SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

TECHNICAL REPORT NO. 645

December 1984
Revised June 1986

CONFIDENCE INTERVALS FOR A BINOMIAL PARAMETER BASED ON MULTISTAGE TESTS

by

Diane E. Duffy & Thomas J. Santner

Key Words: Binomial success probability; Grouped sequential hypothesis test; Clinical trials; Finite horizon.

Abstract

Several methods are developed for the construction of confidence intervals for the binomial success probability p following multistage experimentation. We consider analogues of one-stage methods proposed by Sterne (1954, Biometrika), Crow (1956, Biometrika), and Blyth and Still (1983, Journal of the American Statistical Association). The proposed confidence intervals are implemented and compared with the Clopper-Pearson tail intervals given by Jennison and Turnbull (1982, Technometrics) for the four three-stage testing procedures of Fleming (1982, Biometrics). The new intervals have (i) uniformly shorter total length summed over all outcomes, (ii) nearly uniformly shorter expected length and (iii) closer to nominal probability of coverage. An easily computed Crow-Blyth-Still type construction is particularly attractive for practical application.

1. Introduction

Several authors have constructed multistage sequential tests of H_0: $p \leq p_0$ (given) vs. H_1: $p \geq p_A$ (given) for a binomial parameter p in Phase II clinical trials (see Fleming (1982) and the references therein). There is universal recognition of the benefits of sequential sampling in reducing sample size while permitting legitimate statistical inferences. However as Armitage (1958), Jennison and Turnbull (1982), Tsiatis, Rosner and Mehta (1984), Atkinson and Brown (1985) and others argue, the result of such a trial is often combined with other information such as the extent of undesirable side effects before a final decision is reached. In such cases it is much more useful to determine confidence intervals for p than to merely accept or reject H_0.

The construction of confidence intervals for p after a multistage binomial test involves considerations not present in the single sample case. Section 2 of this paper investigates these issues by describing multistage analogues of various one-sample methods. Two characteristics of the multistage extensions we consider deserve comment. First, all achieve at least their nominal (unconditional) confidence levels. However, the reader should note that the conditional operating characteristics given that the procedure stops at an apriori fixed stage need not attain the nominal level. Second, with the exception of the Clopper-Pearson method, all the other methods described employ heuristics to produce short intervals.

Section 3 applies one of the methods discussed; namely, Sterne's (1954) proposal, to construct confidence intervals for Fleming's (1982) tests. Section 4 considers confidence intervals based on an induced

linear-ordering of the possible outcomes. Of particular interest is a scheme with approximately equal tail probabilities; Blyth and Still (1983) recommend an analogous construction for the one-sample binomial p confidence interval problem. Finally, section 5 compares the intervals of sections 3 and 4 for a particular example with the Jennison and Turnbull (1982) tail intervals according to three criteria: (i) total length of the intervals summed over all outcomes, (ii) expected length, and (iii) coverage probability. While both of the proposed intervals outperform the tail intervals and have competitive operating characteristics, the modified-Sterne intervals require considerably greater computational effort than the ordering-based intervals.

2. Notation and Review of One Sample Binomial Confidence Intervals

A general multistage test of $H_0: p \leq p_0$ vs. $H_1: p \geq p_A$ is specified by (i) K, the maximum number of stages to be used in the experiment, (ii) a set of stopping boundaries $\{(a_1,b_1),\ldots,(a_K,b_K)\}$ with $a_K = b_K-1$, and (iii) the numbers of Bernoulli trials $n_1,\ldots,n_K$ to be performed at each stage. The procedure is carried out as follows. If s_j is the number of successes at the jth stage, $1 \leq j \leq K$, then at stage g sampling continues to the next stage if $a_g < T_g \equiv \Sigma_{j=1}^{g} s_j < b_g$, otherwise it terminates and H_0 is rejected if $T_g \geq b_g$ and H_0 is accepted if $T_g \leq a_g$. Sampling is forced to stop at stage K as $a_K = b_K-1$. Examples are the four $K = 3$ stage plans of Fleming (1982) listed in Table 1. These plans are designed to have size .05 and power .90 (at p_A) for $(p_0,p_A) = (.05,.2)$, $(.1,.3)$, $(.2,.4)$ and $(.3,.5)$, respectively.

Table 1

Fleming (1982) K = 3 stage tests of
H_0: $p \leq p_0$ vs. H_1: $p \geq p_A$.

Plan No.	1	2	3	4
(p_0, p_A)	(.05,.2)	(.1,.3)	(.2,.4)	(.3,.5)
(a_1, b_1, n_1)	(-,4,15)	(0,5,15)	(1,8,15)	(5,12,20)
(a_2, b_2, n_2)	(2,5,15)	(3,6,10)	(7,11,15)	(12,17,15)
(a_3, b_3, n_3)	(4,5,10)	(6,7,10)	(13,14,15)	(20,21,15)

Let G denote the number of stages required for sampling to terminate and $S = \Sigma_{j=1}^{G} s_j$ be the total number of successes observed through stage G. The pair (G,S) is a sufficient statistic; it can take values:

$$
\begin{aligned}
&(1,0),(1,1),\ldots,(1,a_1),\\
&(2,a_1+1),\ldots,(2,a_2),\\
&\qquad\vdots\\
&(K,a_{K-1}+1),\ldots,(K,a_K),\\
&(K,b_K),\ldots,(K,b_{K-1}-1+n_K),\\
&(K-1,b_{K-1}),\ldots,(K-1,b_{K-2}-1+n_{K-1}),\\
&\qquad\vdots\\
&(1,b_1),\ldots,(1,n_1).
\end{aligned}
\tag{1}
$$

We discuss multistage extensions of various methods for one sample binomial p intervals that have been proposed in the literature (Blyth and

Still, (1983)) and outline the important features of the extensions. The symbols X and n denote the number of successes and Bernoulli trials in the description of one sample intervals.

The earliest proposal for constructing one sample $100(1-\alpha)\%$ confidence intervals $(p_L, p_U) = (p_L(X), p_U(X))$ for p are Clopper and Pearson's (1934) tail intervals. They are defined by $p_L(0) = 0$, $p_U(n) = 1$, and otherwise implicitly as the solutions to $P[X \leq x | p = p_U(x)] = \alpha/2$ and $P[X \geq x | p = p_L(x)] = \alpha/2$. The notation $P[\cdot | p]$ means that $X \sim B(n,p)$ in the probability calculation. It is easy to see that this method replaces the requirement $P[p_L(X) \leq p \leq p_U(X) | p] \geq 1-\alpha$ by the stronger requirement $P[p_L(X) > p | p] \leq \alpha/2$ and $P[p_U(X) < p | p] \leq \alpha/2$. The crucial feature of the tail method required to analyze the multistage case is that a linear ordering must be imposed on the outcomes in (1) so that an analogue of the "tail events" $[X \leq x]$ and $[X \geq x]$ can be defined. Several intuitive orderings for the bivariate sufficient statistic (G,S) are possible. Jennison and Turnbull (1983) (JT) give Clopper-Pearson tail intervals for p based on the ordering in (1) (see also Tsiatis et al. (1984)) and state that use of the equally appealing alternative ordering $(S/\Sigma_1^G n_j)$ gives virtually identical intervals.

Sterne (1954) proposed an alternative to the tail method for the one sample problem that consists essentially of constructing a confidence set for p by inverting the family of acceptance regions $\{A(p_0): 0 < p_0 < 1\}$ for the tests $H_0^*: p = p_0$ vs. $H_1^*: p \neq p_0$ defined by $P[X \in A(p_0) | p_0] \geq 1-\alpha$ where $P[X = i | p_0] \geq P[X = j | p_0]$ $\forall$ $i \in A(p_0)$ and $\forall$ $j \notin A(p_0)$.

Thus $A(p_0)$ is of minimum cardinality as it contains the most likely outcomes under p_0. Further, it is obvious that for all p_0, $0 < p_0 < 1$, $A(p_0)$ is an interval of integers, $L(p_0)$ to $U(p_0)$ inclusive, say.

Crow (1956) observed that the confidence sets resulting from inversion of Sterne's $A(p_0)$, $0 < p_0 < 1$, are not necessarily intervals as Sterne's $L(p_0)$ and $U(p_0)$ do not satisfy the following necessary and sufficient condition required for intervals to result from the inversion:

(2) $L(p_0)$ and $U(p_0)$ must be nondecreasing in p_0.

Crow also noted that in addition to (2) above, a family of acceptance regions should satisfy the intuitive property:

(3) Every outcome $x \in \{0,1,\ldots,n\}$ should be in some $A(p_0)$.

Equation (3) is automatically satisfied by Sterne's $A(p_0)$ since $P[X = x|p = x/n] > P[X = j|p = x/n] \quad \forall \quad j \neq x$.

Sterne's confidence sets (which are either intervals or unions of intervals) do have the following heuristic property. They minimize the summed (over all outcomes) Lebesgue measure (length in the case of intervals) of the confidence set among all $100(1-\alpha)\%$ confidence sets for p. This suggests that Sterne confidence sets may also have short expected length. Section 3 describes a multistage method that directly modifies Sterne's acceptance regions $A(p_0)$.

A third class of methods (Crow (1956) and Blyth and Still (1983)) work with $L(p_0)$ and $U(p_0)$ to construct $A(p_0)$ that satisfy (2) and (3) and, with but a few exceptions, produce acceptance regions for H_0^* versus H_1^* containing the same number of points as Sterne's acceptance regions. To develop multistage analogues of these methods one must have available not only a linear ordering among the outcomes but also a rule for choosing among the equal-sized $A(p_0)$ which achieve the nominal level and satisfy (2) and (3). Section 4 considers several possible ordering-and-rule combinations for the multistage problem.

A fourth proposal for studying binomial confidence sets is that of Casella (1984) which is decision-theoretic in nature; he presents an algorithm for constructing a uniformly superior invariant confidence interval from a given invariant confidence interval. The multistage extension of binomial invariance ($[L,U] \rightarrow [1-U,1-L]$ when $S_j \rightarrow n_j - S_j$ for every stage (j) places restrictions on the stopping boundaries of tests. In particular, the stage G at which termination occurs must be invariant under the above transformation. For tests inconsistent with the invariance, such as Fleming's Table 1 tests, invariant confidence intervals are not possible. Extensions of this work are not pursued here.

3. Multistage Modified Sterne Confidence Intervals

Before describing this construction we observe several properties of the probability mass function of (G,S). For a given stopping boundary (a_j,b_j), $j = 1,\ldots,K$, let Q be the set of observable outcomes (g,s) corresponding to stopping. The recursion formulas of Schultz et al. (1973) show the probability that $(G,S) = (g,s)$ is

$$(4) \qquad f((g,s)|p) = Cp^{s}(1-p)^{(\Sigma_1^g n_j)-s}, \quad 0 < p < 1, \quad (g,s) \in Q$$

where $C = C(\underset{\sim}{a},\underset{\sim}{b},s,g)$ is the number of vectors $(s_1,\ldots,s_g)$ of successes for the first g stages satisfying $a_\ell < \Sigma_1^\ell s_j < b_\ell$ for $\ell < g$ and $\Sigma_1^g s_j = s$. In words, C is the number of paths which pass through the first $(g-1)$ "windows" (a_ℓ,b_ℓ), $1 \leq \ell < g$, and terminate with a total numer of s successes at stage g. Equation (4) shows that, as in the one sample binomial case, the probability of observing $(G,S) = (g,s)$ is log concave in p with a maximum at $p = s/\Sigma_1^g n_j$.

To construct $100(1-\alpha)\%$ Sterne intervals, first partition $(0,1)$ by $0 < p_1 < \ldots < p_r < 1$ where the mesh is chosen sufficiently small to provide the desired accuracy for the final intervals. For each fixed p_i calculate $\{f((g,s)|p_i)\colon (g,s) \in Q\}$ and set $A(p_i)$ to be the <u>minimum</u> set of outcomes with <u>largest</u> $f(\cdot|p_i)$ values for which

$$(5) \qquad P[(G,S) \in A(p_i)|p_i] \geq 1-\alpha.$$

In principle, one difficulty can occur during this construction. For a given p_i, outcomes $(g^1,s^1) \neq (g^2,s^2)$ may satisfy

$$(6) \qquad f((g^1,s^1)|p_i) = f((g^2,s^2)|p_i).$$

This causes ambiguity in the definition of $A(p_i)$ if exactly one but not both of (g^1,s^1) and (g^2,s^2) must be included in $A(p_i)$ to satisfy the coverage probability requirement (5). However, (6) occurs only on a set of p_i having (Lebesgue) measure zero and, in practice, this problem did not occur during the construction of intervals for Fleming's tests.

The unimodality of $f((g,s)|p)$ in p insures that when the Sterne construction is applied to the multistage problem, for the majority of outcomes $(g,s) \in Q$, (g,s) is in $A(p_i)$ for a <u>consecutive</u> set of p_i. However, this is not always true and the following cases can occur:
(i) $\{p_i: (g,s) \in A(p_i)\}$ is a union of two or more intervals and
(ii) $(g,s) \notin A(p_i)$ for <u>any</u> p_i.

Phenomenon (i) occurs for the following reason. The (g,s) outcomes in the second and subsequent stages $(g \geq 2)$ which are near the outskirts of the stopping region (i.e. near $(g,a_{g-1}-1)$ and $(g,b_{g-1}-1+n_g)$) have extremely low probability. These (g,s) can flip-flop in and out of $A(p_i)$ as p_i increases rather than remaining in $A(p_i)$ for a consecutive sequence of p_i.

Phenomenon (ii) occurs because, unlike the one sample binomial case, there exist (g,s) for which $f((g,s)|p) < \max\{f((g',s')|p): (g',s') \in Q\}$ for all p. These outcomes need not enter $A(p_i)$ for any p_i. As for (i) above, this phenomenon occurs for (g,s) on the outskirts of the stopping region when $g \geq 2$.

A modification of the direct Sterne construction was used to produce Tables 2a-2d which contain 90% and 95% confidence intervals for each of the four Fleming multistage sampling plans of Table 1. The modification serves

to eliminate (i); i.e. non-interval confidence sets. To illustrate, the direct inversion of Sterne's $A(p_i)$ at the 90% confidence level with $p_i = i/2000$ based on Fleming's Plan 4 boundaries gives the confidence set $(.1700,.3075)\cup(.3195,.3220)$ when $(g,s) = (2,8)$. In four of the eight plan-by-level combinations studied it was possible to construct alternate systems of $A(p_i)$ satisfying the requirements; (a) inversion produces intervals, (b) they contain the same number of points as the Sterne $A(p_i)$ and (c) they achieve their nominal coverage level. Property (c) required checking since the alternate $A(p_i)$ satisfying (a) and (b) resulted from swapping an outcome (g,s) in Sterne's $A(p_i)$ with one not in Sterne's $A(p_i)$ and therefore of lower probability when $p = p_i$. In three of the remaining four cases, either one or two (out of 2000) of the alternate $A(p_i)$ contained exactly one outcome more than the Sterne $A(p_i)$, and in the final case an extra outcome had to be added to $A(p_i)$ for all p_i in the interval $[.5370,.5515]$.

In complicated cases (the worst yielded a Sterne confidence set consisting of the union of five intervals) there were numerous ways to alter the $A(p_i)$ to achieve (a)-(c) above. In one of these cases all possible sets of $A(p_i)$ satisfying (a)-(c) were constructed. The resulting confidence intervals were found to be equivalent in that their coverage probabilities and expected lengths were equal to 3 decimal places for all p. Thereafter, only one alternative set of $A(p)$ was constructed to determine the final intervals.

The modification to produce intervals still yields confidence sets which exhibit (ii); i.e., (g,s) outcomes which do not appear in any

$A(p_i)$. However this is not an acute problem. Figure 1 provides for the second Fleming plan with $(p_0, p_A) = (.1, .3)$ and $\alpha = .10$, a plot of $R(p)$ = probability when p is the true success rate of stopping at an outcome (g,s) not in any $A(p_i)$. For most values of p, $0 < p < 1$, $R(p)$ is quite small; even at its maximum R is less than $.0198(<.2\alpha)$. It should be noted that $R(p)$ is even lower for smaller α values.

The confidence intervals reported in Tables 2a-2d associate these very low probability (g,s) outcomes with the one point confidence interval $(s/\Sigma_{j=1}^{s} n_j)$ which is the value of p which maximizes $f((g,s)|p)$. While this is unsatisfying from a practical point of view, it is favorable for producing a system of intervals which most nearly attains its nominal coverage probability and has short expected length. This convention makes the comparisons of Section 5 particularly sharp.

It should be noted that the problem of one point confidence sets is not confined to the grouped sequential binomial case. In the one sample case, the large sample interval $\hat{p} \pm \hat{p}(1-\hat{p})/n$ with $\hat{p} = X/n$ consists of a point when $\hat{p} = 0$ or 1. An alternative which avoids one point confidence sets in the multistage situation is grouping outcomes (in some arbitrary fashion) before applying the modified Sterne method; this is also dissatisfying.

There is one other anomaly in the Table 2a-2d intervals which deserves comment. In the second and subsequent stages the upper endpoint of the interval can decrease as S increases. This phenomenon occurs only as the observed number of successes S gets quite large; initially as S increases the upper endpoint increases as well. Eventually, the large S

outcomes occur with very low probability. The decrease in the upper endpoint is caused by the fact that rare outcomes S appear in few Sterne acceptance sets and hence have short confidence intervals.

In summary, the modified Sterne intervals have short summed length but require substantial human intervention is required to reconcile inversions which yield unions of intervals as confidence sets. In addition rare outcomes can occur which lead to point confidence sets. The next section describes an alternative method which is relatively easily computed, automatically yields intervals, and whose operating characteristics are nearly as good as those of the modified Sterne method.

4. Multistage Crow-Blyth-Still Confidence Intervals

Given a linearing ordering of the outcomes of a multistage experiment, the basic recipe for determining the $L(p_i)$ and $U(p_i)$ is as follows. First, a partition $0 < p_1 < \ldots < p_r < 1$ of $(0,1)$ is specified as in Section 3. If $(g,s)_1$ is the smallest outcome in the ordering then the set $A(p_1)$ is initialized to be $\{(g,s)_1\}$; equivalently $L(p_1) = U(p_1) = (g,s)_1$ which corresponds to a $100(1-\alpha)\%$ acceptance region for sufficiently small p_1. Subsequent $L(p_i)$ and $U(p_i)$ are determined from previous $L(\cdot)$ and $U(\cdot)$ so as to satisfy (i) $L(p_i) \geq L(p_{i-1})$, (ii) $U(p_i) \geq U(p_{i-1})$, (iii) $A(p_i) = \{L(p_i),\ldots,U(p_i)\}$ has coverage probability $1-\alpha$, and (iv) $A(p_i)$ is as "small" as possible in the sense of cardinality. Property (iv) is incorporated into $A(p_i)$ by first trying to <u>eliminate</u> one or more outcomes from $A(p_{i-1})$ beginning with $L(p_{i-1})$. If this is not possible, one tries to <u>substitute</u> $U(p_{i-1})+1$ for $L(p_{i-1})$

to obtain $A(p_i)$. If one can neither eliminate nor substitute one adds points to $A(p_{i-1})$, beginning with $U(p_{i-1})+1$, to form $A(p_i)$. At this point some convention is often needed to choose among equal-cardinality $A(p_i)$ satisfying (i)-(iii) above.

Note that any interval constructed by such an algorithm automatically has the property that all outcomes must be in at least one $A(p_i)$. For the single sample binomial problem Crow (1956) uses the convention of forcing $L(p_i)$ and $U(p_i)$ to the right as rapidly as possible ["right-handed intervals"] while Blyth and Still (1983) adopt a rule which results in intervals of approximately equal tail probabilities.

Intervals were calculated by the above process for combinations of two linear orderings (ordering (1) ["JT order"] and ordering $S/\Sigma_{j=1}^{G} n_j$ ["ratio order"]) and three rules for determining $L(p_i)$ and $U(p_i)$ ["right-handed", "left-handed", and "equal-tailed" methods]. The latter chooses among minimal cardinality $A(p_i)$ by minimizing the difference between the tail probabilities. These methods were applied to the four $K = 3$ stage plans of Fleming (1982).

Before describing the resulting intervals, we note the following convention for handling one indeterminacy of the method. The ratio ordering is not uniquely defined for all the plans in Table 1 as there exist distinct outcomes with identical $S/\Sigma_{j=1}^{G} n_j$ values. Our investigation indicates that the operating characteristics of intervals based on different ratio-orderings are quite similar. Thus, we consider from now on that ratio-ordering which breaks ties by retaining the order given in (1).

The intervals computed by various combinations of ordering and rule had, in general, shorter total summed length (over all outcomes), shorter expected length, and more nearly nominal coverage probability than the JT tail intervals. The disturbing feature of these methods is the substantial difference in operating characteristics depending on the combination of ordering and rule used, particularly between the right- and left-handed intervals.

The combination of the ratio-ordering $(S/\Sigma_{j=1}^{G} n_j)$ and the choice from among equal cardinality $A(p_i)$ of that set which minimizes the difference between the tail probabilities produced the most attractive intervals in terms of lower expected length and more nearly nominal coverages. Although the imposition of the ratio-ordering is arbitrary, the method above can be fully automated making it feasible in practical applications. Detailed comparisons of its operating characteristics for the Table 1 plans with those of the JT intervals and the modified Sterne intervals are given in the next section.

5. Comparisons

We compare the operating characteristics of the JT tail intervals, the modified Sterne (MS) intervals and the equal tailed version of the ratio-ordered Crow (RC) intervals for the Fleming boundaries introduced in Table 1. We begin by assessing the extent to which the modified Sterne intervals successfully minimize the total summed length over all outcomes. Table 3 lists this value for 90% and 95% intervals based on Plans 1 to 4.

Table 3

Total summed confidence interval length over all outcomes of the JT, MS, and RC intervals and the % decrease of the MS and RC total lengths over the JT total length for the Fleming (1982) K = 3 stage tests of Table 1.

Plan	α	JT	MS (% dec)	RC (% dec)
1	.10	14.029	6.967 (50%)	12.378 (12%)
	.05	15.887	8.376 (47%)	14.533 (9%)
2	.10	13.312	7.177 (46%)	11.018 (17%)
	.05	15.042	8.692 (42%)	13.025 (13%)
3	.10	14.822	8.502 (43%)	13.320 (10%)
	.05	17.169	10.466 (39%)	15.583 (9%)
4	.10	15.708	9.328 (41%)	13.596 (13%)
	.05	18.185	11.435 (37%)	15.974 (12%)

The MS intervals have uniformly shorter total length than the other two families of intervals while the RC intervals have uniformly intermediate summed length. The improvement in summed length of the MS intervals over the JT intervals ranges from a 37% to 50% decrease while the RC intervals exhibit a 9% to 17% decrease over the summed length of the JT intervals. The extent to which the considerable advantage of the MS intervals in terms of summed interval length translates into superior operating characteristics is investigated next.

Figure 2 plots the expected length of the JT, MS, and RC intervals versus $p = 0(.05)1$ for 90% intervals based on Plan 2. The comparisons for the other seven plan-by-level combinations are similar. For small p all three systems have similar (short) expected lengths. However for large p both the MS and RC intervals provide substantial savings over the JT intervals. The reason is as follows. When p is large, Plan 2 will, with

high probability, stop after stage 1. The JT intervals have longer lengths for outcomes $(g,s) = (1,s)$ with s near n_1 than the MS and RC intervals. What is more interesting is the very small difference in expected length between the MS and RC intervals. This shows that the extra summed length of the RC intervals over the MS intervals is due to the inclusion of very low probability outcomes in the acceptance regions which carry little weight in the expected length computation.

The final comparison is of the achieved coverage probabilities. Figure 3 is a typical plot of the achieved coverage probability for the JT, MS, and RC methods versus $p = .05(.05).95$ for the nominal 90% Plan 2 intervals whose expected lengths are given in Figure 2. For the grid of p values considered, the MS and RC intervals have actual coverage probability nearly uniformly closer to nominal than the JT intervals.

In summary, the favorable operating characteristics of the RC (ratio ordered Crow) intervals and the MS (modified Sterne) intervals compared to the tail intervals make them attractive alternatives. The relative ease with which the Crow intervals can be computed favors the former's use in routine applications. A FORTRAN program for calculating the ratio-ordering Crow intervals for other multistage binomial testing plans can be obtained by writing the authors.

ACKNOWLEDGMENT

We would like to thank a referee for suggesting that we investigate equal tailed versions of the Crow Intervals.

The first author's research was supported under a National Science Foundation Graduate Fellowship. The second author's research was partially supported by the U.S. Army Research Office through the Mathematical Science Institute at Cornell University.

REFERENCES

Armitage, P. (1958). Numerical Studies in the Sequential Estimation of a Bionomial Parameter. Biometrika 45, 1-15.

Atkinson, E.N. and Brown, B.W. (1985). Confidence Limits for Probability of Response in Multistage Phase II Clinical Trials. Biometrics 41, 741-744.

Blyth, C.R. and Still, H.A. (1983). Binomial Confidence Intervals. Journal of the American Statistical Association 78, 108-116.

Casella, G. (1984) Refining Binomial Confidence Intervals. Technical Report BU-827-M. Biometrics Unit Series, Cornell University.

Clopper, C.J. and Pearson, E.S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika 26, 404-413.

Crow, E.L. (1956). Confidence Intervals for a Proportion. Biometrika 43, 423-435.

Fleming, T.R. (1982). One Sample Multiple Testing Procedure for Phase II Clinical Trials. Biometrics 38, 143-151.

Jennison, C. and Turnbull, B.W. (1983). Confidence Intervals for a Binomial Parameter Following a Multistage Test With Application to MIL-STD 105D and Medical Trials. Technometrics 25, 49-58.

Schultz, J.R., Nichol, F.R., Elfring, G.L. and Weed, S.D. (1973). Multistage Procedures for Drug Screening. Biometrics 29, 293-300.

Sterne, T.E. (1954). Some Remarks on Confidence of Fiducial Limits. Biometrica 41, 275-278.

Tsiatis, Rosner and Mehta (1984). Exact Confidence Intervals Following a Group Sequential Test. Biometrics 40, 797-803.

Table 2a

Modified Sterne 90% and 95% confidence intervals for Fleming (1982) K = 3 stage test of Table 1 with $(p_0, p_A) = (.05, .20)$

g	s	95% lower limit	90% lower limit	90% upper[1] limit	95% upper[1] limit
1	4	.079	.096	.5	.534
	5	.116	.121	.567	.603
	6	.162	.170	.635	.668
	7	.198	.218	.675	.706
	8	.242	.263	.733	.778
	9	.288	.344	.795	.809
	10	.372	.389	.832	.858
	11	.422	.5	.878	.903
	12	.534	.567	.924	.943
	13	.603	.635	.964	.976
	14	.698	.733	.993	.997
	15	.778	.832	1.0	1.0
2	0	0	0	.121	.122
	1	.002	.004	.17	.191
	2	.012	.018	.218	.244
	5	.061	.075	.314	.341
	6	.091	.104	.344	.372
	7	.122	.135	.374	.402
	8	.145	.169	.389	.422
	9	.191	.233	.4	.448
	10	.244	.333		.431
	11	.367	.367		
	12	.4	.4		
	13	.433	.433		
	14	.467	.467		
	15	.5	.5		
	16	.533	.533		
	17	.567	.567		
	18	.6	.6		
3	3	.028	.037	.152	.186
	4	.038	.048	.228	.242
	5	.051	.063	.257	.265
	6	.071	.086	.263	.289
	7	.106	.152	.233	.288
	8	.186	.2		.198
	9	.225	.225		
	10	.25	.25		
	11	.275	.275		
	12	.3	.3		
	13	.325	.325		
	14	.35	.35		

[1]The two sided confidence interval for any (g,s) for which the upper limit is not listed is the point set consisting of the lower limit. See Section 3 for explanation.

Table 2b

Modified Sterne 90% and 95% confidence intervals for Fleming (1982) K = 3 stage test of Table 1 with $(p_0, p_A) = (.1, .3)$

g	s	95% lower limit	90% lower limit	90% upper[1] limit	95% upper[1] limit
1	0	0	0	.218	.25
	5	.118	.139	.567	.603
	6	.162	.178	.635	.668
	7	.205	.226	.675	.706
	8	.250	.289	.733	.778
	9	.296	.311	.795	.809
	10	.344	.411	.832	.858
	11	.441	.456	.878	.903
	12	.488	.567	.924	.943
	13	.603	.635	.964	.976
	14	.698	.733	.993	.997
	15	.778	.832	1.0	1.0
2	1	.003	.007	.162	.19
	2	.016	.023	.246	.274
	3	.034	.046	.29	.329
	6	.104	.112	.411	.441
	7	.143	.162	.455	.488
	8	.179	.194	.472	.499
	9	.226	.263	.439	.518
	10	.315	.356	.364	.455
	11	.44	.44		
	12	.48	.48		
	13	.52	.52		
	14	.56	.56		
3	4	.057	.084	.194	.226
	5	.066	.072	.269	.296
	6	.081	.097	.319	.344
	7	.095	.123	.331	.358
	8	.131	.155	.356	.39
	9	.19	.218	.344	.375
	10	.274	.286		.315
	11	.314	.314		
	12	.343	.343		
	13	.371	.371		
	14	.4	.4		
	15	.429	.429		

[1]The two sided confidence interval for any (g,s) for which the upper limit is not listed is the point set consisting of the lower limit. See Section 3 for explanation.

Table 2c

Modified Sterne 90% and 95% confidence intervals for Fleming (1982) K = 3 stage test of Table 1 with $(p_0, p_A) = (.2, .4)$

g	s	95% lower limit	90% lower limit	90% upper[1] limit	95% upper[1] limit
1	0	0	0	.218	.268
	1	.003	.007	.326	.366
	8	.248	.272	.733	.778
	9	.305	.326	.795	.809
	10	.357	.386	.832	.858
	11	.413	.442	.878	.903
	12	.471	.551	.924	.943
	13	.584	.611	.964	.976
	14	.658	.733	.993	.997
	15	.778	.832	1.0	1.0
2	2	.024	.036	.105	.174
	3	.036	.048	.237	.255
	4	.052	.066	.281	.305
	5	.071	.087	.343	.357
	6	.092	.105	.377	.405
	7	.116	.119	.41	.457
	11	.199	.218	.538	.552
	12	.229	.237	.551	.584
	13	.255	.281	.597	.611
	14	.274	.313	.611	.65
	15	.328	.343	.633	.658
	16	.366	.41	.655	.692
	17	.423	.458	.675	.706
	18	.49	.6		.675
	19	.633	.633		
	20	.667	.667		
	21	.7	.7		
	22	.733	.733		
3	8	.178	.178		
	9	.171	.2		.248
	10	.141	.164	.301	.328
	11	.149	.153	.359	.386
	12	.159	.177	.386	.413
	13	.174	.192	.419	.438
	14	.187	.208	.442	.471
	15	.215	.224	.458	.483
	16	.243	.252	.478	.497
	17	.268	.301	.466	.507
	18	.313	.377	.446	.49
	19	.405	.422		.423
	20	.444	.444		
	21	.467	.467		
	22	.489	.489		
	23	.511	.511		
	24	.533	.533		
	25	.556	.556		

[1]The two sided confidence interval for any (g,s) for which the upper limit is not listed is the point set consisting of the lower limit. See Section 3 for explanation.

Table 2d

Modified Sterne 90% and 95% confidence intervals for Fleming (1982) K = 3 stage test of Table 1 with $(p_0, p_A) = (.3, .5)$

g	s	95% lower limit	90% lower limit	90% upper[1] limit	95% upper[1] limit
1	0	0	0	.127	.193
	1	.003	.005	.226	.286
	2	.018	.027	.325	.343
	3	.042	.056	.384	.408
	4	.071	.09	.448	.474
	5	.104	.127	.5	.525
	12	.343	.362	.779	.791
	13	.384	.410	.799	.833
	14	.42	.448	.859	.86
	15	.474	.5	.873	.896
	16	.527	.561	.91	.929
	17	.589	.634	.944	.958
	18	.673	.69	.973	.982
	19	.722	.797	.995	.997
	20	.833	.873	1.0	1.0
2	6	.171	.171		
	7	.213	.24	.247	.257
	8	.14	.141	.308	.362
	9	.15	.163	.387	.42
	10	.163	.183	.431	.455
	11	.18	.206	.474	.498
	12	.193	.226	.512	.541
	17	.304	.325	.634	.653
	18	.330	.359	.646	.673
	19	.362	.387	.675	.714
	20	.408	.431	.699	.722
	21	.452	.474	.713	.732
	22	.498	.529	.69	.744
	23	.568	.593		.689
	24	.686	.686		
	25	.714	.714		
	26	.742	.742		
3	13	.26	.26		
	14	.28	.28		
	15	.286	.3		.304
	16	.251	.302	.359	.402
	17	.232	.247	.429	.452
	18	.241	.254	.466	.496
	19	.257	.269	.505	.527
	20	.265	.285	.529	.557
	21	.281	.308	.548	.568
	22	.299	.319	.561	.582
	23	.317	.344	.582	.597
	24	.35	.384	.593	.608
	25	.402	.429	.569	.617
	26	.455	.512	.526	.589
	27	.557	.54		.558
	28	.56	.56		
	29	.58	.58		
	30	.6	.6		
	31	.62	.62		

[1]The two sided confidence interval for any (g,s) for which the upper limit is not listed is the point set consisting of the lower limit. See Section 3 for explanation.

FIGURE 1

Probability of obtaining an outcome corresponding to a one point confidence interval for 90% Plan 2 intervals (Table 2b).

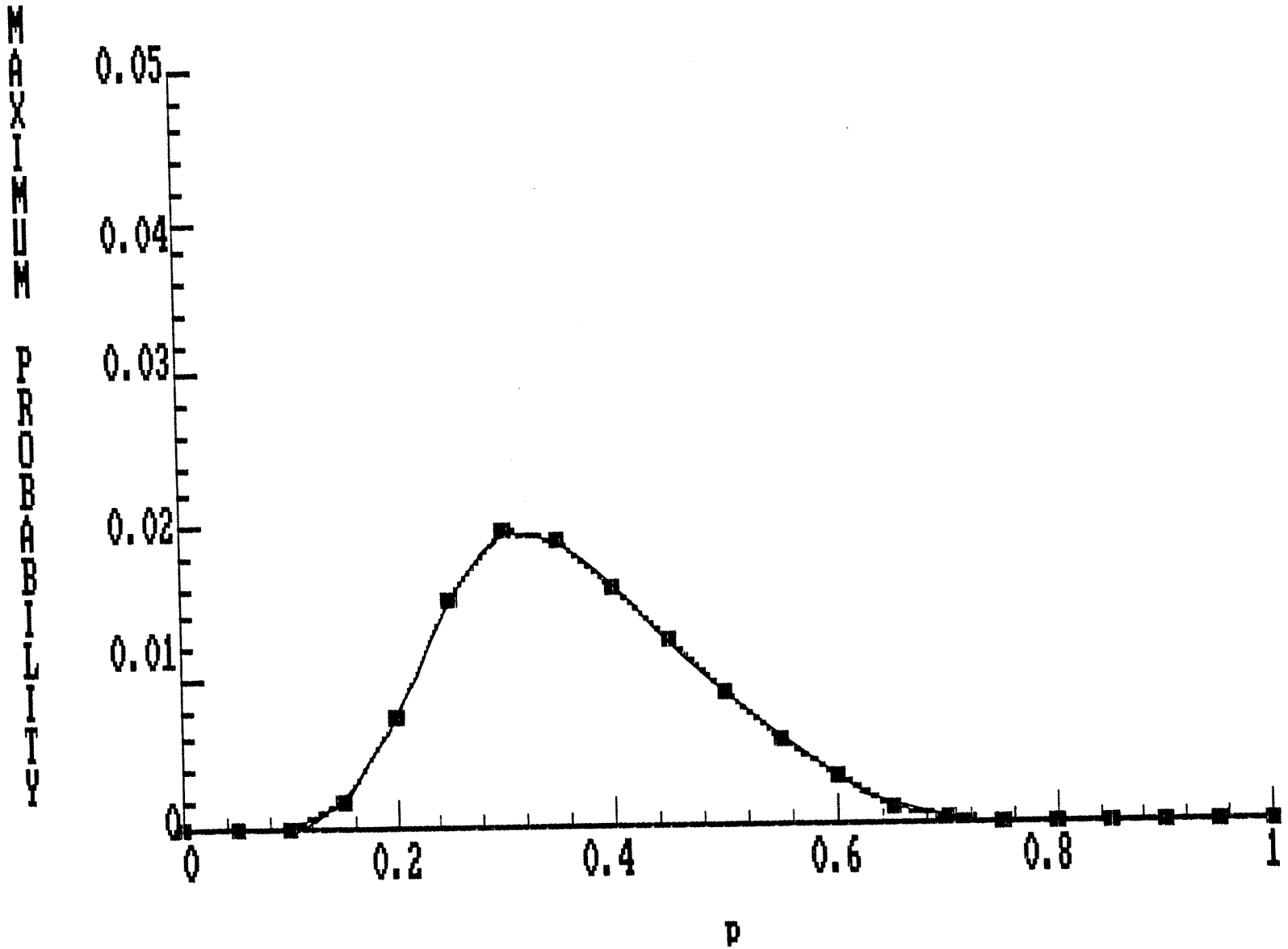

FIGURE 2

Expected lengths of MS (——■——), RC (.....),
and JT (————) intervals vs. p = 0(.05)1.

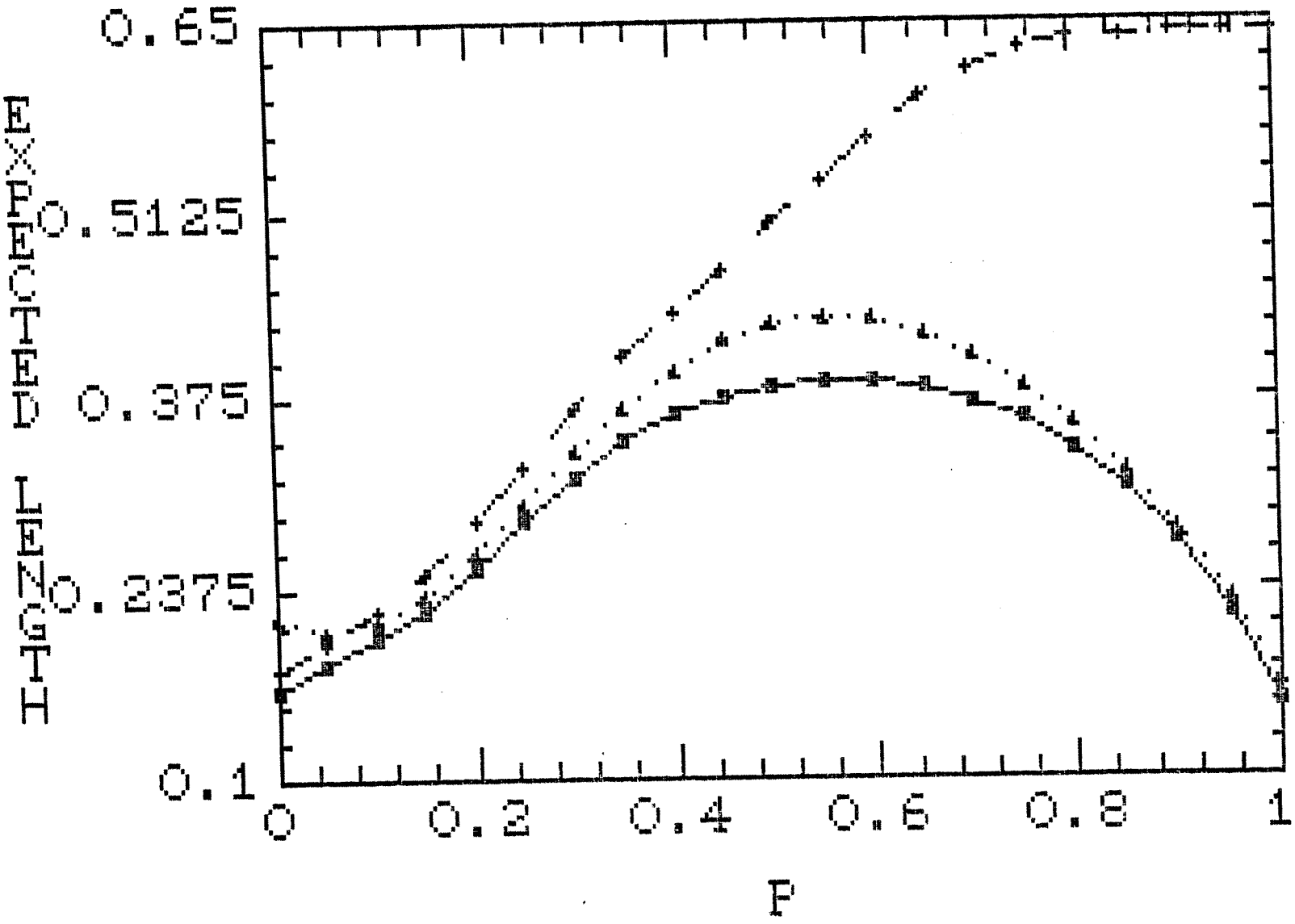

FIGURE 3

Coverage probability of MS (——■——), RC (.....),

and JT (--------) intervals vs. p = .05(.05).95.