Sample Size Re-estimation Using Adaptive Tests and Generalized Likelihood Ratio: A Comparative Study

Shanhong Guan

Merck Research Laboratories

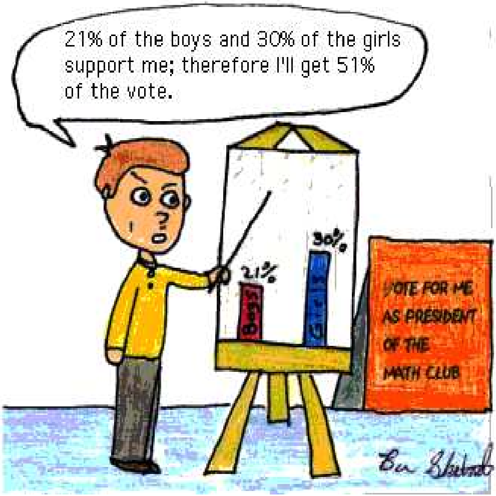
November 8, 2010

1. Outline
2. Introduction
3. Hypothetical study: a common question
4. Combination tests
5. Generalized Likelihood Ratio (GLR) Procedure
6. Numerical study
7. Concluding remarks
8. Backup slides

Sample size calculation

- A standard clinical trial: sample size n is determined by
  - type I error rate α, power $1 - \beta$
  - variance σ^2, effect size $\Delta = \theta_t - \theta_c$, where θ_t and θ_c denote the treatment and control effect respectively
- For effect size Δ
  - Minimum clinically relevant: e.g., minimum effective size (MES)?
  - Realistic: anticipated effect size (AES)?
- Assume variance σ^2 is known, standard formula would be (per arm)

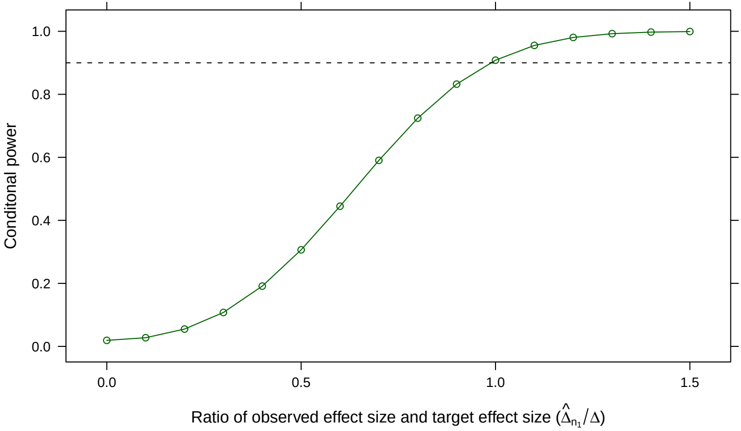
$$n = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2}.$$


21% of the boys and 30% of the girls support me; therefore I'll get 51% of the vote.
21%
Boys
30%
Girls
VOTE FOR ME AS PRESIDENT OF THE MATH CLUB

Effect size Δ

- Suppose $\Delta_{MES} = 3$ and $\Delta_{AES} = 5$, $\sigma = 10$, $\alpha = 5\%$, $\beta = 10\%$.
- How do we want to power the trial?
  - Δ_{MES}: $n = 234$; may overpower
  - Δ_{AES}: $n = 85$; may underpower
- For example: if study is powered at $\Delta = 5$. At some intermediate point (sample size n_1)
  - $\hat{\Delta}_{n_1} = \bar{X}_{t,n_1} - \bar{X}_{c,n_1}$
  - Test statistic $Z_1 = \hat{\Delta}_{n_1}/\sqrt{2\hat{\sigma}^2/n_1}$.
- Suppose $\hat{\Delta}_{n_1} = 3.5$ and $\hat{\sigma} = 15$
  - It is unlikely a significant result will be achieved at the planned end of the trial.
  - However, $\hat{\Delta}_{n_1} = 3.5$ suggests further investigation is warranted.
- Given Z_1, how likely we can detect a significant effect at the end of trial?

Conditional power given $\hat{\Delta}_{n_1}$



Now what?

"...*A well-designed study, poorly analyzed, can be rescued by a reanalysis but a poorly designed study is beyond the redemption of even sophisticated statistical manipulation.*"
- Campbell, Machin and Walters

"...*Not everything that is faced can be changed. But nothing can be changed until it is faced.*"
- James Baldwin

Actual type I error rate

If we adjust second stage sample size n_2 based on $\hat{\Delta}_{n_1}$ to achieve certain conditional power, type I error rate is computed as $\int_{-\infty}^{+\infty} cP(n_2, z_{1-\alpha}|z_1, \Delta = 0)\psi(z_1)dz_1$ (Proschan and Hunsberger, 1995), where

$$cP(n_2, z_{1-\alpha}|z_1, \Delta = 0) = 1 - \Phi\left[\frac{z_{1-\alpha}\sqrt{2(n_1+n_2)} - z_1\sqrt{n_1}}{\sqrt{2n_2}}\right] (1).$$

Inflation of type I error rate

- Note that:

$$\begin{cases} cP(n2, z_{1-\alpha}|z_1, \Delta = 0) = 1 & \text{if} \quad z_1 > z_{1-\alpha} \\ cP(n2, z_{1-\alpha}|z_1, \Delta = 0) \to \alpha & \text{if} \quad z_1 < 0 \\ & \text{and} \quad r \to \infty \\ \max cP(n2, z_{1-\alpha}|z_1, \Delta = 0) = 1 - \Phi(\sqrt{z_{1-\alpha}^2 - z_1^2}) & \text{if} \quad 0 \le z_1 \le z_{1-\alpha} \end{cases}$$

- Therefore, the maximum value of type I error rate (theoretically) could be

$$\alpha_{max} = \alpha + \exp\{-z_{1-\alpha}^2/2\}/4.$$

Table: Type I error inflation

Nominal Type I error rate	0.01	0.02	0.025	0.05	0.1
Actual Type I error rate	0.0267	0.0503	0.0616	0.1146	0.2100

- Typically type I error rate of such procedure is inflated by 30-40% (Cui, 1999)

A brief review of Proschan and Hunsberger method

- To control the type I error rate when design extensions of trial, define *conditional error function* $A(z_1)$

$$\int_{-\infty}^{\infty} A(z_1)\phi(z_1)dz_1 = \alpha$$

  - Any increasing function with range [0,1]
  - PH uses circular conditional error function $A(z_1) = 1 - \Phi(k^2 - z_1^2)$, k is the critical value
  - $A(z_1)$ dictates how much conditional type I error rate to allow at the end of study, *given* $Z_1 = z_1$
- If sample size n_2 for the second stage is derived based on z_1, adjust final critical value from $z_{1-\alpha}$ to c (to maintain type I error rate)
- To derive n_2 and c, define $\delta = (\theta_t - \theta_c)/\sigma$ and note the following relationship (Z_A is shorthand for $Z_{A(z_1)}$):

$$cP_\delta(n_2, c|z_1) = 1 - \Phi\left(z_A - \sqrt{\frac{n_2}{2}\delta}\right).$$

Choose n_2 and final critical bound c

- If we choose conditional power β_2, we have

$$n_2 = \frac{2(z_A + z_{1-\beta_2})^2}{\delta^2}$$

and

$$c = \frac{\delta\sqrt{\frac{n_1}{2}}z_1 + (z_A + z_{1-\beta_2})z_A}{\sqrt{\delta^2\frac{n_1}{2} + (z_A + z_{1-\beta_2})^2}}$$

Conditional error function based SSR

- Due to resource and trial duration, usually there is an upper bound M to the allowable sample size
- SSR based on conditional error function is a two-stage procedure:
  - determine a maximum sample size M
  - specify an initial sample size $m = \xi M, \xi \in (0, 1)$
  - Adjust sample size base on estimate $\hat{\delta}_m$ such that conditional power is $1 - \beta$ and type I error rate is α

Potential issues with PH procedure (1)

- Consider the following scenario constructed based on PH approach:
  - Test $H_0 : \theta \leq 0$ vs. $H_a : \theta > 0$ with $\theta_1 = 0.3$, $\alpha = 0.025$, $\beta = 0.1$
  - Set-up: $m = 40$ for the first stage and maximum sample size $M = 120$
  - PH two-stage procedure (Bartroff and Lai, 2008): futility bound $z_{p^*} = 1.71$ and efficacy bound $k = 2.05$
- Simulation study indicates that even at $\theta_1 = 0.3$, PH test have power less than 0.6
  - In large part, this is due to aggressive stop for futility: if $z_1 < z_{p^*} = 1.71$
  - However, $Pr_{\theta_1}(z_1 < z_{p^*}) = 0.43$, well exceeding the nominal type II error rate

Potential issues with PH procedure (2)

- Use stringent futility stopping is necessary to control the sample size
  - in (unrealistic) extreme case: a 0.025-level PH test that uses $z_p^* = 0$ has expected sample size $> 10^7$
- How to improve the power of SSR procedure under maximum sample size constraints?
  - Use alternative to conditional error function: combination tests
  - Take into account the sampling variability of $\hat{\delta}$ and allow for a possible third stage (sequential procedure)

A brief comment

- What about size re-estimation without unblinding?

Adaptive Design Clinical Trials for Drugs and Biologics (FDA, 2010): *..."Similarly, when a continuous outcome measure is the study endpoint, a blinded examination of the variance of the study endpoint can be made and compared to the assumption used in planning the study. If this comparison suggests the initial assumption was substantially too low and the study is consequently underpowered, an increase in the study sample size can maintain the desired study power..."*

- One concern: confounding between blinded variance estimate and magnitude of assumed treatment effect size
- So is it really well understood?
- Unblinded sample size re-estimation - less well understood: methodology itself (e..g. multiplicity) and regulatory concern (blinding and integrity)

P-value combination test: product of p-values

- Define the test statistic on the p-scale:
  - Stage 1 (sample size n_1): $T_1 = p_1$
  - Stage 2 (sample size n_2): $T_2 = p_1 p_2$,

  where p_i is the stagewise p-value based on sample from stage i.
- To preserve type-I error rate, we solve the following equation for α_2

$$\alpha = \alpha_1 + \alpha_2 \ln \frac{\beta_1}{\alpha_1}.$$

- At the end of stage 2: reject H_0 if $T_2 < \alpha_2$

$$\text{The decision rules are: } \begin{cases} \text{stop and reject} \quad H_0, & \text{if} \quad p_1 \le \alpha_1 \\ \text{stop and accept} \quad H_0, & \text{if} \quad p_1 > \beta_1 \\ \text{continue to stage 2,} & \text{if} \quad \alpha_1 < p_1 \le \beta_1 \end{cases}$$

Conditional type-I error rate

- To see the connection between conditional power and p-value combination test, we define the rejection criterion on the $z-$scale as $z_2 \geq C(\alpha_2, p_1)$.
- For example: for product approach $p_1 p_2 \leq \alpha_2$.
  - $z_2 \geq \Phi^{-1}(1 - \alpha_2/p_1)$,
  - $C(\alpha_2, p_1) = \Phi^{-1}(1 - \alpha_2/p_1)$.
- The conditional power is then given by

$$cP_\delta(n_2, \alpha_2 | p_1) = 1 - \Phi\left[C(\alpha_2, p_1) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right].$$

- Taking $\delta = 0$, conditional probability of making type-I error at the second stage given p_1 is $A(p_1) \overset{\text{def}}{=} 1 - \Phi\left[C(\alpha_2, p_1)\right]$.
- Type-I error rate α is controlled by satisfying the following equation:

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\beta_1} A(p_1) dp_1.$$

Test statistic and conditional error

- Denote by $w_i = \sqrt{n_i/(n_1+n_2)}$
- MSP - method based on sum of p-values
- MPP - method based on product of p-values
- MINP - method based on inverse normal transformation of p-values

Table: Function $C(\alpha_2, p_1)$ for SSR

Method	Test statistic	$C(\alpha_2, p_1)$
MSP	$p_1 + p_2$	$\Phi^{-1}(1 - \max(0, \alpha_2 - p_1))$
MPP	$p_1 p_2$	$\Phi^{-1}(1 - \alpha_2/p_1)$
MINP	$1 - \Phi(w_1 z_{1-p_1} + w_2 z_{1-p_2})$	$\frac{\Phi^{-1}(1-\alpha_2) - w_1\Phi^{-1}(1-p_1)}{w_2}$

Adjusted sample size

- If the trial proceeds $(\alpha_1 < p_1 \leq \beta_1)$ with conditional power β^*, the adjusted sample size is given by

$$n_2 = \left(C(\alpha_2, p_1) - \Phi^{-1}(1 - \beta^*)\right)^2 \frac{2\sigma^2}{\delta^2}$$

Generalized Likelihood Ratio (GLR) Procedure

Wald sequential probability ratio test (SPRT):

- Consider simple hypothesis: $H_0 : \theta = \theta_0$ vs. $H_a : \theta = \theta_1$ with type-I error probability α and type-II error probability β
- Set-up:
  - Observe $X_1, X_2, \cdots$
  - Define $R(X_i) = L(\theta_1,\ X_i)/L(\theta_0,\ X_i)$
  - $S_n = \sum_{i=1}^{n} \log R(X_i)$
- SPRT is specified by two boundary points a and b ($-\infty < a < 0 < b < \infty$) and the following decision rules:
  - $S_n \le a$: accept H_0
  - $S_n \ge b$: accept H_a
  - $a < S_n < b$: continue sampling
- Sample size $N \overset{\text{def}}{=} N(a, b) = \min\{n \ge 1 : S_n \le a \quad or \quad S_n \ge b\}$.

A Fundamental result

- It can be shown that N is finite with probability 1, i.e., $\lim_{n\to\infty} P_\theta(N > n) = 0$
- Among all sequential tests $T(\alpha, \beta) \in \Omega$, SPRT $= \underset{T(\alpha,\beta)\in\Omega}{\text{argmin}} \{n : T(\alpha, \beta)\}$:

Optimal property

SPRT has the smallest expected sample size under H_0 and H_a among all tests with the same type I and II error rates.

- Asymptotically,
  - $E_0(N) \approx \frac{|\log \beta|}{I(\theta_0, \theta_1)}$,
  - $E_a(N) \approx \frac{|\log \alpha|}{I(\theta_1, \theta_0)}$,

  where $I(\theta_0, \theta_1) = E_{\theta_0}[\log\{L(\theta_0)/L(\theta_1)\}]$ is called *Kullback-Leibler* information function.

Kullback-Leibler information function

- For binomial case:

$$I(\theta_0, \theta_1) = (1 - \theta_0) \log \left(\frac{1 - \theta_0}{1 - \theta_1} \right) + \theta_0 \log \left(\frac{\theta_0}{\theta_1} \right).$$

- For normal case:

$$I(\theta_0, \theta_1) = \frac{(\theta_0 - \theta_1)^2}{2}.$$

A three-stage adaptive design: stage 1

Efficient group-sequential tests (Bartroff and Lai, 2009) with maximum sample size M

- First stage sample size $n_1 = \xi M, \xi \in (0, 1)$. Based on n_1 and interim estimate $\hat{\theta}_{n_1}$,

$$\begin{cases} \text{Reject } H_0 & \text{if } n_1 < M, \quad \hat{\theta}_{n_1} > \theta_0, \quad \text{and } n_1 I(\hat{\theta} n_1, \theta_0) \geq b \quad (2) \\ \text{Accept } H_0 & \text{if } n_1 < M, \quad \hat{\theta}_{n_1} < \theta_1, \quad \text{and } n_1 I(\hat{\theta} n_1, \theta_1) \geq \tilde{b} \quad (3) \\ \text{continue} & \text{otherwise} \end{cases}$$

A three-stage adaptive design: stage 2

- The second stage sample size based on SPRT is

$$n_2 = m \vee \{M \wedge [(1+\rho_m)n(\hat{\theta}_m)]\},$$

where $\rho_m > 0$ is a small inflation of $n(\hat{\theta}_m)$ to adjust for the uncertainty in $n(\hat{\theta}_m)$. Based on n_2 and interim estimate $\hat{\theta}_{n_2}$,

$$\begin{cases} \text{Reject } H_0 & \text{if } n_2 < M,\ \hat{\theta}_{n_2} > \theta_0, \text{ and } n_1 I(\hat{\theta} n_2, \theta_0) \geq b \quad (4) \\ & \text{if } n_2 = M,\ \hat{\theta}_M > \theta_0, \text{ and } MI(\hat{\theta} n_M, \theta_0) \geq c \quad (5) \\ \text{Accept } H_0 & \text{if } n_2 < M,\ \hat{\theta}_{n_2} < \theta_1, \text{ and } n_2 I(\hat{\theta} n_2, \theta_1) \geq \tilde{b} \quad (6) \\ & \text{if } n_2 = M,\ \hat{\theta}_M < \theta_0, \text{ or } MI(\hat{\theta} n_M, \theta_0) < c \quad (7) \end{cases}$$

A three-stage adaptive design: stage 3

- The third stage sample size based on SPRT is $n_3 = M$.
- If $n_3 = M, \hat{\theta}_M > \theta_0$, and $MI(\hat{\theta}M, \theta_0) \geq c$, then reject H_0.
- Otherwise, accept H_0.
- The boundary values $b, \tilde{b}, c$ are determined by

$$\begin{aligned}
\mathrm{Pr}_{\theta_1}\{(3) \text{ or } (6)\} &= \tilde{\epsilon}\beta \quad (8)\\
\mathrm{Pr}_{\theta_0}\{(3)\ \&\ (6) \text{ do not occur, } (2)\ \&\ (4) \text{ occur}\} &= \epsilon\alpha \quad (9)\\
\mathrm{Pr}_{\theta_0}\{(2), (3), (4), \&\ (6) \text{ do not occur, } (5)\ \&\ (7) \text{ occur}\} &= (1-\epsilon)\alpha \quad (10)
\end{aligned}$$

- Here
  - ϵ and $\tilde{\epsilon}$ represents the fraction of type I and type II error rate to spend in the first two stages
  - Power and sample size depends minimally on the choice of ϵ and $\tilde{\epsilon}$ (Bartroff and Lai, 2008)
  - Typically [0.2, 0.8]

A numerical study

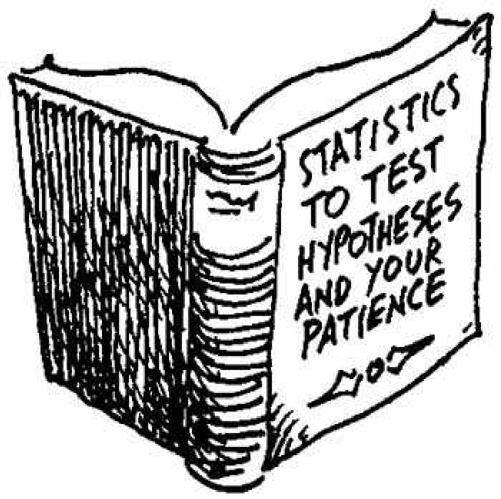
- Disclaimer: it is very *challenging* to evaluate and compare adaptive designs
    - Depends heavily on the choice of design parameters, first stage sample size, and maximum sample size
    - "... The less well-understood adaptive design methods are all based on unblinded interim analyses that estimate the treatment effect(s)." (FDA Draft guidance: Adaptive Design Clinical Trials for Drugs and Biologics, 2010)
- Three approaches are considered here:
    - Proschan and Hunsberger's (PH) conditional power method
    - Combination test approach
    - Generalized likelihood ratio test (GLR)
- For comparison purpose: fixed sample size (FSS) is included

Set-up

- Overall setting:
  - $X \sim N(\theta, \sqrt{0.5})$
  - $\alpha = 0.025$, $\beta = 0.9$
  - $m = 40$, and $M = 120$ (based on non-adaptive design)
- Metrics: Type I error rate, power, average sample size, number of stages, as well as median, $25th$ and $75th$ percentile of sample size
- Proschan and Hunsberger:
  - Let $p^* = 0.0436$, and efficacy boundary $k = 2.05$
- Combination tests:

Table: Combination Test Boundaries

Method	α_1	β_1	α_2
MSP	0.005	0.2	0.2030
MPP	0.01	0.2	0.0032
MINP	0.009	0.185	0.0195


STATISTICS
TO TEST
HYPOTHESES
AND YOUR
PATIENCE

GLR procedure set-up

- Boundaries for intermediate stage (total sample size $< M$): efficacy bound $b = 3.26$ and futility bound $\tilde{b} = 1.99$
- Final efficacy bound $c = 2.05$
- $\epsilon = \tilde{\epsilon} = 1/3$: type I and II error spending fraction
- $\rho_m = 0.1$: small inflation of sample size at the second stage

Results: type I error rate

Table: Power under $H_0 : \theta = 0$

Method	Power (%)	E(N)	$\mathrm{N}_{0.25}$	$\mathrm{N}_{0.5}$	$\mathrm{N}_{0.75}$	S
FSS_{120}	2.5	120	120	120	120	1
PH	2.4	41.1	40	40	40	1.02
MSP	2.5	54.4	40	40	40	1.20
MPP	2.5	54.5	40	40	40	1.20
MINP	2.5	51.2	40	40	40	1.20
GLR	2.5	75.1	40	60	120	1.64

- GLR results are from *Bartroff and Lai (2008)*
- Under H_0, on average GLR has the largest sample size

Results: power

Table: Power under $H_0 : \theta = 0.3$

Method	Power (%)	E(N)	$\mathrm{N}_{0.25}$	$\mathrm{N}_{0.5}$	$\mathrm{N}_{0.75}$	S
FSS_{120}	90.0	120	120	120	120	1
PH	55.2	46.8	40	40	40	1.14
MSP	77.9	76.9	40	76.6	120	1.60
MPP	84.9	75.7	40	67.8	120	1.50
MINP	75.1	63.0	40	48.4	80.4	1.50
GLR	88.8	89.2	40	118	120	1.91

GLR from Bartroff and Lai (2008)

Power under different θ

Table: Power under different θ

H_a		FSS_{120}	PH	MSP	MPP	MINP	GLR
$\theta = 0.15$	Power (%)	37.6	18.7	29.8	31.3	27.9	35.6
	E(N)	120	44.5	73.6	73.6	64.2	98.6
	S	1	1.09	1.50	1.49	1.41	2.05
$\theta = 0.2$	Power	60.0	30.2	47.1	51.0	44.4	57.2
	E(N)	120	45.9	77.7	77.3	66.0	99.4
	S	1	1.11	1.6	1.50	1.50	2.07
$\theta = 0.33$	Power	95.0	63.5	84.2	90.6	81.7	94.0
	E(N)	120	46.7	74.4	73.0	60.8	83.0
	S	1	1.13	1.6	1.50	1.50	1.81

Some observations

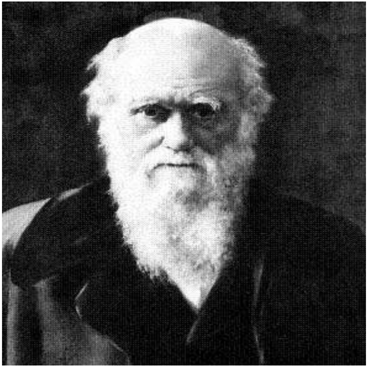
- Given the chosen m and M, conditional power based approaches are underpowered for values of θ considered here, especially PH tests
- Among combination tests, product method (e.g. Fisher's combination test) is more powerful
- GLR procedure has comparable power to FSS and is more powerful than other methods considered here
- However, GLR in general also has (considerably) larger expected sample size even when θ is only a fraction of target effect size
- GLR procedure also could extend to more than two stages, even when there is only marginal treatment effect
- Balance the trade-off between power loss and futility stop

Summary

- Under certain scenarios, conditional error function based approach can have non-trivial power loss
- Combination tests can recover most of the power loss if appropriate test method and boundaries are specified
- GLR achieves power comparable to fixed sample size design and hence are more powerful than PH, especially if the interim results are lukewarm
- Although 3 stages are likely, on average GLR extends to 2 stages
- GLR 3-stage procedure has features of both adaptive and group sequential tests
- Future directions:
  - Compare again variance spending approach with early futility stopping (Shen and Fisher, 1999)
  - Refine GLR procedure such that the number of subjects and stages are reduced under uninteresting θ

Why modify sample size?

"... *It's not the strongest species that survive, nor the most intelligent, but rather the ones most adaptable to change.*"
- Charles Darwin



Key references

- Bartroff, J., Lai, T.L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. in Med.* 27:1593-1611
- Proschan, M., Hunsberger, S. (1995). Designed extension of studies based on conditional power. *Biometrics* 51:1315-1324
- Chang, M. (2006). Adaptive designs based on sum of stagewise p-values. *Stat. in Med.* DOI:10.1002/sim.2755

Combination test and conditional error function

- To see the connection between these two, consider FCT
- Combination function is $C(p_1, p_2) = p_1 p_2$
- The corresponding conditional error function is given by

$$A(p_1) = \begin{cases} 1, & \text{if} \quad p_1 \leq \alpha_1 \\ c/p_1 & \text{if} \quad \alpha_1 < p_1 \leq \beta_1 \\ 0, & \text{if} \quad p_1 > \beta_1 \end{cases}$$