Sample size re-estimation:

Internal pilots and

information monitoring

Christopher Jennison

Dept of Mathematical Sciences,

University of Bath, UK

PSI Conference,

Tortworth Court,

May, 2006

Plan of talk

1. Internal pilots
2. Error-spending group sequential tests
3. Information monitoring
4. Mehta & Tsiatis's group sequential t-tests
5. Conclusions

1. Internal pilots in studies with a single analysis

The sample size needed to satisfy a power requirement often depends on an unknown nuisance parameter.

Examples include:

Normal response: Unknown variance, σ^2.

Binary response: Since the variance depends on p, the sample size needed to detect a specific difference in probabilities $p_1 - p_2 = \delta$ depends on $(p_1 + p_2)/2$.

Survival data: Information is governed by the number of observed deaths, and this depends on the overall failure rate and degree of censoring.

"Over-interpretation of results from a small pilot study, positive or negative, may undermine support for the major investigation" (W. G. Cochran).

Internal pilots: Wittes & Brittain

Wittes & Brittain (*Statistics in Medicine*, 1990) suggest an "internal" pilot.

> Let ϕ denote a nuisance parameter and suppose the sample size required under a given value of this parameter is $n(\phi)$.
>
> From a pre-study estimate, $\hat{\phi}_0$, calculate an initial planned sample size of $n(\hat{\phi}_0)$.
>
> At an interim stage, find a new estimate $\hat{\phi}_1$ from the data obtained so far. Aim for the new target sample size of $n(\hat{\phi}_1)$.

Variations on this are possible, e.g., only allow an increase over the original target sample size.

Internal pilots: properties

Wittes and Brittain's method has a complicated effect on the estimate of variance in the final test statistic.

In general, variance estimates are biased downwards, but results in Jennison & Turnbull (2000, Ch. 14) show the type I error rate is only slightly perturbed.

Binary responses

Two-treatment comparison, H_0: $p_A = p_B$, $\alpha = 0.05$.
Internal pilots are used to achieve power at alternatives $p_B = p_A + \Delta$ for fixed Δ or $p_B = p_A/\rho$ for fixed ρ.

Pilot sample size per treatment, n_0	*Type I error probability*
10	0.057 – 0.059
20	0.051 – 0.061
30	0.049 – 0.057
40	0.051 – 0.053
50	0.049 – 0.053

Internal pilots: properties

Normal data, estimating σ^2

Two-treatment comparison, H_0: $\mu_A = \mu_B$, $\alpha = 0.05$.

Internal pilots are used to achieve power at the alternative $\mu_B - \mu_A = \pm\delta$ for fixed δ.

Degrees of freedom for estimate s_1^2	*Type I error probability*
8	0.052 – 0.065
18	0.050 – 0.057
38	0.052 – 0.053
78	0.051

Blinding: Finding s^2 may reveal the estimated effect, $\widehat{\theta}$.

This is undesirable as it breaks the blinding at what is meant to be an administrative analysis,

adjusting the sample size in the knowledge of $\widehat{\theta}$ can seriously inflate type I error rates.

Blinded variance estimation

Suppose the two treatments A and B have responses $X_{Ai} \sim N(\mu_A, \sigma^2)$ and $X_{Bi} \sim N(\mu_B, \sigma^2)$.

With n observations per treatment, we would usually estimate σ^2 by

$$s^2 = \frac{\sum (X_{Ai} - \bar{X}_A)^2 + \sum (X_{Bi} - \bar{X}_B)^2}{2n - 2},$$

but this requires knowledge of the treatment labels.

However, an estimate based on the Sum of Squares for the pooled data,

$$\begin{aligned} S_P^2 &= \sum (X_{Ai} - \bar{X})^2 + \sum (X_{Bi} - \bar{X})^2 \\ &= (2n - 2)s^2 + \frac{n}{2} (\bar{X}_A - \bar{X}_B)^2, \end{aligned}$$

would not reveal the treatment labels.

Blinded variance estimation

Write the pooled sum of squares as

$$S_P^2 \;=\; (2n-2)s^2 + \frac{n}{2}\,(\bar{X}_A - \bar{X}_B)^2.$$

The first term on the RHS involves the estimate s^2 of σ^2 from unblinded data:

$$(2n-2)s^2 \;\sim\; \sigma^2\,\chi^2_{2n-2}.$$

The second term has a non-central χ^2 distribution

$$(n/2)\,(\bar{X}_A - \bar{X}_B)^2 \;\sim\; \sigma^2\,\chi^2_1\,\{n\,(\mu_A - \mu_B)^2/(2\sigma^2)\}$$

which has expectation $\sigma^2 + n\,(\mu_A - \mu_B)^2/2$.

Ignoring the non-centrality in the second term leads to the variance estimate

$$\widehat{\sigma}^2 \;=\; \frac{S_P^2}{(2n-1)}.$$

Blinded variance estimation

Alternatively, as the mean of the non-central χ^2 term is

$$\sigma^2 + \frac{n(\mu_A - \mu_B)^2}{2},$$

Zucker et al. (*Statist. in Med.*, 1999) subtract the second part of this mean from the pooled sum of squares under the alternative $|\mu_A - \mu_B| = \Delta$.

This yields the "adjusted pooled variance estimate"

$$\frac{S_P^2}{2n-1} - \frac{n\Delta^2}{2(2n-1)}.$$

Friede & Kieser (*Statist. in Med.*, 2001) find this adjusted pooled variance estimate to be:

- simple to evaluate,
- almost as accurate as s^2, the pooled estimate from unblinded data.

2. Error-spending group sequential tests

A two-sided testing problem

Let θ be the treatment effect of a new treatment *vs* a standard, e.g.,

$\theta =$ difference in mean response for normal data, or

$\theta =$ log hazard ratio for survival data.

To look for a difference between the new treatment and standard, test

$$H_0\text{: } \theta = 0 \quad \text{against} \quad \theta \neq 0.$$

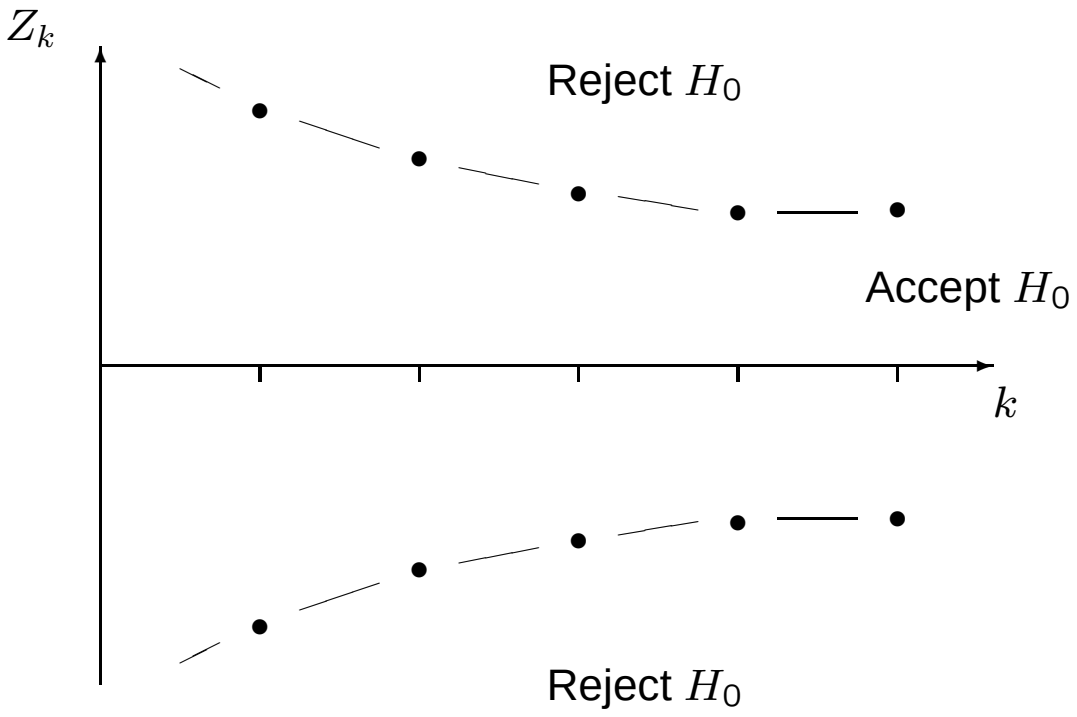
Specify type I error rate $= \alpha$ and power $1-\beta$ at $\theta = \pm\delta$.

Suppose it is desirable to stop early *to reject* H_0 — early stopping for a positive outcome.

Error-spending group sequential tests

In a group sequential test, one monitors the standardised Z statistic at a sequence of interim analyses.

A typical testing boundary has the form:



$E(\text{Sample size})$ can be $\sim$ 70% of the fixed sample size.

(Larger gains are possible in tests with one-sided alternatives and early stopping to accept *or* reject H_0.)

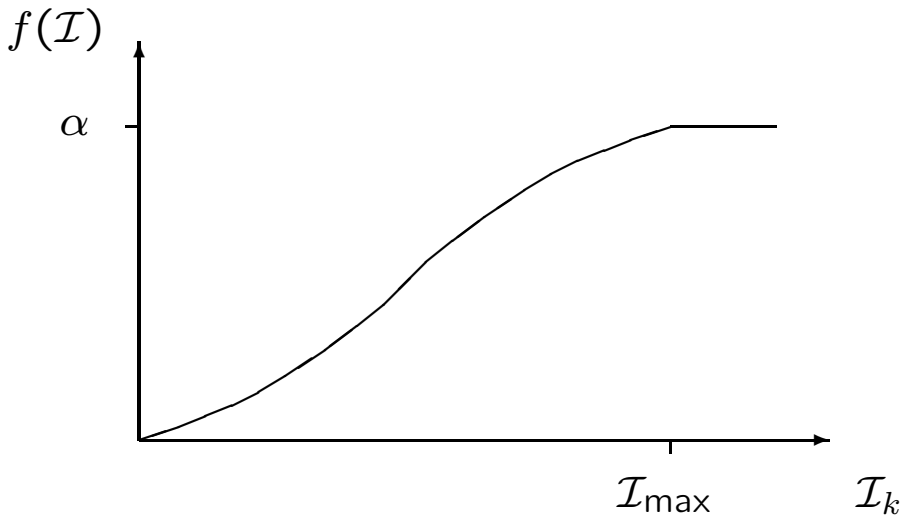
Error-spending group sequential tests

Lan & DeMets (*Biometrika*, 1983) presented tests which "spend" type I error as a function of observed information.

Here, information $\mathcal{I} = 1/\text{Var}(\widehat{\theta})$.

Maximum information design:

Error-spending function $f(\mathcal{I})$

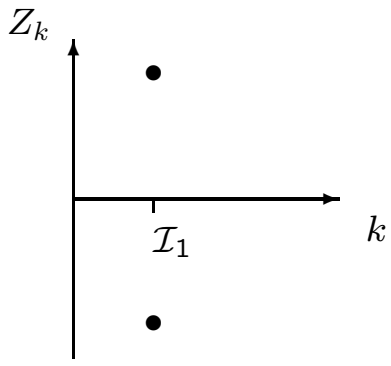


At analysis k, set boundary to give cumulative type I error probability $f(\mathcal{I}_k)$.

Accept H_0 if $\mathcal{I}_{\text{max}}$ is reached without rejecting H_0.

Error-spending group sequential tests

Analysis 1:

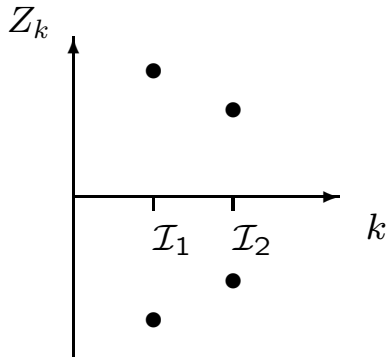


Observed information $\mathcal{I}_1$.

Reject H_0 if $|Z_1| > c_1$ where

$$Pr_{\theta=0}\{|Z_1| > c_1\} = f(\mathcal{I}_1).$$

Analysis 2:

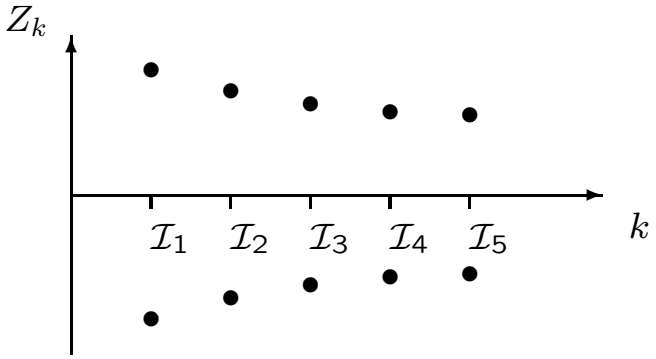


Cumulative information $\mathcal{I}_2$.

Reject H_0 if $|Z_2| > c_2$ where

$$Pr_{\theta=0}\{|Z_1| < c_1, |Z_2| > c_2\}$$
$$= f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

etc, . . .



3. Information monitoring for normal responses

Suppose response distributions on treatments A and B are $X_{Ai} \sim N(\mu_A, \sigma^2)$ and $X_{Bi} \sim N(\mu_B, \sigma^2)$.

With n_A and n_B observations on treatments A and B, information for $\theta = \mu_A - \mu_B$ is

$$\mathcal{I} = \frac{1}{\mathsf{Var}(\widehat{\theta})} = \left\{ \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} \right\}^{-1}.$$

A fixed sample test H_0: $\theta = 0$ against $\theta \neq 0$ with type I error rate α and power $1 - \beta$ at $\theta = \pm\delta$ needs information

$$\mathcal{I}_f = (z_{\alpha/2} + z_\beta)^2/\delta^2.$$

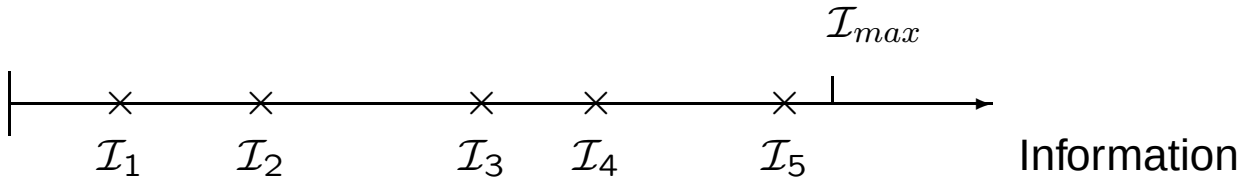
A group sequential test requires maximum information

$$\mathcal{I}_{\mathsf{max}} = R\mathcal{I}_f,$$

where the "inflation factor" R is determined by the boundary shape and number of planned analyses.

Information monitoring for normal responses

Investigators can monitor observed information at interim analyses and modify recruitment to ensure the target $\mathcal{I}_{\max}$ is reached.



The relationship

$$\mathcal{I} = \frac{1}{\mathrm{Var}(\widehat{\theta})} = \left\{ \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} \right\}^{-1}$$

determines the numbers of observations needed to obtain a specified level of information.

Substituting a current estimate of σ^2 gives a present view of the sample size required to reach $\mathcal{I}_{\max}$.

4. Mehta & Tsiatis's group sequential t-tests

Mehta and Tsiatis (Drug Information J., 2001) follow the information monitoring approach.

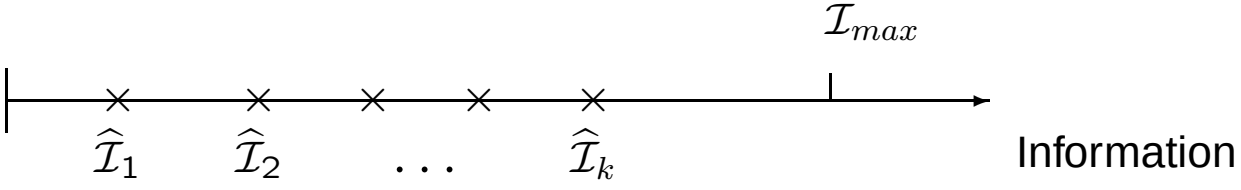
At analysis k, estimate σ^2 by

$$s_k^2 = \frac{\sum(X_{Ai} - \bar{X}_A^{(k)})^2 + \sum(X_{Bi} - \bar{X}_B^{(k)})^2}{n_{Ak} + n_{Bk} - 2}.$$

and estimate observed information by

$$\widehat{\mathcal{I}}_k = \frac{1}{\mathrm{Var}(\widehat{\theta})} = \left\{ \frac{s_k^2}{n_A} + \frac{s_k^2}{n_B} \right\}^{-1}.$$

Use the observed information sequence



to create an error-spending boundary, with cumulative error probability $f(\mathcal{I}_k)$ up to analysis k.

Error spending boundary

Error-spending calculations are really for a sequence of statistics Z_k for normal data with known variance.

To implement the test, define t-statistics

$$T_k = \frac{\bar{X}_A^{(k)} - \bar{X}_B^{(k)}}{\sqrt{s_k^2(1/n_{Ak} + 1/n_{Bk})}},$$

and test at the *significance levels* given by the boundary computed for Z_ks.

Updating the sample size

In a K-group design: at each analysis $k < K$, re-calculate the target for n_{A5} and n_{B5} by solving the equation

$$\left\{\frac{s_k^2}{n_{A5}} + \frac{s_k^2}{n_{B5}}\right\}^{-1} = \mathcal{I}_{\max}$$

and choose the next group size to work towards this target.

Mehta & Tsiatis: updating sample size

Example:

Suppose $\mathcal{I}_{\max} = 140.0$ and an initial estimate of σ^2 is $\hat{\sigma}_0^2 = 0.6$. Solving

$$\left\{ \frac{\hat{\sigma}_0^2}{n_{A5}} + \frac{\hat{\sigma}_0^2}{n_{B5}} \right\}^{-1} = \mathcal{I}_{\max}$$

gives $n_{A5} = n_{B5} = 168$, i.e., initial group sizes of $168/5 = 34$.

Analysis 1. Observe $s_1^2 = 0.42$. Re-estimate target sample size from

$$\left\{ \frac{s_1^2}{n_{A5}} + \frac{s_1^2}{n_{B5}} \right\}^{-1} = \mathcal{I}_{\max},$$

giving $n_{A5} = n_{B5} = 118$.

Aim for this with $(118 - 34)/4 = 21$ observations per treatment arm in group 2.

Mehta & Tsiatis: updating sample size

Analysis 2. Observe $s_2^2 = 0.58$. Re-estimate target sample as $n_{A5} = n_{B5} = 162$.

Take $(162 - 55)/3 = 36$ obs. per arm in group 3.

Analysis 3. Observe $s_3^2 = 0.68$. Re-estimate target sample size as $n_{A5} = n_{B5} = 190$.

Take $(190 - 91)/2 = 50$ obs. per arm in group 4.

Analysis 4. Observe $s_4^2 = 0.72$. Re-estimate target sample size as $n_{A5} = n_{B5} = 202$.

Take $202 - 141 = 61$ obs. per arm in group 5.

Analysis 5. Observe $s_5^2 = 0.69$.

Re-estimating target sample size using s_5^2 gives $n_{A5} = n_{B5} = 193$. We have 202 observations per arm, so the test is most likely a little over-powered.

Mehta & Tsiatis: issues

There are several types of approximation going on:

1. We monitor t-statistics but compute the boundary using the joint distribution of Z-statistics.

This is known to work well in simpler settings (no sample size re-estimation), especially for O'Brien & Fleming type boundaries which are wide early on.

2. The estimates $\widehat{\mathcal{I}}_k$ may decrease as more responses are observed — and this happens much more often than you might expect!

Pragmatic solution:

Do not allow stopping at an analysis k where $\widehat{\mathcal{I}}_k < \widehat{\mathcal{I}}_{k-1}$.

With a fixed total number of analyses, K, if $\widehat{\mathcal{I}}_K < \widehat{\mathcal{I}}_{K-1}$ ($< \mathcal{I}_{\max}$), replace $\widehat{\mathcal{I}}_K$ by $\widehat{\mathcal{I}}_{K-1}$ and spend all remaining error probability.

Mehta & Tsiatis: issues

3. Using estimates of $\mathcal{I}_k$, we mis-specify correlations of the $\{T_k\}$ or of the approximating $\{Z_k\}$.

In fact, Corr$(Z_k, Z_{k+1}) = \sqrt{(\mathcal{I}_k/\mathcal{I}_{k+1})}$, and this ratio does not depend on the unknown σ^2.

So, we can use the precise value of this ratio rather than simply plugging in $\widehat{\mathcal{I}}_k$ and $\widehat{\mathcal{I}}_{k+1}$.

4. Re-estimating sample size based on s^2 produces a downwards bias in s^2, as in the Wittes & Brittain procedure.

We need to investigate whether this leads to inflation of the type I error rate.

5. Mehta & Tsiatis report just one example with a target of over 500 observations per treatment.

Does this indicate problems for smaller sample sizes?

Mehta & Tsiatis: a simulation study

Problem: Two-treatment comparison, normal responses with unknown variance.

To test: H_0: $\theta = 0$ vs $\theta \neq 0$, with type I error probability $\alpha = 0.05$, aiming for power 0.9 at $\theta = \pm\delta$.

True variance is $\sigma^2 = 1$.

We start the procedure with an initial estimate σ_0^2.

Tests are constructed using error-spending function

$$f(\mathcal{I}_k) = \alpha\,(\mathcal{I}_k/\mathcal{I}_{\max})^\rho$$

for various choices of ρ.

Here, $\rho = 1$ gives a similar boundary to Pocock's test (constant significance level)

Boundaries for $\rho = 3$ are close to those of O'Brien & Fleming.

Mehta & Tsiatis: simulation study

Tests with 3 analyses, $\sigma^2 = 1$, $\sigma_0^2 = 1.6$.

δ	*Target degrees of freedom* *	***Type I error rate*** $\rho = 1$	$\rho = 2$	$\rho = 3$
0.5	176	0.051	0.052	0.052
0.7	90	0.052	0.053	0.054
1.0	44	0.054	0.056	0.058
1.5	20	0.054	0.059	0.061
2.0	12	0.057	0.060	0.061

δ	*Target degrees of freedom*	***Power*** $\rho = 1$	$\rho = 2$	$\rho = 3$
0.5	176	0.899	0.897	0.898
0.7	90	0.899	0.897	0.897
1.0	44	0.900	0.898	0.898
1.5	20	0.913	0.906	0.906
2.0	12	0.924	0.924	0.924

*Target for final analysis if $s^2 = \sigma^2$; value is for the case $\rho = 2$, other cases differ by up to $\sim$ 5%.

Mehta & Tsiatis: simulation study

Tests with 5 analyses, $\sigma^2 = 1$, $\sigma_0^2 = 1.6$.

δ	*Target degrees of freedom*	***Type I error rate***		
		$\rho = 1$	$\rho = 2$	$\rho = 3$
0.5	178	0.052	0.053	0.053
0.7	92	0.054	0.056	0.056
1.0	46	0.058	0.062	0.063
1.5	22	0.065	0.069	0.070
2.0	12	0.061	0.066	0.067

Notes on inflation of type I error:

Inflation is greater for higher values of ρ — when boundaries are wide at early analyses, which have low degrees of freedom for estimating σ^2.

Inflation increases with the number of analyses.

Mehta & Tsiatis: simulation study

To understand the source of type I error inflation, consider tests with frequent analyses and very little error spent before the final analysis.

Tests with 20 analyses, $\sigma^2 = 1$, $\sigma_0^2 = 1.6$.

δ	*Target degrees of freedom*	***Type I error rate*** $\rho = 50$
0.5	170	0.052
0.7	86	0.057
1.0	44	0.096
1.5	20	0.101

Conclude:

Repeated re-estimation of sample size is problematic since it enhances the effect of "stopping when the current estimate of σ^2 is unusually low".

(Cf Chow & Robbins, fixed width CI for a normal mean.)

5. Conclusions

1. Sample size can be adapted to estimates of nuisance parameters during the course of a study.

2. This can be done within a group sequential test, particularly when the error-spending approach is used with a “maximum information” design.

3. Frequent re-estimation of sample size may lead to substantial inflation of the type I error rate. A proposed design should be checked by simulation; since the true parameter value (e.g., a normal variance) is unknown, simulations should cover a range of possible values.

4. On occasions, more precise methods are called for, e.g., Denne & Jennison (*Biometrika*, 2000) for normal data with unknown variance.