A Comparison of the Statistical Properties of Confidence Intervals for the Difference in Two Proportions

Michael C. Mosier, Ph.D.   Dawn M. DuBois, M.S.   Debra S. Schleusener, M.S.   Marc A. Meyer, M.S.

Objectives

The objectives of this study were to compare the

- statistical properties,
- software availability, and
- computing demands,

of four competing methods for constructing exact confidence intervals for the difference in two independent proportions $(\hat{p}_1 - \hat{p}_2)$.

Four methods were chosen to investigate:

1. The classic normal approximation (asymptotic Z interval, denoted as (AZ))

This method relies on the central limit theorem, and requires "large samples" in order to achieve the stated confidence level. Most textbooks state that there should be a minimum of 10 successes and 10 failures in each of the two samples. This is not an "exact method," and included here as a benchmark for the following three exact procedures.

2. The Santner and Snell (SS) method (1980)

This method uses the "tail method", which involves inverting two separate, one-tailed tests, using the observed sample difference $T = \hat{p}_1 - \hat{p}_2$ as the test statistic. It was shown to be "exact" in that for any size sample, the coverage probability is at least as large as the stated confidence level. It has since been shown to often be overly conservative.

3. The Chan and Zhang (CZ) method (1999)

This method also uses the "tail method", but with a score statistic $T = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$ where $\hat{p}_1$ and $\hat{p}_2$ are maximum likelihood estimates, from Miettinen and Nurminen (1985). It was shown to also be "exact", but less conservative than the SS method.

4. The Agresti and Min (AM) method (2001)

This method also uses the above score statistic, but inverts a single two-sided test rather than two separate, one-tailed tests. They showed this interval to also be "exact", but even less conservative than the CZ method, therefore providing better performance.

Methods

In the original articles introducing them, each of the three exact methods was studied with only a few combinations of small sample sizes. We compared the intervals using 15 different group size combinations, ranging from small to large samples. Then for each set of sample sizes, we computed intervals over the following combinations of observed sample proportions $(\hat{p}_1, \hat{p}_2)$:

$\hat{p}_1$	0.2	0.2	0.2	0.2	0.2	0.2	0.5	0.5	0.5	0.5	0.5	0.5
$\hat{p}_2$	0.1	0.1	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5	0.3	0.4	0.5

With each of the above 12 sample proportion combinations, times the 15 sample size combinations, this gives 180 cases that were considered for comparison.

Interval Length: The AZ, SS, and CZ methods were run using SAS/STAT® software, Version 9.3 of the SAS® System for Windows. The AM method is not available in SAS®, and therefore was run using StatXact® software with a SAS® interface (StatXact® PROCs 9.0). The 95% confidence intervals and their lengths were determined for each of the 180 cases.

Run Time: In addition, we captured the CPU time required to produce each interval. CPU time is the amount of time the CPU required for the actual computation, and is not affected (or at least minimally) by other processes running in the background. Each was run on the same PC with an Intel Core i5 2.67 GHz processor, running Windows 7, five separate times. The five run times for each interval were then averaged to give an estimate of the time required in our setting. The standard deviation (SD) of the five runs was examined to give us a sense of the variability from run to run. These times are not intended to predict run times on other machines, but instead are intended for relative comparisons of the four methods for constructing the confidence intervals.

The StatXact® PROCs 9.0 software will produce all three of the exact procedures, but the AM method is the only one not also available in SAS/STAT® Version 9.3. Therefore, the run times reflect SAS® for the AZ, SS, and CZ methods, and StatXact® only for AM method.

Results: Summary of Intervals and Lengths

Results are shown in the table at lower left, which spans three panels, beginning with combinations of small samples and moving toward combinations with large samples. Columns are shaded so that the same item has the same color across the four methods.

Comparing the AM method to the SS method shows the AM method to be the clear choice. The confidence intervals for AM are completely contained within the SS intervals for 174/180 cases (96.7%). Likewise, the length of the AM interval is shorter than the SS interval for all 180 cases (100%).

Comparing the AM method to the CZ method again shows the AM method to be the clear choice. While the CZ method does perform better than the SS method, the confidence intervals for AM are still completely contained within the CZ intervals for 148/180 cases (82.2%). The length of the AM interval is shorter than the CZ interval for 168/180 cases (93.3%).

For comparison of the AM with the AZ (normal approx.) method, we consider only cases with sample size/proportion combinations that involve the minimum recommended number of successes and failures (10 of each in both samples). This occurs in 89 of the 180 cases we considered. For situations where this does not hold, of course the AZ is not recommended due to its failure to maintain the stated confidence level.

Among the 89 cases where the AM method is valid, the AM interval is completely contained within the AZ interval only 11 times (12.4%), while the AZ interval is completely contained within the AM interval 24 times (27.0%). The length of the AM interval is shorter than the AZ interval for 29/89 cases (32.6%), indicating that the AZ interval is shorter in 67.4% of cases.

The average interval length across the 89 cases is 0.256 for the AM method, and 0.255 for the AZ method, and it can be seen that in these cases, the two methods produce very similar intervals. Therefore, based on interval lengths, the classic AZ interval appears to be the best choice in cases where the minimum sample size requirements are met.

Detecting a statistically significant difference between groups:

When considering what conclusion may be made regarding a statistically significant difference between groups, we look at whether zero is contained within the interval. If zero is excluded by the interval, then one may conclude there is a significant difference. Over the 180 distinct cases studied, the AZ method excluded zero 97 times (53.9%), while the SS, CZ and AM methods excluded zero 74 (41.1%), 87 (48.3%), and 87 (48.3%) times, respectively. Of the cases where the AZ excludes zero but the CZ or AM methods do not, all but one of those cases involve samples that are too small for AZ to be considered valid.

Interestingly, over the 89 cases with samples large enough for AZ to be valid, the AZ method agrees with the CZ and AM methods 88 times, on the conclusion of a significant difference. For these cases of $n_1 = 100$, $n_2 = 100$, with $\hat{p}_1 = 0.2$ and $\hat{p}_2 = 0.1$, the AZ method excluded zero while the other two methods contained it. However, for this case all three intervals had lower bounds very close to zero, with AZ falling just above zero, and the other two falling just below. Overall, in cases where the AZ method is valid, the AZ method excludes zero 54/89 (60.7%) times, the CZ and AM methods 53/89 (59.6%) times, and the SS method only 42/89 (47.2%) times.

Results: Summary of CPU Run Times

The CPU run times for the four methods show some dramatic differences. By far the shortest run times were seen with the normal approximation based AZ method. Across the 180 different cases, all run times but one were less than or equal to 0.01 seconds, and the largest SD seen for the five runs was 0.014 seconds. Overall, the run times for this method are negligible.

For the SS method, the CPU run times range from 0.03 seconds for small samples, to 32.83 seconds for $n_1 = n_2 = 200$. At the large samples, the highest SD among the five runs was 1.3 seconds. Considering only the 72 cases that have both sample sizes at or below 50, the CPU run times range from 0.03 seconds to 1.83 seconds.

For the CZ method, the CPU run times range from 0.37 seconds for small samples, to 5 min, 20.29 sec for $n_1 = n_2 = 200$. At the large samples, the highest SD among the five runs was 6.56 seconds. Considering only the 72 cases that have both sample sizes at or below 50, the CPU run times range from 0.37 seconds to 18.28 seconds.

For the AM method, using StatXact®, the CPU run times range from 0.72 seconds for small samples, to 41.13 seconds for $n_1 = n_2 = 200$. At the large samples, the highest SD among the five runs was 5.42 seconds. Considering only the 72 cases that have both sample sizes at or below 50, the CPU run times range from 0.72 seconds to 6.78 seconds.

Overall, for very small sample sizes, the actual CPU run times are not long for any of the methods. However, when sample sizes begin to approach 50, the relative differences in run times become large. The CZ method, run in SAS® has the longest run times of the four, while the SS and AM methods are comparable in all but one borderline case.

Methods with long computing times become especially troublesome when many intervals are being produced within a trial, say for multiple populations, endpoints, subgroups, or strata.

Conclusions

Four methods for computing confidence intervals for the difference in two independent proportions $(p_1 - p_2)$ have been compared: the normal approximation method AZ, and three methods for constructing exact confidence intervals (SS, CZ, and AM).

The AZ method should only be used in cases with sample size/proportion combinations that involve the minimum recommended number of successes and failures (10 of each in both samples). When sample sizes are smaller than this, the AM method (Agresti and Min) emerged as the preferred choice since it has the shortest average length while still protecting the stated confidence level.

However, when sample sizes become large enough that the AZ method has coverage probability at the stated confidence level, the intervals are nearly identical to the AM intervals, with dramatically shorter run times, and without the need to purchase additional software (StatXact®) to construct the confidence intervals. The longer run times become especially troublesome when many intervals are being produced within a trial, say for multiple populations, endpoints, subgroups, or strata. Also, when considering the conclusion of a significant difference between groups, these two methods agreed in all but one borderline case.

As such, we recommend the use of the Agresti and Min (AM) method for small sample situations, and the classic Asymptotic Z (normal approx., AZ) method for large samples.

References

Agresti A, Min Y. "On small-sample confidence intervals for parameters in discrete distributions." Biometrics 2001; 57: 963–71.

Chan ISF, Zhang Z. "Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 1999; 55: 1202–9.

Miettinen O, Nurminen M. "Comparative analysis of two rates." Statistics in Medicine 1985; 4: 213–26.

Santer TJ, Snell MK. "Small-sample confidence intervals for p1 - p2 and p1/p2 in 2 x 2 contingency tables." Journal of the American Statistical Association 1980; 75: 386–94.

EMB Statistical Solutions, LLC
A contract research organization for the pharmaceutical and biotechnology industry

Results tables (three panels) giving 95% CI, Length, and Mean Run Time (Min:Sec) for the Asymptotic Z (AZ), Santner & Snell (SS), Chan & Zhang (CZ), and Agresti & Min (AM) methods across combinations of sample sizes (n_1, n_2) and observed sample proportions ($\hat{p}_1$, $\hat{p}_2$). Individual cell values are not legibly reproducible.