Conditional Inference Following Group Sequential Testing

Pamela A. Ohman*and George Casella†
Biometrics Unit
Cornell University
Ithaca, NY 14850
Telephone: (607)255-5488
BUM # 1331

April 18, 1996

*Supported by NSF Training Grant DMS-956682; pao1@cornell.edu
†Supported by NSF Grant DMS-9305547; george@amanita.biom.cornell.edu

Conditional Inference Following Group Sequential Testing

Abstract

The work of Fisher[1] and Buehler[2] discuss the importance of conditioning on recognizable subsets of the sample space. The stopping time is an easily identifiable divider of the sample space when considering group sequential testing. We present confidence intervals which are correct when conditioning on the subset of data such that a trial stopped at a particular analysis. Although these intervals may not be practical, they do have very desirable properties for observations which are highly unusual (given any value of the mean). In addition, they provide insight into how information about the mean is distributed between the two sufficient statistics. Conditional coverage probabilities are used as a way to compare the sample mean, stagewise, and repeated confidence intervals. However, none of these intervals outperforms the others when conditioning on stopping time.

Keywords: recognizable subsets, confidence intervals

1 Introduction

In medical and industrial settings a group sequential approach to hypothesis testing is often used as a method of ending an experiment as soon as significant results are observed[3 4 5 6]. This testing approach results in two sufficient statistics for one parameter of interest, that is, a curved family. Consequently, inference following a sequential trial is not straightforward. Methods of finding point estimates and confidence intervals for the mean, upon termination of a group sequential trial, have proliferated. A review and comparison of both point estimates and confidence intervals is given by Emerson and Fleming[7].

Here, we propose yet another confidence interval. This interval is based on ideas stemming from the work of Kiefer[8] who questions the classical approach of averaging over all possible results and Fisher[1] who stresses consideration of "recognizable subsets". Kiefer[8] and Berger[9] argue that in some cases the classical approach of averaging over all possible results gives incorrect perceptions about the amount of information that is given by the data. Certain characteristics of the data may give us some meaningful information about the parameter of interest which gets lost when averaging over outcomes that have not been observed. Further, Buehler[2], Brown[10], and Olshen[11], in particular, discuss the importance of studying the behavior of confidence statements conditional on subsets of the sample space. Buehler[2] proposed these ideas in the context of betting strategies. Brown[10] demonstrates that the

$t-$interval does not have correct conditional coverage when conditioning on the fact that the sample mean falls within some interval about zero. In fact, the coverage is uniformly below the stated level. Goutis and Casella[12] show how to use this information, about the sample mean falling within an interval about zero, to construct improved t inference from Student's $t-$interval. Similarly, Olshen[11] studied the probability that Scheffé's $S-$method interval covers the parameter of interest conditioned on the fact that a preliminary $F-$test rejects the null hypothesis. Recall that the $S-$method creates simultaneous confidence intervals for a number of parameters with correct coverage probabilities and is often used following an F-test that rejects the null hypothesis that all of the parameters are equal to zero. Here the "recognizable" subset of the sample space is all the possible values of the data such that an $F-$ test would reject the null hypothesis. Olshen found instances where the conditional coverage probability was always less than the unconditional probability, i.e., the stated confidence. In the context of group sequential analysis, the stopping time naturally divides up the sample space into distinct subsets. Additionally, this is a natural conditioning set since, in a sense, conditioning on the stopping time is analogous to conditioning on the sample size.

An overview of the basic structure of group sequential tests is provided in Section 2. Section 3 presents the conditional confidence intervals, giving a summary of their properties. Section 4 compares the conditional coverage probabilities of previously proposed confidence intervals. The questions asked in this section are: Do the

confidence intervals display correct coverage probabilities when conditioning on the fact that we know the stopping time (which indeed we do)? Are these coverages uniformly above or below the stated confidence level? Do any of these intervals outperform the others as determined by conditioning on stopping time? Remaining discussion and conclusions are given in Section 5.

2 Basics

Group sequential allows the investigor to test a single hypothesis at multiple times during the course of the experiment, with the possibility of stopping early when significant results are observed. An overall pre-specified level of significance is maintained. Typically, there is a maximum period of time in which the experiment is to be conducted and the interim analyses are conducted at a few selected times, often equally spaced, during that period of time. Each interim analysis is usually conducted only after a large number of measurements have been accumulated.

For example, consider testing the null hypothesis that some mean is equal to zero, $H_0 : \mu = 0$, against the alternative $H_a : \mu \neq 0$ based on observations which are assumed independently and identically distributed as normal with known variance σ^2. Let X_{ij} represent the i'th observation during the j'th time interval. Further, assume that there are an equal number of observations ($i = 1, \ldots, n$) for each time

interval and the maximum number of analyses allowed is m ($j = 1, \ldots, m$). In a group sequential test, n is usually large enough for σ^2 to be considered known. To further summarize the data, let $Y_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$. Note that the Y_j are independently and identically distributed since they are linear combinations of equal numbers of the independent and identically distributed X_{ij}'s. More precisely, we assume

$$X_{ij} \sim N\left(\mu, \sigma^2\right) \text{ i.i.d.}$$

so that

$$Y_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \sim N\left(\mu, \sigma^2/n\right)$$

Then, to further summarize the data available at each analysis k, let S_k represent the cumulative sum of these observed means at analysis k. These new statistics are no longer independent since each S_k is the sum of the previous statistic S_{k-1} and the newly observed Y_k. The distribution of these cumulative sums can be shown to be multivariate normal such that

$$S_k = \sum_{j=1}^{k} Y_j \sim N\left(k\mu, k\sigma^2/n\right)$$

and the covariance of S_i and S_j, such that $i \neq j$, has the form

$$\text{Cov}\left(S_i, S_j\right) = \min(i, j) \times \sigma^2/n$$

For the remainder of our discussion, we will restrict our attention to the S_k's since they completely summarize the data. Also, without loss of generality, we take $\sigma^2/n = 1$ in all of our calculations.

At each analysis, a Z-test based on the statistic S_k can be used to test the null hypothesis. Equivalently, we can find critical values for each analysis such that if S_k falls outside those values then the null hypothesis is rejected. Since, the test is repeated several times with correlations existing between all the test statistics, corrections must be made to the individual tests with the correlation structure in mind in order to maintain some overall level of significance α.

Lan and DeMets[13] provide a general framework to determine which critical values to use at each analysis. First, one chooses an overall α-level and the rate at which one wants to "spend" the α and write this as a function of the time. For example, this function may be constructed so that the α is spent conservatively, i.e., very little of α is spent at the early analyses. For simplicity, we will continue to assume that the m analyses are equally spaced so that we can replace time by the indices of the analyses, $k = 1, \ldots, m$. Then, taking $\alpha(k)$ to be some function of k such that $\alpha(0) = 0$ and $\alpha(m) = \alpha$, $\pi_k = \alpha(k) - \alpha(k-1)$ is the amount of α that is spent at the k'th analysis. (Note that $\pi_1 + \cdots + \pi_m = \alpha$.) One can then find the appropriate critical values, c_k for each analysis by calculating the probabilities

$$\begin{aligned} \pi_1 &= \Pr\left[|S_1| \geq c_1\right] \\ \pi_2 &= \Pr\left[|S_1| < c_1, |S_2| \geq c_2\right] \\ &\text{etc} \quad \ldots \end{aligned}$$

Alternatively, one could specify some relationship between the critical values, e.g.,

$c_1 = c_2 = \cdots = c_m$, and then determine their values by considering the probability

$$\Pr\left[|S_1| \geq c_1 \; or \ldots or \; |S_m| \geq c_m\right] = \alpha$$

Note, for example, that $|S_2| \geq c_2$ can be observed only if $|S_1| < c_1$, and thus the event that $|S_1| \geq c_1$ is mutually exclusive from the event that $|S_1| < c_1$ and $|S_2| \geq c_2$, and so on. Hence,

$$\begin{aligned}
&\Pr\left[|S_1| \geq c_1 \; or \ldots or \; |S_m| \geq c_m\right] \\
&\quad = \Pr\left[|S_1|\right] + \Pr\left[|S_1| < c_1, |S_2| \geq c_2\right] + \\
&\qquad \cdots + \Pr\left[|S_1| < c_1, |S_2| < c_2, \ldots, |S_m| \geq c_m\right] \\
&\quad = \pi_1 + \pi_2 + \cdots + \pi_m
\end{aligned}$$

(which equals α as noted before). These critical values $(c_1, \ldots, c_m)$ form a set of boundaries for the observed S_k, $k = 1, \ldots, m$. Once S_k goes outside the interval $(-c_k, c_k)$, the null hypothesis is rejected and the experiment ends. Thus the set of critical values $(c_1, \ldots, c_m)$ in a sequential test is sometimes referred to as the boundary conditions for the test.

Various sets of boundary conditions have been proposed and their properties with respect to power, sample size, and practicality have been examined. Two of the most commonly studied are those of Pocock[14] and O'Brien and Fleming[15].

The Pocock boundaries are among the least conservative, in that when using them it is relatively easy to reject at early analyses. These boundaries were originally

formed by letting $c_k = c \times \sqrt{k}$, where c is a constant chosen so that the overall α significance level is maintained. Lan and DeMets[13] found this to correspond to a spending rate of approximately $\alpha(k) = \alpha \times \log\{1 + (e-1) \times k/m\}$. In the case such that $\alpha = .05$ and the maximum number of analyses, m, is 4, $c = 2.361$. Thus the null hypothesis would be rejected at the k'th analysis if the absolute value of S_k is greater than $2.361 \times k$.

The O'Brien-Fleming boundaries correspond to a horizontal boundary, which is to say that $c_1 = c_2 = \cdots = c_m = c$. The spending rate which gives this approximate relationship between the critical values is[13]

$$\alpha(k) = \begin{cases} 0 & \text{if } k = 0 \\ 2 - 2 \times \Phi\left(z_{\alpha/2}\right) \Big/ \sqrt{k/m} & \text{if } k = 1, \ldots, m \end{cases}$$

When $m = 4$, one can calculate c to be 4.048. Here, it is fairly difficult to reject the null hypothesis at the early analyses and thus is a more conservative set of boundaries.

Recall that when using any of these boundary conditions, the trial will be continued until either $|S_k|$ exceeds c_k for some $k = 1, \ldots, m$ or the final analysis m is reached. Thus, for any particular $k > 1$, S_k will be observed if and only if the previous successive sums $S_1, \ldots, S_{k-1}$ all fall inside the critical values. Note that at the end of the trial, only the values of the stopping time T, and the sum at that time, S_T, are recorded. Thus, the density function for (T, S) can be written in an iterative manner.

(The subscript T will be dropped from S_T. except when needed for clarity. Also, we will use (t, s) to denote the observed statistics.) Starting at the first analysis,

$$f_1(s|\mu) = \frac{1}{\sqrt{2\pi\sigma^2/n}} exp\left\{-\frac{1}{2}\frac{(s-\mu)^2}{\sigma^2/n}\right\},$$

where $f_t(s|\mu)$ refers to the density of (T, S) when $T = t$. We can then write the densities for (t, S) such that $t > 1$, using an iterative formulation as follows

$$f_t(s|\mu) = \int_{-c_{t-1}}^{c_{t-1}} \frac{1}{\sqrt{2\pi\sigma^2/n}} exp\left\{-\frac{1}{2}\frac{(s-s_{t-1}-\mu)^2}{\sigma^2/n}\right\} f_{t-1}(s_{t-1})\, ds_{t-1}$$

Observe that the statistics (T, S) are sufficient and form a curved parametric family.

3 Conditional Confidence Intervals

These conditional confidence intervals will be defined conditional on two elements. Suppose that the sequential test is stopped at the t'th analysis, and the value of s is observed. Then given the time of the final analysis t and the direction of s away from the null value, an ordering of the sample space is defined as follows:

Definition 1 *For fixed $T = t$ and $s > \mu_0$, a value (t, s^*) will be considered more extreme than (t, s) if*

- $s^* > s$

For fixed $T = t$ and $s < \mu_0$, a value (t, s^*) will be considered more extreme than (t, s) if

- $s^* < s$

Otherwise, (t, s^*) is less extreme.

This ordering, conditional on the stopping time and the direction of the observation away from the null, leads to the stochastic ordering of the conditional probability with respect to the mean μ, as given in the following theorem for negative values of s.

Theorem 1 *For any constant $c_t > 0$ and $s < -c_t < 0$,*

$$\Pr[S < s | S < -c_t, T = t, \mu] \tag{1}$$

is a decreasing function of μ.

Likewise, for any constant $c_t > 0$ and $s > c_t > 0$,

$$\Pr[S > s | S > c_t, T = t, \mu] \tag{2}$$

is an increasing function of μ.

That is to say, that with respect to μ, these conditional probabilities are stochastically ordered.

For proof of this theorem, see Appendix 1.

Since the conditional probabilities are stochastically ordered with respect to μ under these orderings, $1-\alpha$ confidence intervals can be formed using the guaranteeing an interval method[16]. That is for $s < -c_t$, we can find a μ_L and μ_U that satisfy

$$\Pr[S < s | S < -c_t, \mu_L] = 1 - \alpha/2$$

and

$$\Pr[S < s | S < -c_t, \mu_U] = \alpha/2.$$

The coverage of (μ_L, μ_U) conditioned on the stopping time t and direction of s from μ_0 will then be $1-\alpha$. Of course, these conditional intervals are also correct unconditionally.

As examples, these intervals were found using GAUSS[17] for the set-up previously described using O'Brien-Fleming boundaries and assuming that the trial ended at the first or second analyses. These intervals are given for a range of s/t values in Table 1.

First, note that all previously proposed unconditional orderings, which will be discussed in part in Section 4, reduce to this ordering after conditioning on the stopping time.

Table 1: Conditional 95% Confidence Intervals for $t = 1, 2$ with O'Brien-Fleming boundaries and $m = 4$

s/t t	1	2
-2.5	–	(-4.04, 1.47)
-3	–	(-4.77, -0.93)
-3.5	–	(-5.53, -2.02)
-4	–	(-6.37, -2.84)
-4.5	(-6.28, 3.58)	(-7.26, -3.63)
-5	(-6.93, -0.89)	(-8.18, -4.46)
-5.5	(-7.45, -2.59)	(-9.13, -5.34)
-6	(-7.96, -3.60)	(-10.10, -6.26)

In examining particular characteristics of these intervals, observe that the intervals given a stopping time $t = 1$ are converging to the usual $(s \pm 1.96)$ fixed sample intervals as s gets very large. In addition, they have the property that for the same sample mean, they are different for different values of t. (This is in contrast to Sample Mean intervals discussed in Section 4.)

Also, as s gets very large for $t > 1$, the intervals are increasing in width. The increasing width of the intervals seems to reflect the fact that as s gets very large for analyses $t > 1$, the observed value of the test statistic is becoming more unusual. This might reflect a large true mean. However with a large true mean, one would actually expect the analysis to end at the first analysis. Thus if a very large s is observed with $t > 1$, then the very low probability of observing such an event should lead to very high uncertainty about the value of μ and hence to increasingly wider intervals. The conditional intervals reflect the uncertainty of the information

about μ. This reflection of uncertainty is not present in the unconditional confidence intervals.

A troubling property of these conditional intervals is that for values of s which are close to the critical values, the intervals become extremely wide. A particularly unsettling example is when $-c_1 = -4.048$ and $(t, s) = (1, -4.5)$. In this case, the conditional interval $(-6.28, 3.58)$ has a width of 9.86. By looking at a few numerical examples, it is suspected that as s approaches the boundary, the conditional interval converges to an interval of infinite length. Observe that as s approaches $-c_t$, the numerator and denomenator of the conditional probability are converging to the same value. Thus in order to find the same difference in the conditional probability, one needs greater differences in the value of μ, resulting in a confidence interval which is suspected to converge to an interval of infinite length. The conditioning proposed here leads to conditioning on all of the information that is available about the parameter when s is close to the boundary.

Also, when the null hypothesis that the mean is equal zero has been rejected and the value of s is close to the boundary, the null value will be included in the conditional interval. An example of this can be seen in the numerical example given in the paragraph above.

In the last two mentioned properties of the conditional intervals, what is seen is the shifting of information between the two test statistics s and t. When s is very

close to a critical value, then most of the information about the mean is contained in t. To emphasize this point, observe that when there is continuous monitoring of a sequential experiment, then all of the information about μ is contained in the stopping time since S equals the boundary value when the experiment is stopped. On the other hand, when s is very far from any critical value, then s carries most of the information about μ. In this situation, t behaves more like an ancillary statistic and conditioning on it improves the confidence that can be placed in a statement about the parameter.

Thus, the conditional intervals behave well when an improbable result is observed. However, they become quite wide when s is close to the boundary.

4 Conditional Coverages of Usual Confidence Intervals

Having discovered that it is difficult to control conditional confidence using a direct construction, we next consider conditional properties of previously proposed confidence intervals in the group sequential setting. As discussed in Section 1, we are still interested in studing the validity of these inferences, conditioned on the recognizable subsets defined by stopping time. As in Brown[10] or Olshen[11], if we can show that the coverage probabilities conditioned on any one of the recognizable subsets for any of the intervals fall uniformly below the stated $(1-\alpha)$ level, then

those confidence intervals are tenuous for inference purposes. In the framework of Buehler[2], suppose that one person were to produce a confidence interval for the mean and report that interval, along with the observed stopping time and how that interval was produced. Then another person could bet on whether the mean was contained in that confidence interval. If the second person knew that a particular combination of confidence procedure and stopping time resulted in a coverage that would be below the stated level, then that person could make a bet against the first person, and, on average, he would win. Or in other terms, if a confidence interval is known to have uniformly lower conditional coverage than that which is stated in a publication, the integrity of the published result would be undermined. Thus it is desirable to study the conditional coverages of currently used confidence intervals.

We focus on confidence sets which are formed upon termination of a sequential experiment, such that the trial is terminated early only when the null hypothesis is rejected. These intervals include Stagewise intervals as proposed by Tsiatis, Rosner and Mehta[18 19 20] and Sample Mean intervals as proposed by Emerson and Fleming[7]. We also consider intervals which have been formed in the Repeated Confidence Intervals framework[6] and are then reported as terminal intervals. For details of how these intervals are formed, see the relevant papers. Because of the additional computations needed, Likelihood Ratio confidence sets[21] and Score Test confidence sets[22] are not considered here. In addition, we considered both the O'Brien-Fleming and the Pocock boundary conditions. However, only results for the O'Brien-Fleming

Boundaries are shown in this paper. Results for the Pocock boundary conditions were similar.

4.1 The Calculations

Suppose a particular confidence interval, which is formed at following the t'th analysis is given by $\mathcal{CI}(t) = (S_L(t), S_U(t))$. The conditional coverage probability can be written, applying Baye's rule, as

$$\Pr(\mu \in \mathcal{CI}(t)|T = t) = \frac{\Pr(\mu \in \mathcal{CI},\ T = t|\mu)}{\Pr(T = t|\mu)}$$

where

$$Pr(T = t|\mu) = \begin{cases} \Pr(\text{reject } H_0 \text{ at } T = t|\mu) \text{ when } t < m \\ \Pr(\text{didn't reject } H_0 \text{ at analyses } 1, \ldots, m-1|\mu) \text{ when } t = m \end{cases}$$

More specifically to calculate the conditional coverage probability for each of the intervals, evaluate

$$\begin{aligned} &\Pr(\mu \in \mathcal{CI}(t)|T = t,\ \mu) \\ &\quad = \frac{\Pr(s \in (S_L(t), S_U(t))\,|\mu)}{\Pr(T = t|\mu)} \\ &\quad = \frac{\int_{S_L(t)}^{S_U(t)} f_t(s|\mu)ds}{\Pr(T = t|\mu)} \end{aligned}$$

where

$$Pr(T = t|\mu) = \begin{cases} \int_{-\infty}^{-c_t} f_t(s|\mu)ds + \int_{c_t}^{\infty} f_t(s|\mu)ds & \text{when } t < m \\ \int_{-c_{t-1}}^{c_{t-1}} f_{t-1}(x|\mu)dx & \text{when } t = m \end{cases}$$

For each terminal procedure, with each stopping time and each set of boundary conditions, $S_U(t)$ and $S_L(t)$ and the resulting conditional coverage probability were

found using the programming language GAUSS[17]. This was done for negative values of μ ranging from negative five to zero. The probabilities are symmetric around μ equal to zero. Figure 1 shows the conditional coverage probabilities obtained from evaluating the expression above for the Stagewise intervals. Figure 2 shows the conditional coverages obtained for the Sample Mean intervals. It is comforting to note that conditional on stopping time, none of the coverage probabilities fall uniformly below the stated coverage. However, in all cases, they do show a large amount of variability. For the Stagewise intervals, these coverage probabilities range from exactly zero to one. Given some values of μ and stopping times, the coverage is exactly zero because that value of the parameter is never included in an interval when the trial has stopped at the given analysis. For example, using O'Brien-Fleming boundaries and given that the trial stops at analysis 3, the coverages when $\mu = -0.5$ or $\mu = -3.5$ are zero. On the other hand for some stopping times, there are values of μ which are always included in the Stagewise interval. For example at time 3, the value $\mu = -1$ is always included in the interval, given that the value of s is negative. Thus, the coverage at $\mu = -1$ will be very close to one (not exactly due to the small chance that the trial ends with a positive sample mean). For the Sample Mean intervals, these coverage probabilities range from zero to one as well. When μ is close to zero and $t = 1, 2, 3$, the coverages are exactly zero for the same reasons described for the Stagewise intervals. For $t = 4$, the coverage is greater than $1 - \alpha$ at $\mu = 0$. As μ gets large, the coverage at analysis $t = 1$ converges to $1 - \alpha$. The coverages at analyses $t = 1, 2, 3$ converge to zero. They are never exactly zero due to the construction of the interval relying on probabity calculations from all 4 analyses. The sudden changes in direction for the coverage probabilities for the Sample Mean intervals result from transitions of the quantity $\wedge_k = \min(s/t \times k, -c_k)$ from $-c_k$ to $s/t \times k$. In each case, the Pocock boundaries result in greater coverages for values of the mean μ which are small than when using the O'Brien-Fleming boundaries. Although, even here, the coverages for values of μ very close to zero are zero when

the experiment is terminated early. On the other hand, the Pocock boundaries result in coverages which go to zero much quicker as μ gets large, for $t > 1$.

Now, consider the third type of interval mentioned. Repeated confidence intervals were not designed to be used as terminal intervals. However, one proposal discussed is to use the RCI intervals in conjunction with the appropriate stopping rule[6]. It is likely that under these conditions, the final confidence interval calculated would be reported at the end of the experiment with other results, and thus take on the stature of a terminal interval. For the experiment presented previously, suppose that the study was stopped as soon as the hypothesized null value of the parameter did not fall inside the interval. This would correspond to running a sequential study under a stopping rule that "spends" α in the same way that the Repeated Confidence Intervals do.

These coverage probabilities (see Figure 3) performed in a manner similar to those from the Stagewise and Sample Mean confidence procedures. However, the conditional coverage probabilities did attain at least 95% for a larger interval of values of μ at each stopping time. This is probably due to the fact that the intervals from the Repeated Confidence Interval approach are wider than those from the other approaches.

Note that in general, the unconditional coverages of the RCI intervals are correct when they are used as terminal intervals, i.e., when they are formed upon termination of a sequential trial due to rejecting the null hypothesis based on pre-specified boundaries. When μ equals μ_0, the unconditional coverage is equal to $1 - \alpha$, that is the probability of accepting the null hypothesis. This is apparent by observing that μ_0 is always included in the interval when the null hypothesis is accepted and the null hypothesis is accepted $(1 - \alpha) \times 100\%$ of the time. For other values of μ, the unconditional coverage is greater than $1 - \alpha$. This is shown by Ohman[23].

Ohman[23] examines confidence intervals which are terminated for reasons other than rejection of the null hypothesis, again finding that conditional coverages are unable to differentiate between the possible confidence sets within each setting. That is, none of the previously proposed confidence procedures studied here behaved appreciably better than any others, when judging them conditionally. A way to study this further is to take a partially Bayesian approach, which is carried out in the next section.

4.2 A Bayesian Approach

Suppose that in general, we expect the mean to have a tendency to be close to zero. We could model this prior belief by placing a normal prior distribution (with mean zero and variance τ^2) on the mean of the sampling distribution. The size of the variance for this normal distribution would reflect the strength of our belief that μ is close to zero. If one was quite sure that the mean was fairly close to zero, one would choose a small value for τ^2. If one was relatively unsure, then one would choose a large τ^2.

Here we calculated the Bayesian coverage given stopping time with $\tau^2 = 2.5$, 5, and 10. These coverages were calculated only for the situation where the stopping rule is completely dependent on the data. They were claculated for all three confidence intervals, the Stagewise, the Sample Mean, and the RCI intervals, using the O'Brien-Fleming boundary conditions. The bayesian conditional coverages are found in Table 2.

It appears that as τ^2 gets larger, for analyses two through three, the average coverages get smaller. At analysis one, it is reasonable to assume that the average coverage approaches the stated unconditional coverage of $(1 - \alpha)\%$. This is be-

Table 2: Bayesian Conditional Coverages, $\tau^2 = 2.5, 5, 10$

	τ^2 t	1	2	3	4
Stagewise	2.5	.154	.325	.321	.602
	5	.295	.286	.204	.331
	10	.377	.170	.108	.085
Sample Mean	2.5	.309	.689	.724	.644
	5	.590	.688	.411	.353
	10	.752	.459	.199	.179
RCI	2.5	.456	.454	.396	.329
	5	.478	.415	.258	.185
	10	.483	.271	.140	.095

cause the conditonal coverage approaches $1 - \alpha$ as μ gets very large. Thus a fully non-informative prior would give average coverages that were near or at zero for $t = 2, \ldots, 4$ and near or at $1 - \alpha$ for $t = 1$.

The Sample Mean intervals have average coverages which are much greater than the Stagewise intervals for the all priors considered here. Also, in the same manner, the Sample Mean intervals have better average coverage than the RCI intervals. Thus, given that the normal with variance τ^2 equal to 2.5, 5 or 10 accurately reflects one's prior belief about the mean μ, the Sample Mean intervals would be prefered over the other two kinds terminal intervals.

5 Discussion

Recent work in group sequential analysis has produced a number of confidence intervals to be used in various contexts. The most commonly used procedures and hence, the ones presented and examined here are the Stagewise and Sample Mean

terminal procedures and the Repeated Confidence Intervals procedure. We have shown that the confidence levels of these intervals conditional on the stopping time are extremely erratic. When it is believed that the mean is moderately close to zero, Bayesian calculations showed the Sample Mean interval to be preferable over the Stagewise and RCI intervals.

As an alternative, we proposed intervals which were correct conditional on the stopping time and the direction of the sample mean away from the hypothesized null value. It was shown that these intervals also make more sense intuitively for unlikely results, that is, they seem to make better use of the information which is known at the time the experiment stopped when s is far from the boundary.

Of interest is the observation that all of the orderings that have been proposed, the stage-wise ordering, the sample mean ordering, the likelihood ordering, and the score function ordering all reduce down to the ordering proposed here, when they are conditioned on the stopping time. Thus, the arbitrariness of these orderings is removed by conditioning.

It appears that the amount of information contained in each of the sufficient statistics, t and s, can be thought of as differing depending on the actual value of (t, s). When the observed value is close to the boundary, t appears to contain most of the information about the mean μ. This reasoning stems from the observation that the conditional intervals become extremely wide when s approaches one of the critical values. Also, there seems to be more information about the mean in s when s is far from the boundary. The unconditional intervals, placing too much emphasis on t, do not seem to take full account of this information. In some sense, it seems as though t is acting as an ancillary statistic when s is large. Thus, although the conditional intervals can sometimes be unattractively large, they will always correctly process the sample information. Hence, calculation of these intervals seems an important

aid to any inference.

6 APPENDIX 1

Proof of Theorem 1:

Proof of statement 1 of Theorem 1.

Observe that[24] with $\sigma^2 = 1$,

$$\begin{aligned} f_t(s) &= f(t,s;\mu) \\ &= f(t,s;0) \times e^{(s\mu - t\mu^2/2)} . \end{aligned}$$

Then,

$$f(s|T=t, s<-c_t;\mu) = \frac{f(t,s;0)e^{(s\mu-t\mu^2/2)}}{\int_{-\infty}^{-c_t} f(t,w;0)e^{(w\mu-t\mu^2/2)}dw},$$

and so

$$\begin{aligned} \Pr\left[S<s|T=t, S<-c_t;\mu\right] &= \frac{\int_{-\infty}^{s} f(t,w;0)e^{(w\mu-t\mu^2/2)}dw}{\int_{-\infty}^{-c_t} f(t,w;0)e^{(w\mu-t\mu^2/2)}dw} \\ &= \frac{\int_{-\infty}^{s} f(t,w;0)e^{w\mu}dw}{\int_{-\infty}^{-c_t} f(t,w;0)e^{w\mu}dw}. \end{aligned}$$

The last step follows since $e^{-t\mu^2/2}$ can come out of the integral in the numerator and denominator and cancels out.

We will prove that for all μ, $\frac{\partial}{\partial\mu}\Pr[S<s|T=t,S<-c_t;\mu] < 0$. For ease of notation, in the next equation, let $g(t,w,\mu) = f(t,w;0)e^{w\mu} > 0$.

$$\begin{aligned} &\frac{\partial}{\partial\mu}\Pr[S<s|T=t,S<-c_t;\mu] \\ &= \frac{\int_{-\infty}^{s} wg(t,w,\mu)dw \int_{-\infty}^{-c_t} g(t,w,\mu)dw - \int_{-\infty}^{s} g(t,w,\mu)dw \int_{-\infty}^{-c_t} wg(t,w,\mu)dw}{\left[\int_{-\infty}^{-c_t} g(t,w,\mu)dw\right]^2} \end{aligned} \tag{3}$$

Note that the denomenator is > 0 for all μ and s. Therefore we only need to prove that the numerator is < 0. The numerator is < 0 if and only if

$$\frac{\int_{-\infty}^{s} wg(t,w,\mu)dw}{\int_{-\infty}^{-c_t} wg(t,w,\mu)dw} - \frac{\int_{-\infty}^{s} g(t,w,\mu)dw}{\int_{-\infty}^{-c_t} g(t,w,\mu)dw} > 0 \tag{4}$$

Note that the quantity $\int_{-\infty}^{-c_t} wg(t,w,\mu)dw$ is negative.

For each t and $-\infty < w < -c_t < 0$, define

$$G_\mu(w) = \frac{wg(t,w,\mu)}{\int_{-\infty}^{-c_t} wg(t,w,\mu)}$$

and

$$H_\mu(w) = \frac{g(t,w,\mu)}{\int_{-\infty}^{-c_t} g(t,w,\mu)}.$$

Both are > 0 and $\int_{-\infty}^{-c_t} G_\mu(w) = \int_{-\infty}^{-c_t} H_\mu(w) = 1$.

In the new notation, showing (4) is equivalent to showing

$$\int_{-\infty}^{s} G_\mu(w) - \int_{-\infty}^{s} H_\mu(w) > 0 \tag{5}$$

for any $s < -c_t$.

Taking the derivative of (5) gives

$$\begin{aligned} & G_\mu(s) - H_\mu(s) = \\ & = \frac{g(t,s,\mu)}{\int_{-\infty}^{-c_t} wg(t,w,\mu)dw}\left[s - \frac{\int_{-\infty}^{-c_t} wg(t,w,\mu)dw}{\int_{-\infty}^{-c_t} g(t,w,\mu)dw}\right] \end{aligned}$$

The quantity in []'s is an increasing function of s and the quantity outside is negative. Therefore the only possible sign change of (5) is from positive to negative. Thus, there can only be one interior extremum, and it must be a maximum. Then the only possible maxima are at the extremes. When $s = -c_t$ and when $s = -\infty$, (5) is equal to zero. Thus for all values of $s \in (-\infty, -c_t)$, (5) is > 0. Thus (3) is < 0 and the original probability (1) is decreasing. ∎

Proof of statement 2 of Theorem 1.

Equation 2 is proven by observing that

$$\Pr[S > s | S > c_t, T = t, \mu] = \Pr[S < -s | S < -c_t, T = t, -\mu] \tag{6}$$

From the Theorem, we know that the right hand side of (6) is decreasing as $(-\mu)$ increases and therefore is increasing with μ. Thus the probability (2) is an increasing function of μ. ∎

Bibliography

1. Fisher, R.A. *Statistical Methods and Scientific Inference*, 2nd ed., Hafner, New York, 1959.

2. Buehler, R. J. 'Some validity criteria for statistical inferences'. *Annals of Mathematical Statistics*, **20**, 845-863 (1959).

3. Armitage, P. *Sequential Medical Trials*; 2nd ed., Blackwell, Oxford, 1975.

4. Pocock, S. J. 'Interim analyses for randomized clinical trials: The group sequential approach'. *Biometrics*, **38**, 153-162 (1982).

5. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, Ellis Horwood, Chichester, England, 1983.

6. Jennison, C. and Turnbull, B. W. 'Interim analyses: The repeated confidence interval approach' (with discussion). *J. Roy. Statist. Soc. Ser. B*, **51**, 305-361 (1989).

7. Emerson, S.S. and Fleming, T. R. 'Parameter estimation following group sequential hypothesis testing'. *Biometrika*, **77**, 875-892 (1990).

8. Kiefer, J.C. 'Conditional confidence statements and confidence estimators'. *J. Ameri. Statist. Assoc.*, **72**, 789-827 (1977).

9. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*; 2nd ed., Springer, New York, 1980.

10. Brown, L. 'The conditional level of the t test'. *Annals of Mathematical Statistics*, **38**, 1068-1071 (1967).

11. Olshen, R. A. 'The conditional level of the F-test'. *Journal of the American Statistical Association*, **68**, 692-698 (1973).

12. Goutis c. and Casella, G. 'Increasing the confidence in Students t interval'. *Annals of Statistics*, **20**, 1501-1513 (1992).

13. Lan, K. K. G. and DeMets, D. L. 'Discrete sequential boundaries for clinical trials'. *Biometrika*, **70**, 659-663 (1983).

14. Pocock, S. J. 'Group sequential methods in the design and analysis of clinical trials'. *Biometrika*, **64**, 191-199 (1977).

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

16. Casella, G. and Berger, R. L. *Statistical Inference*, Wadsworth and Brooks/Cole, Belmont, CA, 1990.

17. Aptech Systems. *GAUSS: the GAUSS Sustem Version 3.0*, Aptech Systems, Inc. Maple Valley, WA, 1992.

18. Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. 'Exact confidence intervals following a group sequential test'. *Biometrics*, **40**, 797-803 (1984).

19. Siegmund, D. 'Estimation following sequential tests'. *Biometrika*, **65**, 341-349 (1978).

20. Jennison, C. and Turnbull, B. W. 'Confidence intervals for a binomial parameter follwing a multistage test with application to MIL-STD 105D and medical trials'. *Technometrics*, **25**, 49-58 (1983).

21. Chang, M. N. 'Confidence Intervals for a normal mean following a group sequential test'. *Biometerics*, **45**, 247-254 (1989).

22. Rosner, G. L. and Tsiatis, A. A. 'Exact confidence intervals following a group sequential trial: A comparison of methods'. *Biometrika*, **75**, 723-729 (1988).

23. Ohman, P. A. *Confidence Intervals Following Group Sequential Testing: Conditioning on Stopping Time*, Masters Thesis, Cornell University, Ithaca, NY, 1996.

24. Emerson, S. S. and Fleming, T. R. 'Symmetric group sequential test designs'. *Biometrics*, **45**, 905-923 (1989).

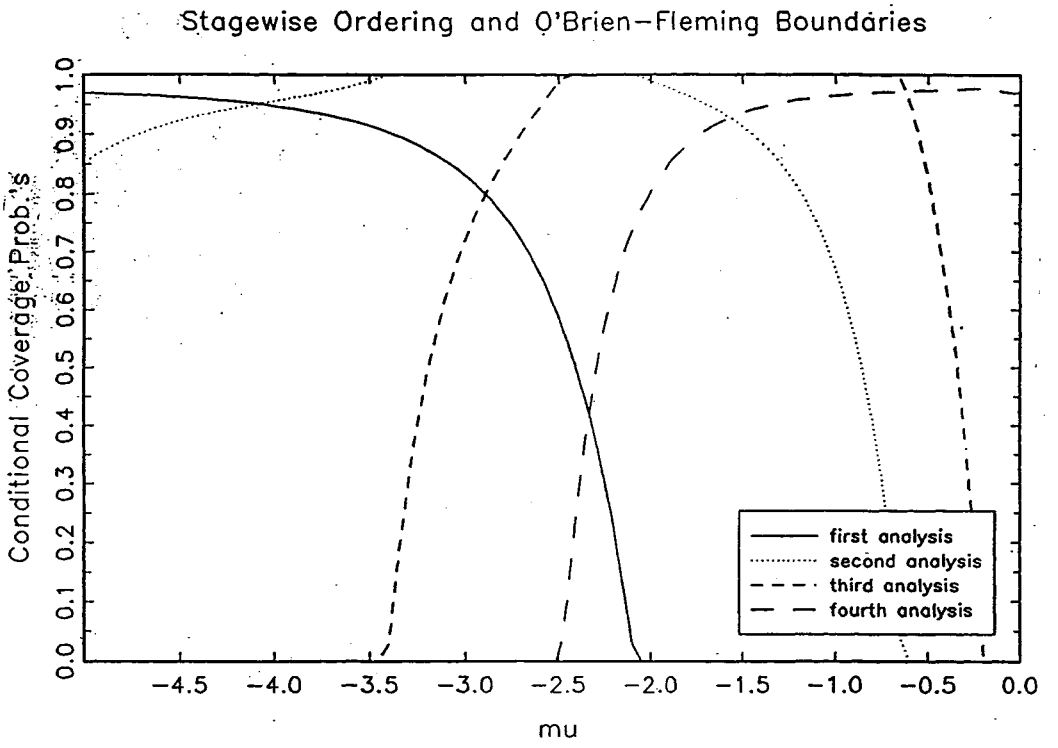

Figure 1: Coverage probabilities conditioned on stopping time of usual Stagewise 95% Confidence Intervals

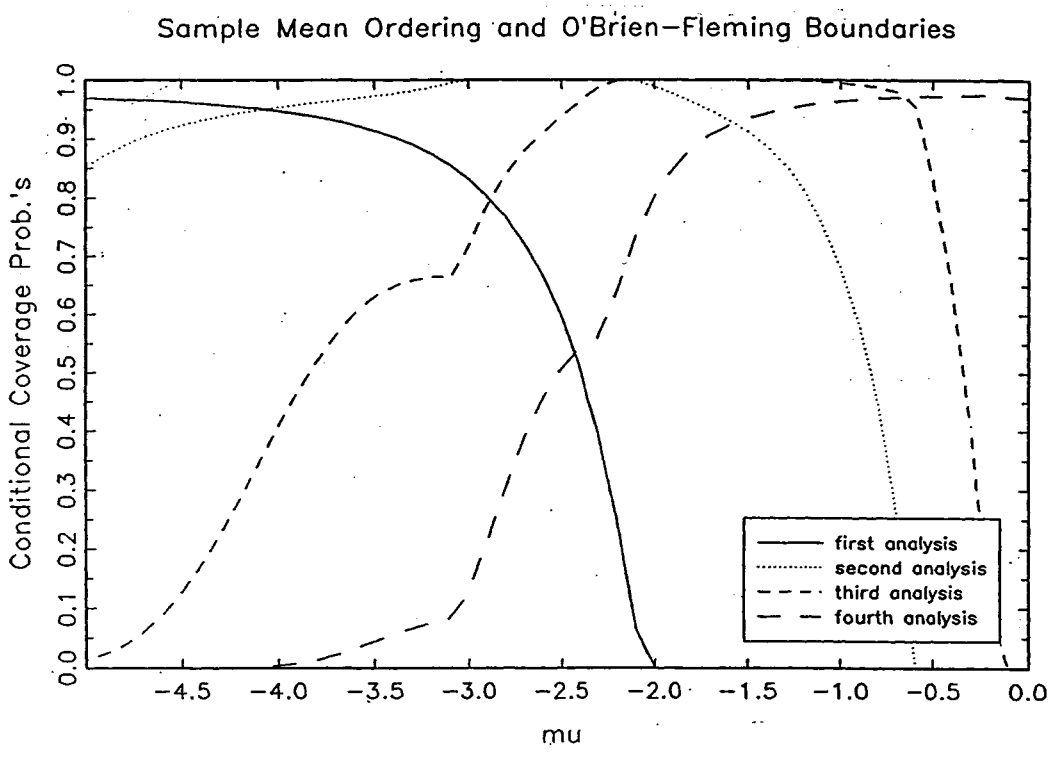

Figure 2: Coverage probabilities conditioned on stopping time of usual Sample Mean 95% Confidence Intervals

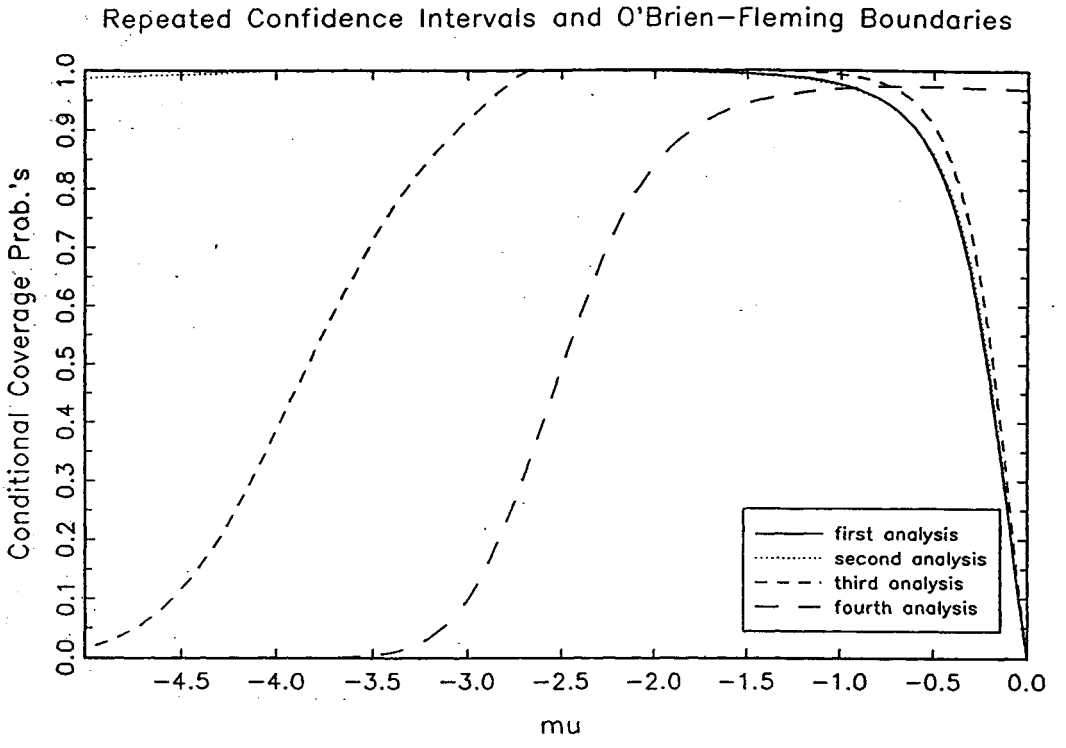

Figure 3: Coverage probabilities conditioned on stopping time of usual RCI 95% Confidence Intervals