Comparison of Test Sizing Approaches for Initial and Follow-On Evaluation of Strategic Weapon Systems

Joseph D. Warfield, Sarah Elise Roberts
Applied Physics Laboratory, Force Projection Sector, Johns Hopkins University, Laurel, Maryland

ABSTRACT Testing programs exist to establish initial estimates of reliability for strategic weapon systems and to determine significant changes in reliability. The methods to determine the appropriate sample sizes are based on one- and two-sample estimates of proportions. There are guidelines that exist that provide specific methods to determine sample sizes for strategic weapons systems. The methods used are considered to be quite conservative, and this article investigates alternative approaches for both initial and follow-on test sizing for weapon system reliability estimates. The methods are compared through Monte Carlo simulation.

KEYWORDS binomial distribution, confidence intervals, coverage probability, proportion, sample size

INTRODUCTION

Testing programs exist to establish initial estimates of reliability for a device and to determine whether a significant decrease in reliability has occurred in relation to an initial value. The type of device being evaluated may be a complex system with many components, assemblies, and subsystems. At the Johns Hopkins University, Applied Physics Laboratory, missile systems are typically analyzed to establish an initial estimate for the reliability of a system and to detect a significant decrease in reliability over time. Though other performance metrics such as accuracy are also evaluated, the focus is on system reliability. Determining the number of tests needed for estimation with confidence and detecting changes in reliability is important when destructive testing of expensive items is involved. The necessary sample size determines the number of items to be tested and helps quantify the cost of the testing program(s).

There are some methodologies appropriate for establishing initial estimates of reliability for weapon systems, whereas other methodologies are more appropriate for detecting changes in reliability. There are various assumptions and approximations possible for both test sizing situations. Guidelines are often established to define the number of full system tests required. These guidelines are based on exact distributional assumptions that typically require more testing than necessary to establish estimates with a specified confidence. Therefore, this
article investigates several methods for initial and follow-on test sizing and provides a comparison of the results using simulations.

The remainder of the article is organized as follows. The first section explains the current initial assessment methodology based on specified guidelines and describes several methods in the literature used for confidence interval estimation of a single proportion. The second section provides details of the simulation study to compare the coverage probabilities and confidence interval widths of the considered methods. The third section outlines the current follow-on assessment methodology based on specified guidelines and describes several other methods in the literature used for confidence interval estimation of two proportions. The fourth section provides the simulation study comparing these methods. The final section provides concluding remarks based on the comparisons and a discussion of other alternatives.

Confidence Interval Methods for a Single Proportion

The standard approach for sample size determination is based on the formulation of a confidence interval for a single proportion. Let \( X \) denote a binomial variable for sample size \( n \) and proportion \( p \). To avoid approximation, the Clopper and Pearson (1934) "exact" confidence interval is typically used. The interval for \( p \) is based on inverting equal-tailed binomial tests of \( H_0 : p = p_0 \). It has endpoints that are the solutions in \( p_0 \) to the equations

\[
\sum_{k = x}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha/2
\]

and

\[
\sum_{k = 0}^{x} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha/2.
\]

In the case when \( x = 0 \) the lower bound is 0, and when \( x = n \) the upper bound is 1. This interval estimator is guaranteed to have coverage probability of at least \( 1 - \alpha \) for every possible value of \( p \). The Clopper-Pearson exact interval is typically treated as the "gold standard" because it is necessarily conservative due to the discreteness of the binomial distribution. For any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level unless \( n \) is quite large. Another issue is that this method only considers integer number of successes. With a highly complex system with many components, one could have partial successes of the system during an operational test.

Confidence intervals that use the asymptotic normality of a maximum likelihood estimator are called Wald confidence intervals (Mukhopadhyay 2000). Let \( \hat{p} = X/n \) denote the sample proportion. Then the \( 100(1 - \alpha)\% \) Wald confidence interval for \( p \) is

\[
\hat{p} \pm z_{1 - \alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}.
\]

The discreteness of the binomial distribution often makes the normal approximation work poorly, resulting in a confidence interval that often has a confidence level lower than stated. There are a considerable number of methods in the literature for forming confidence intervals for a single proportion that try to address the "conservativeness" of the exact confidence interval and the "liberalness" of the normal approximation.

Agresti and Coull (1998) adjusted the Wald interval by adding \( z_{1 - \alpha/2}^2 \) pseudo observations (half successes and half failures), resulting in

\[
\hat{p} = (X + z_{1 - \alpha/2}^2/2)/(n + z_{1 - \alpha/2}^2).
\]

One can think of \( \hat{p} \) as a Bayes estimator with a beta \((z_{1 - \alpha/2}^2/2, z_{1 - \alpha/2}^2)\) prior distribution (see Gelman et al. 2004). The adjusted Agresti-Coull confidence interval for \( p \) is

\[
\hat{p} \pm z_{1 - \alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}.
\]

This simple adjustment to the ordinary Wald interval changes it from a highly liberal to slightly conservative confidence interval. Other adjustments can be made using Jeffreys prior \([\beta(1/2, 1/2)]\) or the uniform prior \([\beta(1, 1)]\) distribution (Hamada et al. 2010). In the simulation study, the two intervals will be denoted as the adjusted Jeffreys and adjusted uniform confidence intervals, respectively.

Instead of using a Bayes estimator, one could calculate a \( 100(1 - \alpha)\% \) equal-tailed Bayesian interval. If \( p \) has a prior distribution \( \beta(a, b) \), then the posterior distribution for \( p \) given \( X \) and \( n \) is distributed \( \beta(X + a, n - X + b) \). The interval is given by the \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles of the beta distribution. Brown et al. (2001) recommended the equal-tailed Bayesian interval using Jeffreys prior (Bayesian Jeffreys) for smaller sample sizes.
Borkowf (2006) proposed a confidence interval to guarantee nominal coverage like the Clopper-Pearson method but is easier to construct and interpret like the Agresti-Coull method. This is done by centering proportions by adding one failure in the construction of the lower confidence bound and one success in the construction of the upper confidence bound. The lower and upper bounds of the Borkowf confidence interval are the following:

\[ LB = p_1 - z_{1-a/2} \sqrt{p_1(1-p_1)/n}; \]
\[ UB = p_2 + z_{1-a/2} \sqrt{p_2(1-p_2)/n} \]

where \( p_1 \) and \( p_2 \), the centering proportions, are defined as \( p_1 = \frac{X}{n+1} \) and \( p_2 = \frac{X+1}{n+1} \). The actual sample size \( n \) is used to construct the lower and upper bounds rather than the augmented sample size \( (n + 1) \).

Newcombe (2001) investigated the logit Wald interval and the Wilson score interval. The logit Wald interval’s major drawback is that it cannot be directly calculated in cases where the sample proportion is very close to 0 or 1, which makes it difficult to recommend for general use. The logit Wald confidence interval is

\[ \log \left( \frac{\hat{p}}{1-\hat{p}} \right) \pm z_{1-a/2} \sqrt{\frac{1}{np(1-p)}} \]

followed by a transformation back to the proportion scale. The Wilson score interval is the inversion of the normal approximation test statistic but uses the null standard error instead of the estimated standard error used in the standard Wald hypothesis test:

\[ \hat{p} \pm \frac{z_{1-a/2} \sqrt{n}}{n + z_{1-a/2}^2} \sqrt{\hat{p}(1-\hat{p}) + \frac{z_{1-a/2}^2}{4n}} \]

where \( \hat{p} \) and \( \hat{p} \) are the same as defined previously (Wilson 1927).

All of the methods considered have some advantages and disadvantages. For example, the Wald-based methods are easy to explain but could give subnominal coverage at particular proportion values. The exact methods guarantee at least nominal coverage but can be conservative under many circumstances, resulting in larger sample sizes (i.e., more cost). The Bayesian approaches provide a slightly different interpretation from the traditional confidence interval. The methods considered are not a complete summary of the literature but include methods that are commonly used.

**Simulation Comparisons of Single Proportions Methods**

This section provides an evaluation of the alternative intervals considered through Monte Carlo simulation. The goal of the study is to determine whether or not there is a confidence interval approach that is uniformly better than other methods. Special attention is given to comparisons against the Clopper-Pearson exact method since it is the gold standard. Simulation results are compared based on two criteria: (1) the coverage probability against a nominal confidence level (i.e., its conservativeness) and (2) the width of the confidence interval.

For a fixed value of a parameter, the actual coverage probability is the proportion of times the confidence interval contains or “covers” the true parameter. If confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter should approximately match the nominal confidence level. A confidence interval with a coverage probability higher than the nominal confidence level is “conservative.”

The simulation is set up as follows:

1. A random sample, \( X \), is drawn from a binomial(\( n, p \)) distribution, where \( n \) is the sample size and \( p \) is the true proportion.
2. All confidence intervals are constructed based on the formulations in the prior section.
3. A value of 1 is given to the interval if it contains the true proportion \( p \) or a value of 0 is given otherwise.
4. The confidence interval width (the upper bound minus the lower bound) is calculated.

The steps are completed 10,000 times for sample sizes ranging from five to seventy-five, proportions from 0.70 to 0.95, and nominal confidence levels from 0.80 to 0.95. The motivation for studying sample sizes up to seventy-five is due to the fact that one is dealing with expensive destructive testing and to consider anything higher would be unachievable. The motivation for the range of proportions considered is that missile systems are typically highly reliable.
Tables A1 through A5 in the Appendix (see the online supplementary information) show the average coverage probabilities and root mean square errors of the confidence interval methods for samples sizes of fifteen through seventy-five. For each sample size and confidence level, the methods are evaluated using a Freeman-Tukey test (Freeman and Tukey 1950). Methods with similar coverage probabilities are indicated by the same superscript. Tables A6 through A10 in the Appendix show the average confidence interval widths and associated standard deviations. Tukey’s studentized range test was used to group the methods based on confidence interval widths (Montgomery 2009). Figures 1 and 2 provide the average coverage probabilities and average confidence interval widths of each method considered for select sample sizes at 0.90 and 0.95 confidence levels.

When the true proportion is between 0.70 and 0.80, most interval procedures provide at least the nominal confidence level. The average coverage probability of the adjusted uniform interval is typically the closest to the desired confidence level but becomes near 1 for very high proportions. Generally, these confidence

![Figure 1](image)

**FIGURE 1.** Average coverage probabilities for true proportion ($p$), confidence level ($c = 0.90, 0.95$), and select samples sizes ($N = 15, 30, 45, 60, 75$).
intervals provide at least the nominal confidence, except in a few cases for smaller sample sizes. In terms of average interval width, the adjusted uniform method runs in the middle of the pack. Therefore, for proportions less than 0.80, it is recommended to use the adjusted uniform method.

For proportions greater than 0.80, the Bayesian Jeffreys and Wilson score methods provide intervals with...
the smallest widths regardless of sample size. Whereas most confidence intervals have coverage probabilities approximately equal to 1 as the true proportion increases, both methods are not as highly conservative as others. For small sample sizes, the coverage probabilities are always at least the nominal confidence level. Whereas the mean coverage probability of the adjusted uniform interval is more conservative for proportions greater than 0.80, the coverage probability of the Bayesian Jeffreys method remains relatively stable. The Bayesian Jeffreys method performs relatively well even for higher true proportions, but it does not provide the same coverage probability as the adjusted uniform. For proportions less than 0.80 both methods provide intervals of the approximately the same width, but for proportions greater than 0.80 the difference in interval widths is statistically significant, in favor of Bayesian Jeffreys. Therefore, the Bayesian Jeffreys confidence interval is recommended for proportions greater than 0.80.

The other confidence interval methods considered become highly conservative with the exception of the Wald and logit intervals. The coverage probabilities of the Wald and logit intervals improve as the sample size increases but fall well below the nominal confidence level for higher proportions. The Clopper-Pearson, Borkowf, and adjusted Agresti-Coull procedures are consistently the most conservative and largest confidence intervals across all sample sizes. When the true proportion is above 0.95, their coverage probabilities are approximately 1. The adjusted Jeffreys method produces intervals of relatively small widths, yet the coverage probability is relatively high for proportions close to 1. This method performs relatively well under certain cases; however, the coverage probabilities are generally not consistently better or worse than the nominal confidence level.

Hypothesis Test Based Methods for Two Proportions

The second critical part of the test and evaluation program for a strategic weapon system is the follow-on test program. The goal of the follow-on test program is to detect a possible decrease in system reliability with a specified probability of detection (power) and a specified false alarm rate (significance level). This naturally leads to the comparison of two independent binomial proportions, which occurs frequently in statistical practice.

Let \( X \) and \( Y \) denote the number of successes in two independent samples from two binomial populations \((n_1, p_1)\) and \((n_2, p_2)\), respectively. The joint probability of a realization is

\[
P(X = x, Y = y) = \binom{n_1}{x} \binom{n_2}{y} \left( p_1 \right)^x \left( 1 - p_1 \right)^{n_1 - x} \left( p_2 \right)^y \left( 1 - p_2 \right)^{n_2 - y}
\]

for \( x = 0, 1, \ldots, n_1 \) and \( y = 0, 1, \ldots, n_2 \). The goal is to test \( H_0: p_1 = p_2 \) against \( H_1: p_1 \neq p_2 \) or to test \( H_0: p_1 = p_2 \) against \( H_1: p_1 > p_2 \). The one-sided test is of interest because one wants to know whether there has been a significant reliability degrade in a given year. To avoid approximation, a quasi-exact method is used to determine the number of tests \( n_2 \) required. The method is described as follows:

\[
\text{Power} = \sum_{x = 0}^{n_1} \sum_{y = 0}^{n_2} \binom{n_1}{x} \binom{n_2}{y} \left( p_1 \right)^x \left( 1 - p_1 \right)^{n_1 - x} p_2^y \left( 1 - p_2 \right)^{n_2 - y} \geq 1 - \beta
\]

\[
\text{Significance Level} = \sum_{x = 0}^{n_1} \sum_{y = 0}^{n_2} \binom{n_1}{x} \binom{n_2}{y} p_1^x \left( 1 - p_1 \right)^{n_1 - x} p_2^y \left( 1 - p_2 \right)^{n_2 - y} \leq \alpha
\]

where the sums are restricted over the rejection region \( \frac{x}{n_1} - \frac{y}{n_2} > C \) for \( C \) between 0 and \( \Delta \), and \( p_2 \) can be defined as \( p_1 - \Delta \). The follow-on samples, \( n_2 \), can be determined for a \( \Delta \) degrade in reliability from the current reliability \( p_1 \) and number of tests \( n_1 \) for a specified power, \( 1 - \beta \), and confidence level \( 1 - \alpha \). As with the initial test program approach, it is also thought that this test procedure is necessarily conservative because of the discreteness of the binomial distribution. Several articles in the literature have focused on assessing confidence interval methods for the difference between two proportions. Newcombe (1998) and Brown and Li (2005) provided comprehensive reviews of several of those methods.

The approach taken here is to evaluate the confidence interval methods with respect to coverage probability and confidence interval width. The candidate confidence interval method will be inverted into the hypothesis testing setup and evaluated against the
quasi-exact method with respect to power and significance level for specified reliability degrades. Berger (1994) evaluated several exact unconditional tests for computing two binomial proportions and concluded that Boschloo’s (1970) test, with the confidence interval modification by Berger and Boos (1994), generally has the best power properties.

Let \( \hat{p}_1 = X/n_1 \) and \( \hat{p}_2 = Y/n_2 \) denote the sample proportions. The standard large-sample Wald confidence interval (Mukhopadhyay 2000) for comparing the difference between two proportions is

\[
\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.
\]

Though this method is straightforward and widely used, it behaves poorly as shown in the one-sample case. The adjustment to the Wald confidence interval by Agresti and Coull (1998) can be extended to the two-sample case (Agresti-Coull extension) by adding \( z^2_{1-\alpha/2} \) pseudo observations (half successes and half failures) to both samples, resulting in \( \hat{p}_1 = (X + z^2_{1-\alpha/2})/(n_1 + z^2_{1-\alpha/2}) \) and \( \hat{p}_2 = (Y + z^2_{1-\alpha/2})/(n_2 + z^2_{1-\alpha/2}) \) and resulting in the following confidence interval:

\[
\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}.
\]

Agresti and Caffo (2000) provided a simple adjustment to the Wald interval by adding one success and one failure to both samples. This adjusted interval will be referred to as the adjusted uniform method in the simulation study. The simple adjustment has advantages in computation and presentation and has shown through simulation to work quite well. However, there has been no theoretical support for their simulation conclusions. Other adjustments could be made using a Jeffreys prior \([\text{Beta}(1/2, 1/2)]\); this adjusted interval will be referred to as the adjusted Jeffreys method.

An alternative way to adjust the Wald confidence interval is to take a fully Bayesian approach (Bayesian Jeffreys). If we assign Jeffreys prior distribution to both \( p_1 \) and \( p_2 \), the joint posterior distribution for \( p_1 \) and \( p_2 \) is proportional to

\[
I(p_1, p_2 | n_1, n_2, x, y) = p_1^x (1-p_1)^{n_1-x} p_2^y (1-p_2)^{n_2-y}
\]

and then one can calculate the probability of various functions of \( p_1 \) and \( p_2 \). Brown et al. (2005) calculated confidence bands based on the fully Bayesian approach and showed that it did not outperform the adjusted Jeffreys method.

Brown et al. (2005) suggested the recentered Wald confidence interval:

\[
\hat{p}_1 - \hat{p}_2 \pm \frac{\tau \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}{1 + \frac{\tau^2}{n_1 + n_2}} - \frac{(\hat{p}_1 - \hat{p}_2)^2}{n_1 + n_2}
\]

where \( \tau = T_1 - a_{1-\alpha} + n_2 - 2 \) and \( \hat{p} \) is a truncated estimate between \( \frac{(\hat{p}_1 - \hat{p}_2)n_2}{n_1 + n_2} \) and \( 1 - \frac{(\hat{p}_1 - \hat{p}_2)n_1}{n_1 + n_2} \). Empirical results showed that using a \( T \)-distribution performed much better than the normal when \( n_1 \) and \( n_2 \) were small.

The Newcombe hybrid score confidence interval (Newcombe 1998) is formed by calculating score intervals for each of the two independent binomial proportions \( p_1 \) and \( p_2 \). Let \( l_1, u_1 \) be the roots of the quadratic equation \( z_{1/2} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \); then the confidence interval has the form

\[
\left( \hat{p}_1 - \hat{p}_2 - z_{1/2} \sqrt{l_1 (1-l_1)/n_1 + u_1 (1-u_1)/n_2}, \hat{p}_1 - \hat{p}_2 - z_{1/2} \sqrt{u_1 (1-u_1)/n_1 + l_2 (1-l_2)/n_2} \right)
\]

Zhou et al. (2004) proposed a confidence interval based on the Edgeworth expansion for the studentized difference between two binomial proportions. Unlike the Wald approximation, this method takes into account the impact of the skewness in the binomial distribution. The resulting two-sided, 100(1 - \( \alpha \))% skewness-corrected interval for \( p_1 - p_2 \) is defined as follows:

\[
\left[ (\hat{p}_1 - \hat{p}_2) - \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} \right. \\
\left. \left( z_{1-\alpha/2} - (n_1 + n_2)^{-1/2} \hat{Q}(z_{1-\alpha/2}) \right), \left( \hat{p}_1 - \hat{p}_2 \right) - \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} \right. \\
\left. \left( z_{1-\alpha/2} - (n_1 + n_2)^{-1/2} \hat{Q}(z_{1/2}) \right) \right]
\]
Comparison Sample Size Methods for System Testing

...
FIGURE 3. Average coverage probabilities for initial reliability, confidence level \( (c = 0.90, 0.95) \), and delta degrade \( (\Delta = 0.1, 0.2, 0.3) \), \( N_1 = 25, N_2 = 6 \).
FIGURE 4. Average coverage probabilities for initial reliability, confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 25, N_2 = 10$. 

*Comparison Sample Size Methods for System Testing* 239
FIGURE 5. Average coverage probabilities for initial reliability, confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 50, N_2 = 6$. 
FIGURE 6. Average coverage probabilities for initial reliability, confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 50, N_2 = 10$. 
FIGURE 7. Average coverage probabilities for initial reliability, confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 75, N_2 = 6$. 
FIGURE 8. Average coverage probabilities for initial reliability, confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 75, N_2 = 10$. 

Comparison Sample Size Methods for System Testing
FIGURE 9. Average confidence interval widths for initial proportion ($p_1$), confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 25, N_2 = 8$. 

---

FIGURE 9. Average confidence interval widths for initial proportion ($p_1$), confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 25, N_2 = 8$. 

---

J. D. Warfield and S. E. Roberts
FIGURE 10. Average confidence interval widths for initial proportion \((p_1)\), confidence level \((c = 0.90, 0.95)\), and delta degrade \((\Delta = 0.1, 0.2, 0.3)\), \(N_1 = 25, N_2 = 10\).
FIGURE 11. Average confidence interval widths for initial proportion \( (p_1) \), confidence level \( (c = 0.90, 0.95) \), and delta degrade \( (\Delta = 0.1, 0.2, 0.3) \), \( N_1 = 50, N_2 = 6 \).
FIGURE 12.  Average confidence interval widths for initial proportion ($p_1$), confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 50, N_2 = 10$. 
FIGURE 13. Average confidence interval widths for initial proportion ($p_1$), confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 75$, $N_2 = 6$. 
FIGURE 14. Average confidence interval widths for initial proportion ($p_1$), confidence level ($c = 0.90, 0.95$), and delta degrade ($\Delta = 0.1, 0.2, 0.3$), $N_1 = 75$, $N_2 = 10$. 

*Comparison Sample Size Methods for System Testing* 249
TABLE 1  
Power and Significance Level for Initial Proportion ($p_1$), Actual Significance Level ($1 - c$), and Delta Degrade ($\delta$), $N_s = 25$

| $p_1$ | Significance level | $n_2$ | Delta | Quasi-exact method | Bayesian Jeffreys |
|-------|-------------------|-------|-------|--------------------|------------------|
|       |                   |       |       | Power              | Actual significance level | Power (standard deviation) |
| 0.80  | 0.10              | 4     | 0.1   | 0.440              | 0.274             | 0.213 (0.0017) |
|       |                   |       | 0.2   | 0.497              | 0.175             | 0.350 (0.0023) |
|       |                   |       | 0.3   | 0.470              | 0.092             | 0.498 (0.0025) |
| 8     | 0.1               | 0.460 | 0.240 | 0.211              | 0.098             | 0.466 (0.0022) |
|       | 0.2               | 0.503 | 0.121 | 0.087             | 0.098             | 0.633 (0.0023) |
|       | 0.3               | 0.657 |       |                   |                   |                  |
| 12    | 0.1               | 0.461 | 0.211 | 0.283              | 0.0020            |                  |
|       | 0.2               | 0.557 | 0.098 | 0.515              | 0.0025            |                  |
|       | 0.3               | 0.782 | 0.098 | 0.728              | 0.0020            |                  |
| 20    | 0.1               | 0.477 | 0.186 | 0.296              | 0.0021            |                  |
|       | 0.2               | 0.628 | 0.098 | 0.574              | 0.0024            |                  |
|       | 0.3               | 0.855 | 0.098 | 0.801              | 0.0016            |                  |
| 0.90  | 0.10              | 4     | 0.1   | 0.440              | 0.274             | 0.436 (0.0025) |
|       |                   |       | 0.2   | 0.553              | 0.218             | 0.588 (0.0024) |
|       |                   |       | 0.3   | 0.706              | 0.218             | 0.728 (0.0020) |
| 8     | 0.1               | 0.462 | 0.241 | 0.454              | 0.0025            |                  |
|       | 0.2               | 0.671 | 0.241 | 0.666              | 0.0022            |                  |
|       | 0.3               | 0.832 | 0.241 | 0.824              | 0.0014            |                  |
| 12    | 0.1               | 0.509 | 0.237 | 0.490              | 0.0025            |                  |
|       | 0.2               | 0.751 | 0.237 | 0.710              | 0.0021            |                  |
|       | 0.3               | 0.903 | 0.237 | 0.873              | 0.0011            |                  |
| 20    | 0.1               | 0.550 | 0.240 | 0.570              | 0.0025            |                  |
|       | 0.2               | 0.802 | 0.240 | 0.791              | 0.0017            |                  |
|       | 0.3               | 0.938 | 0.240 | 0.933              | 0.0006            |                  |
| 0.90  | 0.25              | 4     | 0.1   | 0.494              | 0.275             | 0.282 (0.0020) |
|       |                   |       | 0.2   | 0.459              | 0.131             | 0.443 (0.0025) |
|       |                   |       | 0.3   | 0.549              | 0.073             | 0.596 (0.0024) |
| 8     | 0.1               | 0.450 | 0.179 | 0.285              | 0.0020            |                  |
|       | 0.2               | 0.520 | 0.077 | 0.505              | 0.0025            |                  |
|       | 0.3               | 0.732 | 0.077 | 0.688              | 0.0021            |                  |
| 12    | 0.1               | 0.458 | 0.152 | 0.377              | 0.0023            |                  |
|       | 0.2               | 0.640 | 0.087 | 0.619              | 0.0024            |                  |
|       | 0.3               | 0.848 | 0.087 | 0.805              | 0.0016            |                  |
| 20    | 0.1               | 0.491 | 0.135 | 0.363              | 0.0023            |                  |
|       | 0.2               | 0.771 | 0.097 | 0.661              | 0.0022            |                  |
|       | 0.3               | 0.933 | 0.097 | 0.875              | 0.0011            |                  |
| 0.90  | 0.25              | 4     | 0.1   | 0.494              | 0.275             | 0.518 (0.0025) |
|       |                   |       | 0.2   | 0.569              | 0.209             | 0.673 (0.0022) |
|       |                   |       | 0.3   | 0.710              | 0.209             | 0.795 (0.0016) |
| 8     | 0.1               | 0.492 | 0.200 | 0.518              | 0.0025            |                  |
|       | 0.2               | 0.729 | 0.200 | 0.722              | 0.0020            |                  |
|       | 0.3               | 0.879 | 0.200 | 0.867              | 0.0012            |                  |
| 12    | 0.1               | 0.492 | 0.226 | 0.565              | 0.0025            |                  |
|       | 0.2               | 0.812 | 0.226 | 0.789              | 0.0017            |                  |
|       | 0.3               | 0.935 | 0.226 | 0.923              | 0.0007            |                  |
| 20    | 0.1               | 0.636 | 0.231 | 0.664              | 0.0022            |                  |
|       | 0.2               | 0.878 | 0.231 | 0.885              | 0.0010            |                  |
|       | 0.3               | 0.971 | 0.231 | 0.968              | 0.0003            |                  |

(Continued on next page)
probabilities improve for all three intervals, but the same general pattern still holds. The Bayesian Jeffreys and Newcombe hybrid methods provide intervals with the smallest widths regardless of sample size. For initial proportions between 0.70 and 0.85, the Newcombe hybrid method provides the smallest intervals, whereas the Bayesian Jeffreys method produces smaller intervals for initial proportions greater than 0.85. Therefore, the Newcombe hybrid method is recommended for initial proportions less than 0.85 and the Bayesian Jeffreys method is recommended for initial proportions greater than 0.85.

For small sample sizes, the coverage probabilities of the Edgeworth expansion, Wald, and recentered Wald methods fall well below the nominal confidence level. The coverage probabilities of these three methods improve for larger sample sizes but fall short compared to the other methods further discussed. The Agresti-Coull and adjusted uniform methods have coverage probabilities exceeding 0.95 in almost all cases, even when the nominal confidence level is 0.90.

Table 1 provides the results comparing the Bayesian Jeffreys method to the quasi-exact method with respect to power. The actual significance level is listed for the quasi-exact method because the method cannot always achieve the desired significance level. Overall the methods are comparable, with the quasi-exact method being slightly more powerful as the sample size increases.

**CONCLUSIONS**

Testing programs exist to establish initial estimates of reliability for systems and to determine whether a significant decrease in reliability has occurred in relation to the initial value. For initial test programs, the methodology is based on confidence intervals for a single proportion. Several methods were investigated through a simulation study. Based on the simulation results, the Bayesian Jeffreys method performs the best in the majority of situations, with the adjusted uniform performing best for proportions less than 0.80. For follow-on test programs, the methodology is based on hypothesis testing for two proportions. Several confidence interval methods for the difference between two
proportions were investigated through a simulation study. Based on the simulation study results, the Newcombe hybrid score method and the Bayesian approach with Jeffreys prior show the most promise. However, the Bayesian method provided very little gain in power when compared to the quasi-exact method.

ABOUT THE AUTHORS

Joseph D. Warfield is a senior research statistician and section supervisor of the Statistical Sciences Section of the Force Projection Sector at the Johns Hopkins University Applied Physics Laboratory. His current areas of research focus on optimal design of experiments and reliability related methods. He received a B.S. degree in Mathematics from Loyola University, an M.S. degree in Statistics from Virginia Tech, and a Ph.D. in Statistics from the University of Maryland, Baltimore County.

Sarah Elise Roberts is a research statistician in the Force Projection Sector at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. She received a B.S. degree in Mathematics from the Pennsylvania State University and an M.S. degree in Statistics from the University of Maryland, Baltimore County.

SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed on the publisher’s website.

REFERENCES

Agresti, A., and B. A. Coull. 1998. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52: 119–26.

Berger, R. L. 1994. Power comparison of exact unconditional tests for comparing two binomial proportions. Technical Report 2266, Department of Statistics, North Carolina State University, NC, USA.

Berger, R. L., and D. D. Boos. 1994. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 89: 1012–16.

Borkowf, C. B. 2006. Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. *Statistics in Medicine* 25: 3679–95. http://dx.doi.org/10.1002/sim.2469

Boschloo, R. D. 1970. Raised confidence level of significance for the 2×2-table when testing the equality of two probabilities. *Statistica Neerlandica* 24: 1–35. http://dx.doi.org/10.1111/j.1467-9574.1970.tb00104.x

Brown, L. D., T. T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science* 16: 101–33.

Brown, L. D., and X. Li. 2005. Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference* 130: 359–75. http://dx.doi.org/10.1016/j.jspi.2003.09.039

Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404–13. http://dx.doi.org/10.1093/biomet/26.4.404

Freeman, M. F., and J. W. Tukey. 1950. Transformations related to the angular and the square root. *Annals of Mathematical Statistics* 21: 607–11. http://dx.doi.org/10.1214/aoms/1177729756

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall.

Montgomery, D. C. 2009. *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons.

Mukhopadhyay, N. 2000. *Probability and statistical inference*. New York: Marcel Dekker.

Newcombe, R. G. 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 17: 857–72. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Newcombe, R. G. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 17: 873–90. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Newcombe, R. G. 2001. Logit confidence intervals and the inverse sinh transformation. *The American Statistician* 55: 200–203. http://dx.doi.org/10.1198/000313001317098167

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22: 209–212. http://dx.doi.org/10.1080/01621459.1927.10502953

Zhou, X. H., M. Tsao, G. S. Qin. 2004. New intervals for the difference between two independent binomial proportions. *Journal of Statistical Planning and Inference* 123: 97–115. http://dx.doi.org/10.1016/S0378-3758(03)00146-0