Santner:

A Note on Teaching Binomial Confidence Intervals

Sonderforschungsbereich 386, Paper 87 (1997)

Online unter: http://epub.ub.uni-muenchen.de/

A Note on Teaching Binomial Confidence Intervals

Thomas J. Santner[1]
Department of Statistics
Ohio State University
Columbus, OH, 43210

[1]Written while visiting Sonderforschungsbereich 386, Ludwig-Maximilians-Universität München, 80539 München

1 Introduction

In a recent article, Simon (1996) advocates teaching one of two methods for constructing confidence intervals for a binomial proportion p. His recommendation is based on the comparison of the closeness of the achieved coverage of these intervals to their nominal value. Letting $S.E.(\hat{p}) = \sqrt{\hat{p} \times (1-\hat{p})/n}$ denote the standard error of $\hat{p}$, Simon compares the 95% z-interval, $\hat{p} \pm 1.96 \times S.E(\hat{p})$, with a t-based interval, $\hat{p} \pm t \times S.E.(\hat{p})$, where t is the two-sided upper-0.05 critical point of the t-distribution with $n-1$ degrees of freedom, and with the continuity-corrected normal interval

$$\hat{p} \pm 1.96 \times \{S.E.(\hat{p}) + 1/2n\}\,. \tag{1.1}$$

Based on a set of comparisons of achieved coverage for sample sizes 5 to 40, Simon concludes "the t-based interval achieves better coverage than the z-based interval," and furthermore continuity-corrected intervals are an attractive alternative to t-intervals.

This paper addresses this same important question of which binomial confidence interval method should be the standard method taught in elementary courses. We argue that a third, easily motivated, variant of the z-interval should be the standard asymptotic method presented in elementary books. We also recommend that an alternative method be simultaneously presented for use in small sample applications; this method produces intervals that *achieve at least the nominal coverage no matter what the sample size and true p.*

2 The Methods

The t and continuity-corrected intervals will be referred to as t-intervals and c-intervals, respectively, throughout this article. They are calculated as described in Section 1. We shall

show that both of these methods are inferior to an easily explained, asymptotic method and a second small-sample method that attains at least its nominal coverage for all sample sizes and all true p.

Since students know from their study of the Central Limit Theorem that

$$\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}$$

is approximately normal distributed,

$$P_p\left\{n(\hat{p}-p)^2 \leq p(1-p)z^2\right\} \approx 1-\alpha \tag{2.2}$$

where z is the two-sided upper-α critical point of the standard normal distribution. The n for which (2.2) is accurate depends on p. The zeroes of the (concave) quadratic equation in p defined by the event in (2.2) are the endpoints of the interval. We denote the system of intervals implicitly defined by (2.2) as q-intervals ("quadratic").

Note that the only difference between q-intervals and z-intervals is that the population standard error of $\hat{p}$ is used to define the former while this standard error is *estimated* in the latter. It is exactly the "lumpiness" of the estimated standard error that causes the z-intervals to have inferior coverage relative to those defined by (2.2). Ghosh (1979) presented a detailed study comparing the small sample properties of z- and q-intervals (see Santner and Duffy, Section 2.1.B).

A set of (small-sample) intervals that attain at least their nominal confidence level can be motivated for elementary students by considering the family of point null hypotheses

$$H_0\colon\ p = p_0 \text{ versus } H_A\colon\ p \neq p_0$$

corresponding to each p_0 with $0 < p_0 < 1$. Equipped with a family of size α rejection regions, the confidence interval corresponding to data $Y = j$ is constructed, in principle, as follows. All those p_0 for which the null hypotheses $H_0\colon\ p = p_0$ is "accepted" (not rejected) for $Y = j$ are placed in the confidence set. If the family of rejection regions has certain monotonicity properties in p_0, the confidence set will be a confidence interval.

Blyth and Still (1983) showed how to use an idea of Sterne (1954) to find a set of short intervals that achieve at least the nominal level for all sample sizes and all p by constructing acceptance regions to contain *as few points as possible* and have the required monotonicity properties relative to one another. Intuitively, the resulting intervals should be short. These requirements, by themselves, do not necessarily lead to intervals when the acceptance regions are inverted. By adding additional invariance and monotonicity requirements for the intervals, they were able, using a combination of computer work and manual intervention, to produced a table of intervals for sample sizes from 1 to 30 that satisifies their desired properties.

For arbitrary families of multi-stage hypothesis tests, Duffy and Santner (1987) give an algorithm for computing binomial intervals that have most of the properties that Blyth and Still describe. A FORTRAN program is available that implements their method. Applied to single-stage data, the program produces intervals that attain at least their nominal confidence level for the binomial problem. Throughout we will use the notation D/S-intervals to denote this system of intervals.

Table 1: Number of times each system of intervals is the unique method with coverage closest to 0.95 nominal level and number of times it is tied with one or more other methods for being closest to nominal level

Method	Number of Times Unique Winner	Number of Times Shared Winner
t-interval	0	557
c-interval	0	652
q-interval	0	1182
D/S-interval	0	874

3 Comparisons

Initially we consider the same criteria of closeness of achieved confidence level to nominal confidence level that was used by Simon (1996). We construct the same set of comparisons of achieved coverage as he does for the nominal 95% t-, c-, q-, and D/S-intervals. In these comparisons the achieved coverage,

$$P_p\left\{\hat{p}_{lower} < p < \hat{p}_{upper}\right\} = \sum_{i=0}^{n} \binom{n}{i} p^s (1-p)^{n-i} \times I\left[\hat{p}(i)_{lower} < p < \hat{p}(i)_{upper}\right],$$

was computed for each method and for all $n = 5(1)40$ and all $p = 0.05(.01)0.50$. Here $I[\cdot]$ is the indicator function which takes values 1 or 0 as the event in the brackets is true or not. The 36 different n and the 46 different p values yield 1,656 (n, p) combinations for which the achieved coverage is calculated for each method.

Tables 1 and 2 are calculated from the set of absolute values of the difference between the nominal level and the achieved level corresponding to each of the 1,656 (n, p) combinations. Table 1 summarizes the simultaneous comparisons among the four systems of intervals. Because of the highly competitive nature of these systems of intervals, in no (n, p) case did any one system have strictly lower absolute difference than all the other three systems; there were many ties.

In paired comparisons of the absolute difference of achieved and nominal values, the q- and D/S-intervals are strictly closer to the target far more often than the t or c intervals. Compared with each other, q-intervals have stricly smaller absolute difference from the 0.95 target level than do D/S-intervals for 517 (n, p) combinations; the reverse is true 183 times, and they are tied for 956 $(= 1656 - 517 - 183)$ (n, p) combinations.

The comparisons in Tables 1 and 2 ignore the actual amounts by which the achieved values are above or below the target value–if the closest achieved level among the four methods is only 0.5, for example, then the performance may still be unacceptable. Because the effect of sample size is important in determining this effect, we investigate this issue separately for three different groups of sample sizes. Figure 1 displays, using common y-axis scaling, boxplots of the set of achieved coverages for the four methods grouped by sample size: (A) $n = 5$ to 15, (B) $n = 16$ to 30, and (C) $n = 31$ to 40. In the discussion below, we refer to these as "small," "medium," and "moderate" sample size groups, respectively.

Table 2: Number of times that each pair of methods is strictly closer to nominal level

	Loser			
	t-interval	c-interval	q-interval	D/S-interval
Winner t-interval	—	106	302	471
Winner c-interval	477	—	394	490
Winner q-interval	1035	938	—	517
Winner D/S-interval	964	750	183	—

[Figure 1 about here]

For all sample sizes t-intervals can have great deficiency in the achieved level. The c-intervals can also have substantial deficiency in the achieved level for small or medium sample sizes but tend to be median unbiased for the nominal level when the sample size is medium or moderate. The asymptotic q-intervals never have achieved probability below .90 for sample sizes greater that 15 and are nearly median unbiased for the nominal level for sample sizes greater than 30. As advertised, the Duffy/Santner intervals meet or exceed the nominal confidence level for all sample sizes and probability levels (due to rounding, 6 of the cases had achieved level of 0.949) and the median of the achieved levels for the three sample size groups is 0.972 for $n = 5$ to 15, 0.963 for $n = 16$ to 30, and 0.962 for $n = 31$ to 40. The unbiased nature of the distribution of achieved levels of the q-intervals is the primary reason that the absolute value of the difference between the achieved and nominal coverages favors the q-intervals.

4 Conclusion

The conclusions are these.

- The q-interval is simple to justify, easy to program and has superior coverage to both the t-interval and the continuity corrected interval. When $n \geq 15$, the distribution of achieved confidence level of the q-interval for the (n, p) cases studied here is approximately median unbiased for the nominal level.
- In cases where the sample size is small or one must be certain that the nominal coverage is attained, use D/S intervals.

References

Blyth, Colin, and Still, Harold, (1983). Binomial Confidence Intervals. *Jour. Amer. Stat. Assoc.* **78**, 108-116.

Duffy, Diane E., and Santner, Thomas J., (1987). Confidence Intervals for a Binomial Parameter Based on Multistage Tests, *Biometrics* **43**, 81-93.

Ghosh, B.K., (1979). A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter. *Jour. Amer. Statist. Assoc.* **74**, 894-900.

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

Simon, Gary, (1996). z or t for Binomial Confidence Intervals, *Teaching Statistics* **18**(3), 90-91.

Sterne, T., (1954). Some Remarks on Confidence or Fiducial Limits. *Biometrika* **41**, 275–278.

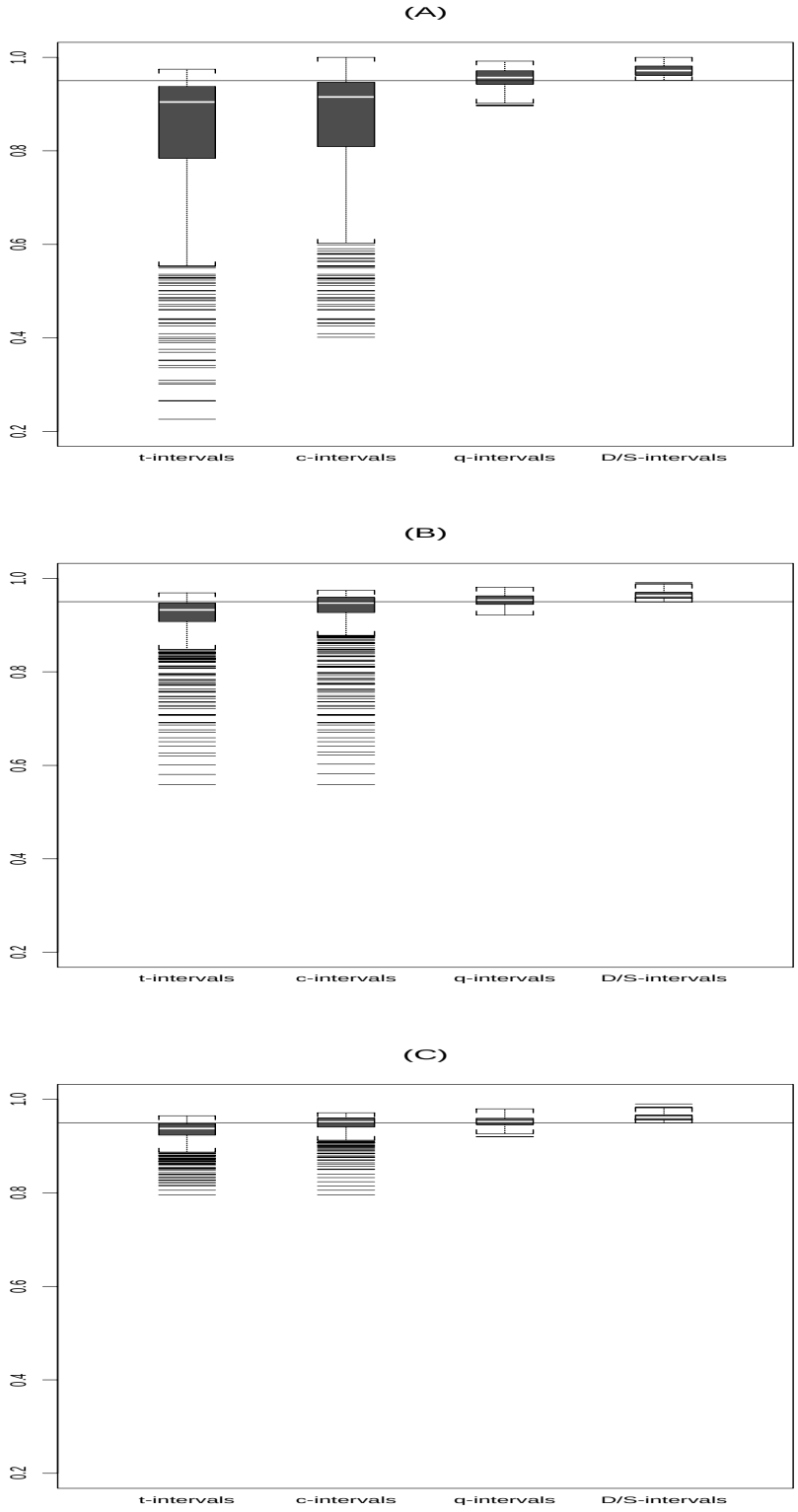

Figure 1: Distributions of achieved coverages for 95% nominal t, c, q, and D/S binomial limits that are grouped into small, medium, and moderate sample size classes. Panel (A) is for $n \in \{5, \ldots, 15\}$ (506 points in each boxplot), Panel (B) $n \in \{16, \ldots, 30\}$ (690 points in each boxplot), and Panel (C) $n \in \{31, \ldots, 40\}$ (460 points in each boxplot). The horizontal line is at the 0.95 nominal level.