On Sample Size Guidelines for Teaching Inference about the Binomial Parameter in Introductory Statistics

Alan Agresti and Yongyi Min

Department of Statistics

University of Florida

Gainesville, Florida 32611-8545

e-mail AA@STAT.UFL.EDU

August 15, 2002

On Sample Size Guidelines for Teaching Inference about the Binomial Parameter in Introductory Statistics

ABSTRACT

Textbooks for introductory statistics courses use a sample size of 30 as a lower bound for large-sample inference about the mean of a quantitative variable. For binary data, there is no consensus bound, and rarely is the student told how to handle small samples. We suggest a guideline for interval estimation that ties in with the rule of 30 for quantitative variables and also gives direction for smaller samples: Form the usual confidence interval if at least 15 outcomes of each type occur, and otherwise use that interval after adding two successes and two failures.

KEY WORDS: Binomial distribution; Confidence interval for proportion; Score test; Small sample; Wald test.

1 INTRODUCTION

Textbooks for introductory statistics courses commonly use a sample size of 30 as a lower bound for large-sample inference about the mean of a quantitative variable. Motivation for this is the similarity when $n \geq 30$ between the standard normal distribution and the t distribution, with a normal population. Obviously, such a simple guideline cannot apply well to all cases and may fail for highly nonnormal populations. The case usually treated differently is a binary population, for which there is no consensus guideline.

For instance, for interval estimation of a binomial parameter, nearly all such texts present the confidence interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \tag{1}$$

where $\hat{p} = x/n$ denotes the sample proportion based on a binomial outcome of x in n independent Bernoulli trials. The most common guideline for using (1) is the occurrence of at least a certain number of observations of each type ("successes" and "failures"). Most common is the lower bound of 10 for each (e.g., Moore and McCabe 1998), although many texts use 5 instead (e.g., Triola 2000). When such guidelines are not satisfied, the student is usually told that more complex methods are needed that are beyond the scope of the text. We propose an alternative guideline that relates to the $n \geq 30$ bound for quantitative variables and tells the student how to proceed if the bound is not satisfied.

2 A NEW SAMPLE SIZE GUIDELINE

Considerable research has shown that interval (1) performs poorly when p is near 0 or 1, even for large n (e.g., Brown et al. 2001). Even when p is near .5, coverage probabilities can be rather low for moderate n. Based on this evidence, we believe that the guideline of $\min(x, n-x) \geq 5$ or 10 to use (1) is too liberal. We suggest the alternative bound of $\min(x, n-x) \geq 15$. When this bound is not satisfied, we suggest applying formula (1) after adding two outcomes of each type (Agresti and Coull 1998).

Figure 1 illustrates the performance of this hybrid strategy of using the ordinary interval if $\min(x, n-x) \geq 15$ and otherwise using the adjusted "add 2 successes and 2 failures" interval. It shows the coverage probability for the ordinary interval (1) and for the hybrid strategy as a function of p, for n = 10, 20, 30, 40, 50, and 60. The hybrid interval can be quite conservative for p near 0 or 1, but this is preferable to the very low coverage probabilities the ordinary interval can give in those regions. With the hybrid strategy, as n increases the region gradually expands around $p = .5$ in which the ordinary interval (1) is used with probability close to 1.

When the hybrid strategy instead uses a smaller lower bound such as $\min(x, n-x) \geq$ 5 or 10, the coverage curve drops unacceptably with p near .5 and smaller n to that for the ordinary interval, and the coverage curves actually can show increasing conservativeness elsewhere. Figure 2 illustrates, showing the performance when $n = 50$, using 5, 10, and 15 in each category as the lower bound.

The interval (1) inverts the Wald test. It consists of the set of p_0 values for which $|\hat{p} - p_0|/\sqrt{\hat{p}(1-\hat{p})/n} < z_{\alpha/2}$. The confidence interval that inverts the score test, giving the set of p_0 values for which $|\hat{p} - p_0|/\sqrt{p_0(1-p_0)/n} < z_{\alpha/2}$, is well known to perform much better. The score interval behaves well even for small n except when p is very near 0 or 1, where it is too liberal. Its computations are too complex for most elementary courses, however. An adjusted interval that adds $z^2_{\alpha/2}/2$ successes and failures before computing the Wald interval (1) has the same midpoint as the score interval but is wider, being conservative for p near 0 and near 1 (Agresti and Coull 1998). With 95% confidence, this is what motivates the adjusted "add two successes and two failures" interval (i.e. $z^2_{.025}/2 \approx 2$). In the elementary course, it is simplest to use this adjustment regardless of α. The consequence is that for p near 0 or 1 this adjusted interval can be a bit conservative for 95% and 99% confidence and a bit liberal for 90% confidence, but it still performs much better than the ordinary interval (1) (Agresti and Caffo 2000).

For $n < 30$, the hybrid strategy requiring $\min(x, n-x) \geq 15$ always uses the adjusted interval. Figure 1 shows that it does surprisingly well for small n, being somewhat conservative for p near 0 or 1 where the Wald method fails. Because of this, Agresti and Caffo (2000) suggested that for teaching in elementary courses it is adequate to use this adjusted interval for all n. In fact, this is our preference, because the adjusted interval often gives coverage probabilities somewhat closer to the nominal level than the hybrid interval. Specifically, it has somewhat higher coverage probabilities for p near .5

when the hybrid strategy is likely to use the Wald interval and somewhat lower coverage probabilities for values of p for which the hybrid strategy might use either the Wald or the adjusted interval. Figure 2 illustrates for $n = 50$, showing the adjusted interval as the hybrid interval for a boundary of ∞.

One text (Samuels and Witmer 1999) does recommend using an adjusted interval for all n. However, for large n realistically most instructors would prefer to teach the well-known Wald interval (1), showing students the simple idea of point estimate plus and minus a multiple of the standard error, which they also see for quantitative data. The hybrid strategy then seems like a reasonable compromise: The guideline has a connection with the $n \geq 30$ guideline for quantitative variables (as 30 is the minimum sample size for which one would ever use the Wald interval), yet students receive guidance on what to do when the bound is not achieved. The instructor can remind students that the t distribution is inappropriate here, since a binary population is far from normal. In fact, t methods perform poorly, partly because for binary data the denominator of the t (or Wald) test statistic is a function of the numerator rather than independent of it as in the normal response case. The instructor might also demonstrate the severe skew of the binomial when p is close to 0 or 1, show the degenerate Wald interval that results when $\hat{p} = 0$ or 1, or use simulation software or an applet to show its poor performance when p is near 0 or 1 even when $n > 30$.

3 COMMENTS

Teaching the Wald confidence interval (1) has the disadvantage of the lack of duality between that interval and the usual large-sample test for a proportion taught in the elementary course, which uses the score statistic $z = (\hat{p} - p_0)/\sqrt{p_0(1 - p_0)/n}$. This test behaves much better than the corresponding Wald test. Figure 3 shows the actual size of a two-sided score test having nominal size of .05 when $n = 10$, 20, 30. In examples and exercises in introductory statistics courses, the test most commonly has $p_0 = .5$. Thus, it is simple and reasonable to tell students they can use this large-sample two-sided test for small n also.

To achieve duality with the interval estimation proposed above, one could adopt the hybrid strategy of using the Wald test when $\min(x, n-x) \geq 15$ and otherwise using that test after adding two successes and two failures. Figure 2 also shows the actual size of that hybrid test. Achieving duality using the score test and score confidence interval is more attractive, however, as a single standard approach then performs relatively well for all n. The score confidence interval is rarely taught in introductory courses, since at that level it is desireable to keep things as simple as possible. However, we recommend that instructors with more sophisticated students consider the score method for both inferences, explaining to students the principle of forming a confidence interval from the set of p_0 not rejected in the test. With appropriate software, an instructor might use the score inference duality even in the elementary course, if it is not considered necessary to provide a formula for the confidence interval. However, the score confidence interval is not available in standard software (e.g., Minitab, Excel) for that course.

Those who worry about poor performance of the score method when p is near 0 or 1 may prefer a more conservative duality using the binomial distribution for small n. Our evaluations showed good performance of a hybrid strategy using the score test when $\min[np_0, n(1-p_0)] \geq 15$ and a binomial test otherwise (inverting results to construct a confidence interval). The related small-sample binomial test and confidence interval are conservative, having actual size no greater than the nominal size; e.g., see Blaker (2000) for a sensible implementation that is less conservative than the Clopper-Pearson approach. However, discussion of such an approach, and the attendant issues such as discreteness, are beyond the scope of the elementary course.

REFERENCES

Agresti, A., and Caffo, B. (2000), "Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures," *The American Statistician*, 54, 280-288.

Agresti, A., and Coull, B. A. (1998), "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, 52, 119-126.

Blaker, H. (2000), "Confidence curves and improved exact confidence intervals for discrete distributions," *Canadian Journal of Statistics* **28**, 783-798.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval estimation for a binomial proportion." *Statistical Science*, 16, 101-133.

Moore, D., and McCabe, G. (1998), *Introduction to the Practice of Statistics*, 3rd ed., W. H. Freeman.

Samuels, M. L., and Witmer, J. W. (1999), *Statistics for the Life Sciences*, 2nd ed., Prentice Hall.

Triola, M. F. (2000), *Elementary Statistics*, 8th ed., Addison-Wesley.

Figure 1. Coverage probabilities for the binomial parameter p with the 95% Wald confidence interval and the hybrid interval that adjusts it when the number of outcomes of either type is less than 15

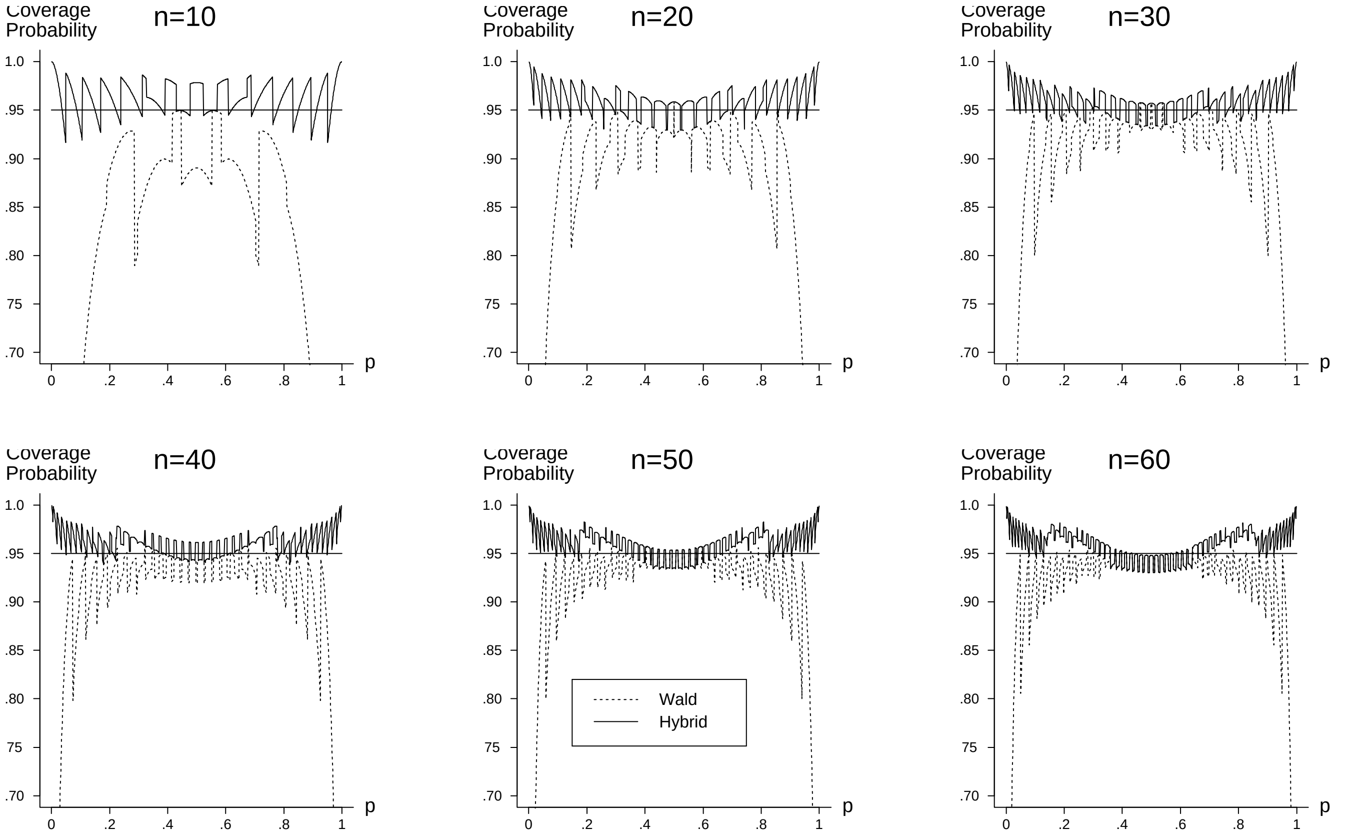


Figure 2. Coverage probabilities for the binomial parameter p with four versions of the lower bound for using the Wald interval with the hybrid method, when n = 50

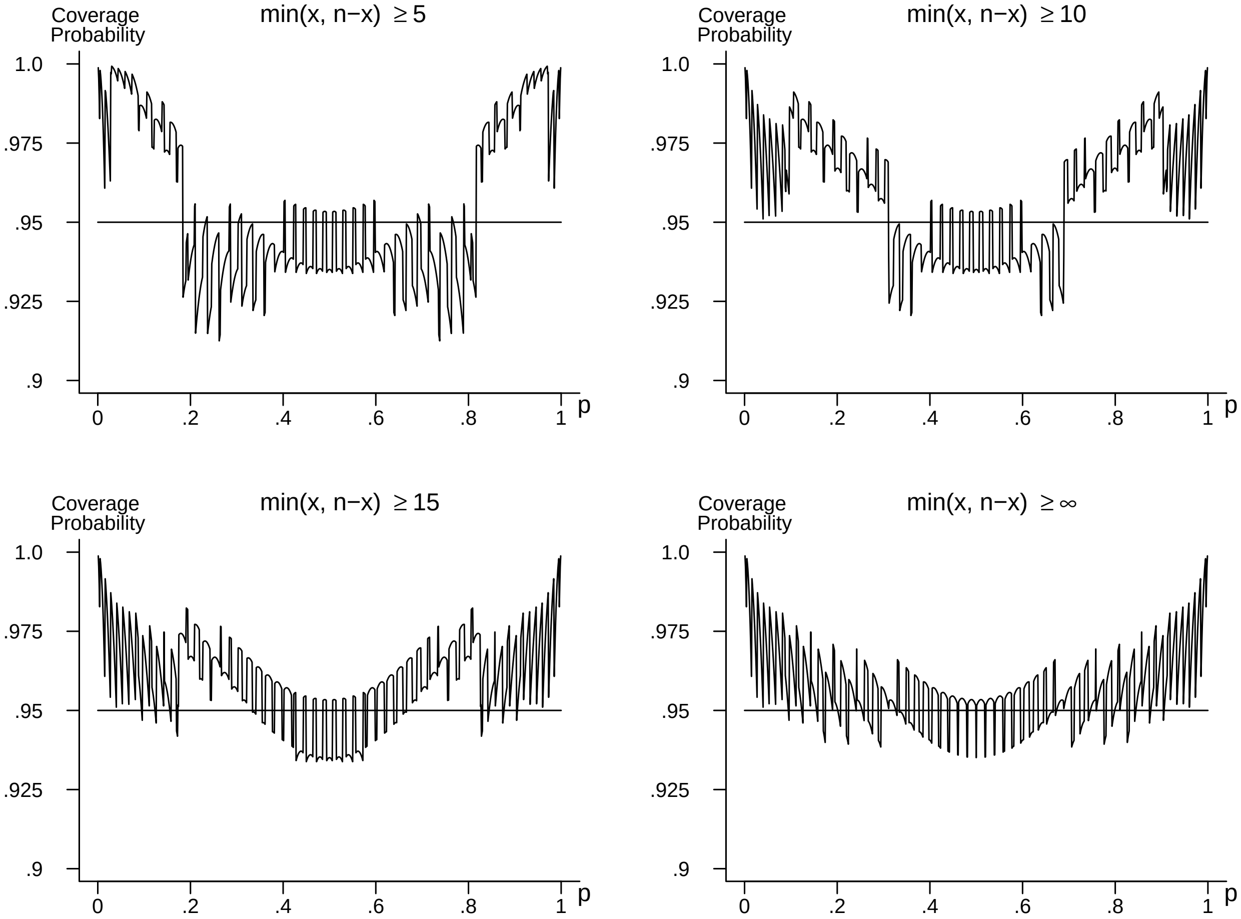
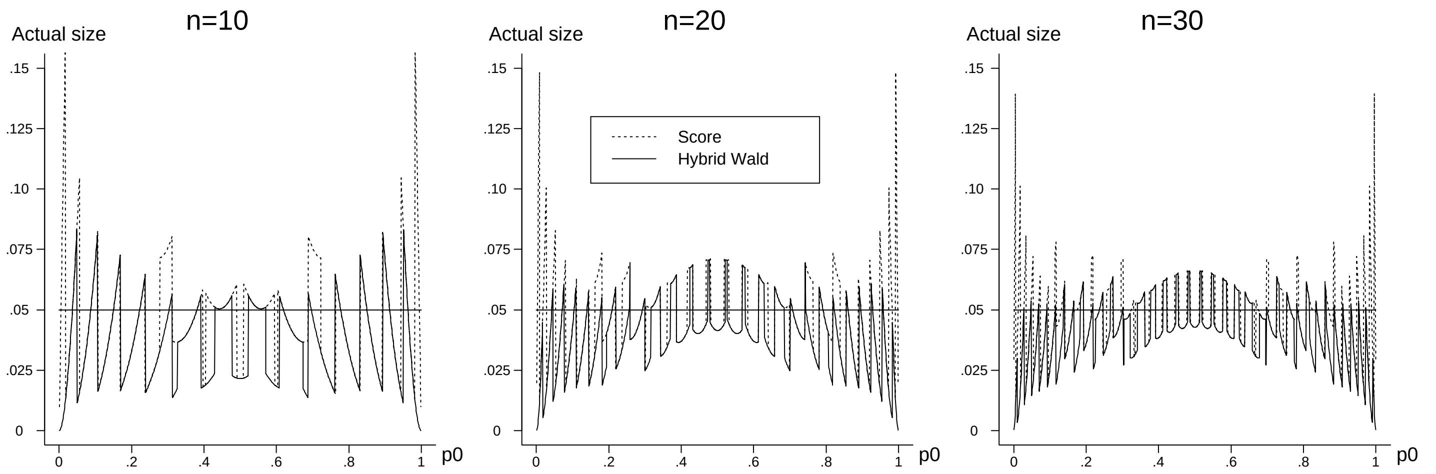


Figure 3. Actual size of score test and hybrid Wald test, for nominal size .05.


n=10
Actual size
p0
n=20
Actual size
Score
Hybrid Wald
n=30
Actual size