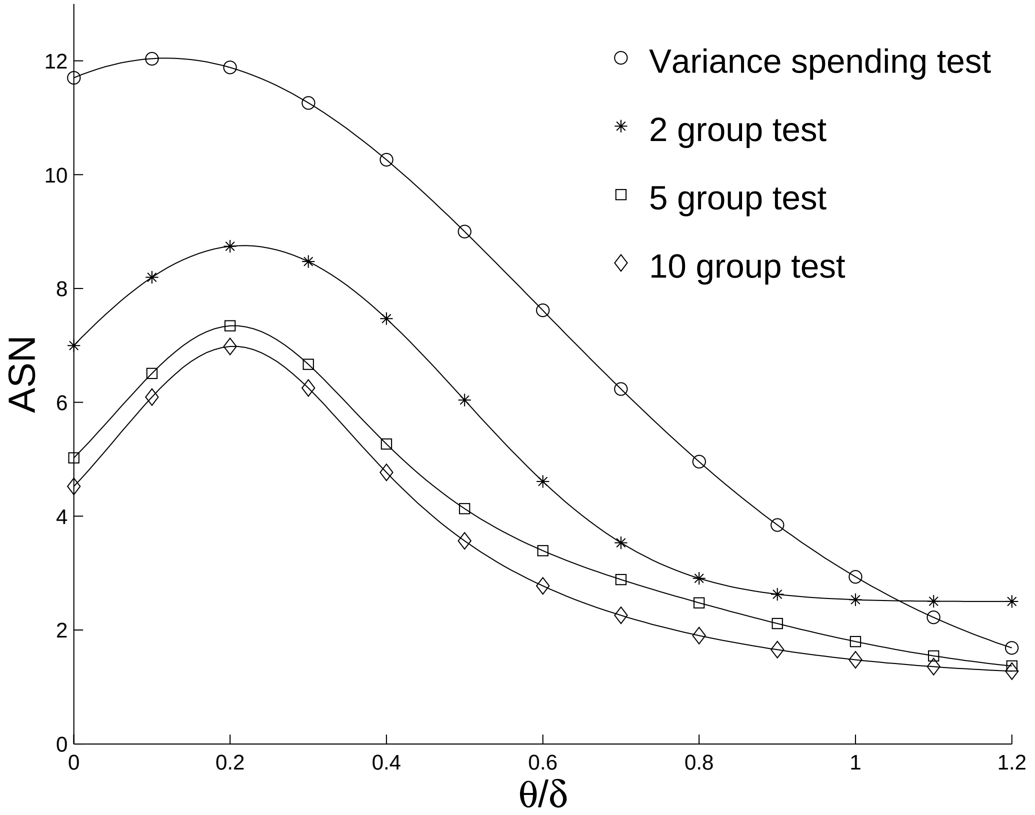


Figure 6. ASN curves of Variance Spending test and 2, 5 and 10 Group tests with power 0.9 at $\theta = 0.33\,\delta$.



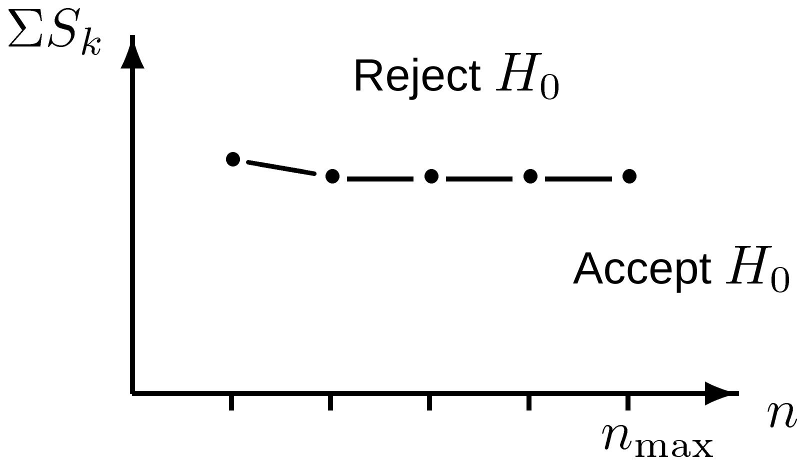
ASN scale is in multiples of the original fixed sample size, n.

§6 Example 2: A Cui, Hung & Wang (1999) design

Original group sequential design:

To test H_0: $\theta = 0$ with Type I error rate 0.025 and power 0.9 at $\theta = \delta$.

Observations taken in 5 groups; early stopping allowed to *reject* H_0.



$n_{\max} = 10.8/\delta^2$, cf fixed sample size, $n = 10.5/\delta^2$.

Design modification

Cui et al suggest adjusting the design at just one interim analysis.

Changing design at stage 3:

Group 4

Original plan: $S_4 =$ sum of $n/5$ terms $(X_{Ai} - X_{Bi})$

Revised plan: $S_4' =$ sum of $\gamma\, n/5$ terms $(X_{Ai} - X_{Bi})$

Use $\gamma^{-1/2}\, S_4'$ in place of S_4, preserving the null distribution.

Group 5 — similarly.

Example

As in Example 1, aim for the effective sample size needed in the original test to attain power 0.9 at $\theta = \widehat{\theta}_1$.

At the 3rd analysis of 5, fraction of the total sample size is $r = 0.6$. Set $\tilde{\xi} = \delta / \widehat{\theta}_1$ truncated to the range $(0.6, 3)$.

Then

$$\gamma(\widehat{\theta}_1) \; = \; \frac{\{\tilde{\xi}(\widehat{\theta}_1) - 0.6\}^2}{(1 - 0.6)^2}.$$

Hence $\gamma \; \in \; (0, 36)$ and total sample size $\in \; (0.6n, \; 15n)$.

Figure 7. Power functions of Cui et al 5 Group Adaptive test and Fixed Sample test with power 0.9 at $\theta = \delta$.

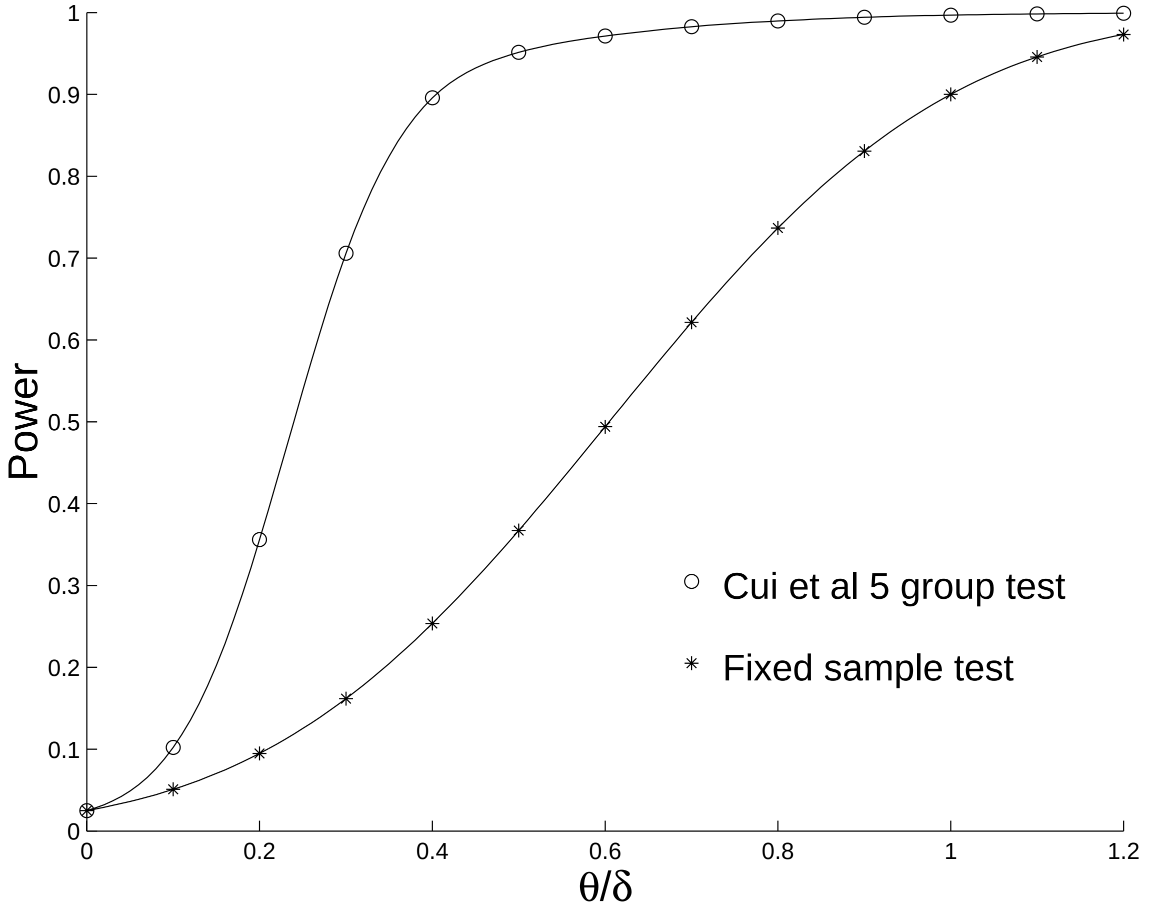
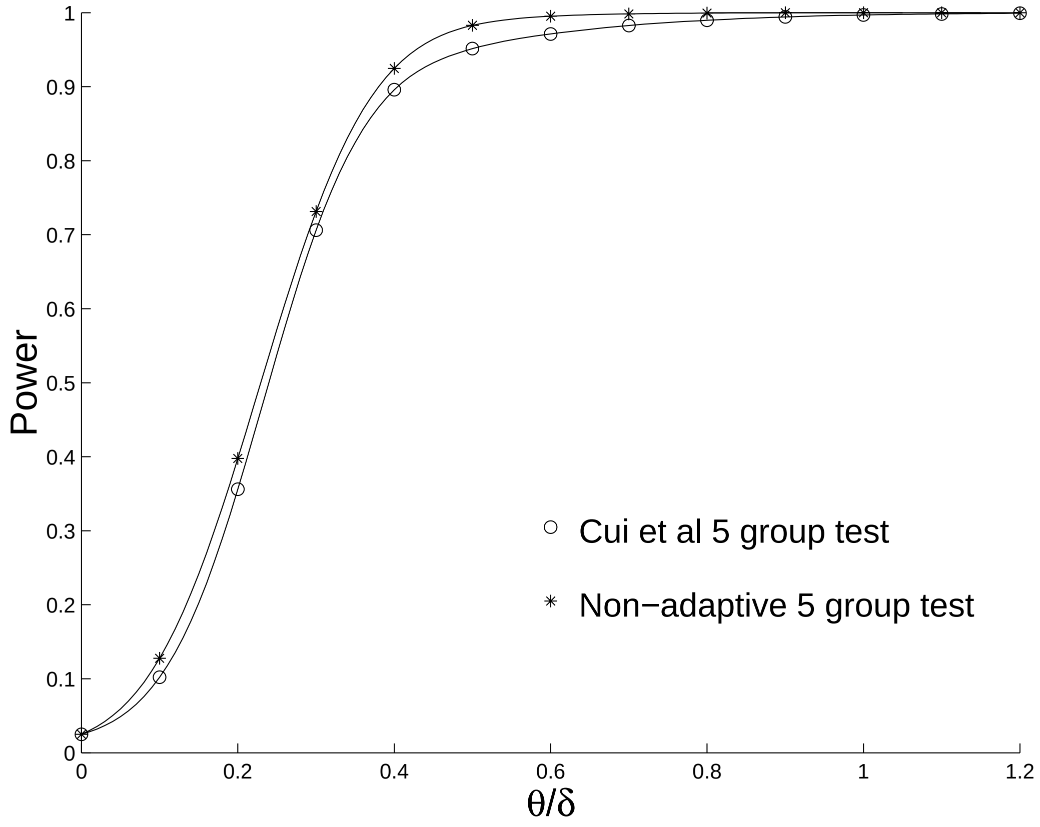
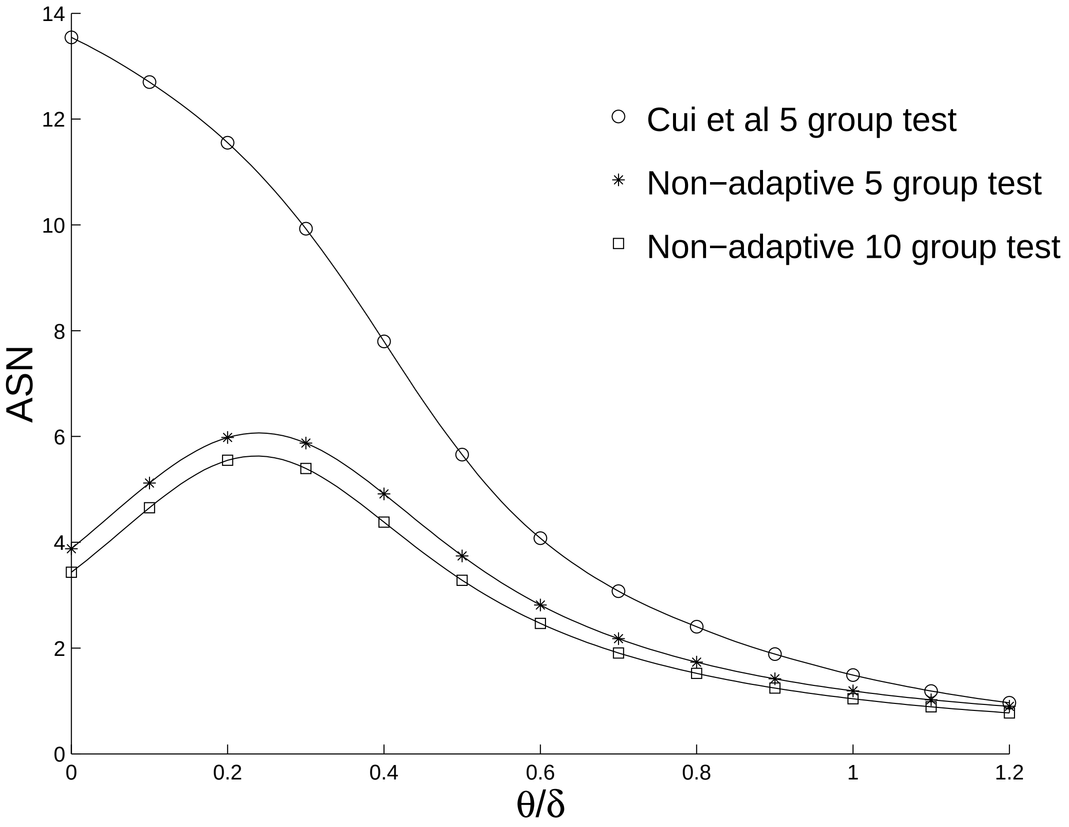


Figure 8. Power functions of Cui et al 5 Group Adaptive test and Non-Adaptive 5 Group test with power 0.9 at $\theta = 0.38\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 0.75$.

Figure 9. ASN curves of Cui et al 5 Group Adaptive test and Non-Adaptive 5 and 10 Group tests with power 0.9 at $\theta = 0.38\,\delta$.



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

Example 3: Shen & Fisher (*Biometrics,* 1999) designs

Testing H_0: $\theta = 0$ with Type I error rate α.

Initially calculate the total sample size N_1 giving power $1 - \beta$ at $\theta = \delta$.

Collect observations in blocks of pre-specified size,

e.g., $B_1 = N_1/2$, $B_2 = B_3 = \ldots = N_1/6$.

Data in block j provide $Z_j \sim N(0, 1)$ under H_0.

Allocate block j a weight w_j, dependent on data in blocks $1, \ldots, j-1$.

When $\sum_1^m w_j^2 = 1$, the sum $\sum_1^m w_j Z_j \sim N(0, 1)$ under H_0,

so reject H_0 if $\sum_1^m w_j Z_j \geq z_\alpha$.

Shen & Fisher designs

Weights and Stopping Rule

Before sampling block j, compute target additional sample size N_j

if $B_j \geq N_j$, make this the last block, setting

$$w_j^2 \;=\; 1 - \sum_{i=1}^{j-1} w_i^2,$$

otherwise, set (say)

$$w_j^2 \;=\; \frac{B_j}{N_j} \left(1 - \sum_{i=1}^{j-1} w_i^2 \right).$$

Shen & Fisher designs

Stopping to accept H_0

Stop for "futility" after block j if $\widehat{\theta}_j$ is low.

Version (1): compare $\widehat{\theta}_j$ with δ.

Version (2): compare $\widehat{\theta}_j$ with $\tilde{\delta}$ ($\tilde{\delta} < \delta$).

Figure 10. Power functions of Shen & Fisher Adaptive test (1) and Fixed Sample test with power 0.9 at $\theta = \delta$.

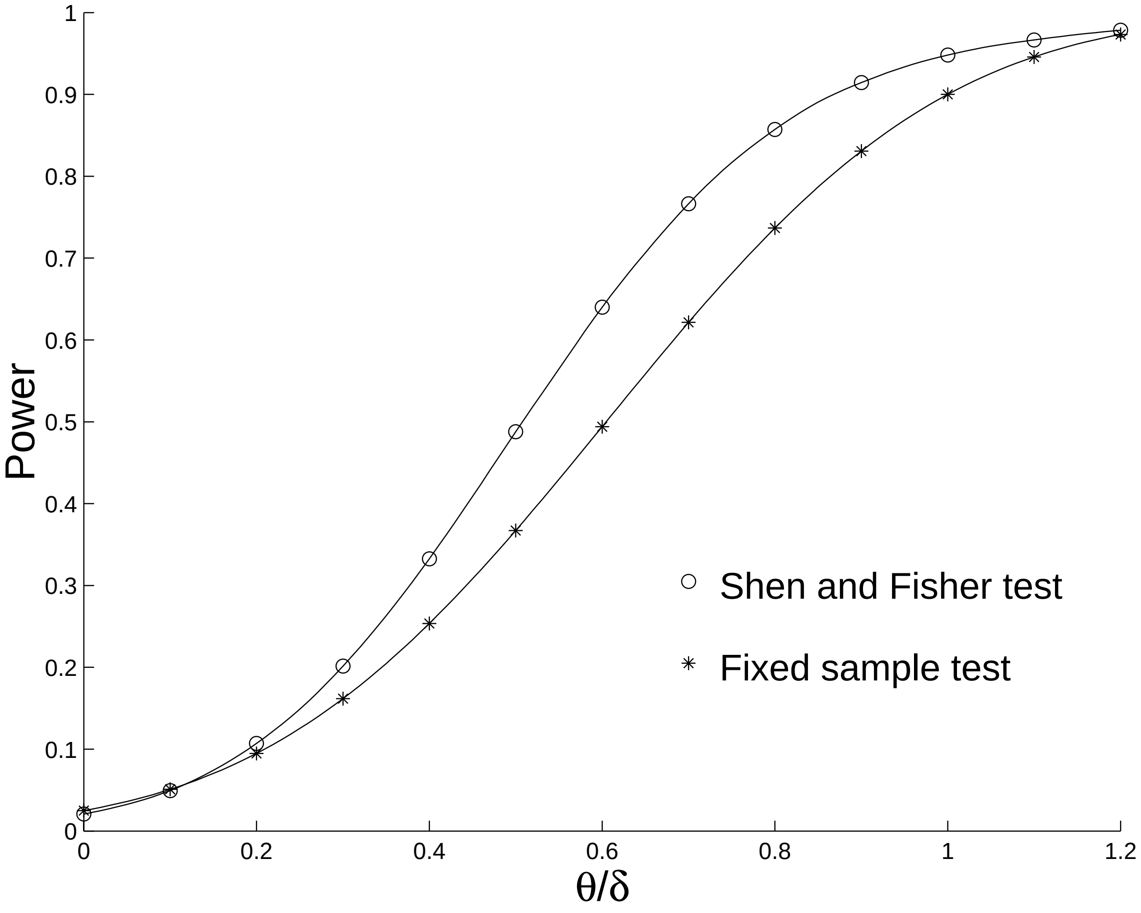
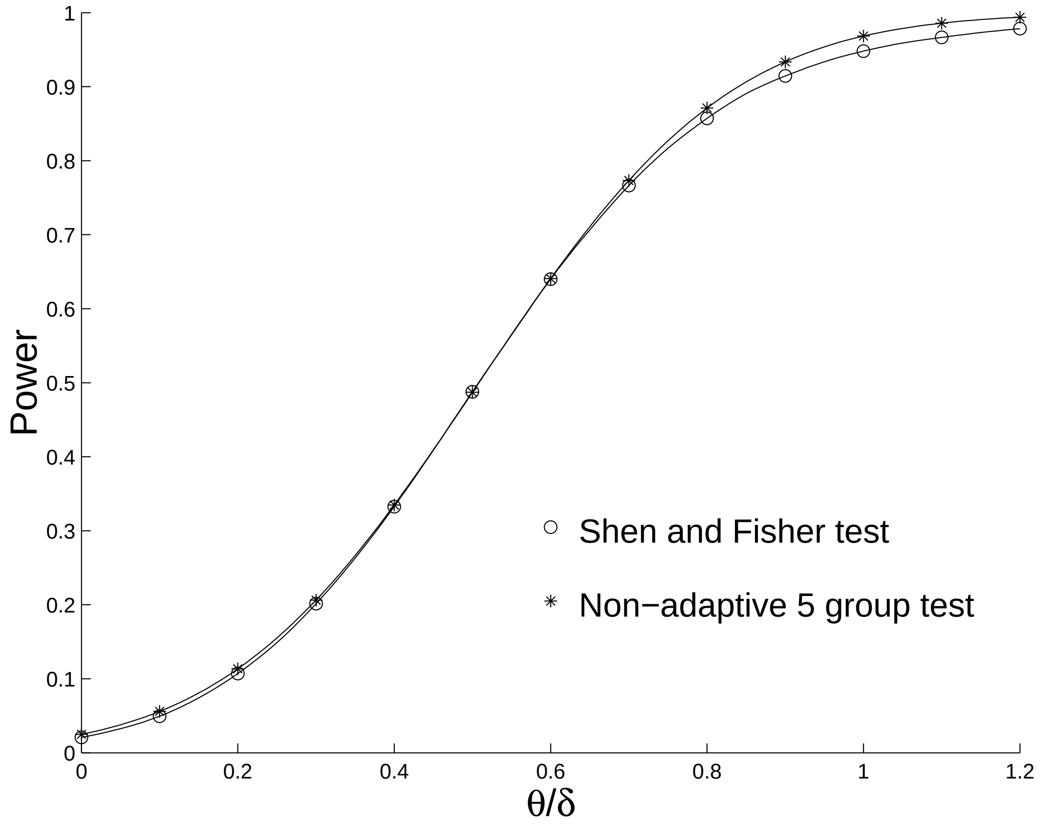
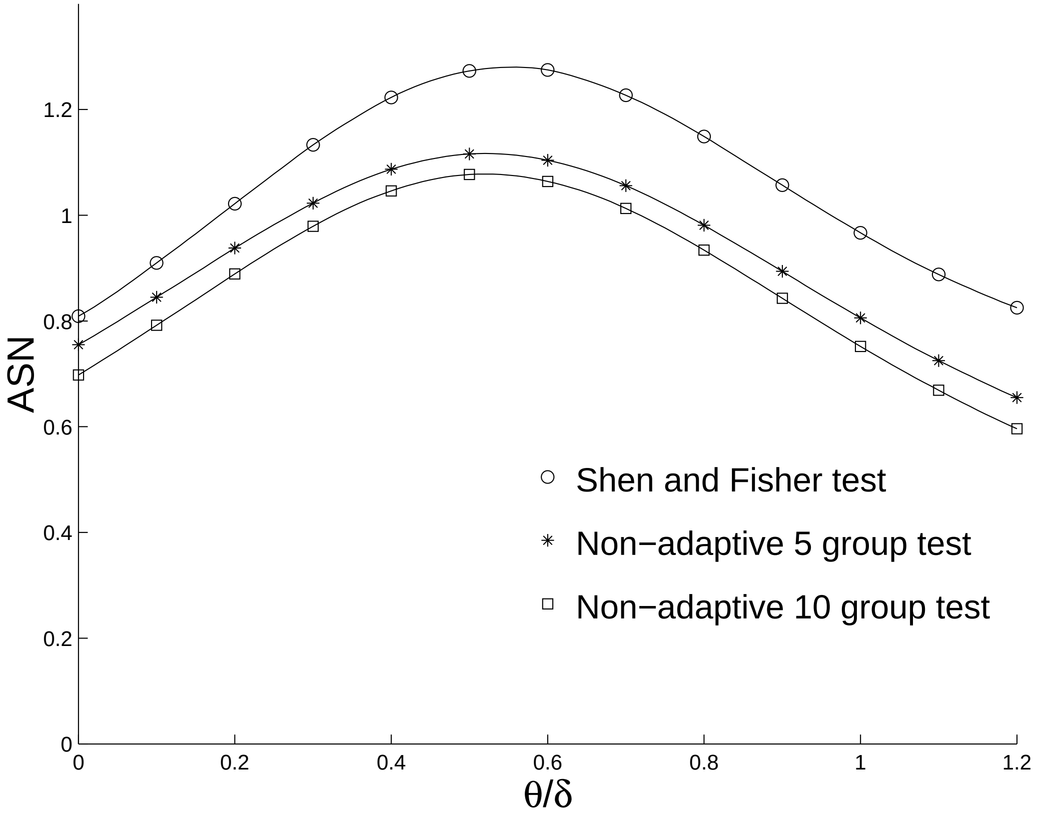


Figure 11. Power functions of Shen & Fisher Adaptive test (1) and Non-Adaptive 5 Group test with power 0.9 at $\theta = 0.84\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 1$.

Figure 12. ASN curves of Shen & Fisher Adaptive test (1) and Non-Adaptive 5 and 10 Group tests with power 0.9 at $\theta = 0.84\,\delta$.



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

Figure 13. Power functions of Shen & Fisher Adaptive test (2) and Fixed Sample test with power 0.9 at $\theta = \delta$.

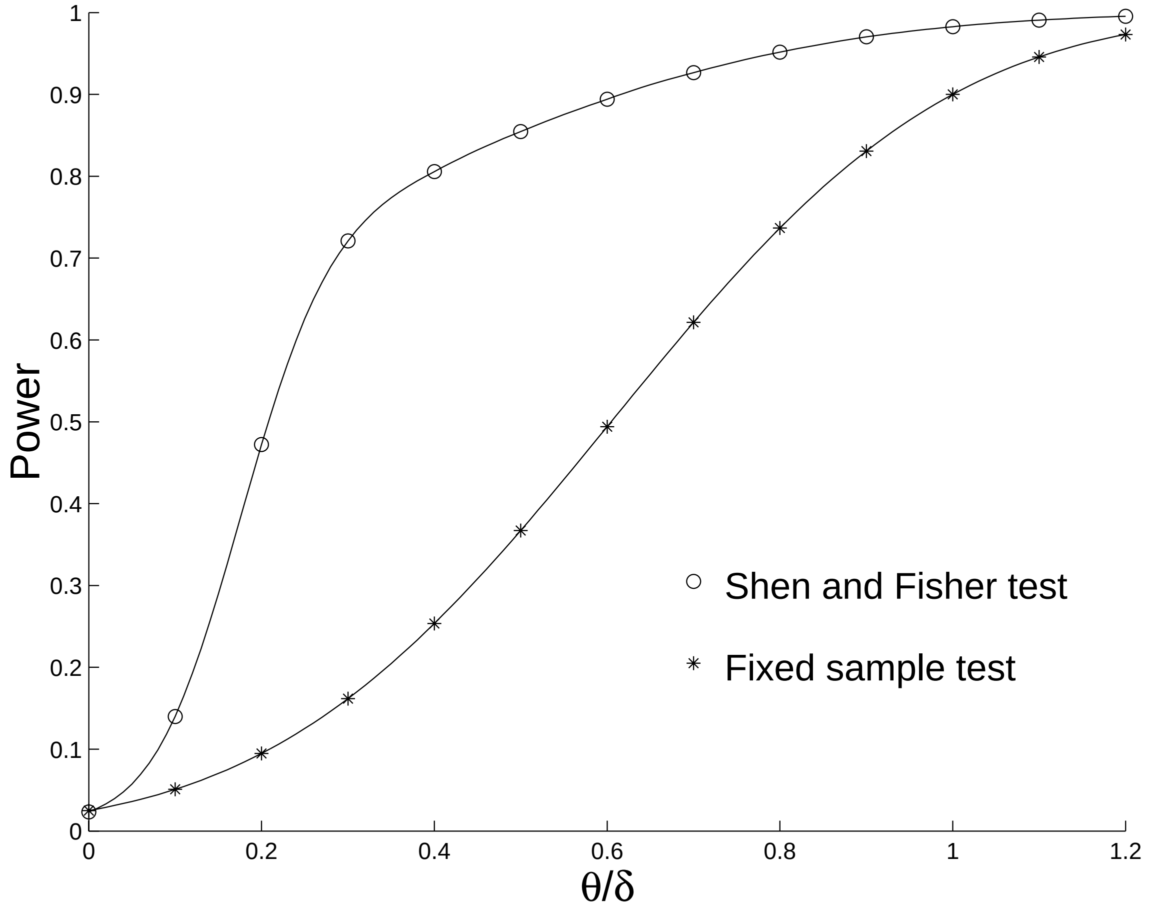
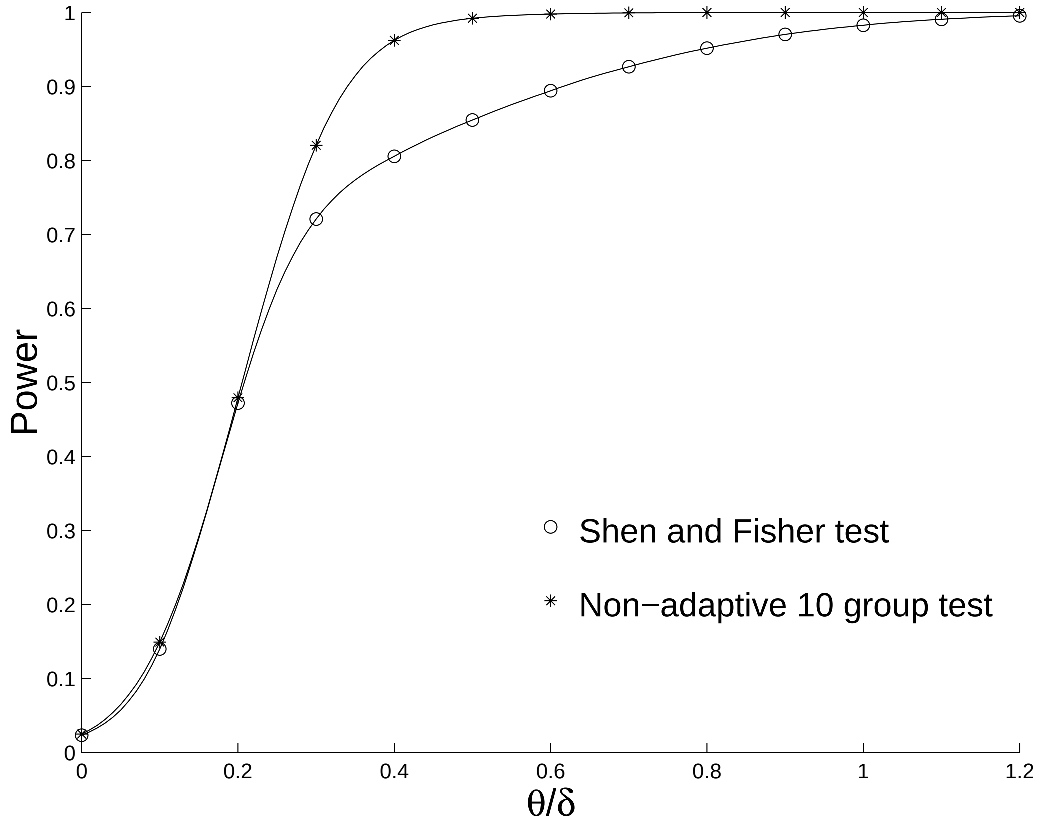
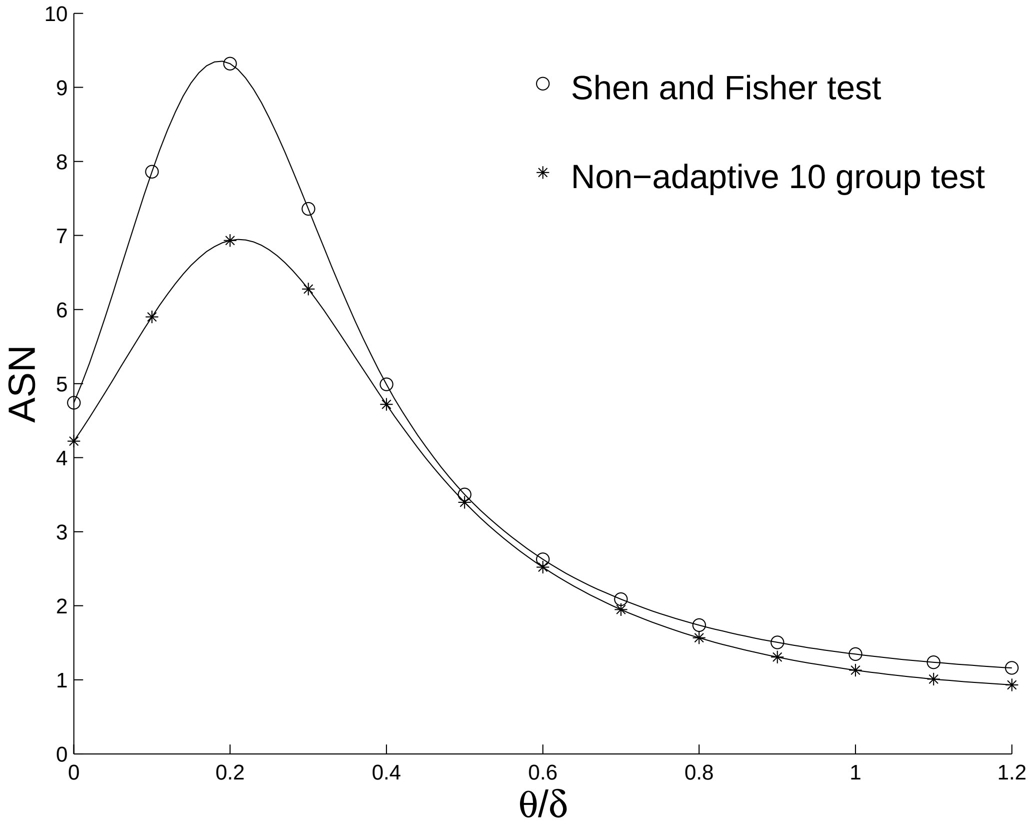


Figure 14. Power functions of Shen & Fisher Adaptive test (2) and Non-Adaptive 10 Group test with power 0.9 at $\theta = 0.34\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 0.75$.

Figure 15. ASN curves of Shen & Fisher Adaptive test (2) and Non-Adaptive 10 Group test with power 0.9 at $\theta = 0.34\,\delta$.

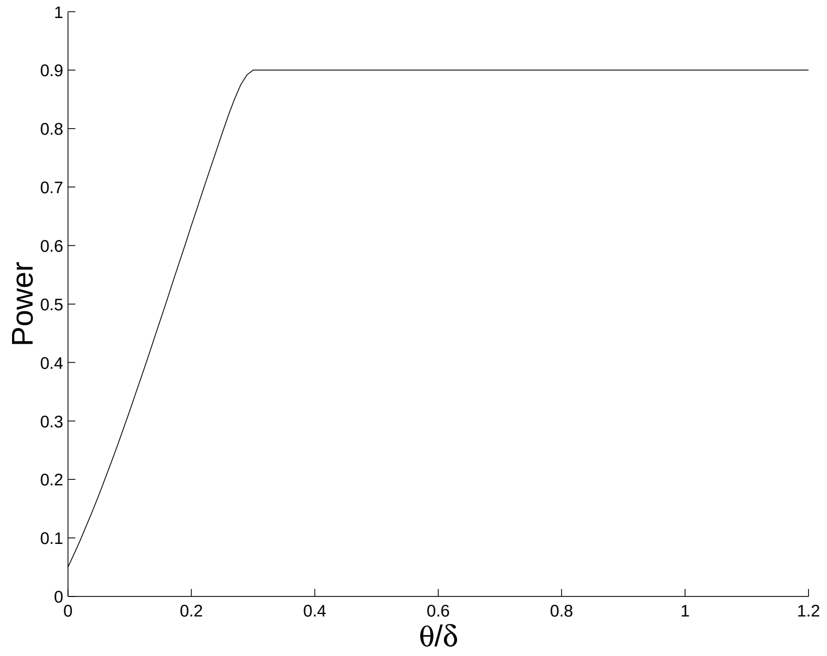


ASN scale is in multiples of the original fixed sample size, n.

Setting power: 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 (!)



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!