Adaptive Confidence Intervals of Desired Length and Power for Normal Means

Joachim Hartung[1] **and Guido Knapp**

Department of Statistics, Dortmund University of Technology, Dortmund, Germany

Abstract: In all empirical or experimental sciences, it is a standard approach to present results, additionally to point estimates, in form of confidence intervals on the parameters of interest. The length of a confidence interval characterizes the accuracy of the whole findings. Consequently, confidence intervals should be constructed to hold a desired length. Basic ideas go back to Stein (1945) and Seelbinder (1953) who proposed a two-stage procedure for hypothesis testing about a normal mean. Tukey (1953) additionally considered the probability or power a confidence interval should possess to hold its length within a desired boundary. In this paper, an adaptive multi-stage approach is presented that can be considered as an extension of Stein's concept. Concrete rules for sample size updating are provided. Following an adaptive two-stage design of O'Brien and Fleming (1979) type, a real data example is worked out in detail.

Keywords: Power of a confidence interval, Length of a confidence interval, Adaptive sample size planning, Multi-stage confidence interval, Group sequential trial

1 Introduction

The outcome of an experiment may be described by a normally distributed random variable X with unknown mean μ and unknown variance σ^2. Based on n independent replications of the experiment, the parameters μ and σ^2 are estimated and, being of main interest, a confidence interval on μ is derived. For a predefined confidence level, the length of the confidence interval stands for the accuracy of the whole estimation process. Therefore, it is an old problem to construct confidence intervals of a desired length. Stein (1945) provided a two-stage procedure, where the sample size of the second stage is based on the results of the first stage. Given some prior information on σ^2, Seelbinder (1953) showed how to choose the sample size of the first stage.

A question now is what is the probability to achieve such a confidence interval planned for a desired length. This problem was considered, for instance, in Hsu (1989) and already

[1] Address correspondence to Joachim Hartung, Department of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany; E-mail: hartung@statistik.tu-dortmund.de

Tukey (1953), mentioned by Hsu (1989), proposed to construct confidence intervals of given confidence level which had the desired length with a certain probability. Brown (1995) discussed the confidence level of the confidence interval on σ^2 from a first stage with regard to its use in the sample size planning for a second stage.

In the present paper, we use the dual relation between hypotheses testing and confidence intervals in order to provide confidence intervals of predefined confidence level which will have a length within some desired boundary with a required probability or power. Extending the two-stage concept of Stein (1945), we consider a multi-stage approach based on adaptive group sequential designs, see Hartung (2006). In doing so, the information from previous stages is used not only for planning the sample size for the next stage but also for computing the confidence intervals in the present stage. The confidence intervals are determined implicitly by combining parameterized p-values, see Cox and Hinkley (1974), obtained in the several stages. As combination method for the p-values, we apply the inverse normal method well known in meta-analysis, see for instance Hartung, Knapp, and Sinha (2008).

The outline of the present paper is as follows: In Section 2, one-stage confidence intervals of desired length and power for a normal mean are presented when a reliable estimate of the variance is known in advance. In Section 3, an adaptive group sequential approach is described which yields multi-stage confidence intervals for a normal mean of predefined level. These intervals are nested, so that their lengths are decreasing when the number of stages increases. In Section 4, maximum likelihood estimators of μ and σ^2 are presented. These estimators are median unbiased. In Section 5, adaptive planning is considered with respect to the desired length and power. We give concrete rules for sample size updating. In Section 6, the homogeneity of the means underlying the different stages of the trial is tested. In Section 7, a real data example, following an adaptive two-stage design of O'Brien and Fleming (1979) type, is worked out in detail. Further, some approximate formulas, helpful for computational purposes, are presented in connection with the example. Some additional comments are given in Section 8.

2 A Confidence Interval of Desired Length and Power when a Reliable Estimate of σ^2 is Known

Let X be a normally distributed random variable with mean μ and variance $\sigma^2 > 0$, $\bar{X}$ be the sample mean of $n \geq 2$ independent and identically distributed random variables, and S^2 be the sample variance, succinctly $X \sim \mathcal{N}(\mu, \sigma^2)$, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, where χ^2_ν stands for the χ^2-distribution with ν degrees of freedom.

Denote μ_0 a comparison value and $\Delta > 0$ an equivalence margin that is used here as an accuracy parameter for the length of the confidence interval. The length of the interval should be less than 2Δ. We are interested in hypotheses testing for noninferiority with regard to μ_0, that is,

$$\mathrm{H}_{0,\mathrm{L}} : \mu = \mu_0 - \Delta \quad \text{versus} \quad \mathrm{H}_{1,\mathrm{L}} : \mu > \mu_0 - \Delta, \tag{1}$$

and for nonsuperiority with regard to μ_0, that is,

$$\mathrm{H}_{0,\mathrm{U}} : \mu = \mu_0 + \Delta \quad \text{versus} \quad \mathrm{H}_{1,\mathrm{U}} : \mu < \mu_0 + \Delta. \tag{2}$$

Each test will be performed at level α, $0 < \alpha < 1/2$.

It holds

$$T_0(\mu) = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t_{n-1}, \tag{3}$$

that is, for the *true* parameter μ, the pivotal statistic $T_0(\mu)$ follows a central t-distribution with $n-1$ degrees of freedom. Let $t_{n-1;1-\alpha}$ denote the $(1-\alpha)$-quantile of the t_{n-1}-distribution, then the lower $(1-\alpha)$-confidence interval on μ is given as

$$\mathrm{I}_{0,\mathrm{L}}(\mu) = [\mu_\mathrm{L}, \infty)\,, \quad \mu_\mathrm{L} = \bar{X} - S\; t_{n-1;1-\alpha}/\sqrt{n}, \tag{4}$$

where μ_L solves $T_0(\mu_\mathrm{L}) = t_{n-1;1-\alpha}$. The null hypothesis $\mathrm{H}_{0,L}$ is rejected in favor of the alternative $\mathrm{H}_{1,L}$ in (1), if

$$T_0(\mu_0 - \Delta) > t_{n-1;1-\alpha} = T_0(\mu_\mathrm{L}), \quad \text{or equivalently} \quad \mu_\mathrm{L} > \mu_0 - \Delta. \tag{5}$$

Note that $T_0(\mu)$ is a (strictly) monotone decreasing function in μ.

With the upper $(1-\alpha)$-confidence interval on μ, that is,

$$\mathrm{I}_{0,\mathrm{U}}(\mu) = (-\infty, \mu_\mathrm{U}]\,, \quad \mu_\mathrm{U} = \bar{X} + S\; t_{n-1;1-\alpha}/\sqrt{n}, \tag{6}$$

we reject $\mathrm{H}_{0,\mathrm{U}}$ in favor of $\mathrm{H}_{1,\mathrm{U}}$ in (2), if

$$T_0(\mu_0 + \Delta) < -t_{n-1;1-\alpha} = T_0(\mu_\mathrm{U}), \quad \text{or equivalently} \quad \mu_\mathrm{U} < \mu_0 + \Delta. \tag{7}$$

Let us assume that a reliable estimate, say $s_0^2 > 0$, of σ^2 is known in advance. Then, conditioned on this assumption, the power of the test at $\mu = \mu_0$ should be $1-\beta$, $0 < \beta < 1$, in test problem (1) or (2). Note that $\mu - (\mu_0 - \Delta)$ is equal to Δ for $\mu = \mu_0$ in test problem (1). Then for testing $\mathrm{H}_{0,\mathrm{L}}$ versus $\mathrm{H}_{1,\mathrm{L}}$ using $T_0(\mu_0 - \Delta)$ from (3), the sample size n should be chosen as

$$n \geq n_0 = f_0(\alpha, \beta) := \frac{\left[\max\left\{0, \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right\}\right]^2}{\Delta^2/s_0^2}, \tag{8}$$

where Φ^{-1} is the inverse of the standard normal distribution function Φ. Note that we have used the normal sample size spending function in (8) for ease of presentation.

In the same way, we can proceed in test problem (2). Note that $(\mu_0 + \Delta) - \mu$ is equal to Δ for $\mu = \mu_0$ in this test problem and we obtain the same sample size formula as in (8) for testing $\mathrm{H}_{0,\mathrm{U}}$ versus $\mathrm{H}_{1,\mathrm{U}}$.

For $n \geq n_0$ and conditioned on $s_0^2 = \sigma^2$, both null-hypotheses will be rejected if $\mu = \mu_0$ with probability or power $1 - 2\beta$, $0 < \beta < 1/2$, implying

$$\mu_0 - \Delta < \mu_\mathrm{L} \leq \mu_\mathrm{U} < \mu_0 + \Delta \tag{9}$$

with μ_L and μ_U from (4) and (6). This means, the two-sided $(1-2\alpha)$-confidence interval $\mathrm{I}_0(\mu) = [\mu_L, \mu_U]$ has length $\mu_U - \mu_L < 2\Delta$ with (conditional) power $1 - 2\beta$ for $n \geq n_0$. Note that μ_0 has not to be explicitly known for constructing the interval $\mathrm{I}_0(\mu)$.

3 Multi-stage Confidence Intervals

The trial, with the underlying outcome variable X, is carried out consecutively in a number of independent stages, say K. In the i-th stage, $i = 1, \ldots, K$, we observe the sample mean $\bar{X}_i$ of the $n_i \geq 2$ responses, the sample variance S_i^2, and define the pivotal t-statistic

$$T_i(\mu) = \sqrt{n_i}\,\frac{\bar{X}_i - \mu}{S_i} \sim t_{n_i - 1}. \tag{10}$$

For combining the pivotal statistics, we apply the inverse normal method known from meta-analysis, see for instance Hartung, Knapp, and Sinha (2008).

Let F_{t_ν} denote the cumulative distribution function of a t-variable with ν degrees of freedom, then it holds, for the $1-p$-value,

$$F_{t_{n_i - 1}}\left(T_i(\mu)\right) \sim U(0,1), \qquad i = 1, \ldots, K, \tag{11}$$

where $U(0,1)$ stands for the uniform distribution in the unit interval. Consequently, we have

$$\Phi^{-1}\left[F_{t_{n_i-1}}\left(T_i(\mu)\right)\right] \sim \mathcal{N}(0,1), \quad i = 1, ..., K. \tag{12}$$

Since the stages are assumed to be independent, we can define the combining pivotal statistic

$$Z_j(\mu) = \sum_{i=1}^{j} \Phi^{-1}\left[F_{t_{n_i-1}}\left(T_i(\mu)\right)\right] \sim \sqrt{j}\,\mathcal{N}(0,1), \quad j = 1, \ldots, K. \tag{13}$$

Generally, let $Y_1, \ldots, Y_K$ be mutually independent $\mathcal{N}(0,1)$-distributed random variables. Then, given level α, $0 < \alpha < 1/2$, positive critical values $cv_1, \ldots, cv_K$ may be defined by the following probability condition:

$$P\left(\sum_{i=1}^{j} Y_i \leq cv_j \text{ for all } j = 1, ..., K\right) = 1 - \alpha, \tag{14}$$

see Hartung (2006).

Using critical values cv_j defined by (14), we get the following probability statements for the combination statistic (13),

$$P_\mu\left(Z_j(\mu) \leq cv_j \text{ for } j = 1, \ldots, k \leq K\right) \begin{cases} \geq 1 - \alpha & \text{for } k < K, \\ = 1 - \alpha & \text{for } k = K. \end{cases} \tag{15}$$

Using (15), we define the lower confidence sets on μ as

$$\mathrm{CI}_{k,\mathrm{L}}(\mu) = \left\{\tilde{\mu} \mid Z_j(\tilde{\mu}) \leq cv_j \text{ for } j = 1, ..., k\right\}, k = 1, \ldots, K, \tag{16}$$

and the confidence coefficient of $\mathrm{CI}_{k,\mathrm{L}}(\mu)$ is at least $1 - \alpha$ and exactly $1 - \alpha$ for $k = K$.

The functions $\Phi^{-1}(\cdot)$ and $F_{t_{n_i-1}}(\cdot)$ used in $Z_j(\mu)$ are (strictly) monotone increasing in their arguments. The pivotal statistic $T_i(\mu)$ from (10) is monotone decreasing in μ, implying that $\Phi^{-1}[F_{t_{n_i-1}}(T_i(\mu))]$ is monotone decreasing in μ. Consequently, the functions $Z_j(\mu), j = 1, \ldots, K$, are monotone decreasing in μ. Thus, for $k = 1, \ldots, K$, $\mathrm{CI}_{k,\mathrm{L}}(\mu)$ can be represented as an interval,

$$\mathrm{CI}_{k,\mathrm{L}}(\mu) = [\mu_{k,\mathrm{L}}, \infty) \tag{17}$$

where $\mu_{k,\mathrm{L}} = \max\{\mu_{\mathrm{L}}(1), ..., \mu_{\mathrm{L}}(k)\}$ and

$$\mu_L(j) \text{ solves } Z_j\left(\mu_L(j)\right) = cv_j, \quad j = 1, ..., k. \tag{18}$$

Note that the solutions in (18) are unique and can easily be found iteratively using standard statistical software.

The null-hypothesis $\mathrm{H}_{0,\mathrm{L}}$ from (1) will be rejected, at level of at most α, in favor of $\mathrm{H}_{1,\mathrm{L}}$ at stage j, $j \leq k \leq K$, if

$$Z_j(\mu_0 - \Delta) > cv_j = Z_j\left(\mu_L(j)\right), \quad \text{or equivalently,} \quad \mu_0 - \Delta < \mu_L(j) \leq \mu_{k,L}. \tag{19}$$

Thereby, we used that $Z_j(\mu)$ is monotone decreasing in μ. Note that μ_0 is not assumed to be known.

In a similar way, we define the upper confidence sets

$$\mathrm{CI}_{k,\mathrm{U}}(\mu) = \left\{\tilde{\tilde{\mu}} \mid -cv_j \leq Z_j(\tilde{\tilde{\mu}}) \text{ for } j = 1, \ldots, k\right\}, \quad k = 1, \ldots, K, \tag{20}$$

which have confidence coefficients of at least $1 - \alpha$ and exactly $1 - \alpha$ for $k = K$. For $k = 1, \ldots, K$, the interval representation is given by

$$\mathrm{CI}_{k,\mathrm{U}}(\mu) = (-\infty, \mu_{k,\mathrm{U}}], \tag{21}$$

where $\mu_{k,\mathrm{U}} = \min\{\mu_\mathrm{U}(1), ..., \mu_\mathrm{U}(k)\}$ and

$$\mu_\mathrm{U}(j) \text{ solves } Z_j(\mu_\mathrm{U}(j)) = -cv_j, \quad j = 1, \ldots, k. \tag{22}$$

The null-hypothesis $\mathrm{H}_{0,\mathrm{U}}$ from (2) will be rejected, at level of at most α, in favor of $\mathrm{H}_{1,\mathrm{U}}$ at stage j, $j \leq k \leq K$, if

$$Z_j(\mu_0 + \Delta) < -cv_j = Z_j(\mu_\mathrm{U}(j)), \quad \text{or equivalently,} \quad \mu_0 + \Delta > \mu_\mathrm{U}(j) \geq \mu_{k,\mathrm{U}}. \tag{23}$$

The two-sided confidence intervals on μ, defined as the intersection of the intervals (17) and (21), that is,

$$\mathrm{CI}_k(\mu) = \mathrm{CI}_{k,\mathrm{L}}(\mu) \cap \mathrm{CI}_{k,\mathrm{U}}(\mu) = [\mu_{k,\mathrm{L}},\ \mu_{k,\mathrm{U}}], \tag{24}$$

are nested, that is,

$$\mathrm{CI}_{k+1}(\mu) \subset \mathrm{CI}_k(\mu), \quad k = 1, ..., K-1, \tag{25}$$

and have confidence coefficients of at least $1 - 2\alpha$, $0 < \alpha < 1/2$. Note that the length of CI_k decreases when k increases, $k = 1, ..., K$.

Moreover, if both null-hypotheses in (1) and (2) are rejected at some stages $j_1, j_2 \leq k \leq K$ and $\mu_{k,\mathrm{L}} \leq \mu_{k,\mathrm{U}}$, then by (19) and (23), it holds

$$\mu_0 - \Delta < \mu_{k,\mathrm{L}} \leq \mu_{k,\mathrm{U}} < \mu_0 + \Delta. \tag{26}$$

Consequently, the length of the two-sided interval $\mathrm{CI}_k(\mu)$ is

$$\mu_{k,\mathrm{U}} - \mu_{k,\mathrm{L}} < 2\Delta. \tag{27}$$

Depending on the choice of α, it may occur that $\mu_{k,\mathrm{U}} < \mu_{k,\mathrm{L}}$, so that the interval CI_k is empty. For interpreting such an event, let us refer to Section 6.

4 Point Estimation of μ and σ^2

The combination statistic $Z_j(\mu)$ from (13) is $\mathcal{N}(0,j)$-distributed with mode and median 0. The maximum likelihood (ML)-estimator $\hat{\mu}_{\mathrm{ML}}(j)$ of the mean μ at stage j is given by:

$$\hat{\mu}_{\mathrm{ML}}(j) \quad \text{solves} \quad Z_j(\hat{\mu}_{\mathrm{ML}}(j)) = 0, \quad j = 1, ..., K. \tag{28}$$

The solution in (28) is unique. Sometimes, such an estimator is also called pseudo ML-estimator.

The *global* p-value at stage j is

$$p_G(j) = 1 - \Phi\left(Z_j(\mu)/\sqrt{j}\right), \quad j = 1, \ldots, K, \tag{29}$$

and solving (29) for μ such that $p_G(j) = 1/2$ yields $\hat{\mu}_{\mathrm{ML}}(j)$ as solution. Since $Z_j(\mu)$ is monotone in μ,

$$\hat{\mu}_{ML}(j) \text{ is median unbiased,} \quad j = 1, \ldots, K, \tag{30}$$

see Cox and Hinkley (1974, p. 273), that is, the ML-estimator $\hat{\mu}_{ML}(j)$ lies with equal probability as well below the parameter μ as above μ.

Recall that the variance estimator S_i^2 at stage i, $i = 1, \ldots, K$, is a scaled χ^2-distributed random variable, that is, $(n_i - 1)\ S_i^2/\sigma^2 \sim \chi^2_{n_i-1}$. In analogy to (11), we obtain

$$F_{\chi^2_{n_i-1}}\left((n_i - 1)\ \frac{S_i^2}{\sigma^2}\right) \sim U(0,1), \quad i = 1, \ldots, K,$$

where $F_{\chi^2_{n_i-1}}$ denotes the distribution function of a $\chi^2_{n_i-1}$-variable. Like in (13), it holds for the combining statistic up to stage j

$$Z_j^V(\sigma^2) = \sum_{i=1}^{j} \Phi^{-1}\left[F_{\chi^2_{n_i-1}}\left((n_i - 1)\ \frac{S_i^2}{\sigma^2}\right)\right] \sim \mathcal{N}(0,j), \quad j = 1, \ldots, K. \tag{31}$$

Moreover, $Z_j^V(\sigma^2)$ is monotone decreasing in $\sigma^2 > 0$. Consequently, the ML-estimator $\widehat{\sigma^2_{ML}}(j)$ of σ^2 at stage j is given by:

$$\widehat{\sigma^2_{ML}}(j) \quad \text{solves} \quad Z_j^V\left(\widehat{\sigma^2_{ML}}(j)\right) = 0, \quad j = 1, ..., K. \tag{32}$$

Again, $\widehat{\sigma^2_{ML}}(j)$ is a median unbiased estimator for σ^2.

5 Adaptive Sample Size Planning to Attain the Desired Power

Let $f_j(\alpha, \beta)$ denote the sample size spending function from (8) at stage j, $1 \leq j \leq K-1$, when s_0^2 is replaced by some estimate $S(j)^2$ of σ^2 based on information of the stages $0, 1, \ldots, j$, where stage 0 stands for prior information. For example, $S(j)^2$ may be chosen as the ML-estimate $\widehat{\sigma^2_{ML}}(j)$ from (32), or as the pooled estimate up to stage j given by

$$\widehat{\sigma^2_{Pool}}(j) = \frac{1}{\sum_{h=1}^{j} n_h - j} \sum_{i=1}^{j} (n_i - 1)\; S_i^2.$$

Recall now from (14) the event

$$A := \left\{ \sum_{i=1}^{h} Y_i \leq cv_h \text{ for all } h = 1, ..., K \right\},$$

and let us consider the event for an arbitrary, but fixed stage j

$$B := \left\{ \sum_{i=1}^{h} Y_i \leq cv_h \text{ for all } h = 1, ..., j-1 \text{ and } \sum_{i=1}^{j-1} Y_i + \sum_{i=j}^{K} Y_i \leq cv_K \right\}.$$

Clearly, the probability of event B is larger than of event A. Moreover, collapsing $\sum_{i=j}^{K} Y_i$, which is $\mathcal{N}(0, K-(j-1))$-distributed, to Y_j and giving all the remaining weight to Y_j, we obtain

$$\left\{ \sum_{i=1}^{h} Y_i \leq cv_h \text{ for } h = 1, ..., j-1, \text{ and } \sum_{i=1}^{j-1} Y_i + \sqrt{(K-(j-1))}\; Y_j \leq cv_K \right\}$$
$$\supset \left\{ \sum_{i=1}^{h} Y_i \leq cv_h \text{ for all } h = 1, ..., K \right\}. \tag{33}$$

Consequently, if we decide after stage $(j-1)$ to omit the interim analyses j up to $K-1$, we can assign the remaining weight $\sqrt{K-(j-1)}$ to the next final stage and build the next test statistic according to (13) as

$$Z_{j,K}(\mu_0 - \Delta) = Z_{j-1}(\mu_0 - \Delta) + \sqrt{(K-j+1)}\; \Phi^{-1}\left[F_{t_{n_j - 1}}\left(T_j(\mu_0 - \Delta)\right)\right], \tag{34}$$

where $Z_{j,K}(\mu_0 - \Delta) \sim \sqrt{K}\; \mathcal{N}(0,1)$ under $\mathrm{H_{0,L}}$ from (1), $j = 1, ..., K$, and $Z_0 = 0$. The test statistic $Z_{j,K}(\mu_0 - \Delta)$ has to be compared with the K-th critical value cv_K in testing $\mathrm{H_{0,L}}$ from (1).

Note that the p-value of testing $\mathrm{H}_{0,\mathrm{L}}$ at stage i by use of $T_i(\mu_0 - \Delta)$ is given as

$$p_i = p_i(\mu_0 - \Delta) = 1 - F_{t_{n_i - 1}}\left(T_i(\mu_0 - \Delta)\right), \quad i = 1, ..., K. \tag{35}$$

Assume that, after stage $(j-1)$, in a next stage we want to reach cv_K by use of the final test statistic

$$\hat{Z}_{j,K}(\mu_0 - \Delta) = Z_{j-1}(\mu_0 - \Delta) + \sqrt{(K - j + 1)}\ \Phi^{-1}\left(1 - \hat{p}_{j,K}(\mu_0 - \Delta)\right), \tag{36}$$

then the projected p-value $\hat{p}_{j,K}(\mu_0 - \Delta)$ of the next trial part should be

$$\hat{p}_{j,K}(\mu_0 - \Delta) = 1 - \Phi\left[\left(cv_K - Z_{j-1}(\mu_0 - \Delta)\right)/\sqrt{(K - j + 1)}\right]. \tag{37}$$

Conditioned an $S(j-1)^2$, an estimate of σ^2, a power of $1 - \beta$ in testing $\mathrm{H}_{0,\mathrm{L}}$ from (1) is reached for $\mu = \mu_0$ when the sample size of the next final stage is chosen at least as

$$M_{j,\mathrm{L}}(\mu_0 - \Delta) := f_{j-1}\left(\hat{p}_{j,K}(\mu_0 - \Delta), \beta\right), \tag{38}$$

where $f_{j-1}\left(\hat{p}_{j,K}(\mu_0 - \Delta), \beta\right)$ is the sample size from (8) with α replaced by the projected p-value $\hat{p}_{j,K}(\mu_0 - \Delta)$.

Similarly, we derive for testing $\mathrm{H}_{0,\mathrm{U}}$ from (2) the projected p-value

$$\hat{p}^*_{j,K}(\mu_0 + \Delta) = 1 - \Phi\left[\left(-cv_K - Z_{j-1}(\mu_0 + \Delta)\right)/\sqrt{(K - j + 1)}\right]. \tag{39}$$

Whereas $H_{0,L}$ from (1) will be rejected when the α-level of the next final stage, say $\alpha_{j,K}$, satisfies $\alpha_{j,K} \leq \hat{p}_{j,K}(\mu_0 - \Delta)$, the null-hypothesis $H_{0,U}$ from (2) will be rejected when $\alpha_{j,K} \leq 1 - \hat{p}^*_{j,K}(\mu_0 + \Delta)$. So conditioned on $S(j-1)^2$, a power of $1 - \beta$ in testing $\mathrm{H}_{0,\mathrm{U}}$ from (2) is reached for $\mu = \mu_0$ when the sample size of the next final stage is chosen at least as

$$M_{j,\mathrm{U}}(\mu_0 + \Delta) := f_{j-1}\left(1 - \hat{p}^*_{j,K}(\mu_0 + \Delta), \beta\right). \tag{40}$$

Consequently, both null-hypotheses in (1) and (2) will be rejected with (conditional) power of at least $1 - 2\beta$, $0 < \beta < 1/2$, for $\mu = \mu_0$ if the sample size of the next final stage is chosen at least as,

$$M_j(\mu_0) = \max\left\{M_{j,\mathrm{L}}(\mu_0 - \Delta), M_{j,\mathrm{U}}(\mu_0 + \Delta)\right\}. \tag{41}$$

In case we do not want to finish the trial in this way and have in mind the originally planned $K - (j-1)$ further stages, we will choose the sample size of stage j proportionally as

$$n_j = n_j(\mu_0) = \frac{M_j(\mu_0)}{K - j + 1}, \quad j = 1, ..., K, \tag{42}$$

which is a (slightly) conservative choice according to (33), and use cv_j in (18) and $-cv_j$ in (22) to compute the confidence interval $\mathrm{CI}_j(\mu)$ from (24). Note that the sample size in each stage should be at least 2.

Especially for $j = 1$, we obtain the projected p-values $\hat{p}_{1,K} = 1 - \Phi(cv_j/\sqrt{K})$ and $1 - \hat{p}^*_{1,K} = \Phi(-cv_j/\sqrt{K}) = \hat{p}_{1,K}$. Consequently we get the starting sample size of our trial as

$$n_1 = M_1/K \tag{43}$$

where, see (8),

$$M_1 = \left(\frac{cv_K}{\sqrt{K}} + \Phi^{-1}(1-\beta)\right)^2 s_0^2/\Delta^2,$$

with $0 < \beta < 1/2$ and $s_0^2 > 0$ is a prior guess of σ^2.

In applications, we use the following algorithm in a trial planned for at most K stages: We start with n_1 observations, n_1 from (43), and compute the first confidence interval CI_1. When the length of CI_1 is below 2Δ, we finish the trial. Otherwise, we apply the above proceeding for the stages $j \geq 2$ until that stage k when the length of CI_k is the first time below 2Δ. Then we can finish the trial because all confidence intervals computed so far possess a confidence coefficient of at least $1 - 2\alpha$, see Section 3. Not later than stage $k = K$, we will receive a two-sided confidence interval $\mathrm{CI}_k(\mu)$ with confidence coefficient of at least $1-2\alpha$, see (24), which with (conditional) probability or power of at least $1-2\beta$, $0 < \beta < 1/2$, will have the desired length below 2Δ, see (27).

In the sample size planning for stage j, $j \geq 2$, we use the median unbiased ML-estimate $\hat{\mu}_{ML}(j-1)$ of μ from stage $(j-1)$, see (28). These estimators are used for calculating the projected p-values $\hat{p}_{j,K}$, see (37), and $\hat{p}^*_{j,K}$, see (39), so that in (42), we finally get $n_j = n_j(\hat{\mu}_{ML}(j-1))$, $j = 2, \ldots, K$. The power in (38) and (40) will then be conditioned on $S(j-1)^2$ and $\hat{\mu}_{ML}(j-1)$ for $j = 2, \ldots, K$. Note that these estimates are used only for planning the sample sizes, but not for computing the confidence intervals.

Further, we may formally define the p-values, see (11), as suiting to the null-hypothesis that μ is the *true* parameter, see Cox and Hinkley (1974, p. 221). So we may apply the general result that under the null-hypothesis p-values preserve their distribution and independence (for continuous null-distributions) when sample sizes are chosen adaptively in a consecutive way, see Brannath, Posch, and Bauer (2002). Since all our procedures are based on these p-values, all our statements remain valid when sample sizes are chosen adaptively as demonstrated in this section, see also Hartung (2006).

6 Homogeneity of the Means in the Different Stages

Let us consider the extended model that we allow the parameters μ and σ^2 to be different in each stage , say μ_i and σ_i^2 in stage i, so that it holds for the sample means

$$\bar{X}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{n_i}\right), \; i = 1, \ldots, K. \tag{44}$$

Since $\bar{X}_i$ and the sample variance S_i^2 are stochastically independent unbiased estimators of μ_i und σ_i^2, respectively, the pivotal t-statistic $T_i(\mu_i)$, see (10), is t_{n_i-1}-distributed, $i = 1, \ldots, K$. Further, the combination statistic up to stage j, see (13),

$$Z_j(\mu_1, \ldots, \mu_j) = \sum_{i=1}^{j} \Phi^{-1}\left[F_{t_{n_i}-1}\left(\sqrt{n_i}\,\frac{\bar{X}_i - \mu_i}{S_i}\right)\right], \; j = 1, \ldots, K, \tag{45}$$

is $\mathcal{N}(0, j)$-distributed.

Denote $m' = (m_1, \ldots, m_k)$ the transposed of a vector m in $\mathbb{R}^k$, then by (14), the k-dimensional confidence region, $k = 1, \ldots, K$,

$$\mathrm{CR}_k = \{m \in \mathbb{R}^k \mid -cv_j \leq Z_j(m_1, \ldots, m_j) \leq cv_j, \text{ for } j = 1, \ldots, k\} \tag{46}$$

covers $(\mu_1, \ldots, \mu_k)'$ with probability of at least $1 - 2\alpha$, $0 < \alpha < 1/2$. Note that CR_k is not empty for all $\alpha \in (0, 1/2)$, since the observed vector $(\bar{x}_1, \ldots, \bar{x}_k)'$ lies always in CR_k.

When we assume that the parameters μ_i are really identical, say $\mu_i = \mu$ for $i = 1, \ldots, k$, then the k-dimensional parameter $(\mu, \ldots, \mu)'_k$ is covered by CR_k, or, in other words, $(\mu, \ldots, \mu)'_k \in \mathrm{CR}_k$ with probability of at least $1 - 2\alpha$. But this is equivalent to $\mu \in \mathrm{CI}_k$ with probability of at least $1 - 2\alpha$, where CI_k is introduced in (24). Thus, if CI_k is empty for a common confidence level $1 - 2\alpha$, this will speak against the assumption of an identical mean μ over the first k stages. This can formally be stated as a test on homogeneity of the means.

In testing

$$\mathrm{H}_{0,\mathrm{hom}}(k): \; \mu_1 = \ldots = \mu_k \quad \text{versus} \quad \mathrm{H}_{1,\mathrm{hom}}(k): \; \mu_{i_1} \neq \mu_{i_2} \tag{47}$$

for some $i_1, \; i_2 \in \{1, \ldots, k\}, \; k = 2, \ldots, K$, the homogeneity hypothesis $\mathrm{H}_{0,\mathrm{hom}}(k)$ will be rejected at level of at most 2α, if the two-sided confidence interval $\mathrm{CI}_k(\mu)$ from (24) is empty. If $\mathrm{H}_{0,\mathrm{hom}}(k^*)$ is rejected, then also $\mathrm{H}_{0,\mathrm{hom}}(k)$ will be rejected for $k^* \leq k \leq K$. On the other hand, under the model assumption of an identical mean μ underlying the different stages of the trial, the probability to obtain an empty confidence interval CI_k is bounded by 2α. The same test principle is used by Hartung and Knapp (2003) in deriving a test on homogeneity of variances of random treatment-by-sample interactions.

7 A Real Data Example

Let us consider an application one of the authors was concerned with. The effect of a drug for treating patients with asthma bronchiale is analysed with respect to a lung function parameter called FEV$_1$, that is, forced expiratory volume in 1 second, measured in liter (ℓ), and an underlying approximate normal distribution of the outcome can be assumed.

From a small pre-study with selected patients, we have the rough estimates of 2.5ℓ for the mean and $s_0 = 0.6\ell$ for the standard deviation. A study was planned 'to determine, with a safety of 90%, the mean with a reliability of 95% within an accuracy of $\pm 0.2\ell$.' This means in our setting: $\alpha = 0.025$, $\beta = 0.05$, and $\Delta = 0.2$.

Since larger variances were expected with an extended spectrum of patients, the decision was made in favor of an adaptive two-stage plan of O'Brien and Fleming (1979) type, see Hartung (2006). Using the combination statistic (13), we get the constant critical values $cv_1 = cv_2 = 2.797$ in (14), see Hartung (2006) or Jennison and Turnbull (2000).

By formula (43), we get the starting sample size of the trial as $n_1 = 60$ using the prior guess $s_0 = 0.6$. In the first stage, we observed for the mean $\bar{x}_1 = 2.67\ell$ and the standard deviation $s_1 = 0.87\ell$, so that equating

$$Z_1(\mu) = \Phi^{-1}\left[F_{t_{59}}\left(\sqrt{60}\,\frac{2.67-\mu}{0.87}\right)\right]$$

to 2.797 and to -2.797, see (18) and (22), yields the first confidence interval on the mean as

$$\mathrm{CI}_1 = [2.3437\ell,\ 2.9963\ell].$$

Replacing μ_0 by $\hat{\mu}_{ML}(1) = \bar{x}_1$, we compute

$$Z_1(\mu_0 - \Delta) = Z_1(\bar{x}_1 - 0.2) = \Phi^{-1}\left[F_{t_{59}}\left(\sqrt{60}\,\frac{0.2}{0.87}\right)\right] = 1.7500$$

and thus the projected p-value, see (37),

$$\hat{p}_{2,2}(\mu_0 - \Delta) = 1 - \Phi(2.797 - 1.7500) = 0.1476,$$

and the projected p-value from (39),

$$\hat{p}^*_{2,2}(\mu_0 + \Delta) = 1 - \Phi(-2.797 + 1.7500) = 0.8524,$$

with $Z_1(\mu_0 + \Delta) = -1.7500$.

Since $\hat{p}_{2,2}(\mu_0 - \Delta) = 1 - \hat{p}^*_{2,2}(\mu_0 + \Delta)$, the sample size of the second and final stage should be at least, see (42) and (8),

$$n_2 = n_2(\mu_0) = f_1(0.1476, 0.05) = \frac{\left[\Phi^{-1}(1-0.1476) + \Phi^{-1}(1-0.05)\right]^2}{\Delta^2/s_1^2} = 137.111.$$

With $n_2 = 138$ patients in the second stage, we observed the estimates $\bar{x}_2 = 2.70\ell$ and $s_2 = 0.81\ell$, and equating

$$Z_2(\mu) = Z_1(\mu) + \Phi^{-1}\left[F_{t_{137}}\left(\sqrt{138}\,\frac{2.7-\mu}{0.81}\right)\right]$$

to 2.797 and to -2.797 yields the final confidence interval on the mean as

$$\mathrm{CI}_2 = [2.5681\ell, 2.8081\ell].$$

Note that the actual length of CI_2 (0.24ℓ) is below the desired accuracy or length of $2\Delta = 0.4\ell$.

Instead of solving nonlinear equations to determine the boundaries of the confidence intervals, we provide some simple approximate solutions in the following. Let us approximate the central t-distributions involved in the combination statistics by normal distributions with the same first two moments.

The variance of the t_{n_i-1}-variate is $(n_i-1)/(n_i-3)$ and the variance of $\bar{X}_i$ is estimated by S_i^2/n_i. Let us define weights

$$w_i = \sqrt{\frac{(n_i-3)\ n_i}{(n_i-1)\ S_i^2}}, \quad n_i \geq 4, \quad i = 1, \ldots, K, \tag{48}$$

and thus, the combination statistic $Z_j(\mu)$ from (13) can be approximated by

$$Z_j^A(\mu) = \sum_{i=1}^{j} w_i(\bar{X}_i - \mu),\ j = 1, \ldots, K. \tag{49}$$

Equating $Z_j^A(\mu)$ to the critical value cv_j and to $-cv_j$ and solving for μ yield the approximate boundaries, see (18) and (22), for $j = 1, \ldots, K$,

$$\mu_L^A(j) = \sum_{i=1}^{j} \frac{w_i\ \bar{X}_i}{\sum_{h=1}^{j} w_h} - \frac{cv_j}{\sum_{h=1}^{j} w_h} \quad \text{and} \quad \mu_U^A(j) = \sum_{i=1}^{j} \frac{w_i\ \bar{X}_i}{\sum_{h=1}^{j} w_h} + \frac{cv_j}{\sum_{h=1}^{j} w_h}. \tag{50}$$

Furthermore, the approximate median unbiased ML-estimator at stage j, see (28), is given by

$$\hat{\mu}_{ML}^A(j) = \sum_{i=1}^{j} \frac{w_i\ \bar{X}_i}{\sum_{h=1}^{j} w_h},\ j = 1, \ldots, K. \tag{51}$$

Note that, in combining the means, the inverse estimated standard errors are used in the weights and not the inverse estimated variances of the means as known from the standard estimator of the overall mean in meta-analysis, see for instance Hartung, Knapp, and Sinha (2008). Weighted means like $\hat{\mu}^A_{ML}(j)$ from (51) are used in the generalized Cochran-Wald statistics considered by Hartung, Böckenhoff, and Knapp (2003).

In the example, we obtain the weights $w_1 = 8.7512$ and $w_2 = 14.3966$, so that the approximate confidence intervals on the mean are given as $\text{CI}^A_1 = [2.3504\ell,\ 2.9896\ell]$ and $\text{CI}^A_2 = [2.5678\ell,\ 2.8095\ell]$. Note that the approximate confidence intervals are nearly identical to the exact confidence intervals, especially the final intervals CI_2 and CI^A_2.

The midpoints of the approximate intervals are approximate ML-estimates of the mean, that is, $\hat{\mu}^A_{ML}(1) = 2.67\ell$ in the first stage and $\hat{\mu}^A_{ML}(2) = 2.6887\ell$ in the second stage. Whereas the exact $\hat{\mu}_{ML}(1)$ is identical to $\hat{\mu}^A_{ML}(1)$, the exact ML-estimate at the second stage, by equating $Z_2(\mu) = 0$ and solving for μ, takes on the value $\hat{\mu}_{ML}(2) = 2.6886\ell$. Again, the approximate solution nearly exactly coincides with the exact solution.

The ML-estimates of the variance parameter σ^2 are $\hat{\sigma}^2_{ML}(1) = 0.8749^2$ and $\hat{\sigma}^2_{ML}(2) = 0.8367^2$, which are the solutions of, see (32),

$$
\begin{aligned}
Z^V_1(\sigma^2) &= \Phi^{-1}\left[F_{\chi^2_{59}}\left(59\,\frac{0.87^2}{\sigma^2}\right)\right] = 0, \text{ and} \\
Z^V_2(\sigma^2) &= Z^V_1(\sigma^2) + \Phi^{-1}\left[F_{\chi^2_{136}}\left(136\,\frac{0.81^2}{\sigma^2}\right)\right] = 0.
\end{aligned}
$$

8 Final Remarks

In Section 3, we defined positive one-sided critical values $cv_j,\ j = 1, \ldots, K$, by the probability condition (14). For a fixed number of stages K and an overall significance level α, we get an O'Brien and Fleming (1979) design with constant critical values in (14), say $cv_j = \text{cons}_{OBF}(K, \alpha)$, and a Pocock (1977) design with monotone increasing critical values given as $cv_j = \sqrt{j}\ \text{cons}_{PO}(K, \alpha),\ j = 1, \ldots, K$, see Hartung (2006), where also some of these one-sided critical values are tabulated. Designs with intermediate values of the critical values are considered, for instance, in Jennison and Turnbull (2000).

Usually, the two-sided critical values at level 2α for the correspondent symmetric two-sided tests are tabulated in literature. For $K \geq 2$, these two-sided critical values are slightly smaller than the one-sided critical values at level α. At least for $\alpha \leq 0.05$, these two-sided critical values may be used for practical applications, see Jennison and Turnbull (2000, p. 192).

We defined the two-sided confidence interval CI_k as the intersection of the one-sided intervals $\mathrm{CI}_{k,L}$ and $\mathrm{CI}_{k,U}$, see (24), and the confidence coefficient of CI_k is at least $1-2\alpha$. If we use the critical values of the correspondent two-sided tests at level 2α, we get a two-sided confidence interval, say CI_k^0, that is slightly narrower than CI_k for $K \geq 2$, but has a confidence coefficient being at least $1-2\alpha$ as well. Moreover, CI_K^0 reaches a confidence coefficient of exactly $1-2\alpha$. However, using now the boundaries of CI_k^0 in our testing considerations (13) and (23), the test level α cannot be guaranteed. Indeed, no severe differences are expected for practical applications at least for $\alpha \leq 0.05$, see above.

In our presentation, sample sizes n are computed using a normal approximation for applying a t_{n-1}-variate. Nearly exact values are achieved by correcting the sample size n with the variance of a t_{n-1}-variate, that is, replacing n by $n_{\mathrm{corr}} = n(n-1)/(n-3),\ n \geq 4$. The idea behind is the same as in replacing a t-variate by a normal variate with identical variance.

References

[1] Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 236–244.

[2] Brown, R. H. (1995). On use of pilot sample for sample size determination. *Statistics in Medicine* **14**, 1933–1940.

[3] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, New York.

[4] Hartung, J. (2006). Flexible designs by adaptive plans of generalized Pocock- and O'Brien-Fleming-type and by Self-designing clinical trials. *Biometrical Journal* **48**, 521–535.

[5] Hartung, J., Böckenhoff, A., and Knapp, G. (2003). Generalized Cochran-Wald statistics in combining of experiments. *Journal of Statistical Planning and Inference* **113**, 215–237.

[6] Hartung, J. and Knapp, G. (2003). Confidence regions on the variance components in an extended ANOVA model for combining information. *Acta Applicandae Mathematicae* **78**, 207–221.

[7] Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications.* Wiley, New York.

[8] Hsu, J. C. (1989). Sample size computation for designing multiple comparison experiments. *Computational Statistics and Data Analysis* **7**, 79–91.

[9] Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. CRC Press Inc., Boca Raton.

[10] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

[11] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

[12] Seelbinder, B. M. (1953). On Stein's two-stage sampling scheme. *Annals of Mathematical Statistics* **24**, 640–649.

[13] Stein, Ch. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258

[14] Tukey, J. W. (1953). *The Problem of Multiple Comparisons.* Unpublished manuscript, cited in Hsu (1989).