Statistica Sinica **8**(1998), 393-409

THE EVALUATION OF CONFIDENCE SETS WITH APPLICATION TO BINOMIAL INTERVALS

Michael D. deB. Edwardes

Royal Victoria Hospital

Abstract: An expected volume coefficient (EV) is defined and proposed to displace volume and selectivity as criteria for the evaluation of confidence sets, and a proposal for evaluation is given. This proposal addresses anomolies that occur with sets based on discrete probability distributions; for example, that classical exact confidence intervals are wider than approximate ones. Options for the other key criterion, coverage, range from attaining average coverage, a liberal (i.e., leading to smaller sets) criterion, to attaining coverage for all values of the unknown parameter and all sample sizes, a very conservative criterion for sets based on discrete distributions. Use of EV is demonstrated with two-sided confidence intervals for the binomial probability parameter, leading to new recommendations; in particular, a Wald logit interval with negative continuity correction.

Key words and phrases: Average coverage, binomial proportion, confidence intervals, continuity correction, coverage, expected volume, expected width, logit, Neyman shortness, selectivity.

1. Introduction

The problem of constructing a confidence interval based on a (binomial) proportion would seem at first consideration to be a relatively simple matter. Many textbooks have given that impression. The topic is actually quite controversial and has generated much literature. One is led to believe that the exact confidence interval, that is, the one based directly on binomial probabilities, is the gold standard against which all approximate intervals are to be judged. However, there are at least four alternative exact binomial confidence intervals, and they can be very different. For example, Blyth and Still (1983) report that the Clopper-Pearson (1934) exact interval can be 12.6% longer than their exact interval. Different good properties of intervals are enforced in the construction of alternative exact intervals. These properties, such as equal probability tails and monotonicity in sample size, go beyond the definition of a confidence interval. Evaluation is usually attempted on the basis of the definition. However, the requirement of *attaining coverage* in the standard definition is controversial for binomial intervals.

The success of a given confidence set is usually judged by its being in some sense the smallest set that attains a given probability of including the true value of the parameter of interest. Ghosh (1979) evaluated binomial confidence intervals using three basic criteria: coverage, width and Neyman shortness. Neyman shortness, known also as selectivity, is the propensity of a confidence set to cover all false values of the parameter with relatively small probabilities. Adequate coverage is to selectivity as Type I error is to Type II error. If selectivity is ignored, confidence intervals may be chosen that are too wide in probability. Due to the Ghosh-Pratt identity (Ghosh (1961), Pratt (1961)) for a given true parameter value, the probability of false coverage by a set equals the expected volume of the set. This fact was used by Cohen and Strawderman (1973) in proposing admissibility criteria and by Brown, Casella and Hwang (1995) in optimizing confidence sets. Selectivity is usually ignored in confidence interval proposals and evaluations; for example, by Blyth and Still (1983) and Vollset (1993). This may be partly due to the fact that when it is considered, it is shown awkwardly as a table of a few selected probabilities of false coverage, as in Ghosh (1979) and Edwardes (1994). A coefficient (EV), which is a weighted average over exact expected volumes, is introduced below in order to summarize in one number what is incompletely shown in separate tables or graphs of volume (width) and Neyman shortness (selectivity).

In §3, a proposal for the evaluation of confidence sets is given, based on average coverage, expected volume (EV) and, for discrete intervals, minimum coverage. In §4, the proposal is applied to previously recommended and new binomial confidence intervals and the exact interval properties are discussed.

2. Definitions

It is possible for a confidence set to be disjoint, even when the parameter space is not. As discussed by Blyth and Still (1983), it is possible for a disjoint set to attain coverage and have smaller volume than the shortest binomial interval that attains coverage. Non-disjointness is not required for our arguments, but is assumed for the binomial application.

For simplicity, all confidence set bounds are assumed to be either within Θ, the parameter space for the vector θ, or at limits of infinite sequences within Θ. In the evaluation of confidence sets, therefore, the criteria that I propose are to be applied to the sets after having been truncated so that they lie (almost) within Θ.

Let X have distribution function $F(\cdot|\theta)$. Say that a $100 \times (1-\alpha)\%$ nonrandomized confidence set $C(x)$ is required for θ, given X$=x$ is observed. The corresponding inclusion probability is $P(\theta \epsilon C(x)|x)$, which is simply $I(\theta \epsilon C(x))$, an indicator function equal to one or zero, for a nonrandomized set. The volume of $C(x)$ with respect to Lebesgue measure is given by $vol(C(x)) = \int_\Theta I(t \epsilon C(x))dt$. The expected volume is

$$E_\theta(vol(C(\mathrm{X}))) = \int_\Omega vol(C(x))dF(x|\theta), \tag{1}$$

given θ as the true value, where Ω is the sample space of X. This is an obvious measure of size, given θ. The *coverage* probability of $C(\mathrm{X})$ is

$$P_\theta(\theta\epsilon C(\mathrm{X})) = \int_\Omega I(\theta\epsilon C(x))dF(x|\theta). \tag{2}$$

Applying the Ghosh-Pratt identity, the expected volume (1) may be re-expressed as

$$E_\theta(vol(C(\mathrm{X}))) = \int_\Theta P_\theta(\theta'\epsilon C(\mathrm{X}))d\theta'. \tag{3}$$

As an example, consider the problem of forming a confidence interval $C(\mathrm{X})$ for the binomial parameter p, the fixed probability of an event with parameter space [0, 1], given a random sample of size n with x observed events. Then $C(x)$ is a formula in terms of x and n for the upper and lower limits of the interval, $U(x,n)$ and $L(x,n)$, respectively. (For a one-sided interval, either $U = 1$ or $L = 0$.) The volume is the width of the interval, and so the expected volume is

$$\sum_{x=0}^{n}[U(x,n) - L(x,n)]\binom{n}{x}p^x(1-p)^{n-x}, \tag{4}$$

given n and p. The coverage probability is

$$P_p[L(\mathrm{X},n) \leq p \leq U(\mathrm{X},n)] = \sum_{x:p\epsilon C(x)}\binom{n}{x}p^x(1-p)^{n-x}.$$

Applying the Ghosh-Pratt identity, the expected volume (4) may be re-expressed as

$$\int_0^1 \sum_{x:p'\epsilon C(x)}\binom{n}{x}p^x(1-p)^{n-x}dp' \ ,$$

thus illustrating the correspondence between expected volume and selectivity. Of course, formula (4) is computationally easier.

The standard definition is that $C(\mathrm{X})$ be the smallest interval satisfying the standard requirement $\min_\theta P_\theta(\theta\epsilon C(\mathrm{X})) \geq 1-\alpha$, called attaining coverage. Given $P_\theta(\theta\epsilon C(\mathrm{X})) \geq 1-\alpha$, Brown, Casella and Hwang (1995) seek to minimize (1) at a selected value of θ, but their application is an unusual situation where such a value is selected a priori. A problem with ordinary confidence sets with F discrete is that the standard requirement leads to the ideal confidence interval being selected on the basis of an unknown parameter. That is, for each θ, there is a shortest $C(\mathrm{X})$. An alternative coverage requirement (Santner and Duffy (1989)) is to *attain average coverage*

$$AC(C(\mathrm{X})) = \int_\Theta P_\theta(\theta\epsilon C(\mathrm{X}))dH(\theta) \geq 1-\alpha,$$

where H is a somewhat arbitrary d.f. (distribution function) that gives equitable weight to all values θ. If Θ is bounded, H may be the uniform d.f. over Θ.

The solution for expected volume and selectivity proposed here is to integrate (1) or (3) over Θ to obtain an Expected Volume coefficient (EV):

$$EV(C(\mathrm{X})) = \int_{\Theta}\int_{\Theta} P_{\theta}(\theta' \epsilon C(\mathrm{X})) d\theta' dH(\theta),$$

where H is as described for AC. Although $EV(C(\mathrm{X}))$ is an improvement in summarizing, it still varies with fixed parameters such as sample size and α. For the binomial parameter, it is more precisely written $EV(C(\mathrm{X}), \alpha, n)$. (I write $EV(C(\mathrm{X}), n)$ and $AC(C(\mathrm{X}), n)$ below for a specific n.) Thus, any evaluation of binomial confidence intervals must still consider alternative values of n and α. Unlike p, however, these are known parameters.

Choice of H. The expected volume criterion suggested by Brown, Casella and Hwang (1995), formula 16 is $EV(C(\mathrm{X}))$, described as a Bayesian criterion since H can be an a priori parameter distribution. It is proposed here that H can be simply a device to give equitable weight to all θ. Whenever H can be the uniform d.f. over Θ (i.e., when Θ is completely bounded or compact), it should be. For example, I use $dH(p) = dp$ for the binomial p.

When Θ is not completely bounded, H should still somehow be equivalent to uniform. My initial preference is to choose bounds within Θ in order to create a bounded, compact sub-space and have AC and EV based on H uniform in the sub-space. The criteria AC and EV may then be calculated at different boundary values as the sub-space converges to Θ, in order to show either convergence towards a limit or else the relationship between diverging criteria and the changing boundary values.

The choice of H here has much in common with the on-going search for the perfect *non-informative prior* (Bernardo and Smith (1994)), except that in this application it is not sensible to require H to be invariant to transformation. If θ is transformed, then a different confidence interval is the goal.

3. Proposal

Proposal for Confidence Set Evaluation: Over a family of alternative formulas $C(\mathrm{X})$ for sets $C(x)$ in Θ, the most *accurate* confidence set formula is the one with smallest $EV(C(\mathrm{X}))$ for which $AC(C(\mathrm{X})) \geq 1 - \alpha$ and $\min_{\theta} P_{\theta}(\theta \epsilon C(\mathrm{X})) \geq 1 - \alpha - k\alpha$, with $k = 1/2$.

Since the most accurate $C(\mathrm{X})$ can be different when fixed parameters (e.g., α and n) differ, a thorough tabulation should be done of all three criteria by fixed parameters. For example, I calculate the binomial $MC(n) = \min_p P_p(p\epsilon C(\mathrm{X}))$ for $\alpha = 0.01, 0.05, 0.1$ and ranges of n from 1 to $10,000$. Calculations are done for every n in that range because F is discrete. This contrasts with the idea explored in Blyth and Still (1983) of just looking for $\lim_{n\to\infty} MC(n)$ (which they show does not converge to $1-\alpha$, contrary to previous opinions). That just may not be relevant when $30 < n < 100$. The criteria AC and EV change smoothly as n varies for the binomial problem, so these need not be calculated for every n. The most accurate approximate $C(\mathrm{X})$ for large n may well be different from that for small n. A recommended interval, if it is to be the same $C(\mathrm{X})$ over $\alpha = 0.01$, 0.05, 0.1, may have to be a compromise, one whose properties are close to those of the most accurate interval as α and n vary.

For F binomial or Poisson, I regard the requirement $k = 1/2$ to be conservative. Choosing $k = 1$ is liberal, though certainly not as liberal as simply attaining average coverage. Choosing $k = 0$ is very conservative, and works against the desirable minimization of expected volume. The value of $k > 0$ is irrelevant when $P_\theta(\theta \epsilon C(\mathrm{X})) = 1 - \alpha$ is always achievable, as when F is continuous, since then $AC(C(\mathrm{X})) = 1 - \alpha$.

For a thorough and definitive evaluation, the calculations should be exact. When $P_\theta(\theta \epsilon C(\mathrm{X}))$ is difficult to compute, however, the average coverage and the average volume over (uniform or H-based) simulations will likely be accurate estimates of AC and EV, respectively. For discrete F, $\min_\theta P_\theta(\theta \epsilon C(\mathrm{X}))$ is required, but the minimum based on monte-carlo i.e., randomly selected values of θ and of fixed parameters can over-estimate the true minimum. For a rough comparison of confidence sets a monte-carlo minimum may suffice, but I cannot say at this point what conditions are required for a simulation to yield an acceptable estimate.

4. Applications to Binomial Intervals

4.1. Four exact intervals–with discussion

The current standard for an exact, two-sided nonrandomized binomial confidence interval is Blyth and Still's (1983) shortest exact interval. This is the interval of minimum sum of $n+1$ widths (i.e., from $x = 0$ to $x = n$) that attains coverage for all values p, while satisfying four "desirable" properties (below). Minimizing expected width seems a more pragmatic criterion, since it would give less weight to less likely binomial data. Nevertheless, an interval which minimizes the sum of widths will likely have a relatively small expected width. The reason that the shortest exact interval is a notable improvement in being shorter than the previous standard—the ubiquitous Clopper-Pearson (1934) exact interval—is that it avoids the imposition of an extra requirement; stated roughly, equal probability above and below the acceptance interval. That extra but unnecessary requirement made the mathematics easier.

Another requirement avoided by Blyth and Still is *unbiasedness*, which is the property $P_p(p' \epsilon C(\mathrm{X})) \leq P_p(p \epsilon C(\mathrm{X}))$ for all p, p'. Diagonal entries in a selectivity table, such as ‘ Table 1, where $p = p'$, may be used to detect departure from unbiasedness.

Table 1. Neyman shortness (selectivity) comparison of three confidence intervals: exact Binomial probabilities for $1-\alpha = 0.99$ and $n = 100$

	True p											
	0.05			0.1			0.3			0.5		
False p'	β_L^*	β_S	β_{S1}	β_L	β_S	β_{S1}	β_L	β_S	β_{S1}	β_L	β_S	β_{S1}
0.001	.0371	.0371	.0371	.0003	.0003	.0003	.0000	.0000	.0000	.0000	.0000	.0000
0.01	.4360	.4360	.4360	.0237	.0237	.0237	.0000	.0000	.0000	.0000	.0000	.0000
0.05	.9957	.9957	.9957	.7030	.7030	.7030	.0000	.0000	.0000	.0000	.0000	.0000
0.1	.9941	.9629	.9629	.9954	.9951	.9951	.0045	.0045	.0045	.0000	.0000	.0000
0.15	.4152	.3840	.5640	.9427	.9424	.9763	.1136	.1136	.1631	.0000	.0000	.0000
0.2	.0282	.0631	.0631	.5487	.6791	.6791	.5491	.6331	.6331	.0000	.0001	.0001
0.25	.0005	.0005	.0015	.1239	.1239	.1982	.9201	.9201	.9469	.0033	.0033	.0060
0.3	.0000	.0000	.0000	.0046	.0100	.0100	.9915	.9939	.9939	.0666	.0666	.0666
0.35	.0000	.0000	.0000	.0001	.0003	.0003	.9520	.9711	.9711	.3086	.3822	.3822
0.4	.0000	.0000	.0000	.0000	.0000	.0000	.7036	.7756	.7756	.6913	.7579	.7579
0.5	.0000	.0000	.0000	.0000	.0000	.0000	.0530	.0799	.0799	.9880	.9934	.9934
0.6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0003	.6913	.7579	.7579

* $\beta_L = P_p(p' \epsilon WL(-0.57))$, $\beta_S = P_p(p' \epsilon SBS(r))$, where $0.72 \le r \le 0.78$ and $\beta_{S1} = P_p(p' \epsilon SBS(0.86))$.
Results for $p = 0.5$ and $p' = 0.5 + d$ are equal to the results for $p' = 0.5 - d$.

I propose that two of Blyth and Still's four "desirable" properties should be enforced for binomial intervals and the other two should not be enforced. The first two are (1) that the interval be equivariant and (2) not be disjoint. The debatable two are (3) that the interval have monotonicity in n and (4) monotonicity in x. The first property retains the symmetry of the binomial problem. Equivariance means that any joint substitution $p \to 1-p$ and $x \to n-x$ changes no values or formulas. To not do so would be to give differential weight to different values of p since the binomial F is unchanged with this substitution. The second property I accept because it is an interval that is required, not just any type of region. Monotonicity in n means that $U(C(x), n+1) < U(C(x), n)$ and $L(C(x), n+1) < L(C(x), n)$. Monotonicity in x means that $U(C(x), n) < U(C(x+1), n)$ and $L(C(x), n) < L(C(x+1), n)$. Although both of these properties are attractive, their enforcement restricts the choice of formulas $C(\mathrm{X})$ so that a most accurate $C(\mathrm{X})$ may be excluded. Casella's (1986) algorithm for creating exact confidence intervals will not work if inequality is not strict for properties (3) and (4). Is it more important that $C(\mathrm{X})$ be accurate over all p and x or that it be beautiful for every p and x? Blyth and Still (1983) chose to impose (3) and (4), but not unbiasedness or equal probability tails. They state that their shortest exact interval is *approximately* unbiased and the tails have *approximately* equal probability. I propose that (3) and (4), as well as unbiasedness and equal property tails, be regarded as attractive properties rather than necessary requirements. They will tend to be satisfied anyway for more accurate $C(\mathrm{X})$. It is useful to enforce such properties when forming exact intervals, in order to reduce the choice of solutions based on a set of requirements, as in Blyth and Still (1983) and Casella (1986). Evaluation is another matter, however, which should be done on the basis of overall accuracy.

Despite the preceding, the most accurate $(0.58 > k \geq 0)$ binomial confidence interval shown in this paper is Blyth and Still's shortest exact interval (labelled *BSEXACT*). This may be because the other intervals are all based on symmetric tails. Unfortunately, this interval is not available at the 90% level or for $n > 30$.

Also shown are results for the Clopper-Pearson interval, labelled *CPEXACT*, and the mid-P exact interval (Stone (1969), Miettinen (1985), Vollset (1993)), labelled *MPEXACT*. The interval usually labelled the exact binomial interval and found in tables such as Fisher and Yates (1948) and Lentner (1982) is *CPEXACT*. The interval herein labelled *MPCXACT* is *MPEXACT* for $0 < x < n$ and *CPEXACT* when $x = 0$ or $x = n$, as advocated by Vollset (1993). The *MPCXACT* interval is recommended specifically for $\alpha = 0.01$ by Vollset (1994). Except at $x = 0, 1$ or $x = n-1, n$, these three intervals cannot be expressed in closed form. The bound L is the minimum $p \leq x/n$ for which $cp_x + \sum_{j=x+1}^{n} p_j \geq \alpha/2$ and the bound U is the maximum $p \geq x/n$ for which $cp_x + 1 - \sum_{j=x}^{n} p_j \geq \alpha/2$, where $p_j = (n!/[j!(n-j)!])p^j(1-p)^{n-j}$, $j = 0,\ 1, \ldots, n$. For *CPEXACT*, $c = 1$. For *MPEXACT*, $c = 1/2$. The computations can be done easily with software that computes F distribution tails (Ling (1992)).

Thus, we have four "exact" intervals! This is an anomoly that is due to imposing differing requirements on the tails which are not part of the definition of a confidence interval. The *MPEXACT* and *MPCXACT* intervals actually violate the standard requirement, as seen in Table 2. By the standard confidence set definition *MPEXACT* and *MPCXACT* are not confidence sets!

Table 2. Minimum coverages* for 99% Binomial intervals listed in order of overall minimum

99% intervals	MC_{m0}	MC_{m1}	MC_{m2}	MC_{m3}	MC_{m4}	n at min	$MC(1000)\%$	$MC(10000)\%$
$SBS(0.5)$	96.61	97.60	97.59	97.56	97.55	1	97.55	97.55
$WL(-0.3)$	97.72	97.63	97.63	97.62	97.62	275	97.62	97.62
$SBS(0.60)$	97.80	98.08	98.07	98.05	98.04	1	98.04	98.04
$SBS(0.62)$	98.01	98.17	98.16	98.14	98.13	1	98.13	98.13
$WL(-0.39)$	98.04	98.03	98.04	98.04	98.05	14	98.06	98.06
$MPEXACT$	98.40	98.25	98.36	98.25	98.27	16	98.66	98.61
$MPCXACT$	98.62	98.34	98.37	98.38	98.40	16	98.66	98.63
$WL(-0.5)$	98.44	98.44	98.47	98.49	98.51	8	98.51	98.52
$WL(-0.52)$	98.51	98.51	98.55	98.56	98.58	8	98.58	98.59
$SBS(0.72)$	98.62	98.57	98.56	98.54	98.54	10000	98.54	98.54
$WL(-0.57)$	98.66	98.66	98.72	98.61	98.65	45	98.65	98.64
$WL(-0.95)$	99.00	98.77	98.77	98.67	98.74	43	98.86	98.86
$SBS(0.78)$	98.83	98.79	98.77	98.76	98.75	10000	98.75	98.75
$SBS(0.86)$	99.00	99.00	99.00	99.00	99.00	10000	99.00	98.99
$BSEXACT$	99.00	99.00	99.00					
$CPEXACT$	99.19	99.03	99.02	99.00	99.00	67	99.00	99.00

* Minima found by search through values of p near confidence bounds, confirmed by systematic search at 5 decimal places. The n at min(imum) is over $10001 > n > 0$, but some values of $p > 0.12$ were not checked for $n > 3000$. $MC(n)\% = \min_p P_p(p \epsilon C(\mathrm{X})) \times 100$. $MC_{mi} = \min_{n \epsilon A_i} MC(n)\%$ for $i = 0, 1, 2, 3, 4$, with $A_0 = \{n : 0 < n < 10\}$, $A_1 = \{n : 9 < n < 21\}$, $A_2 = \{n : 20 < n < 31\}$, $A_3 = \{n : 30 < n < 101\}$ and $A_4 = \{n : 100 < n < 1001\}$.

4.2. Approximate binomial intervals

I describe here two very good formulas for the nonrandomized two-sided binomial confidence interval: a score interval and a logit interval, both based on a Normal approximation with corrections for continuity. The simplistic binomial interval $x/n \pm z(\hat{p}(1-\hat{p})/n)^{1/2}$ has been shown by several authors, including Ghosh (1979) and Vollset (1993) to be extremely inaccurate, even with continuity correction (Blyth and Still (1983)). It is not even uniformly robust (Lehmann and Loh (1990)). Therefore, I mention it no further except to say that usually $AC < 1-\alpha$, $k > 1$ and it is still the most widely taught and used binomial interval.

The score (or test-based) interval (Wilson (1927)) has been recommended often over all alternative approximate intervals for the binomial parameter p. Blyth and Still (1983), Santner and Duffy (1989) and Vollset (1993) specifically recommend a modified score interval with continuity correction 1/2. The modification is the substitution of exact limits at $x = 0,\ n$ and (Blyth and Still) at $x = 1,\ n-1$. The modified score interval proposed by Vollset differs from that proposed by Blyth and Still in the limits for $x = 0,\ 1,\ n-1$ and n. Edwardes (1994) shows that these intervals do not come close to attaining coverage for $\alpha = 0.01$, even though they do so for $\alpha = 0.05$ and 0.1. Because of that problem, Blyth and Still proposed for $\alpha = 0.01$ the substitution of exact limits for $x = 0, 1, \ldots, 14, n-14, n-13, \ldots, n$. However, the exact limits, if they are to be from the shortest exact interval, have not been tabulated for $n > 30$. Some exact lower limits have a simple, closed form at $x = 0, 1, n-1, n$, and so I incorporate those.

A modified score $100\times(1-\alpha)\%$ confidence interval with continuity correction c_c and Normal critical point z is:

$$SBS(c_c)\ :\ \frac{(x \pm c_c) + \frac{z^2}{2} \pm z\left[(x \pm c_c) - \frac{(x\pm c_c)^2}{n} + \frac{z^2}{4}\right]^{1/2}}{n + z^2},$$

except that for $x = 0$, the lower limit is zero; for $x = n$, the upper limit is one; for $x = 1$, the lower limit is $1-(1-\alpha)^{1/n}$ and for $x = n-1$, the upper limit is $(1-\alpha)^{1/n}$. The value z is the usual point exceeded by $\alpha/2$ of the standard Normal probability (e.g., $z = 2.5758$ for $\alpha = 0.01$). The modified score interval as defined by Blyth and Still is $SBS(0.5)$. I consider alternative values for c_c.

A Wald logit interval (Rubin and Schenker (1987), Vollset (1993)) has been used to date either with or without a continuity correction $c_c = 1/2$. The modified Wald logit $100\times(1-\alpha)\%$ confidence interval with continuity correction $c_c > -1$

and Normal critical point z is:

$$WL(c_c)\ :\ 1-\Big[1+\Big(\frac{x+c_c}{n-x+c_c}\Big)\exp\Big\{\pm z\Big(\frac{n+2c_c}{(x+c_c)(n-x+c_c)}\Big)^{1/2}\Big\}\Big]^{-1},$$

for $0 < x < n$, except that the lower limit is $1-(1-\alpha)^{1/n}$ for $x = 1$ and the upper limit is $(1-\alpha)^{1/n}$ for $x = n-1$. For $x = 0$, the lower limit is zero and the upper limit is $1-(\alpha/2)^{1/n}$. For $x = n$, the upper limit is one and the lower limit is $(\alpha/2)^{1/n}$. Thus, the Clopper-Pearson exact limits are used for $x = 0$ and $x = n$, as suggested by Vollset.

All modified logit and score intervals for values of c_c examined in this paper satisfy Blyth and Still's list of four desirable properties, with one exception: the 99% upper limits for $WL(c_c)$ with $-0.6 < c_c < -0.2$ at $x = 1$ are larger than the upper limits at $x = 2$, which violates the monotonicity in x property. This also happens for $WL(-0.95)$, for which also the 99% upper limit at $x = 2$ is larger than the upper limit at $x = 3$ for $n > 30$, with the consequence of a longer than desirable interval for $x = 1, 2, n-2, n-1$.

Computing the width of the score interval reveals that a larger value of c_c yields a wider confidence interval, a property not shared by the logit intervals. For score intervals, therefore, minimizing the continuity correction minimizes the width.

4.3. Computation

All calculations in Tables 1 to 5 are exact, not based on monte-carlo, and computation is double precision (i.e., 16 digit accuracy). For both AC and EV, numerical quadrature of curves with very uneven surfaces is involved. Four digit accuracy was accomplished for sample sizes up to 1,000 using subroutine NDIMRI of Davis and Rabinowitz (1984) by integrating over formula (2) for AC with [0, 1] divided into 1113 sub-intervals and over formula (4) for EV with 4 sub-intervals. Computation time for EV at $n = 1000$ varied from 10 seconds to 68 seconds for logit and the score intervals, using Micro-Soft Fortran, version 4.10 on a 66 MHz. IBM 486 clone. Computation time of AC at $n = 1000$ varied from 36 seconds to 20 minutes, but did not exceed 44 seconds for three digit accuracy.

Tables 2 and 4 give $MC(n) = \min_p P_p(p\epsilon C(\mathrm{X}))$ for value ranges of n up to $10,000$. Finding $MC(n)$ is done quickly by taking advantage of the fact, pointed out indirectly by Blyth and Still, that the Binomial coverage minima must be found near $p = U(\mathrm{X}, n),\ L(\mathrm{X}, n)$. For the intervals in our examples, the minima for each n occur specifically at one of $p = U + \epsilon$ or $p = L - \epsilon$ for $\epsilon = 0$ or a very small number. I used $\epsilon = 10^{-9}$.

Table 3. Average coverage and expected volume for 99% Binomial confidence intervals*

					Sample size				
99% intervals	10	18	19	20	30	45	100	200	1000
$MPEXACT$	0.995	0.993	0.993	0.993	0.992	0.992	0.991	0.990	0.990
	0.581	0.452	0.441	0.431	0.358	0.295	0.200	0.142	0.0639
$MPCXACT$	0.996	0.994	0.994	0.994	0.993	0.992	0.991	0.991	0.990
	0.589	0.455	0.444	0.434	0.359	0.296	0.200	0.142	0.0639
$BSEXACT$	0.995	0.994	0.994	0.994	0.993				
	0.578	0.458	0.448	0.439	0.363				
$WL(-0.30)$	0.996	0.994	0.994	0.994	0.993	0.993	0.992	0.991	0.990
	0.618	0.481	0.469	0.458	0.376	0.307	0.204	0.144	0.0640
$WL(-0.50)$	0.997	0.995	0.995	0.995	0.994	0.993	0.992	0.991	0.990
	0.632	0.490	0.478	0.466	0.381	0.310	0.205	0.144	0.0640
$WL(-0.52)$	0.997	0.995	0.995	0.995	0.994	0.993	0.992	0.991	0.990
	0.634	0.491	0.479	0.467	0.382	0.310	0.205	0.144	0.0640
$WL(-0.57)$	0.997	0.995	0.995	0.995	0.994	0.993	0.992	0.991	0.990
	0.639	0.495	0.481	0.470	0.384	0.311	0.205	0.144	0.0640
$CPEXACT$	0.998	0.996	0.996	0.996	0.995	0.995	0.993	0.993	0.991
	0.617	0.480	0.468	0.457	0.378	0.310	0.208	0.147	0.0648
$SBS(0.5)$	0.996	0.995	0.995	0.995	0.994	0.994	0.993	0.992	0.991
	0.602	0.476	0.464	0.454	0.377	0.311	0.209	0.147	0.0649
$SBS(0.72)$	0.997	0.997	0.996	0.996	0.996	0.995	0.994	0.993	0.992
	0.625	0.492	0.480	0.470	0.389	0.319	0.213	0.149	0.0653
$SBS(0.86)$	0.998	0.997	0.997	0.997	0.997	0.996	0.995	0.994	0.992
	0.639	0.502	0.491	0.479	0.396	0.324	0.216	0.150	0.0656
$WL(-0.95)$	0.998	0.996	0.996	0.996	0.995	0.994	0.992	0.991	0.990
	0.700	0.546	0.532	0.519	0.423	0.341	0.221	0.153	0.0657

* Each double line is a row of average coverages, $AC(C(\mathrm{X}), n)$, where $C(\mathrm{X})$ is the confidence interval, followed by a row of $EV(C(\mathrm{X}), n)$. $BSEXACT$ is Blyth & Still's shortest exact interval as given in Table 2.1.3 of Santner & Duffy (1989). The intervals are listed in order of $EV(C(\mathrm{X}), 1000)$ (except by $EV(C(\mathrm{X}), 30)$ for $BSEXACT$). Shown ties in EV are only due to rounding, thus $EV(WL(-0.52), n) > EV(WL(-0.5), n)$ at $n = 1000$.

Table 4. Minimum coverages* listed in order of minimum over all $n > 5$

90% intervals	MC_{m0}	MC_{m1}	MC_{m2}	MC_{m3}	MC_{m4}	n at min	$MC(1000)\%$	$MC(10000)\%$
$MPEXACT$	83.4	84.1	83.3	82.5	85.8	58	85.8	85.7
$MPCXACT$	83.4	84.7	84.2	85.7	85.8	9	85.8	85.7
$WL(-0.5)$	87.0	84.2	85.4	85.9	86.6	13	87.2	87.1
$SBS(0.14)$	79.9	84.8	86.4	86.7	86.6	1	86.6	86.6
$SBS(0.15)$	80.4	85.1	86.5	86.8	86.7	1	86.7	86.6
$WL(0.11)$	85.6	85.3	85.2	85.1	85.1	3138	85.1	85.1
$WL(-0.43)$	86.9	85.2	85.5	86.0	86.7	18	87.7	87.6
$WL(-0.12)$	86.6	85.7	86.1	86.5	86.8	17	87.5	87.7
$WL(0.0)$	86.5	86.0	86.2	86.6	86.3	17	86.3	86.3
$SBS(0.27)$	85.3	87.4	88.2	88.1	88.7	1	89.0	89.0
$SBS(0.5)$	90.0	90.0	90.0	90.0	90.0	2	90.0	90.0
$CPEXACT$	91.1	90.1	90.0	90.0	90.0	10000	90.0	90.0
95% intervals								
$MPEXACT$	91.2	91.3	92.5	92.1	91.7	6	91.7	91.7
$MPCXACT$	92.8	92.1	92.5	92.1	92.6	35	92.8	92.7
$SBS(0.24)$	87.9	92.5	92.5	92.5	92.5	1	92.5	92.5
$WL(-0.13)$	92.6	92.5	92.5	92.5	92.6	25	92.6	92.6
$WL(-0.62)$	94.6	92.7	92.8	92.8	93.5	16	93.8	94.0
$WL(-0.5)$	94.6	92.9	92.8	92.8	93.6	21	94.1	93.9
$WL(-0.27)$	93.8	93.3	93.1	93.2	93.7	25	93.7	93.7
$SBS(0.29)$	89.4	93.1	93.1	93.1	93.1	1	93.1	93.1
$WL(-0.25)$	93.6	93.3	93.2	93.4	93.7	22	93.7	93.7
$SBS(0.31)$	89.9	93.3	93.3	93.3	93.3	1	93.3	93.3
$SBS(0.41)$	92.5	94.5	94.4	94.4	94.4	1	94.4	94.4
$SBS(0.5)$	94.5	95.0	95.0	95.0	95.0	1	95.0	95.0
$BSEXACT$	95.0	95.0	95.0					
$CPEXACT$	95.3	95.1	95.1	95.0	95.0	10000	95.0	95.0

* See footnote of Table 2. Order is for $n > 5$ since SBS coverage is minimized at $n = 1$ and is low for 90% intervals at $n = 5$. For listed 95% SBS intervals, the minimum over $1 < n < 10001$ is at $n = 10000$.

Table 5. Average coverage and expected volume *

			Sample size			
90% intervals	10	20	30	100	200	1000
$MPEXACT$	0.929	0.917	0.913	0.905	0.902	0.901
	0.393	0.283	0.233	0.129	0.0912	0.0408
$MPCXACT$	0.937	0.922	0.916	0.906	0.903	0.901
	0.403	0.286	0.234	0.129	0.0912	0.0408
$WL(-0.50)$	0.943	0.926	0.919	0.906	0.903	0.901
	0.412	0.290	0.236	0.129	0.0912	0.0408
$WL(-0.43)$	0.942	0.926	0.919	0.907	0.904	0.901
	0.410	0.289	0.236	0.129	0.0912	0.0408
$WL(0.0)$	0.935	0.922	0.916	0.907	0.904	0.901
	0.402	0.287	0.235	0.129	0.0913	0.0408
$WL(0.11)$	0.933	0.920	0.915	0.906	0.904	0.901
	0.400	0.287	0.235	0.129	0.0914	0.0409
$SBS(0.15)$	0.932	0.924	0.919	0.911	0.908	0.903
	0.396	0.289	0.238	0.131	0.0924	0.0411
$SBS(0.27)$	0.944	0.934	0.929	0.917	0.912	0.906
	0.414	0.299	0.245	0.133	0.0936	0.0413
$CPEXACT$	0.963	0.950	0.943	0.926	0.920	0.909
	0.448	0.317	0.257	0.137	0.0957	0.0418
$SBS(0.5)$	0.962	0.950	0.943	0.927	0.920	0.910
	0.446	0.318	0.258	0.138	0.0958	0.0418
95% intervals						
$MPEXACT$	0.968	0.961	0.958	0.953	0.952	0.950
	0.461	0.335	0.276	0.153	0.109	0.0486
$MPCXACT$	0.972	0.963	0.960	0.953	0.952	0.950
	0.470	0.338	0.277	0.153	0.109	0.0486
$WL(-0.13)$	0.972	0.965	0.961	0.955	0.953	0.951
	0.477	0.344	0.282	0.154	0.109	0.0487
$BSEXACT$	0.972	0.967	0.962			
	0.474	0.345	0.283			
$WL(-0.50)$	0.977	0.968	0.964	0.955	0.953	0.951
	0.492	0.349	0.284	0.154	0.109	0.0487
$WL(-0.62)$	0.979	0.969	0.964	0.955	0.953	0.951
	0.501	0.353	0.286	0.154	0.109	0.0487
$SBS(0.24)$	0.971	0.966	0.964	0.958	0.956	0.953
	0.470	0.346	0.285	0.157	0.111	0.0491
$SBS(0.41)$	0.979	0.973	0.970	0.963	0.959	0.955
	0.493	0.359	0.295	0.160	0.112	0.0494
$CPEXACT$	0.984	0.977	0.973	0.965	0.961	0.955
	0.508	0.366	0.299	0.161	0.113	0.0496
$SBS(0.5)$	0.982	0.976	0.973	0.965	0.961	0.955
	0.504	0.366	0.300	0.162	0.113	0.0496

* See footnote of Table 3. The 95% WL results are in reverse order: $EV(WL(-0.62),n) < EV(WL(-0.5),n) < EV(WL(-0.13),n)$ for $n > 199$.

4.4. Comparison of evaluation tools

For illustration, in Figure 1 and Tables 1 to 3 are results for 99% binomial intervals ($\alpha = 0.01$). In Table 1 and Figure 1 are tools conventionally used for comparison of intervals. Note that Table 1 is computed only for $n = 100$ and most of Figure 1 is also for $n = 100$. In Tables 2 and 3 is the proposed evaluation. Table 2 deals with all n up to 10,000. Table 3 is based on 9 values of n from 10 to 1,000.

A good use of coverages graphs is illustrated in the comparisons among $WL(c_c)$ of Figure 1. As c_c varies, clearly the variation in coverage is for small p in a region that gets smaller for larger n. The value $c_c = -0.57$ is the point of compromise at $n = 100$ where the minimum for very small p equals the minimum at $p = 0.05$ and at $p = 0.3$, so that the bottom level of coverage is levelled across the entire range of p. For $SBS(c_c)$, clearly the bottom level of coverage moves up for most of the range of p as c_c increases. This is likely true for all n since, as Vollset (1993) demonstrates, rough coverage patterns hold as n varies; but leveling coverage at $n = 100$ certainly does not lead to level coverage for $WL(-0.57)$ at $n = 10$.

Although coverage graphs can be compelling, it is not convenient to show these for a variety of n. The shortcomings of the Neyman Shortness table are worse, since it can only include a few values of θ and θ', as in Table 1. A table of volume (or width) is also awkward for evaluation. One way to summarize volume calculations is to sum over all volumes, but this disregards unequal probabilities of outcomes. The expected volume, (1), weights the sum by probability, but then a value of θ must be selected. The EV summarizes expected volume by averaging over all θ.

From a Bayesian point of view, $EV(C(\mathrm{X}))$ is the expected volume assuming θ has distribution H. From a non-Bayesian viewpoint, with H uniform, it is a weighted average of all possible expected volumes, where weight is uniformly applied over the entire range of Θ. Given no prior idea of the value of θ, this is a fair summary measure. A possible competitor which does not require H is the MS (Maximum Selectivity criterion):

$$\mathrm{MS}(C(\mathrm{X})) = \max_{\theta} \int_{\Theta} P_{\theta}(\theta' \epsilon C(\mathrm{X}))d\theta' \quad .$$

This may seem like a fair criterion at first glance. When applied to the binomial parameter, however, it is quite useless. This is because most binomial confidence intervals (cf. Figure 1) are highly erratic in their coverage behaviour for extreme values of $\theta = p$. The maximum (and the minimum) is usually found at p near 0 or 1, so that $\mathrm{MS}(C(\mathrm{X}))$ gives little weight to all moderate values of p.

An important point is illustrated in the comparison of $WL(-0.95)$ to $SBS(0.86)$ in Table 3. Usually, as $EV = EV(C(\mathrm{X}), n)$ increases or decreases then $AC = AC(C(\mathrm{X}), n)$ increases or decreases. Consequently, minimizing AC is a potential replacement for minimizing EV in the evaluation proposal. However, in Table 3 is shown $AC(WL(-0.95)) < AC(SBS(0.86))$ but $EV(WL(-0.95)) > EV(SBS(0.86))$ for most values of n, illustrating that AC is no substitute for EV. Furthermore, given the choice between a smaller AC and a smaller EV, the latter choice is clearly preferable, given $AC \geq 1 - \alpha$.

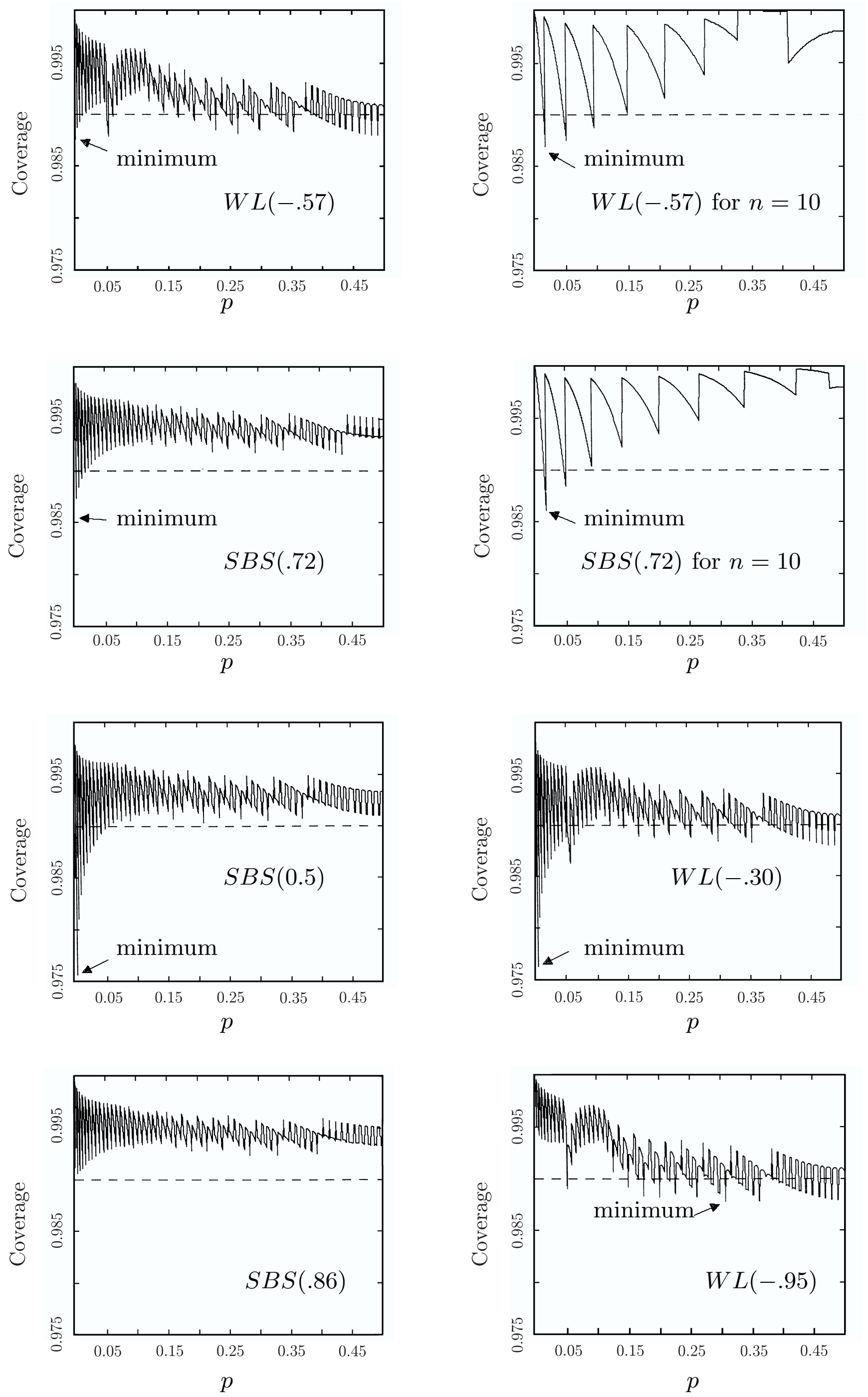

Figure 1. Coverage probability $P_p(p\epsilon C(\mathrm{X}))$ for 99% binomial confidence intervals with $n = 100$, except as indicated.

4.5. Results

Tables 2 and 3 are results for 99% binomial intervals ($\alpha = 0.01$). Tables 4 and 5 are results for 90% and 95% intervals. Given the proposal, it is quite possible that an exact interval based on a discrete F can be less accurate than an approximate interval.

Looking at the exact intervals first, *BSEXACT* is clearly superior when available. For $0 \leq k < 0.58$, *BSEXACT* is the most accurate interval for $n < 31$ at the 95% and 99% levels. Tables 3 and 5 show *MPEXACT* and *MPCXACT* to be often slightly shorter than *BSEXACT*, but Tables 2 and 4 show that neither interval satisfies the $k = 1/2$ minimum coverage criterion, though they both satisfy the liberal ($k = 1$) criterion. Both *BSEXACT* and *CPEXACT* satisfy the very conservative $k = 0$ criterion, but the EV of *BSEXACT* is smaller, and seems likely to be closer to that of *MPCXACT* than that of *CPEXACT* for $n > 30$, and even at $n = 1000$, the average length of *CPEXACT* is $(.0418 - .0408)/.0408 = 2.5\%$ wider than that of *MPCXACT* at $\alpha = 0.1$. Blyth and Still's recommendation of $SBS(0.5)$ for $n > 30$ as a replacement for *BSEXACT* is questionable, now that we can average expected width. A closer replacement appears to be $WL(-0.5)$.

For the approximate intervals, continuity correction values c_c chosen for logit and score intervals in Tables 3 and 5 include those, to the closest hundreth, that minimize EV at $n = 10,\ 100$ or $1,000$ while $P_p(p\epsilon C(\mathrm{X})) \geq 1-\alpha-\alpha/2$ holds (i.e. $k = 1/2$) for all n. For 99% WL (logit) intervals, the latter property is satisfied for $-1.0 < c_c \leq -0.52$. For 99% SBS (score) intervals, it is satisfied for $0.72 \leq c_c$. Table 2 shows that $WL(-0.52)$ and $SBS(0.72)$ satisfy the $k = 1/2$ minimum coverage criterion and that $WL(-0.57)$ satisfies a more conservative $k = 2/5$ criterion, but $SBS(0.5)$ does not even satisfy the liberal $k = 1$ criterion. Only $SBS(0.5)$ satisfies the $k = 0$ criterion at the 90% and 95% levels, but $c_c > 0.86$ is required at the 99% level. These $k = 0$ score intervals are clearly wider than the competition, however. Roughly speaking, when $c_c = c_1,\ c_2$ are chosen so that $\min_n MC(n)$ is about the same for $SBS(c_1)$ and $WL(c_2)$, $WL(c_2)$ is shorter for $n > 17$ and $SBS(c_1)$ is shorter for $n < 11$. (Two alternative versions of SBS are S and SC in Edwardes (1994). Vollset favoured $S(0.5)$. By the evaluation proposal, these are inferior to SBS, except at the 90% level.)

For the logit intervals, AC and EV converge quickly as n increases, so that, for example, $WL(0.11)$ is the most accurate 90% interval overall, but $WL(-0.5)$ becomes as accurate quickly. If one simple value need be chosen for all three levels, $WL(-0.5)$ will do. It satisfies the $k = 1/2$ minimum coverage criterion for all n at the 95% level, for $n > 18$ at the 90% level and $n > 60$ at the 99% level. Table 2 shows that it satisfies $k = 0.56$ for all n at the 99% level, and $WL(-0.5)$ is very close to $WL(-0.52)$, the most accurate 99% logit interval.

5. Conclusions and Discussion

Tables such as Tables 2 and 3, with EV in particular, are useful for evaluating confidence intervals or sets. These summarize succinctly the information conventionally obtained from coverage graphs and Neyman shortness tables and they can consider the entire range of levels and sample sizes. They are therefore less subject to omission bias due to editing. For F continuous, calculating minimum coverage is not required and estimates of AC and EV from simulations suffice, which is what is usually done.

When a confidence set is based on a discrete distribution, choosing the very conservative point of view (from the standard definition, in fact, of a confidence set) that the minimum confidence coefficient should attain the desired coverage level should be tempered by the knowledge that coverage could be excessive for most of the range of θ. Indeed, the minimum could occur at such a sharp spike in the coverage that it is virtually a point, and coverage can be much higher for all other values of θ, as suggested by Figure 1. Does a confidence set which attains coverage for all but a few points of Θ merit no consideration even when $AC(C(\mathrm{X})) \geq 1-\alpha$ and its expected volume is smaller than its competitors? The view of several authors, such as Vollset (1993), is that such a confidence set can be recommended, which contrasts sharply from the traditional view of Blyth and Still (1983), who enforce the standard requirement. After all, $P_\theta(\theta \epsilon C(\mathrm{X})) = 1-\alpha$ is the ideal requirement, usually attainable when F is continuous. No direction is implicit in the ideal. The direction of inequality in $P_\theta(\theta \epsilon C(\mathrm{X})) \geq 1-\alpha$ may be acceptable to those who assume that one should be conservative to be safe when uncertain. The classical (Clopper-Pearson 1934) exact binomial and Poisson confidence intervals are larger on average than approximate intervals, because they are required to attain coverage everywhere but attaining coverage is never enforced for approximate intervals. Is a larger confidence set necessarily the safe solution for all applications? Surely, a too-large interval may result in decision inaction in applications. Is inaction correlated with safety? No!

Table 3 shows the values of AC to be high for small n. As long as $AC \geq 1-\alpha$, lowering AC has smaller priority than lowering EV. Within the framework of our evaluation proposal, however, there seems to be room for a better exact interval. One way is to ignore minimum coverage altogether, in which case the arbitrary choice of k would not be required, thus replacing consideration of minimum coverage with $AC(C(\mathrm{X})) \geq 1-\alpha$. This may be too liberal. What if $P_{p_1}(p_1 \epsilon C(\mathrm{X})) < 0.5$ with $\alpha < 0.2$ for one or more binomial values p_1? This can easily happen when $AC \geq 1-\alpha$. Some statisticians would not be perturbed by this, but many would. Since $P_p(p \epsilon C(\mathrm{X})) = 1-\alpha$ is intended, departure from that ideal should be minimized. The proposal $k = 1/2$ of §3 is a compromise.

Two ideas that require further development are: (1) The choice of a non-uniform H when Θ is not bounded may be avoided by having H uniform for progressively receding imposed bounds. (2) In the presence of nuisance parameters, one has the choice of computing AC and EV for various nuisance values or of computing new coefficients by averaging AC and EV over the entire nuisance range.

If one has an idea in advance that some values of θ are more likely than others, a Bayesian approach (Bernardo and Smith (1994)) should be considered for an interval based on a posterior distribution, as done by Rubin and Schenker (1987). Of course, AC and EV can accomodate any prior distribution $H(\theta)$, such as when greater weight should be given to, say, smaller values of $\theta = p$, in the evaluation of a $C(\mathrm{X})$.

For a Binomial confidence interval, giving equal weight to all p, the most accurate $(0 \leq k < 0.58)$ exact interval for $n < 31$ is Blyth and Still's. Regrettably, it is presently unavailable for $n > 30$ or at the 90% level. Vollset (1994) stated that the problem is the lack of an unambiguous algorithm for its computation. Casella (1986) showed an explicit algorithm which yields a family of exact intervals, including *BSEXACT*, which Casella recommends as the solution that best corresponds to no a priori information. If one is prepared to accept $k = 1$, then *MPEXACT* is the most accurate exact interval (except that *BSEXACT* is still shorter at $n < 5$ for $\alpha = 0.05$ and $n < 11$ for $\alpha = 0.01$). The easily computed approximate interval $WL(-0.5)$ is a good compromise choice at all levels for $n > 18$, and compares favourably to score intervals such as $SBS(0.0)$ or $SBS(0.5)$ which previous authors have chosen as a best closed form interval.

Computer programs written in FORTRAN for computing coefficients are available from the author, for a nominal fee. The coverage graphs were done with the GAUSS computer package.

Acknowledgement

The author is supported by an Équipe grant from Fonds de la recherche en santé du Québec (FRSQ).

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Wiley, New York.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *J. Amer. Statist. Assoc.* **78**, 108-116.

Brown, L. D., Casella, G. and Hwang, J. T. G. (1995). Optimal confidence sets, bioequivalence and the limaçon of Pascal. *J. Amer. Statist. Assoc.* **90**, 880-889.

Casella, G. (1986). Refining binomial confidence intervals. *Canad. J. Statist.* **14**, 113-129.

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.

Cohen, A. and Strawderman, W. (1973). Admissibility implications for different criteria in confidence estimation. *Ann. Statist.* **1**, 363-366.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration,* 2^{nd} *ed.* Academic Press, Orlando, Florida.

Edwardes, M. D. (1994). Letter to the editor on Vollset's Confidence intervals for a binomial proportion. *Statist. Med.* **13**, 1693-1698.

Fisher, R. A. and Yates, F. (1948). *Statistical Tables for Biological, Agricultural and Medical Research,* 3^{d} *ed.* Oliver and Boyd, London.

Ghosh, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *J. Amer. Statist. Assoc.* **74**, 894-900.

Ghosh, J. K. (1961). On the relation among shortest confidence intervals of different types. *Calcutta Statist. Assoc. Bull.* **10**, 147-152.

Lehmann, E. L. and Loh, W.-Y. (1990). Pointwise versus uniform robustness of some large-sample tests and confidence intervals. *Scand. J. Statist.* **17**, 177-187.

Lentner, C. (ed.) (1982). *Geigy Scientific Tables.* Ciba-Geigy, Basel.

Ling, R. F. (1992). Just say no to Binomial (and other discrete distributions) tables. *Amer. Statistician* **46**, 53-54.

Miettinen, O. S. (ed.) (1985). *Theoretical Epidemiology.* Wiley, New York.

Pratt, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56**, 549-567.

Rubin, D. B. and Schenker, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. In *Sociological Methodology* **17** (Edited by Clogg, C. C.), 131-144. American Sociological Association, Washington, D. C.

Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York.

Stone, M. (1969). The role of significance testing. Some data with a message. *Biometrika* **56**, 485-493.

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statist. Med.* **12**, 809-824.

Vollset, S. E. (1994). Author's reply to Edwardes (1994). *Statist. Med.* **13**, 1698.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* **22**, 209-212.

Division of Clinical Epidemiology, Royal Victoria Hospital, 687 Pine W., Ross 4.06, 687 Pine Ave. West, Montréal, Québec, H3A 1A1 Canada.

E-mail: medward@rvhmed.lan.mcgill.ca

(Received March 1996; accepted April 1997)