Exact Repeated Confidence Intervals for Bernoulli Parameters in a Group Sequential Clinical Trial

Paul R. Coe, Ph.D. and Ajit C. Tamhane, Ph.D.

Roosevelt University, Chicago, Illinois (P.R.C.) and Northwestern University, Evanston, Illinois (A.C.T.)

ABSTRACT: This paper presents methods for constructing exact repeated confidence intervals (RCIs) for the success probability, p, of a single Bernoulli treatment and for the difference of success probabilities, $\Delta = p_1 - p_2$, of two independent Bernoulli treatments in the context of a group sequential clinical trial. These RCIs calculated at each interim analysis are useful for evaluating the data in light of all the information available rather than relying on rigid stopping criteria used by repeated significance tests. Extensions to construction of RCIs for the relative risk $\rho = p_1/p_2$ and odds ratio $\psi = p_1(1 - p_2)/p_2(1 - p_1)$ are indicated.

KEY WORDS: *Repeated confidence intervals, interim analyses, binomial distribution, Berkson's simple difference, relative risk, odds ratio*

INTRODUCTION

In order to detect early evidence of treatment differences or harmful side effects, periodic reviews of the accumulating data (called interim analyses) are often performed in clinical trials. Many sequential methods have been developed for this purpose, most taking the form of a repeated significance test. These methods have not been readily embraced in practice, however, because the rigid stopping criteria that they require are often inappropriate in clinical settings [1,2]. More appropriate in these settings is the use of repeated confidence intervals (RCIs), which allow the study results to be evaluated flexibly at each review in light of all the information available including the data on efficacy, safety, and concurrent findings of other research groups [3].

RCIs were first derived independently by Lai [4] and Jennison and Turnbull [5] for normally distributed responses. These authors also indicated how the large sample normal approximation theory can be used to derive RCIs for the parameters of interest in problems involving nonnormal data such as Bernoulli and survival. For a recent review of the work on this topic, see an article by Jennison and Turnbull [6].

Address reprint requests to: Prof. Ajit C. Tamhane, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60208-4070.

Received September 27, 1991; revised August 13, 1992.

Controlled Clinical Trials 14:19–29 (1993)

655 Avenue of the Americas, New York, New York 10010

0197-2456/93/$6.00

Many clinical trials are carried out with rather small numbers of patients. This is especially true when the accumulating data are analyzed on an interim basis at initial stages of a group sequential trial. Therefore, large sample normal approximations may not be valid when the data are nonnormal. In this paper we present exact methods for constructing RCIs when the response variable is dichotomous in nature, eg, success or failure, and follows a Bernoulli distribution for each treatment group (referred to as a Bernoulli treatment). We first consider a single Bernoulli treatment and show how RCIs can be constructed for its success probability p using a single-stage method due to Blyth and Still [7] which is a modification of the Sterne [8] method. The basic idea underlying the method developed for p is extended in the next section to construct RCIs for the difference between two success probabilities, $\Delta = p_1 - p_2$, associated with two independent Bernoulli treatments. We illustrate this method for Δ by a numerical example based on leukemia trial data given by O'Brien and Fleming [9]. Next we briefly indicate how RCIs can also be constructed for the relative risk $\rho = p_1/p_2$ and odds ratio $\psi = p_1(1 - p_2)/p_2(1 - p_1)$. We conclude the paper by comparing the Sterne-type approach followed here to an alternative approach [10] for constructing an exact confidence interval for a Bernoulli parameter p.

REPEATED CONFIDENCE INTERVALS FOR A SINGLE BERNOULLI PARAMETER

Let X be a binomially distributed random variable (rv) with parameters n and p (henceforth denoted by $X \sim B\ (n,\ p)$). Several methods have been proposed for constructing an exact (small sample) confidence interval (CI) for p. A review can be found in Ref. 7.

Sterne's [8] method begins by dividing the range of p (ie, the interval [0,1]) into a fine grid and then constructs minimal cardinality subsets of the sample space of X with probability content at least $1 - \alpha$ for each value $p = p_0$ in the grid. Each such subset is a $(1 - \alpha)$ level acceptance region for testing H_0: $p = p_0$, and is an interval $[L(p_0), U(p_0)]$ satisfying

$$P\{L(p_0) \leqslant X \leqslant U(p_0)|p_0\} \geqslant 1 - \alpha$$

Clearly, this minimal cardinality acceptance region is obtained by including in it the most likely outcomes X under $p = p_0$. By inverting these acceptance regions in the usual manner, CIs with the shortest width can be obtained. Unfortunately, these CIs are not always interval-valued. Blyth and Still [7] modified the Sterne method so that both $L(p_0)$ and $U(p_0)$ are nondecreasing in p_0, which makes the CIs interval-valued. Earlier Crow [11] proposed an alternative modification of the Sterne method to serve the same purpose. We use the Blyth and Still modification of the Sterne method as the basic building block of our method for constructing RCIs for p.

In the group sequential setting, let K be the maximum number of stages and $n_1, n_2, \ldots, n_K$, the number of observations (patients) in these K stages. Let X_k be the number of successes (positive responses) in the stage k ($X_k \sim B(n_k, p)$) and $S_k = X_1 + X_2 + \cdots + X_k$, the cumulative number of successes in the first k stages in a total of $N_k = n_1 + n_2 + \cdots + n_k$ observations. Based on these data we wish to construct RCIs $\{[p_{kL}, p_{kU}], k = 1,2, \ldots, K\}$, so that they have an overall coverage probability at least $1 - \alpha$, ie,

$$P\{p_{kL} \leqslant p \leqslant p_{kU} \quad \text{for all } k = 1,2, \ldots, K\} \geqslant 1 - \alpha \tag{1}$$

In our algorithm we need to choose a sequence of error rates $\alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_K = \alpha$; several possible choices are given in Ref. 12. For the sake of simplicity, we will assume that the number of stages K, the sample sizes $n_1, n_2, \ldots, n_K$, and the error rates $\alpha_1, \alpha_2, \ldots, \alpha_K$ are fixed, although this need not be the case (see Refs. 4 and 12).

We now describe our algorithm. First partition the interval [0,.5] into a sufficiently fine grid on p. (For $p > .5$, the acceptance regions can be obtained by symmetry.) We used a grid consisting of 500 points, beginning at .0005 and with a step size of .0010, ie, the grid $p = .0005, .0015, .0025, \ldots, .4995$. This grid gives a three-decimal-place accuracy for the CIs. At stage 1, construct a $(1 - \alpha_1)$ level acceptance region $[L_1(p_0), U_1(p_0)]$ as in (1) for each p_0 value in the grid by applying the Blyth and Still method. Invert these acceptance regions to obtain a first stage CI, $[p_{1L}, p_{1U}]$, for p for any given outcome $S_1 = 0, 1, \ldots, N_1$.

In general, at each stage k construct a $(1 - \alpha_k)$ level minimal cardinality acceptance region, $[L_k(p_0), U_k(p_0)]$, in the sample space of S_k for each p_0 value in the grid using the Blyth and still method subject to the following condition:

$$P\{L_i(p_0) \leqslant S_i \leqslant U_i(p_0) \quad \text{for } i = 1,2, \ldots, k|p_0\} \geqslant 1 - \alpha_k \tag{2}$$

This ensures that $S_1, S_2, \ldots, S_k$ are all in their acceptance regions,

$$[L_1(p_0), U_1(p_0)], [L_2(p_0), U_2(p_0)], \ldots, [L_k(p_0), U_k(p_0)]$$

which are computed sequentially. Since $(1 - \alpha_k)$ is decreasing in k, the existence of these acceptance regions is guaranteed. Invert the acceptance regions $[L_k(p_0), U_k(p_0)]$ to obtain a kth stage CI, $[p_{kL}, p_{kU}]$, for any given outcome $S_k = 0,1, \ldots, N_k$.

That the resulting sequence of CIs has an overall coverage probability at least $1 - \alpha$ follows because for any value of p (strictly speaking, only in the grid),

$$\begin{aligned} P\{p_{kL} \leqslant p \leqslant p_{kU} \quad \text{for } k = 1,2, \ldots, K|p\} &= P\{L_k(p) \leqslant S_k \leqslant U_k(p) \\ &\quad \text{for } k = 1,2, \ldots, K|p\} \\ &\geqslant 1 - \alpha_K \\ &= 1 - \alpha \end{aligned} \tag{3}$$

Note that the S_i values are the sums of the independent binomially distributed X_j values which facilitates the evaluation of the probability in (2). For example,

$$\begin{aligned} &P\{L_i(p_0) \leqslant S_i \leqslant U_i(p_0) \quad \text{for } i = 1,2\} \\ &= \sum_{x_1 = L_1(p_0)}^{U_1(p_0)} P\{X_1 = x_1\} \sum_{x_2 = L_2}^{U_2} P\{X_2 = x_2\} \end{aligned}$$

where $L_2 = \max\{0, L_2(p_0) - x_1\}$ and $U_2 = \min\{n_2, U_2(p_0) - x_1\}$. See Ref. 13 for further computational details.

A BASIC program to implement this algorithm is available from the first author. This program was used to compute 90% RCIs for p for a three-stage group sequential plan with $(n_1,n_2,n_3) = (15,10,10)$ (ie, $(N_1,N_2,N_3) = (15,25,35)$. In computing these RCIs we used the linear error rate use function $(\alpha_1,\alpha_2,\alpha_3) = (.0333,.0667,.1000)$. Other choices such as the convex use function are possible but the linear use function represents an "intermediate" choice (Ref. 125 p. 662) between the O'Brien and Fleming [9] and Pocock [14] approaches, and has the advantage of ease of application. The results are presented in Table 1.

These RCIs are not directly comparable with anything presently available. But it might be useful to know how much is lost (in terms of, say, the CI

Table 1 90% RCIs for a Three-Stage Sampling Plan of Fleming [15]

Stage k	N_k	S_k	CI for p	Stage k	N_k	S_k	CI for p
1	15	0	(.000,.222)	3	35	0	(.000,.089)
1	15	1	(.002,.326)	3	35	1	(.003,.133)
1	15	2	(.019,.418)	3	35	2	(.015,.174)
1	15	3	(.048,.500)	3	35	3	(.032,.216)
1	15	4	(.085,.567)	3	35	4	(.051,.258)
1	15	5	(.127,.613)	3	35	5	(.070,.289)
1	15	6	(.174,.674)	3	35	6	(.088,.325)
1	15	7	(.222,.738)	3	35	7	(.104,.346)
1	15	8	(.262,.778)	3	35	8	(.133,.387)
1	15	9	(.326,.826)	3	35	9	(.142,.414)
1	15	10	(.387,.873)	3	35	10	(.173,.438)
1	15	11	(.433,.915)	3	35	11	(.193,.475)
1	15	12	(.500,.952)	3	35	12	(.216,.500)
1	15	13	(.582,.981)	3	35	13	(.241,.525)
1	15	14	(.674,.998)	3	35	14	(.258,.562)
1	15	15	(.778,1.00)	3	35	15	(.289,.586)
				3	35	16	(.313,.607)
2	25	0	(.000,.134)	3	35	17	(.343,.646)
2	25	1	(.003,.199)	3	35	18	(.354,.657)
2	25	2	(.017,.258)	3	35	19	(.393,.687)
2	25	3	(.038,.300)	3	35	20	(.414,.711)
2	25	4	(.062,.354)	3	35	21	(.438,.742)
2	25	5	(.089,.400)	3	35	22	(.475,.759)
2	25	6	(.116,.437)	3	35	23	(.500,.784)
2	25	7	(.134,.482)	3	35	24	(.525,.807)
2	25	8	(.167,.518)	3	35	25	(.562,.827)
2	25	9	(.199,.563)	3	35	26	(.586,.858)
2	25	10	(.220,.600)	3	35	27	(.613,.867)
2	25	11	(.258,.646)	3	35	28	(.654,.896)
2	25	12	(.300,.675)	3	35	29	(.675,.912)
2	25	13	(.325,.700)	3	35	30	(.711,.930)
2	25	14	(.354,.742)	3	35	31	(.742,.949)
2	25	15	(.400,.780)	3	35	32	(.784,.968)
2	25	16	(.437,.801)	3	35	33	(.826,.985)
2	25	17	(.482,.833)	3	35	34	(.867,.997)
2	25	18	(.518,.866)	3	35	35	(.911,1.00)
2	25	19	(.563,.884)				
2	25	20	(.600,.911)				
2	25	21	(.646,.938)				
2	25	22	(.700,.962)				
2	25	23	(.742,.983)				
2	25	24	(.801,.997)				
2	25	25	(.866,1.00)				

width) because of using these *repeated* CIs as opposed to using *one-time* CIs. In a group sequential setting, a one-time CI might be used when a decision is made to stop the trial based on some fixed stopping criterion. The three-stage plan that we have used here was in fact presented along with a stopping boundary by Fleming [15] for testing H_0: $p = .1$ vs. $H_1 = .3$ with type I and type II error rates approximately equal to .05 and .10, respectively. Jennison and Turnbull [15] and Duffy and Santner [17] have given methods for computing such one-time CIs upon exceeding the stopping boundary. Table 2 presents a comparison of our RCIs with these one-time CIs (both having a nominal 90% confidence level) for the outcomes outside Fleming's [15] stopping boundary. Also included in the table are one-time 90% *fixed-sample* CIs

Table 2 Comparison of 90% RCIs With Other CIs for a Three-Stage Sampling Plan of Fleming [15] to Test H_0: $p = .10$ vs. H_1: $p = .30$ (Stopping Boundaries Considered)

Stage k	N_k	S_k	Decision	Fixed Sample Size	Jennison & Turnbull	Duffy & Santner	Repeated (RCI)
1	15	0	Accept	(.000,.168)	(.000,.181)	(.000,.165)	(.000,.222)
1	15	5	Reject	(.168,.567)	(.142,.577)	(.165,.558)	(.127,.613)
1	15	6	Reject	(.205,.635)	(.191,.640)	(.203,.640)	(.174,.674)
1	15	7	Reject	(.267,.675)	(.244,.700)	(.260,.675)	(.222,.738)
1	15	8	Reject	(.325,.733)	(.300,.756)	(.321,.744)	(.262,.778)
1	15	9	Reject	(.365,.795)	(.360,.809)	(.389,.795)	(.326,.826)
1	15	10	Reject	(.433,.832)	(.423,.858)	(.436,.847)	(.387,.873)
1	15	11	Reject	(.500,.878)	(.489,.903)	(.523,.879)	(.433,.915)
1	15	12	Reject	(.567,.924)	(.560,.943)	(.558,.925)	(.500,.952)
1	15	13	Reject	(.635,.964)	(.637,.976)	(.639,.964)	(.582,.981)
1	15	14	Reject	(.733,.993)	(.721,.997)	(.744,.993)	(.674,.998)
1	15	15	Reject	(.832,1.00)	(.819,1.00)	(.846,1.00)	(.778,1.00)
2	25	1	Accept	(.004,.179)	(.003,.199)	(.007,.204)	(.003,.199)
2	25	2	Accept	(.021,.221)	(.016,.238)	(.023,.225)	(.017,.258)
2	25	3	Accept	(.045,.280)	(.034,.284)	(.049,.268)	(.038,.300)
2	25	6	Reject	(.118,.420)	(.108,.412)	(.127,.390)	(.116,.437)
2	25	7	Reject	(.158,.460)	(.127,.445)	(.151,.436)	(.134,.482)
2	25	8	Reject	(.179,.500)	(.137,.470)	(.165,.447)	(.167,.518)
2	25	9	Reject	(.221,.540)	(.141,.489)	(.194,.559)	(.199,.563)
2	25	10	Reject	(.246,.580)	(.142,.501)	(.203,.559)	(.220,.600)
2	25	11	Reject	(.280,.611)	(.142,.508)	(.260,.640)	(.258,.646)
2	25	12	Reject	(.320,.640)	(.142,.510)	(.320,.675)	(.300,.675)
2	25	13	Reject	(.360,.680)	(.142,.511)	(.321,.675)	(.325,.700)
2	25	14	Reject	(.389,.720)	(.142,.511)	(.389,.745)	(.354,.742)
3	35	4	Accept	(.051,.243)	(.057,.286)	(.046,.260)	(.051,.258)
3	35	5	Accept	(.071,.272)	(.066,.297)	(.081,.290)	(.070,.289)
3	35	6	Accept	(.084,.302)	(.081,.315)	(.097,.321)	(.088,.325)
3	35	7	Reject	(.110,.343)	(.095,.335)	(.112,.322)	(.104,.346)
3	35	8	Reject	(.128,.371)	(.104,.351)	(.123,.340)	(.133,.387)
3	35	9	Reject	(.154,.400)	(.107,.363)	(.148,.390)	(.142,.414)
3	35	10	Reject	(.171,.429)	(.108,.369)	(.164,.437)	(.173,.438)
3	35	11	Reject	(.201,.457)	(.108,.372)	(.165,.437)	(.193,.475)
3	35	12	Reject	(.220,.486)	(.108,.373)	(.194,.559)	(.216,.500)
3	35	13	Reject	(.243,.514)	(.108,.374)	(.203,.559)	(.241,.525)
3	35	14	Reject	(.272,.543)	(.108,.374)	(.203,.559)	(.258,.562)
3	35	15	Reject	(.300,.571)	(.108,.374)	(.260,.640)	(.289,.586)

for the same outcomes computed using the Blyth and Still [7] method. The RCIs in Table 2 are the same those in Table 1 because even though the sample points listed here for stages 1 and 2 correspond to the trial being terminated early due to exceeding the stopping boundary, any unused error probability cannot be applied to the last CI. Such a modification of the error rate use function is valid only if it is done on the basis of variables independent of the response variable, which is clearly not the case here; see Refs. 18 and 19.

The average CI widths for the four methods given in Table 2 are as follows: fixed sample (.293), Jennison and Turnbull (.313), Duffy and Santner (.306), and RCIs (.328). (The Duffy–Santner intervals given here are different from those given in their paper and were provided to us by Professor Santner. They represent an improvement over their original intervals in that the interval end points at each stage i are nondecreasing in S_i, there are no single-point intervals, and the intervals do not require any manual adjustment and can be machine-generated). We conclude that the extra price for computing RCIs compared to fixed sample intervals is about 10%.

REPEATED CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO BERNOULLI PARAMETERS

Let $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ be two independent binomial rv's. Methods have been proposed for constructing an exact CI for $\Delta = p_1 - p_2$ by several authors, most recently by Santner and Yamagami [20] and Coe and Tamhane [21]. Santner and Yamagami (SY) extended the Crow [11] method for constructing an exact CI for a single p to construct an exact CI for Δ, while Coe and Tamhane (CT) extended the Blyth and Still method for the same purpose. The CT intervals are shorter than the SY intervals for more outcomes in the sample space of (X, Y) and these outcomes lie around the line $\hat{p}_1 - \hat{p}_2 = x/n_1 - y/n_2 = 0$ (which corresponds to Δ near zero); also the average and maximum interval widths are shorter for the CT method than for the SY method. Furthermore, the CT method is easier to extend to the multistage setting for constructing RCIs for Δ. Therefore we used the CT method as a basic building block in the algorithm for RCIs for Δ just as we used the Blyth and Still method as a basic building block in the algorithm for RCIs for p. Before presenting our algorithm for the multistage setting, we first give a brief outline of the CT method.

The CT method proceeds as follows: First partition the parameter space, ie, the interval $\{-1,1]$, into a grid.

$$-1 \leqslant \Delta_{-M} < \Delta_{-M}+_{1} < \cdots < \Delta_{-1} < \Delta_0 < \Delta_1 < \cdots < \Delta_M \leqslant 1$$

where $\Delta_{-i} = \Delta_i$ for $i = 1,2, \ldots, M$ and $\Delta_0 = 0$. One must then construct a $(1 - \alpha)$ level acceptance region $A(\Delta_i)$ for each value $\Delta = \Delta_i$, $i = 1,2, \ldots,$ M. (For $-1 \leqslant \Delta_i \leqslant 0$, the acceptance regions can be obtained by symmetry.) The acceptance regions must have the property:

$$P\{(X, Y) \text{ is in } A(\Delta_i)|p_1,p_2 = p_1 - \Delta_i\} \geqslant 1 - \alpha \qquad \text{for all } p_1 \text{ and for all } i.$$

These acceptance regions can then be inverted to give CIs for Δ.

To construct an acceptance region $A(\Delta_i)$ for given Δ_i, partition the range of the nuisance parameter p_1, ie, the interval $[\Delta_i,1]$, into a grid $\Delta_i = p_{i1} < p_{i2} < \cdots < p_{iN_i} = 1$ symmetrically about the midpoint $(1 + \Delta_i)/2$. For each $j = 1,2, \ldots, N_i$, construct an acceptance region $A(\Delta_i,p_{ij})$ such that

$$P\{(X, Y) \text{ is in } A(\Delta_i,p_{ij})|p_1 = p_{ij}, p_2 = p_1 - \Delta_i\} \geqslant 1 - \alpha$$

and $A(\Delta_i,p_{ij})$ contains as few points as possible. This is done by including in $A(\Delta_i,p_{ij})$ those points (x, y) that have the highest probabilities under $p_1 = p_{ij}, p_2 = p_1 - \Delta_i$ until the total probability becomes at least equal to $1 - \alpha$. Let

$$A(\Delta_i) = A(\Delta_i,p_{i1}) \cup A(\Delta_i,p_{i2}) \cup \cdots \cup A(\Delta_i,p_{iN_i}) = \bigcup_{j=1}^{N_i} A(\Delta_i,p_{ij})$$

This acceptance region $A(\Delta_i)$ often is larger then necessary; ie, there are points (x^*, y^*) in $A(\Delta_i)$ such that

$$P\{(X, Y) \text{ is in } A(\Delta_i) - (x^*, y^*)|p_1, p_2 = p_1 - \Delta_i\} \geqslant 1 - \alpha$$

for all p_1. Such points must be eliminated from $A(\Delta_i)$ subject to the condition that the final acceptance regions must be in some sense monotone in Δ_i so that when they are inverted, the resulting CIs for Δ will be interval-valued. For more details, see Refs. 21 and 22.

Returning to the multistage setting, we first introduce some notation. Let $X_1, X_2, \ldots, X_K$ and $Y_1, Y_2, \ldots, Y_K$ be two mutually independent sequences of binomial rv's such that $X_k \sim B(n_{1k},p_1)$ and $Y_k \sim B(n_{2k},p_2)$ for $k = 1,2, \ldots, K$; X_k and Y_k denote the number of successes on the two treatments in the kth stage. Let $S_k = X_1 + X_2 + \cdots + X_k$ and $T_k = Y_1 + Y_2 + \cdots + Y_k$ be the cumulative number of successes on the two treatments in the first k stages in a total of $N_{1k} = n_{11} + n_{12} + \cdots + n_{1k}$ and $N_{2k} = n_{21} + n_{22} + \cdots + n_{2k}$ observations, respectively. Based on these data we wish to construct RCIs $\{[\Delta_{kL},\Delta_{kU}], k = 1,2, \ldots, K\}$ so that they have an overall coverage probability at least $1 - \alpha$, ie,

$$P\{\Delta_{kL} \leqslant \Delta \leqslant \Delta_{kU} \quad \text{for all } k = 1,2, \ldots, K\} \geqslant 1 - \alpha$$

As before, choose a sequence $\alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_K = \alpha$. For the sake of simplicity, we will assume that the number of stages K, the sample sizes $(n_{11}, n_{21}), \ldots, (n_{1K}, n_{2K})$, and the error rates $\alpha_1, \alpha_2, \ldots, \alpha_K$ are fixed, although this need not be the case.

The algorithm for Δ is a logical extension of the algorithm for p. First partition the interval $[0,1]$ into a sufficiently fine grid on Δ. (For $-1 < \Delta < 0$, the acceptance regions can be obtained by symmetry.) At stage 1, construct a $(1 - \alpha_1)$ level acceptance region, $A_1(\Delta_i)$, in the two-dimensional sample space of (S_1, T_1) by applying the CT method, ie,

$$P\{(S_1,T_1) \text{ is in } A_1(\Delta_i)|p_1,p_2 = p_1 - \Delta_i\} \geqslant 1 - \alpha_1 \qquad \text{for all } p_1$$

In general, at each stage k, apply the CT method as above, to obtain $(1 - \alpha_k)$ level acceptance regions $A_k(\Delta_i)$ subject to the following condition:

$$P\{(S_j,T_j) \text{ is in } A_j(\Delta_i) \quad \text{for } j = 1,2,\ldots,k|p_1,p_2 = p_1 - \Delta_i\} \geqslant 1 - \alpha_k$$

This ensures that (S_1, T_1), (S_2, T_2), . . . , (S_k, T_k) are all in their respective acceptance regions, $A_1(\Delta_i)$, $A_2(\Delta_i)$, . . . , $A_k(\Delta_i)$, which are computed sequentially. Since $1 - \alpha_k$ is decreasing in k, the existence of these acceptance regions is guaranteed. Invert these acceptance regions to obtain a kth stage CI, $[\Delta_{kL},\Delta_{kU}]$, for Δ for any given outcome (S_k, T_k), $S_k = 0,1, \ldots, N_{1k}; T_k = 0,1, \ldots, N_{2k}$. That the resulting sequence of CIs has an overall coverage probability $1 - \alpha$ follows by the same argument as in (3).

A FORTRAN program to implement this algorithm is available from the first author. This program was used to compute 90% RCIs for Δ for a three-stage group sequential plan with $(n_1,n_2,n_3) = (15,10,10)$, $i = 1,2$. The linear error rate use function $(\alpha_1,\alpha_2,\alpha_3) = (.333, .0667, .1000)$ as in Table 1 was chosen. The detailed results are not presented here, but we note that the average CI width at the third stage was .346, which may be compared with .310, which is the average CI width for 90% intervals for Δ using the CT method for a fixed sample with $n_1 = n_2 = 35$. Thus the cost of making two previous intervals is about an extra 10% in average interval width.

A NUMERICAL EXAMPLE

O'Brien and Fleming [9] describe a clinical trial carried out at the Mayo Clinic during the years 1958–1973 to compare two drugs, prednisone (treatment 1) vs. prednisone + vincristine (treatment 2) in the treatment of leukemia. The response was remission (success), which in this case occurs relatively soon after the treatment or not at all (failure). The study had three stages with $n_{1k} = 7$, $n_{2k} = 14$ for $k = 1, 2, 3$. The cumulative number of successes (S_k, T_k) were (5, 12), (9, 25), and (14, 38) for $k = 1, 2, 3$, respectively.

O'Brien and Flemming [9] consider three methods for testing H_0: $\Delta = p_1 - p_2 = 0$, all based on normal approximation (specifically, Pearson's χ^2 statistic): a fixed-sample method, Pocock's [14] group sequential method with three stages, and O'Brien and Fleming's [9] group sequential method with three stages. Both the fixed sample method and the O'Brien–Fleming method reject H_0 against a two-tailed alternative at a signifcance level $\alpha = .02$, but the Pocock method rejects only at $\alpha = .05$; both the O'Brien–Fleming and Pocock methods reject at third stage. (Note that the significance levels reported in O'Brien and Fleming [9] are one-tailed.)

The normal approximation methods seem inapplicable for these data because of the small number of nonremissions in each treatment group. Specifically, there are only (2, 3, 2) and (2, 1, 1) nonremissions in the three stages for treatment 1 and treatment 2, respectively. In fact, the exact two-tailed p value for the results of this trial (assuming a fixed sample trial) using Fisher's exact test is .032. Therefore we computed exact 95% RCIs for Δ for these data with $(\alpha_1, \alpha_2, \alpha_3) = (.0167, .0333, .0500)$. They are $(-.59, .28)$, $(-.56, .05)$, $(-.48, .00)$, respectively, for the three stages. Using these RCIs would lead to rejection of H_0 in favor of a two-tailed alternative at stage 3 at $\alpha = .05$, but these RCIs could also be weighed with other information before drawing a final conclusion. Again, other choices of the α_i are possible, resulting in slightly different RCIs.

REPEATED CONFIDENCE INTERVALS FOR THE RELATIVE RISK AND ODDS RATIO

The basic idea behind the construction of exact RCIs for Δ easily extends to other functions of p_1 and p_2 such as the relative risk $\rho = p_1/p_2$ and the odds ratio $\psi = p_1(1 - p_2)/p_2(1 - p_1)$. What one needs is a Sterne-type method to construct a one-time fixed-sample exact CI for the given parameter that can be used as a basic building block for the multistage algorithm. Such methods are given by several authors but we used the ones proposed by Coe and Tamhane [21]. The CT method for ρ is very similar to that for Δ; it also involves finding the minimum probability content (over the nuisance parameter p_1) of the acceptance region for each value of $\rho = \rho_0$ in the grid, but note that here the range of ρ is the infinite interval $[0, \infty]$. The CT method for ψ is a conditional method (as are most other methods for ψ) based on the conditional distribution of X given the total number of successes $X + Y = m$. This conditional distribution is independent of the nuisance parameter p_1 and reduces the sample space to one dimension. Thus the computations are much easier in this case. FORTRAN programs for computing RCIs for ρ and ψ are available from the first author.

DISCUSSION

A referee has noted the following drawbacks associated with the Sterne type approach used in the present paper:

1. It gives short two-tailed CIs by trading error probabilities in the two tails, which ignores the possibility of different costs for over- and under-estimating the parameter of interest.
2. The error probabilities in the two tails are not individually bounded. Therefore, if such a CI is used to test a one-tailed alternative hypothesis then the type I error probability will only be known to lie between 0 and α. On the other hand, if a Clopper–Pearson [10] type of approach is used, then the type I error probability will be controlled at $\alpha/2$.

Drawback (2) can be readily overcome by modifying the method proposed in the present paper to construct one-tailed (upper or lower) RCIs if they are intended to be used to test a one-tailed alternative hypothesis. The resulting test would in general be more powerful since it will control the type I error probability at α rather than $\alpha/2$. The two-tailed RCIs derived in the present paper would result in a more powerful test of a two-sided alternative hypothesis compared to a test based on Clopper–Pearson–type RCIs because the latter would be longer on average due to the separate bounds on the individual tail probabilities. Of course, one must remember that the very reason for resorting to RCIs is to get away from the rigid framework of repeated hypothesis testing.

Drawback (1) cannot be similarly overcome. A Clopper–Pearson type of approach must be used if the lower and upper tail error probabilities are separately specified to be, say, α_L and α_U ($\alpha_L + \alpha_U = \alpha$), respectively. However, as we show below, this would put severe restrictions on how these error probabilities can be allocated over the successive stages of a group sequential trial, which would result in unnecessarily wide RCIs. We will illustrate this for the case of a single Bernoulli parameter p.

To extend the Clopper–Pearson approach, it is necessary to construct two sequences of RCIs $\{[p_{kL}, 1], k = 1, 2, \ldots, K\}$ and $\{[0, p_{kU}], k = 1, 2, \ldots, K\}$ such that

$$P\{p_{kL} \leqslant p \quad \text{for all } k = 1, 2, \ldots, K\} \geqslant 1 - \alpha_L$$

and

$$P\{p \leqslant p_{kU} \quad \text{for all } k = 1, 2, \ldots, K\} \geqslant 1 - \alpha_U$$

These two inequalities taken together imply the overall coverage probability condition 1. To apply the method, one needs to choose two error rate sequences: $\alpha_{1L} \leqslant \alpha_{2L} \leqslant \cdots \leqslant \alpha_{KL} = \alpha_L$ and $\alpha_{1U} \leqslant \alpha_{2U} \leqslant \cdots \leqslant \alpha_{KU} = \alpha_U$. Then, as explained earlier, a grid of values of p must be chosen and for each value $p = p_0$ in the grid, two sequences of lower and upper acceptance limits, $\{L_k(p_0), k = 1, 2, \ldots, K\}$ and $\{U_k(p_0), k = 1, 2, \ldots, K\}$, must be determined satisfying

$$P\{L_i(p_0) \leqslant S_i \leqslant U_i(p_0) \quad \text{for } i = 1, 2, \ldots, k-1;\ L_k(p_0) \leqslant S_k | p_0\} \geqslant 1 - \alpha_{kL},\ k = 1, 2, \ldots, K$$

and

$$P\{L_i(p_0) \leqslant S_i \leqslant U_i(p_0) \quad \text{for } i = 1, 2, \ldots, k-1;\ S_k \leqslant U_k(p_0) | p_0\} \geqslant 1 - \alpha_{kU},\ k = 1, 2, \ldots, K$$

Combining these two and putting $\alpha_{kL} + \alpha_{kU} = \alpha_k$, it follows that the acceptance limits $[L_k(p_0), U_k(p_0)]$ satisfy condition 2. But, in order for $L_k(p_0)$ and $U_k(p_0)$ to exist, one must have

$$\max(1 - \alpha_{kL}, 1 - \alpha_{kU}) \leqslant 1 - (\alpha_{k-1,L} + \alpha_{k-1,U}) = 1 - \alpha_{k-1} \quad \text{for } k = 2, \ldots, K$$

which is equivalent to

$$\min(\alpha_{kL}, \alpha_{kU}) \geqslant \alpha_{k-1} \qquad \text{for } k = 2, \ldots, K$$

Thus, if one chooses $\alpha_{kL} = \alpha_{kU}$ at each stage, then one must have

$$\alpha_{kL} = \alpha_{kU} = \alpha_k/2 \leqslant \alpha/(2 \times 2^{K-k}) \qquad \text{for } k = 1, 2, \ldots, K$$

As an example, for $\alpha = .05$ and $K = 3$, one must have $\alpha_{1L} = \alpha_{1U} \leqslant .0125/2$ and $\alpha_{2L} = \alpha_{2U} \leqslant .025/2$. Thus the α values for initial stages are very small making the corresponding RCIs too wide. This problem is further compounded by the discreteness of the probabilities being computed which typically causes the inequalities to be strict. Therefore, it is our view that the Clopper–Pearson type of RCIs, although computable, will be extremely conservative and thus of much less practical value.

ACKNOWLEDGMENTS

We thank the two referees who pointed out an error in an earlier version of the paper.

REFERENCES

1. DeMets DL: Stopping guidelines vs. stopping rules: A practitioner's point of view. Commun Stat Series A 13:2395–2418, 1984
2. Meier P: Statistics and medical experimentation. Biometrics 31:511–529, 1975

3. Meier P: Terminating a trial—the ethical problem. Clin Pharmaco Therapeut 25:633–640, 1979
4. Lai TL: Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. Commun Stat Series A 13:2355–2368, 1984
5. Jennison C, Turnbull BW: Repeated confidence intervals for group sequential clinical trials. Controlled Clin Trials 5:33–45, 1984
6. Jennison C, Turnbull BW: Interim analyses: The repeated confidence intervals approach (with discussion). J Roy Stat Soc Series B 51:305–361, 1989
7. Blyth CR, Still HA: Binomial confidence intervals. J Am Stat Assoc 78:108–116, 1983
8. Sterne TE: Some remarks on confidence or fiducial limits. Biometrika 41:275–278, 1954
9. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. Biometrics 35:549–556, 1979
10. Clopper CJ, Pearson ES: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 47:381–381, 1934
11. Crow E: Confidence intervals for a proportion. Biometrika 43:423–435, 1956.
12. Lan KKG, DeMets DL: Discrete sequential boundaries for clinical trials. Biometrika 70:659–663, 1983
13. Schultz JR, Nichol FR, Elfring, GL, Weed SD: Multiple-stage procedures for drug screening. Biometrics 65:341–349, 1973
14. Pocock SJ: Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191–199, 1977
15. Fleming TR: One-sample multiple testing procedures for Phase II clinical trials. Biometrics 38:143–151, 1982
16. Jennison C, Turnbull BW: Confidence intervals for a binomial parameter following a multistage test with applications to MIL-STD 105D and medical trials. Technometrics 25:49–58, 1983
17. Duffy DE, Santner TJ: Confidence intervals for a binomial parameter based on multistage tests. Biometrics 43:81–93, 1987
18. Fleming TR, Harrington DP, O'Brien PC: Design of group sequential tests. Controlled Clin Trials 5:348–361, 1984
19. Lan KKG, DeMets DL: Changing frequency of interim analysis in sequential monitoring. Biometrics 45:1017–1020, 1989
20. Santner TJ, Yamagami S: Invariant small sample confidence intervals for the difference of two success probabilities. Submitted for publication, 1991
21. Coe PR, Tamhane AC: Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. Submitted for publication, 1991
22. Coe PR: Exact repeated confidence intervals for binomial parameters in group sequential experiments. Doctoral dissertation, Northwestern University, 1989