Group Sequential Selection Procedures with Elimination and Data-Dependent Treatment Allocation

Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

http://www.bath.ac.uk/∼mascj

Manchester, March 2005

Motivation

In a drug development programme:

- Many compounds will be tested.
- Dose finding studies (Phase IIb) are conducted for compounds showing some efficacy.
- A Phase III study compares the new treatment at the selected dose level against the appropriate control, active or placebo.

To approve a new treatment, regulatory bodies (MHRA, FDA) usually require two positive Phase III studies plus supporting evidence from the preceding development process.

Efficiency

There are opportunities for efficient study design in:

- selecting the best of several versions of the treatment in Phase IIb,
- comparing one or more treatments vs control in Phase III.

One may wish to employ

- mid-study elimination of poorly performing treatments,
- an overall sequential stopping rule.

One may also wish to merge the Phase IIb and Phase III trials.

Plan of talk

1. Selection procedures with no control treatment

2. Selection methods with testing against a control

3. Seamless transition from Phase IIb to Phase III

1. Selection procedures with no control treatment

Suppose for each "population" or "treatment" $i = 1, \ldots, k$,

$$X_{i1},\, X_{i2},\, \ldots \;\sim\; N(\theta_i,\, \sigma^2).$$

Aim: To select the population i with the largest mean θ_i.

The equivalent of *power* is a requirement on the probability of correct selection under certain sets of means $(\theta_1, \ldots, \theta_k)$.

Methods will include:

- Early elimination of weak treatments.
- Response-dependent treatment allocation to reduce the inferior treatment number and total sample size.

Earlier work

Paulson (*Ann. Math. Statist.*, 1964)

Elimination procedures based on *continuous* sequential comparisons of 2 populations at a time.

Robbins and Siegmund (*JASA*, 1974)

Adaptive sampling for a 2 population comparison with continuous monitoring.

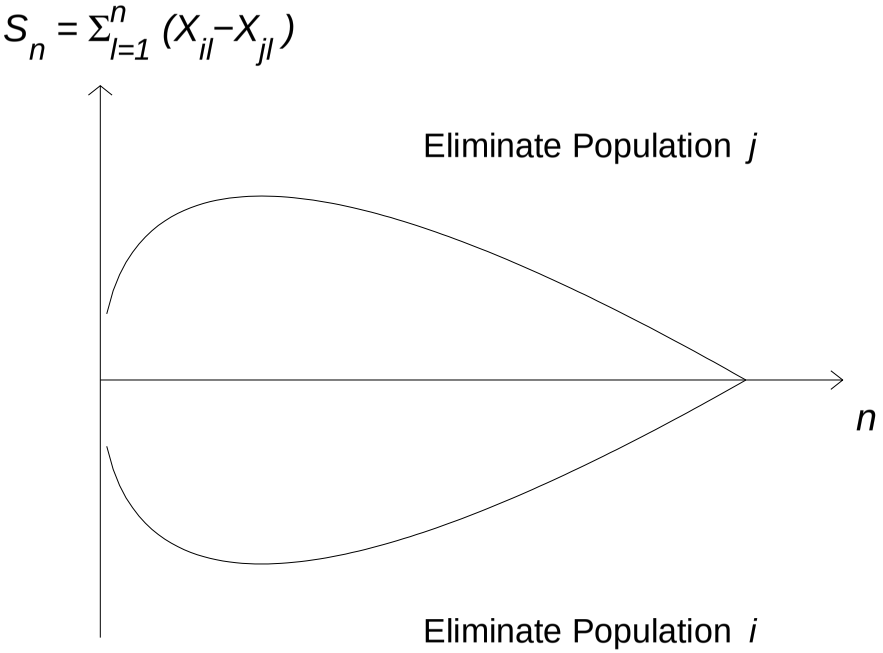
Jennison, Johnstone and Turnbull (*Purdue Symposium*, 1982)

Combining the above.

Update: To take advantage of group sequential tests, error spending, modern computation.

Paulson's procedure

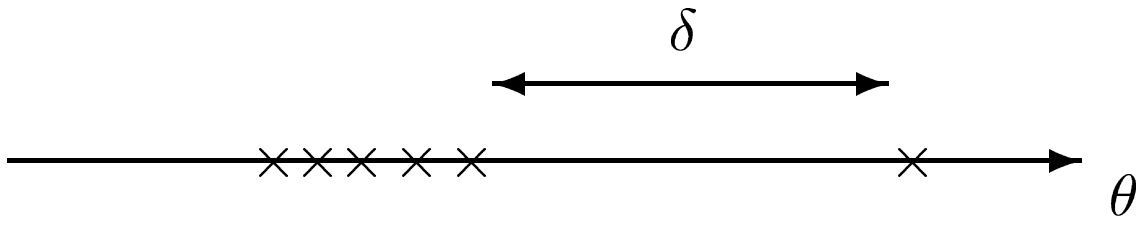
Compare all pairs Treatment i *vs* Treatment j.



If $\theta_i = \theta_j - \delta$, then $Pr\{\text{Pop. } i \text{ eliminates Pop. } j\} = \alpha/(k-1)$.

Paulson's procedure: Probability of correct selection

Indifference zone formulation



Suppose $\theta_i \leq \theta_k - \delta$ for $i = 1, \ldots, k-1$.

Then

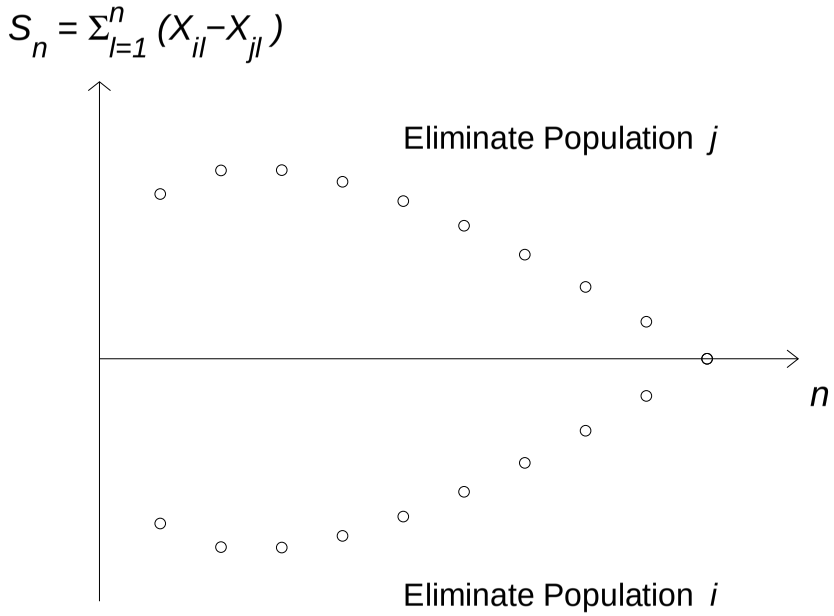
$$Pr\{\text{Population } k \text{ is eliminated at some stage}\}$$

$$\leq \sum_{i=1}^{k-1} Pr\{\text{Pop. } i \text{ eliminates Pop. } k \text{ at some stage}\}$$

$$\leq (k-1)\frac{\alpha}{k-1} = \alpha.$$

Paulson's procedure: Group sequential monitoring

Compare treatments at regular interim analyses.



Choose a group sequential boundary with error rate $\alpha/(k-1)$ at $\theta_i - \theta_j = \pm\delta$ and good early stopping under likely $(\theta_1, \ldots, \theta_k)$.

Adaptive sampling in Paulson's procedure

Motivation

Observations on the leading population are used in $k-1$ comparisons.

Allocating more observations to the leader can

- Reduce total sample size
- Reduce observations on inferior treatments
  - ethical for medical studies
  - we learn more about better treatments.

Need:

Theory to support adaptive sampling in each pair-wise comparison.

Adaptive sampling in a group sequential test

Jennison and Turnbull (*Sequential Analysis*, 2001)

For a 2-treatment comparison with

$$X_{1i} \sim N(\theta_1,\, \sigma^2) \quad i = 1, 2, \ldots ,$$

$$X_{2i} \sim N(\theta_2,\, \sigma^2) \quad i = 1, 2, \ldots .$$

At analysis m out of M, with n_{1m} observations on population 1 and n_{2m} on population 2,

$$\widehat{\theta}_{1m} - \widehat{\theta}_{2m} = \bar{X}_{1m} - \bar{X}_{2m} \sim N(\theta_1 - \theta_2,\, \sigma^2\{\tfrac{1}{n_{1m}} + \tfrac{1}{n_{2m}}\})$$

$$\sim N(\theta_1 - \theta_2,\, \mathcal{I}_m^{-1}), \quad \text{say.}$$

Adaptive sampling

The score statistic

$$S_m \;=\; \mathcal{I}_m(\widehat{\theta}_{1m} - \widehat{\theta}_{2m}) \;\sim\; N(\{\theta_1 - \theta_2\}\,\mathcal{I}_m,\, \mathcal{I}_m).$$

Without adaptive sampling, $\{S_1, S_2, \ldots\}$ is distributed as a Brownian motion with drift $\theta_1 - \theta_2$ observed at $\mathcal{I}_1, \mathcal{I}_2, \ldots$.

This remains true if group sizes $n_{1m} - n_{1,m-1}$ and $n_{2m} - n_{2,m-1}$ depend on $\widehat{\theta}_{1,m-1} - \widehat{\theta}_{2,m-1}$ — but sampling *cannot* depend more generally on $(\widehat{\theta}_{1,m-1},\, \widehat{\theta}_{2,m-1})$.

Theory generalises to normal linear models containing θ_1 and θ_2.

This extends Robbins and Siegmund (1974) to the group sequential case.

Adaptive sampling: Problem 1

Problem 1

With $k \geq 3$, sampling rules of interest do not satisfy

"m th group sizes for populations 1 and 2 depend

only on $\widehat{\theta}_{1,m-1} - \widehat{\theta}_{2,m-1}$".

Solution

- Fix sampling ratios at the start of each group,
- estimate $\theta_i - \theta_j$ within each group of data,
- combine estimates with weights $\propto$ variance^{-1}.

This equates to fitting a linear model with additive "stage" effects

— also recommended to avoid bias from time trends.

Adaptive sampling: Problem 1

JT (2001) assess performance of 2-treatment tests:

With stage effects in the model, one cannot compensate later on for sub-optimal sampling ratios in early stages. Savings in Inferior Treatment Numbers are reduced by about a half.

- Fitting stage effects to avoid bias from a time trend is reasonable.
- But, if such a trend is not really present, data are being used inefficiently — ethically questionable for medical studies

JJT (1982) took a “heuristic” line, running simulations of their methods without stage effects and there was no apparent harm to error rates.

Adaptive sampling: Problem 2

Problem 2

Information levels for comparing populations i and j

$$\mathcal{I}_{ij,1},\ \mathcal{I}_{ij,2},\ \mathcal{I}_{ij,3},\ \ldots\ ,$$

depend on the sampling rule, which involves $S_{ij,1},\ S_{ij,2},\ S_{ij,3},\ \ldots$.

Standard group sequential designs, including error spending tests, do not allow such a dependence.

Solution A

Reported studies of such "data-dependent analysis times" show only minor effects on error probabilities — trust these studies and ignore the problem!

Adaptive sampling: Problem 2

Solution B

Recent designs which "adapt" to observed data offer a precise solution:

Denne (*Statistics in Medicine*, 2001),

Müller and Schäfer (*Biometrics*, 2001).

Procedure

- Set up an error spending test for anticipated $\{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$
- Recursively for $m = 1, 2, \ldots,$
  - At analysis m, compute conditional error probabilities given S_m
  - Run stages $m+1$ to M as an error spending test with this conditional error.

A sampling rule (JJT, 1982)

In comparing $N-1$ populations with a control, the most efficient allocation is

$\sqrt{N-1}$ observations on the control to

1 observation on each other population.

Adaptive rule:

At stage m, with N_m non-eliminated populations, sample

$\sqrt{N_m-1}$ observations on the leading population to

1 observation on each other population.

A group sequential Paulson procedure with adaptive sampling

Eliminate populations using Paulson's pair-wise comparisons.

Run these comparisons as error spending group sequential tests.

a) Base tests on overall population means (cf JJT, 1982)

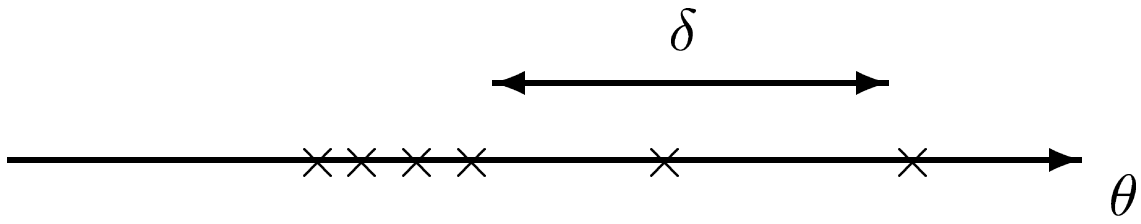
Sample in stage m to achieve ratios $\sqrt{N_m - 1} : 1 : \ldots : 1$ of total observations on the N_m surviving populations.

b) Combine stage-wise estimates of each $\theta_i - \theta_j$

Sample in ratios $\sqrt{N_m - 1} : 1 : \ldots : 1$ within stage m.

Problem 1 is dealt with properly in (b); Problem 2 is ignored (Solution A!)

Beyond the indifference zone



What if there is a θ_i within δ of the highest θ_j?

It should be OK to select a population within δ of the best. But can a non-optimal population eliminate the best, then be eliminated itself?

Kao and Lai (*Comm. Statist. Th. Meth.*, 1980) provide a solution, raising the boundary for any pair-wise elimination before the final decision.

This method works for Paulson's procedure with adaptive sampling and can be extended to choosing the best s populations out of k.

Ideas to take forward to comparisons with a control

- Paulson's scheme offers a simple approach to sequential elimination of treatments.
- Pair-wise comparisons plus "Bonferroni" is not badly conservative.
- Using efficient group sequential tests in each pair-wise comparison leads to good overall performance.
- Adaptive treatment allocation can help reduce sample size.
- Dealing with power is not simple when you need to consider a procedure's behaviour under all possible vectors of treatment means.

2. Selection methods with testing against a control

Aim: Conduct a single study to select a treatment (e.g., dose level) and test for superiority to a control.

Two-stage procedures are proposed by:

Thall, Simon and Ellenberg (*Biometrika*, 1988)

Schaid, Wieand and Therneau (*Biometrika*, 1990)

Stallard and Todd (*Statist. in Medicine*, 2003)

Stage 1:

Compare k experimental treatments and 1 control.

Stage 2:

If appropriate, continue with selected treatments vs the control.

Selection and testing

The 3 papers consider 3 different response types (binary, survival, general) but generic normal test statistics are used in each case.

We shall look at the TSE procedure in detail:

Index control treatment by 0, experimental treatments by 1,. . . ,k.

Stage 1

Take n_1 observations per treatment and control.

Denote standardised statistic for comparing treatment j against control by $T_{j,1}$ and let the maximum of these be $T_{j^*,1}$.

If $T_{j^*,1} > C_1$, select treatment j^* and proceed to Stage 2,

if $T_{j^*,1} \leq C_1$, stop and accept H_0: $\theta_0 = \theta_1 = \ldots = \theta_k$.

Selection and testing

Stage 2

Take n_2 further observations on selected treatment, j^*, and control.

Combine data from both stages in the standardised statistic $T_{j^*,2}$.

If $T_{j^*,2} > C_2$, reject H_0 and conclude $\theta_{j^*} > \theta_0$,

if $T_{j^*,2} \leq C_2$, accept H_0.

Values of n_1, n_2, C_1 and C_2 need to be chosen to satisfy type I error and power conditions.

There is additional freedom to tune the procedure's performance, e.g., minimise expected sample size in certain situations.

Type I error and power

The experimental treatment j^* is said to be "chosen" if

treatment j^* is selected at the end of Stage 1, and

H_0 is rejected in favour of $\theta_{j^*} > \theta_0$ at Stage 2.

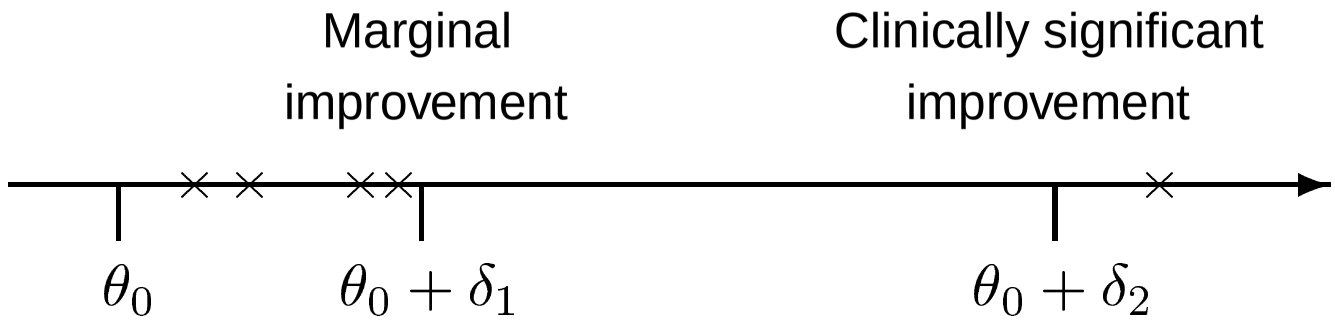
The type I error rate is

$$Pr_{\boldsymbol{\theta}} \{\text{Any experimental treatment is chosen}\}$$

under H_0: $\theta_0 = \theta_1 = \ldots = \theta_k$.

Power depends on the full vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$.

Type I error and power



Any treatment with $\theta_j \geq \theta_0 + \delta_2$ is said to be "acceptable".

Consider cases of $\boldsymbol{\theta}$ where:

at least one treatment is acceptable,

no θ_j lies in the interval $(\theta_0 + \delta_1,\ \theta_0 + \delta_2)$.

The power function is

$$1 - \beta(\boldsymbol{\theta}) \ = \ Pr_{\boldsymbol{\theta}} \{\text{An acceptable choice is made}\}.$$

Type I error and power

TSE show that, over cases as described above, $1 - \beta(\boldsymbol{\theta})$ is minimised under the *least favourable configuration:*

$$\theta_1 = \ldots = \theta_{k-1} = \theta_0 + \delta_1 \quad \text{and} \quad \theta_k = \theta_0 + \delta_2.$$

They call this configuration $\boldsymbol{\theta}^*$ and specify a value for $1 - \beta(\boldsymbol{\theta}^*)$ as their power condition.

Numerical integration is feasible under H_0 and $\boldsymbol{\theta}^*$. Hence, parameters $n_1,\ n_2,\ C_1$ and C_2 satisfying the type I error and power conditions can be found.

Tests minimising expected sample size averaged over these two cases are found by searching feasible parameter combinations.

Thall, Simon and Ellenberg's procedure

Comments on the TSE two-stage procedure

Inclusion of the control treatment in Stage 1 is important: it allows results from that stage to be pooled with the data on treatment j^* vs the control in Stage 2.

The type II errors under $\boldsymbol{\theta}^*$ comprise

mostly: failure to reject H_0,

to a smaller degree: choosing a sub-optimal treatment as superior to the control.

Schaid, Wieand and Therneau (*Biometrika*, 1990)

Schaid et al allow more options at the end of Stage 1:

stop to accept H_0,

stop and choose an experimental treatment as superior to the control.

More than one experimental treatment may continue to Stage 2. This is appropriate for a survival study where differences may appear in longer term survival.

Type I error and power properties are found by pairwise comparisons with the control, combined by Bonferroni's inequality.

Stallard and Todd (*Statist. in Medicine*, 2003)

Stallard and Todd select just one treatment at the end of Stage 1.

They allow further interim analyses during Stage 2 at which termination may occur either to accept or to reject H_0.

These analyses are defined as a group sequential test with a specified error spending function.

Computations are based on the null distribution of the maximum score for an experimental treatment against the control (at the end of Stage 1), followed by increments in this score according to the usual stochastic process.

3. Seamless transition from Phase IIb to Phase III

The proposals of TSE, SWT and ST can be used with Stage 1 taking the place of a Phase IIb trial and Stage 2 the ensuing Phase III trial.

These proposals can be extended. One could:

(i) Combine all the various ingredients of the 3 methods

(ii) Allow sequential monitoring and elimination of inferior treatments throughout

(iii) Introduce unequal/data-dependent treatment allocation.

It would be difficult to compute properties of such complex designs — let alone find “optimal” versions.

Modelling the dose-response curve

When comparing dose-levels, it is natural to expect efficacy to change fairly smoothly as dose increases.

In their ASTIN trial, Krams et al (2003) adopted a simple nonparametric dose-response model and developed a Bayesian approach to design, monitoring and analysis.

The resulting adaptive experimental design contains the elements of the TSE, SWT and ST proposals.

Computation is quite a task, but a close-to-optimal adaptive sampling scheme can be found. Frequentist properties of the design are found by simulation and parameters tuned to give a specified Type I error probability.

Combining two or more Phase III studies

There is efficiency in using a single control arm to assess several active treatments.

But, when a treatment proves successful, the remaining lifetime of its patent is paramount to the developers — so, savings in trial costs are unhelpful if they cause a delay in reaching a positive outcome.

Whatever you decide to do:

Problem formulation is crucial — what do you really wish to achieve?

Keeping some simplification can help to:

make computations feasible,

give an easily interpretable method.

References

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645–2660.

Jennison, C., Johnstone, I. M. and Turnbull, B. W. (1982) Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III, Vol. 2*, (eds S. S. Gupta and J. Berger) New York: Academic Press **2**, 55–86.

Jennison, C. and Turnbull, B. W. (2000) *Group Sequential Tests with Applications to Clinical Trials*, Chapman and Hall/CRC, 390 pages.

Jennison, C. and Turnbull, B. W. (2001) Group sequential tests with outcome-dependent treatment assignment. *Sequential Analysis*, **20**, 209–234.

References, continued

Kao, S. C. and Lai, T. L. (1980). Sequential selection procedures based on confidence sequences for normal populations. *Comm. Statist. Theory and Methods*, **A9(16)**, 1657–1676.

Krams, Lees, Hacke, Grieve, Orgogozo and Ford (2003). Acute stroke therapy by inhibition of neutrophils (ASTIN). *Stroke*, **34**, 2543–2548.

Müller, H-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **57**, 886–891.

Paulson, E. (1964) A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.*, **35**, 174–180.

References, continued

Proschan, M. A., Follmann, D. A and Geller, N. L. (1994). Monitoring multi-armed trials. *Statistics in Medicine*, **13**, 1441–1452.

Robbins, H. and Siegmund, D. O. (1974). Sequential tests involving two populations. *J. Amer. Statist. Assoc.*, **69**, 132–139.

Schaid, D. J., Wieand, S. and Therneau, T. M. (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika*, **77**, 507–513.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*, **22**, 689–703.

Thall, P. F., Simon, R. and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, **75**, 303–310.