Adaptive Confidence Intervals of Desired Length and Power for Normal Means

Guido Knapp and Joachim Hartung

Department of Statistics
TU Dortmund University
44221 Dortmund, Germany
guido.knapp@tu-dortmund.de
joachim.hartung@tu-dortmund.de

Abstract. In all empirical or experimental sciences, it is a standard approach to present results in form of confidence intervals on the parameters of interest additionally to point estimates. 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) 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 and a real data example is worked out in detail.

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

1 Introduction

Let us consider a normally distributed random variable X with unknown mean μ and unknown variance σ^2. Based on n independent replications, a confidence interval on μ is derived and the length of the confidence interval, given a predefined confidence coefficient, stands for the accuracy of the whole estimation process. 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. A crucial point in studying the overall performance of any statistical inference is the choice of the sample size of the first stage. Given some prior information on σ^2, Seelbinder (1953) showed how to choose the sample size of the first stage. Moshman (1958) also made some attempts to formulate an approach to determine the sample size of the first stage. Recently, Mukhopadhyay (2005) introduced an easy-to-implement criterion through the Fisher information for the determination of a pilot sample size.

The question arises what is the probability to achieve a confidence interval planned for a desired length. Already Tukey (1953), mentioned in Hsu (1989), proposed to construct confidence intervals of given confidence level which have the desired length with a certain probability. 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). Using all the available data from previous stages, we perform the sample size planning for the next stage and the computation of the confidence intervals after each 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 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 Section 3, an adaptive group sequential approach is described which yields multi-stage confidence intervals for a normal mean of predefined level. In Section 4, adaptive planning is considered with respect to the desired length and power. We give concrete rules for sample size updating. In Section 5, a real data example, following an adaptive two-stage design of O'Brien and Fleming (1979) type, is worked out in detail.

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

Let be $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, where $\bar{X}$ is the sample mean of n independent and identically distributed normal random variables, S^2 the sample variance , and χ^2_ν denotes a χ^2-distributed random variable with ν degrees of freedom.

Denote μ_0 a comparison value and $\Delta > 0$ an equivalence margin used here as the accuracy parameter for the length of the confidence interval. The length should be less than 2Δ. Let us consider the test problem for noninferiority with regard to μ_0, that is,

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

and for nonsuperiority with regard to μ_0, that is,

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

Note that, for the *true* parameter μ, we have

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

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

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

and the upper $(1-\alpha)$-confidence interval as

$$\mathrm{I}_{0,\mathrm{U}}(\mu) = (-\infty, \mu_\mathrm{U}]\,, \quad \mu_\mathrm{U} = \bar{X} + S\,t_{n-1;1-\alpha}/\sqrt{n}. \tag{5}$$

We reject $\mathrm{H}_{0,\mathrm{L}}$ in (1) at level α iff $\mu_\mathrm{L} > \mu_0 - \Delta$ and $\mathrm{H}_{0,\mathrm{U}}$ in (2) at level α iff $\mu_\mathrm{U} < \mu_0 + \Delta$.

Assume a reliable estimate of σ^2, say $s_0^2 > 0$, is given and the power of the test at $\mu = \mu_0$ should be $1-\beta$, $0 < \beta < 1$, in (1) and (2). Then, the sample size n should be chosen in both test problems as

$$n \geq n_0 = f_0(\alpha, \beta) := \frac{\left[\max\left\{0, \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right\}\right]^2}{\Delta^2/s_0^2}, \tag{6}$$

where Φ^{-1} denotes the inverse of the standard normal distribution function Φ.

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

$$\mu_0 - \Delta < \mu_\mathrm{L} \leq \mu_\mathrm{U} < \mu_0 + \Delta. \tag{7}$$

Consequently, 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 power $1-2\beta$ for $n \geq n_0$.

3 Multi-stage Confidence Intervals

Let us consider a trial which is carried out consecutively in a number of independent stages, say K. In the i-th stage, $i = 1, \ldots, K$, let be $\bar{X}_i$ the sample mean of $n_i \geq 2$ independent and identically distributed normal random variables, and S_i^2 the sample variance. Consider the pivotal t-quantity

$$T_i(\mu) = \sqrt{n_i}\,\frac{\bar{X}_i - \mu}{S_i} \sim t_{n_i-1}. \tag{8}$$

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,

$$F_{t_{n_i-1}}\left[T_i(\mu)\right] \sim U(0,1), \quad i = 1, \ldots, K, \tag{9}$$

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

$$\Phi^{-1}\left[F_{t_{n_i-1}}\left(T_i(\mu)\right)\right] \sim \mathcal{N}(0,1), \quad i = 1, ..., K. \tag{10}$$

Since the stages of the trial are independent, we define the combining pivotal quantity

$$Z_j(\mu) = \sum_{i=1}^{j} \Phi^{-1}\left[F_{t_{n_i-1}}\left(T_i(\mu)\right)\right] \sim \sqrt{j}\,\mathcal{N}(0,1), \quad j = 1, \ldots, K. \tag{11}$$

Let us consider critical values cv_j such that following probability statements hold

$$P_\mu\Big(Z_j(\mu) \le cv_j \text{ for } j = 1, \ldots, k \le K\Big) \begin{cases} \ge 1-\alpha & \text{for } k < K, \\ = 1-\alpha & \text{for } k = K. \end{cases} \tag{12}$$

Since $Z_j(\mu)$ is a multiple of the standard normal distribution, the critical values cv_j can be borrowed from classical group sequential trials, see Hartung (2006), for $j = 1, \ldots, K$. Using (12), the lower confidence sets on μ are then defined as

$$\text{CI}_{k,\text{L}}(\mu) = \{\tilde{\mu} \mid Z_j(\tilde{\mu}) \le cv_j \text{ for } j = 1, ..., k\}, k = 1, \ldots, K, \tag{13}$$

and the upper confidence sets as

$$\text{CI}_{k,\text{U}}(\mu) = \left\{\tilde{\tilde{\mu}} \mid -cv_j \le Z_j(\tilde{\tilde{\mu}}) \text{ for } j = 1, \ldots, k\right\}, \quad k = 1, \ldots, K. \tag{14}$$

The confidence coefficients of $\text{CI}_{k,\text{L}}(\mu)$ and $\text{CI}_{k,\text{U}}(\mu)$ are at least $1-\alpha$ and exactly $1-\alpha$ for $k = K$.

Since the functions $Z_j(\mu), j = 1, \ldots, K$, are monotone decreasing in μ, $\text{CI}_{k,\text{L}}(\mu)$ from (13) can be represented as an interval, namely,

$$\text{CI}_{k,\text{L}}(\mu) = [\mu_{k,\text{L}}, \infty) \tag{15}$$

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

$$\mu_L(j) \text{ solves } Z_j\left(\mu_L(j)\right) = cv_j, \quad j = 1, ..., k. \tag{16}$$

By analogy, $\text{CI}_{k,\text{U}}(\mu)$ from (14) can be represented as an interval, namely,

$$\text{CI}_{k,\text{U}}(\mu) = (-\infty, \mu_{k,\text{U}}], \tag{17}$$

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

$$\mu_\text{U}(j) \text{ solves } Z_j(\mu_\text{U}(j)) = -cv_j, \quad j = 1, \ldots, k. \tag{18}$$

The two-sided confidence intervals on μ, defined as the intersection of the intervals (15) and (17), that is,

$$\text{CI}_k(\mu) = \text{CI}_{k,\text{L}}(\mu) \cap \text{CI}_{k,\text{U}}(\mu) = [\mu_{k,\text{L}},\ \mu_{k,\text{U}}], \tag{19}$$

are nested, that is,

$$\text{CI}_{k+1}(\mu) \subset \text{CI}_k(\mu), \quad k = 1, ..., K-1. \tag{20}$$

The confidence coefficient of each interval $\text{CI}_k(\mu)$ is at least $1-2\alpha$, $0 < \alpha < 1/2$. Moreover, if both null-hypotheses in (1) and (2) are rejected at some stages $j_1, j_2 \le k \le K$ and $\mu_{k,\text{L}} \le \mu_{k,\text{U}}$, it holds

$$\mu_0 - \Delta < \mu_{k,\text{L}} \le \mu_{k,\text{U}} < \mu_0 + \Delta. \tag{21}$$

Consequently, the length of the two-sided interval $\text{CI}_k(\mu)$ is $\mu_{k,\text{U}} - \mu_{k,\text{L}} < 2\Delta$.

4 Adaptive Sample Size Planning to Attain the Desired Power

Let $f_j(\alpha, \beta)$ denote the sample size spending function from (6) at stage j, $1 \leq j \leq K-1$, when s_0^2 is replaced by some estimate $S(j)^2$ of σ^2. The estimate $S(j)^2$ is based on information from all the previous stages $0, 1, \ldots, j$, where stage 0 stands for prior information. A possible choice of $S(j)^2$ is the pooled variance estimate up to stage j given by

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

Assume that we decide after stage $(j-1)$ that the interim analyses j up to $K-1$ should be omitted. Then, we can assign the remaining weight $\sqrt{K-(j-1)}$ to the next and final stage and build the test statistic according to (11) as

$$Z_{j,K}(\mu_0 - \Delta) = Z_{j-1}(\mu_0 - \Delta) + \sqrt{(K-j+1)}\, \Phi^{-1}\left[F_{t_{n_j-1}}\left(T_j(\mu_0 - \Delta)\right)\right], \tag{22}$$

where $Z_{j,K}(\mu_0 - \Delta) \sim \sqrt{K}\, \mathcal{N}(0,1)$ under $\mathrm{H}_{0,\mathrm{L}}$ 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 $\mathrm{H}_{0,\mathrm{L}}$ 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

$$p_i = p_i(\mu_0 - \Delta) = 1 - F_{t_{n_i-1}}\left(T_i(\mu_0 - \Delta)\right), \quad i = 1, ..., K. \tag{23}$$

Assume that in the next and final stage, the final test statistic

$$\hat{Z}_{j,K}(\mu_0 - \Delta) = Z_{j-1}(\mu_0 - \Delta) + \sqrt{(K-j+1)}\, \Phi^{-1}\left(1 - \hat{p}_{j,K}(\mu_0 - \Delta)\right), \tag{24}$$

coincides with the critical value cv_K, then the projected p-value $\hat{p}_{j,K}(\mu_0 - \Delta)$ of the final stage must be

$$\hat{p}_{j,K}(\mu_0 - \Delta) = 1 - \Phi\left[\left(cv_K - Z_{j-1}(\mu_0 - \Delta)\right) / \sqrt{(K-j+1)}\right]. \tag{25}$$

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

$$M_{j,\mathrm{L}}(\mu_0 - \Delta) := f_{j-1}\left(\hat{p}_{j,K}(\mu_0 - \Delta), \beta\right), \tag{26}$$

where $f_{j-1}\left(\hat{p}_{j,K}(\mu_0 - \Delta), \beta\right)$ is the sample size from (6) with α replaced by the projected p-value $\hat{p}_{j,K}(\mu_0 - \Delta)$.

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

$$\hat{p}^*_{j,K}(\mu_0 + \Delta) = 1 - \Phi\left[\left(-cv_K - Z_{j-1}(\mu_0 + \Delta)\right) / \sqrt{(K-j+1)}\right]. \tag{27}$$

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 at $\mu = \mu_0$ when the sample size of the next and final stage is chosen at least as

$$M_{j,\mathrm{U}}(\mu_0 + \Delta) := f_{j-1}\left(1 - \hat{p}^*_{j,K}(\mu_0 + \Delta), \beta\right). \tag{28}$$

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

$$M_j(\mu_0) = \max\left\{M_{j,\mathrm{L}}(\mu_0 - \Delta), M_{j,\mathrm{U}}(\mu_0 + \Delta)\right\}. \tag{29}$$

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, the sample size of stage j is then proportionally chosen as

$$n_j = n_j(\mu_0) = \frac{M_j(\mu_0)}{K-j+1}, \quad j = 1, ..., K. \tag{30}$$

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

$$n_1 = M_1/K \tag{31}$$

where, see (6),

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

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 (31), 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 $\text{CI}_k(\mu)$ with confidence coefficient of at least $1-2\alpha$, see (19), which will have the desired length below 2Δ, see (21), with (conditional) probability or power of at least $1-2\beta$, $0 < \beta < 1/2$.

Note that we also use estimates of the mean μ based on all data from previous stages to determine the sample size for the next stage, see the example in the next section.

5 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.

A small pre-study yielded the rough estimates of 2.5ℓ for the mean and $s_0 = 0.6\ell$ for the standard deviation. The 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 $\Delta = 0.2\ell$. An adaptive two-stage plan of O'Brien and Fleming (1979) type was planned, see Hartung (2006). Using (11), the constant critical values are $cv_1 = cv_2 = 2.797$ satisfying (12).

By (31), the starting sample size of the trial is $n_1 = 60$ using the prior guess $s_0 = 0.6$. In the first stage, we observed the mean $\bar{x}_1 = 2.67\ell$ and the standard deviation $s_1 = 0.87\ell$. Equating

$$Z_1(\mu) = \Phi^{-1}\left[F_{t_{59}}\left(\sqrt{60}\,\frac{2.67-\mu}{0.87}\right)\right] \tag{32}$$

to 2.797 and to -2.797 and solving for μ yields the confidence interval on the mean as

$$\text{CI}_1 = [2.3437\ell,\ 2.9963\ell]. \tag{33}$$

Replacing μ_0 through $\bar{x}_1$, we compute

$$Z_1(\mu_0 - \Delta) = Z_1(\bar{x}_1 - 0.2) = \Phi^{-1}\left[F_{t_{59}}\left(\sqrt{60}\,\frac{0.2}{0.87}\right)\right] = 1.7500 \tag{34}$$

and thus the projected p-values

$$\hat{p}_{2,2}(\mu_0 - \Delta) = 1 - \Phi(2.797 - 1.7500) = 0.1476, \tag{35}$$

and

$$\hat{p}^*_{2,2}(\mu_0 + \Delta) = 1 - \Phi(-2.797 + 1.7500) = 0.8524, \tag{36}$$

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

$$n_2 = n_2(\mu_0) = f_1(0.1476, 0.05) = \frac{\left[\Phi^{-1}(1 - 0.1476) + \Phi^{-1}(1 - 0.05)\right]^2}{\Delta^2/s_1^2} = 137.111. \tag{37}$$

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

$$Z_2(\mu) = Z_1(\mu) + \Phi^{-1}\left[F_{t_{137}}\left(\sqrt{138}\,\frac{2.7 - \mu}{0.81}\right)\right] \tag{38}$$

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

$$\text{CI}_2 = [2.5681\ell, 2.8081\ell], \tag{39}$$

whose actual length is below the desired accuracy or length of $2\Delta = 0.4\ell$.

References

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

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.

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

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

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

Moshman, J. (1958). A method for selecting the size of the initial sample in Stein's two-sample procedure. *Annals of Mathematical Statistics*, **29**, 667–671.

Mukhopadhyay, N. (2005). A new approach to determine the pilot sample size in two-stage sampling. *Communications in Statistics—Theory and Methods*, **34**, 1275–1295.

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

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

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.

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