A Self–Designing Rule
for Clinical Trials with Arbitrary Response Variables

JOACHIM HARTUNG
Department of Statistics*, University of Dortmund
D-44221 Dortmund, Germany

Address:

Department of Statistics
University of Dortmund
D-44221 Dortmund, Germany

Phone:  ++49 231 755 3163
Fax:  ++49 231 755 5304
Email:  Hartung@statistik.uni-dortmund.de

*and project ’Meta–Analysis in Biometry and Epidemiology’ (SFB 475) of the German Research Community (DFG)
A Self–Designing Rule for Clinical Trials with Arbitrary Response Variables

Abstract. For testing one–sided but also two–sided hypotheses concerning several treatment arms in group sequentially performed clinical trials with arbitrary outcome variables, a general learning method is considered that allows for a complete self–designing of the study. All information available prior to a stage is used for estimating the sample size and the weight for the next step. In ’using up’ the variance, the test statistic is built in a bounded finite but random number of stages to test just once the null–hypothesis on rejecting.

1 Introduction

In a recent paper L. Fisher [1] introduces in a general setting for normal variables with known variances self–designing trials, for which Shen and L. Fisher [2] with regard to a one–sided hypothesis give a concrete proposal for building the test statistic. There the sequence of sample sizes is fixed prior to the beginning of the study, although in [1] there is already pointed out that this can be chosen adaptively using information prior to the respective stage. One continues to assign groups of subjects until the variance of the test statistic is ’used up’. Related for two–stage procedures or relative updating within one stage are [3], [4], [5], [6], [7], [8]. The aim of the self–designing procedures [1], [2] is not to test the null–hypothesis on a rejection after each stage like for instance in the classical group sequential trials, cf. [9], [10] and references given there.

An adaptive procedure designed for up to two interim analyses is given by Bauer and Köhne [11] for a general setting by use of p–values related to the tests carried out at each stage, which then are combined by R. A. Fisher’s method, cf. [12] and also [13].

In the present paper we employ the inverse normal transformation of the p–values suiting so under the null–hypothesis to the assumptions of L. Fisher’s [1] main
result. In [14] this p-value transformation is taken in connection with the classical
group sequential trials.

By a reformulation of the original hypothesis also two-sided cases can be considered
for arbitrary response variables in several treatment arms.

A general learning rule for completely self-designing trials is presented below which
at each stage adaptively estimates the sample size and the weight associated to that
stage upon all prior data knowledge.

2 The basic procedure

In a clinical trial let corresponding to a medication \(i\) be \(x_i\) an outcome variable
with mean \(\mu_i = \text{Ex}_i, \ i = 1, \ldots, I\). Denote
\[
\theta = \sum_{i=1}^{I} \left( \mu_i - \frac{1}{I} \cdot \sum_{j=1}^{I} \mu_j \right)^2,
\]
then the two-sided test problem \(H_{0, \theta} : \mu_1 = \ldots = \mu_I\) vs. \(H_{1, \theta} : \mu_{i_1} \neq \mu_{i_2}, \ i_1 \neq i_2,\)
for at least two \(i_1, i_2 \in \{1, \ldots, I\}\), becomes equivalent to the one-sided testing of
\(H_0 : \theta = 0\) vs. \(H_1 : \theta > 0,\)

for instance with the known homogeneity tests. In the case of \(I = 2\) and e.g. \(i = 1:\)
verum, \(i = 2:\) placebo, for a one-sided comparison one puts \(\theta = \mu_1 - \mu_2\). Note that
in the general formulation here \(\mu_i\) may represent a probability if the trial deals with
binary variables. The study is formally divided into an infinite number of disjoint
study parts: \(\text{stp}(1), \ldots, \text{stp}(k), \ldots,\) and it is the aim of a designing rule, that of
those only a finite number, say \(K\), has to be carried out really.

In \(\text{stp}(k)\) \(n_k\) patients are randomized across the \(I\) treatment groups, each con-
stisting so of \(n_{ik}\) patients, \(\sum_{i=1}^{I} n_{ik} = n_k\). Upon their responses \(x_{ik,1}, \ldots, x_{ik,n_{ik}}\),
\(i = 1, \ldots, I\) we test in \(\text{stp}(k)\) \(H_0\) vs. \(H_1\) by a, - with respect to \(H_0\) and \(H_1\) one-
sided -, test statistic \(T_k\), where large values of \(T_k\) may lead to a rejection of \(H_0.\)
Under \(H_0\) let \(T_k\) have a continuous distribution function \(F_{k,0}\) (otherwise the results
are known to tend to be usually somewhat conservative), then the p-values
\[
p_k = 1 - F_{k,0}(T_k)
\]
are uniformly distributed on the interval \((0, 1)\), such that

\[ z_k = \Phi^{-1}(1 - p_k) \]

is standard normally distributed, \(z_k \overset{H_0}{\sim} N(0, 1)\), where \(\Phi^{-1}\) denotes the inverse of the \(N(0, 1)\)-distribution function \(\Phi\), cf. [12].

For a quantity \(a\) to be used in or for \(\text{stp}(k)\) let us introduce the notation

\[ a = \hat{a}\{k - 1\}, \text{ i.e. } a = \hat{a}\{\text{stp}(0), \text{stp}(1), \ldots, \text{stp}(k - 1)\}, \]

to indicate that \(a\) is determined or estimated upon all the knowledge obtained in the previous study parts before the beginning of \(\text{stp}(k)\), where \(\text{stp}(0)\) may denote the prior information, implying \(\hat{a}\{0\}\) to be in any case a constant in the present trial.

Defining now an infinite sequence of nonnegative weights \(w_1, \ldots, w_k, \ldots\), such that with probability one under \(H_0\) there exists a finite (random) \(K\) with

\[ \sum_{k=1}^{K} w_k^2 = \sum_{k=1}^{\infty} w_k^2 = 1, \text{ where } w_k = \hat{w}\{k - 1\}, \]

then by theorem 1 of L. Fisher [1] we can deduce, that under \(H_0\) the statistic

\[ Z = \sum_{k=1}^{\infty} w_k \cdot z_k \]

is standard normally distributed,

\[ Z = \sum_{k=1}^{K} w_k \cdot z_k \overset{H_0}{\sim} N(0, 1). \]

That means, at given size \(\alpha_G\) the null-hypothesis \(H_0\) is rejected, if \(Z > \Phi^{-1}(1 - \alpha_G)\).

Furthermore, if the sample sizes \(n_k\) are determined upon data–knowledge from the previous study parts, this does not influence under \(H_0\) the distribution of the \(p\)-values \(p_k\), or of \(z_k\), and even not the independence of \(p_{k_1}, p_{k_2}, \ldots, z_{k_1}, z_{k_2}, k_1 \neq k_2\); cf. also the respective extensive discussions by L. Fisher [1], Proschan and Hundsberger [8], Bauer and Köhme [11]. Hence the statements above remain valid, if we allow: \(n_k = \hat{n}\{k - 1\}\).

In the case of multiple endpoints, i.e. with \(x_i\) a random vector, a multivariate extension is provided by putting \(\theta = \sum_{i=1}^{n} b_i^T b_i \text{ with } b_i = (\hat{q} - (1/I) \cdot \sum_{j=1}^{I} \theta_j) \) and \(b_i^T\) the transpose of the vector \(b_i\).
3  A general rule for completely self–designing

The distributions of $x_1, \ldots, x_I$ might depend on further parameters $\psi_1, \ldots, \psi_I$, for example variances (and correlations in a multivariate setting), and denote $\psi = (\psi_1, \ldots, \psi_I)$, analogously $\vartheta = (\vartheta_1, \ldots, \vartheta_I)$. For given Type I and II error rates $\alpha$ and $\beta$ let in dependence on the involved test statistic $T$ a sample size spending function $S = S_T$ be defined, such that by

$$n = S(\alpha, \beta \mid \vartheta, \psi, \theta \neq 0)$$

the smallest, finite $n$ is delivered such that in a sample of size $n$ the test of $H_0$ by $T$ has at least level $\alpha$ and power $1 - \beta$. For example in the simple originally two–sided normal case of $H_{0,\vartheta}$ with $I = 2$ this is the well known formula

$$n = n_1 + n_2 = 2 \cdot \left\{ \Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta) \right\} \cdot \sqrt{2} \cdot \sigma/(\vartheta_1 - \vartheta_2)^2$$

with $\sigma^2$ the common variance of $x_1, x_2$, respectively $n = [n_1 + 1] + [n_2 + 2]$ with $n_1 = n_2$, where $[m]$ denotes the largest natural number less than $m$.

The self–designing rule $R$ is characterized now by the sept–tuple

$$R = R(\alpha_G, \beta_G; n_1, w_1; \beta_g, \epsilon; \alpha_L),$$

that consists of the global Typ I and II error rates $\alpha_G$ and $\beta_G$, e.g. $\alpha_G = 0.05$, $\beta_G = 0.1$, the starting configuration $n_1 = \hat{n}\{0\}, w_1 = \hat{w}\{0\} \leq 1$ for stp(1), the Type II error rate $\beta_g \geq \beta_G$ for generating the sequential sample sizes $n_k = \hat{n}\{k - 1\}$, e.g. $\beta_g = 0.2$ or larger, where $\beta_g$ can also be defined in dependence of $k$, $\beta_g(\beta_g)$, further $\epsilon > 0$ is a lower bound for the weights $w_k$, $\epsilon < w_1$, e.g. $\epsilon = 0.1$ or $\sqrt{0.1}$, and $\alpha_L$ defines by $\Phi^{-1}(\alpha_L)$ a lower bound for $\sum_{j=1}^{k} z_j / \sqrt{k}$, i.e. if that statistic falls below the bound, $H_0$ is early accepted, e.g. $\alpha_L = 0.6$, or $\alpha_L = \alpha_L(k)$, increasing with $k$ starting e.g. even in zero, but cf. [2] for a detailed discussion of that point.

Extending and modifying the basic idea of Shen and L. Fisher [2], given for the one–sided normal case with known variances and the whole sequence of sample sizes fixed prior to the beginning of the trial, the rule $R$ procedure is derived as follows: Let $w_j, p_j, z_j$ be given for $j = 1, \ldots, k-1$, with $Z_{k-1} = \sum_{j=1}^{k-1} w_j \cdot z_j$, then if for
stp(k) in the equation

\[ P \left( Z_{k-1} + \sqrt{1 - \sum_{j=1}^{k-1} w_j^2 \cdot \tilde{z}_k > \Phi^{-1}(1 - \alpha_G)} \mid \hat{\theta} = \hat{\psi}\{k-1\} \right) = 1 - \beta \]

we would claim \( \beta = \beta_G \), so by putting \( w_k = w_{k,G} = \sqrt{1 - \sum_{j=1}^{k-1} w_j^2} \) we would have \( \sum_{j=1}^{k-1} w_j^2 + w_{k,G}^2 = 1 \) with \( Z_{k,G} = Z_{k-1} + w_{k,G} \cdot z_k(\beta_G) \) our final statistic, that would hold level \( \alpha_G \) and power \( 1 - \beta_G \), conditionally under \( \hat{\psi}\{k-1\}, \hat{\psi}\{k-1\}, \hat{\psi}\{k-1\} > 0 \); note that \( z_k(\beta_G) \) is obtained upon \( n_k(\beta_G) \) observations in stp(k), see below. In this way by letting \( \beta_g(k) \) go to \( \beta_G \) if \( k \) goes to some \( K \), the termination of the study can be accelerated.

Now usually we choose \( \beta_g > \beta_G \) in order to give the parameter estimates more chances to stabilize. Putting

\[ 1 - \hat{p}_k = \Phi(\tilde{z}_k) = \Phi \left( \frac{\Phi^{-1}(1 - \alpha_G) - Z_{k-1}}{\sqrt{1 - \sum_{j=1}^{k-1} w_j^2}} \right) \]

and

\[ S_k(\alpha, \beta) = S(\alpha, \beta \mid \hat{\psi}\{k-1\}, \hat{\psi}\{k-1\}, \hat{\psi}\{k-1\} > 0), \]

we define potential sample size numbers \( m_k \) and \( M_k \) for stp(k) by:

\[ m_k = S_k(\hat{p}_k, \beta_g), \quad M_k = S_k(\hat{p}_k, \beta_G). \]

At the power \( 1 - \beta_G \) these sample sizes would in stp(k) lead to the levels \( \hat{\alpha}_k(m_k) \) and \( \hat{\alpha}_k(M_k) \), respectively, given by the following implicit equations:

\[ m_k = S(\hat{\alpha}_k(m_k), \beta_G), \quad M_k = S(\hat{\alpha}_k(M_k), \beta_G), \quad m_k, M_k \text{ given.} \]

Introducing now recursively the weight function

\[ W(k) = \sqrt{1 - \sum_{j=1}^{k-1} W(j)^2} \cdot \frac{\Phi^{-1} \left( 1 - \frac{\hat{\alpha}_k(m_k)}{2} \right)}{\Phi^{-1} \left( 1 - \frac{\hat{\alpha}_k(M_k)}{2} \right)}, \quad W(1) = w_1 \leq 1 \text{ given,} \]
we set the control parameters weight \( w_k \) and sample size \( n_k \) for stp\((k)\) as follows:

\[
w_k = \begin{cases} 
W(k) & , \text{if } W(k) \geq \epsilon, \\
\sqrt{1 - \sum_{j=1}^{k-1} w_j^2} & , \text{if } W(k) < \epsilon,
\end{cases}
\]

and

\[
n_k = \begin{cases} 
m_k & , \text{if } W(k) \geq \epsilon, \\
M_k & , \text{if } W(k) < \epsilon.
\end{cases}
\]

If \( W(k) < \epsilon \), then put \( k = K \) and the trial stops after stp\((K)\), i.e. \( w_{K+j} = 0 \), for \( j = 1,2,\ldots \).

Then we get with the \( n_k \) patients in stp\((k)\) by the test statistic \( T_k \) the \( p \)-value \( p_k \), yielding the intermediate result \( Z_k = Z_{k-1} + w_k \cdot \Phi^{-1}(1 - p_k) \), or for \( k = K \) the final result \( Z_K \).

A generally longer running sequence of study parts we obtain if in the formula for \( W(k) \) we replace the second factor by \( \sqrt{m_k/M_k} \), or by the ratio \( \Phi^{-1}(F_{k,0}(E \ T_k(m_k))) / \Phi^{-1}(F_{k,0}(E \ T_k(M_k))) \), where \( E \ T_k(n) \) denotes the expectation of \( T_k \) with \( n \) patients in stp\((k)\) under \( \vartheta = \hat{\vartheta}\{k-1\} \) and \( \psi = \hat{\psi}\{k-1\} \).

Note that the random final number \( K \) of study parts to be really performed is bounded in any way by \( K < 2 + (1 - w_1^2) / \epsilon^2 \).

### 4 An illustrative example

Let us consider two medications with binary outcome variables, and \( \vartheta_1, \vartheta_2 \) be for instance the expected cure rates. We are interested in the two–sided test problem:

\[
H_{0,\vartheta}: \vartheta_1 = \vartheta_2 \quad \text{vs.} \quad H_{1,\vartheta}: \vartheta_1 \neq \vartheta_2,
\]

becoming one–sided by:

\[
H_0: \theta = (\hat{\vartheta}_1 - \vartheta_2)^2 = 0 \quad \text{vs.} \quad H_1: \theta > 0.
\]

Having at the beginning no real information, only a guess that not less than 200 patients would be involved, we choose the starting configuration as \( n_1 = n_{11} + n_{21} = 40 \)
patients, to be equally randomized, as in the following stages, too, across the two treatment groups, and the weight \( w_1^2 = 40/200 = 0.2 \). The other chosen control parameters can be seen in the rule

\[
R = R(\alpha_C = 0.05, \beta_1 = 0.1; 40, \sqrt{0.2}; \beta_2 = 0.25, \epsilon = 0.1; \alpha_L = 0.6).
\]

To keep the representation here more self-contained, we use the well known Arcus-Sinus Formula as approximate sample size \( n \) spending function \( S \),

\[
n = (\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta))^2/\delta^2,
\]

where \( \delta = \sin^{-1}(\sqrt{\theta_1}) - \sin^{-1}(\sqrt{\theta_2}) \), yielding for given \( n, \beta, \delta \) the explicit representation for the solution \( \tilde{\alpha}(n) \) by

\[
\Phi^{-1}(1 - \tilde{\alpha}(n)/2) = \sqrt{n \cdot \delta^2} - \Phi^{-1}(1 - \beta).
\]

Now instead of defining somehow the \( p \)-value of a two-sided test-statistic for testing \( H_{0,\theta} \), we take the \( \chi^2 \)–test statistic for \( (2 \times 2) \)–tables with \( n_k \) subjects as test statistic \( T_k \) in \( \text{stp}(k) \), being one–sided with respect to \( H_0 \) vs. \( H_1 \).

Hence we get in \( \text{stp}(k) \):

\[
1 - p_k = \chi^2_1(T_k) = 2 \cdot \Phi(\sqrt{T_k}) - 1, \quad \text{and} \quad z_k = \Phi^{-1}(1 - p_k).
\]

Denote \( \hat{\theta}_{ik} = \text{(number of cured patients in \( \text{stp}(k)/n_{ik} \)) the estimate of \( \theta_i, i = 1, 2 \) in \( \text{stp}(k) \), for simplicity here assumed to be constantly across all \( \text{stp}'s \) equal to: \( \hat{\theta}_1 = 0.7 \) and \( \hat{\theta}_2 = 0.5 \). So with \( n_1 = 40, w_1^2 = 0.2 \) we get in \( \text{stp}(1) \): \( T_1 = 1.67, 1 - p_1 = 0.8, z_1 = 0.84, Z_1 = 0.376 \), and therefore for \( \text{stp}(2) \): \( 1 - \hat{\rho}_2 = \Phi(1.43), 1 - \hat{\rho}_2/2 = 0.96, m_2 = 164, M_2 = 256, w_2 = 0.89 \cdot 1.28/1.92 = 0.59 \). Hence with \( n_2 = 164 \) patients in \( \text{stp}(2) \) we obtain: \( T_2 = 6.83, 1 - p_2 = 0.99, z_2 = 2.33, Z_2 = 1.75, \) and therefore for \( \text{stp}(3) \): \( 1 - \hat{\rho}_3 = \Phi(-0.15), m_3 = (0.6 + 0.036 = 46, M_3 = (0.6 + 1.28/0.036 = 98, \) and \( W(3) = 0.67 \cdot 0.035/0.598 = 0.039 < \epsilon \), so \( w_3 = 0.67, n_3 = M_3 = 98 \), giving in \( \text{stp}(3) \) the result: \( T_3 = 4.1 \) or \( z_3 = 1.75 \).

Thus the final test statistic takes on the form: \( Z_3 = 0.45 \cdot 0.59 \cdot 2.33 + 0.67 \cdot 1.77 = 2.94 = \Phi^{-1}(0.9984) \), achieved with \( 40 + 164 + 98 = 302 \) patients. At first we may wonder about the weights in relation to the respective number of patients. But this is a characteristic of the learning scheme: after \( \text{stp}(1) \) the algorithm is cautious in giving a weight to the second stage, however after the big \( \text{stp}(2) \) it has learned that
the parameter estimates remained as expected and so the next stage can get the rest of possible weight.

Secondly we see that we have not much payed for the learning in form of the number of patients to be enclosed: in a fixed sample size plan we would have calculated, with the same approximate formula, 292 necessary patients, only 10 less than above, if the parameter difference $\delta$ would have been known in advance. So we can get the impression that our learning algorithm is on the one hand cautious in spending high weights too early, and on the other hand it uses patients sparingly.

5 Conclusion

We propose a group sequential method for a complete self-designing of clinical trials as basically introduced by L. Fisher [1]. The termination of the whole trial is steered by a sequentially built weighting function until the variance of the test statistic is 'used up'. Based on the non-parametric character of the p-values our method applies to a wide range of situations.

References


