ESTIMATING COMPLETION RATES FROM SMALL SAMPLES USING BINOMIAL CONFIDENCE INTERVALS: COMPARISONS AND RECOMMENDATIONS

Jeff Sauro
Oracle
Denver, CO USA
jeff.sauro@oracle.com

James R. Lewis
IBM
Boca Raton, FL
jimlewis@us.ibm.com

The completion rate – the proportion of participants who successfully complete a task – is a common usability measurement. As is true for any point measurement, practitioners should compute appropriate confidence intervals for completion rate data. For proportions such as the completion rate, the appropriate interval is a binomial confidence interval. The most widely-taught method for calculating binomial confidence intervals (the "Wald Method," discussed both in introductory statistics texts and in the human factors literature) grossly understates the width of the true interval when sample sizes are small. Alternative "exact" methods over-correct the problem by providing intervals that are too conservative. This can result in practitioners unintentionally accepting interfaces that are unusable or rejecting interfaces that are usable. We examined alternative methods for building confidence intervals from small sample completion rates, using Monte Carlo methods to sample data from a number of real, large-sample usability tests. It appears that the best method for practitioners to compute 95% confidence intervals for small-sample completion rates is to add two successes and two failures to the observed completion rate, then compute the confidence interval using the Wald method (the "Adjusted Wald Method"). This simple approach provides the best coverage, is fairly easy to compute, and agrees with other analyses in the statistics literature.

Introduction

Estimating completion rates with small samples is an important and challenging task. Confidence intervals are taught as an appropriate way to qualify results from small samples. The addition of confidence intervals to completion rate estimates helps both the engineer and readers of usability reports understand the variability inherent in small samples. While the importance of adding confidence intervals is widely agreed upon, the best method for computing them is not.

Most practitioners interpret a 95% confidence interval to indicate that in 95 out of 100 experiments, the interval constructed from the sample will contain the true value for the population. The extent to which this is the case for any given method of computing intervals is the "coverage" for that method.

The Wald method is the most commonly presented formula for calculating binomial confidence intervals (see Figure 1 below).

Figure 1: Wald Confidence Interval

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

Task completion rates are often modeled using a binomial distribution because the outcome of a task attempt is usually a binomial value (complete / didn't complete). The Wald interval is simple to compute, has been around for some time (Laplace, 1812) and is presented in most introductory statistics texts and some writings in the human factors literature (e.g., Landauer, 1988). Unfortunately, it produces intervals that are too narrow when samples are small, especially when the completion rate is not near 50%. Under these conditions its average coverage is approximately 60%, not 95% (Agresti and Coull, 1998). This is a real problem considering that HF practitioners rely on confidence intervals to have true coverage that is equal to nominal coverage in the long run.

To improve the poor average coverage of the Wald interval, advanced statistics texts often present a more complicated method called the Clopper-Pearson or "Exact" method (see Figure 2 below).

Figure 2: "Exact" / "Clopper-Pearson" Interval

$$\left[1+\frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1+\frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1}$$

The Exact method provides more reliable confidence intervals with small samples (Clopper and Pearson, 1934) and has also been discussed in the HF literature (e.g., Lewis, 1996, and Sauro, 2004). In actual practice, however, the Exact interval produces overly conservative confidence intervals with true coverage closer to 99% when the nominal confidence is 95%. It is especially vulnerable to this overly conservative nature when samples sizes are small (n <15) (Agresti and Coull, 1996). Thus, Exact intervals are too wide and Wald intervals are too narrow.

A third method called the "Score" interval (Wilson, 1927) is not overly conservative, and provides average coverage near 95% for nominal 95% intervals. Unfortunately, its computation is as cumbersome as the Exact method (see Figure 3 below), and it has some serious coverage problems for certain values when the completion rate is near 0 or 1 (Agresti and Coull, 1998).

Figure 3: "Score" / Approximate Interval

$$\left(\hat{p}+\frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p})+z^2_{\alpha/2}/4n]/n}\right)/(1+z^2_{\alpha/2}/n).$$

Another alternative method, named the Adjusted Wald method by Agresti and Coull (1998, based on work originally reported by Wilson, 1927), simply requires, for 95% confidence intervals, the addition of two successes and two failures to the observed completion rate, then uses the Wald formula to compute the 95% binomial confidence interval. Its coverage is as good as the Score method for most values of p, and is usually better when the completion rate approaches 0 or 1. The method is astonishingly simple, and has been recommended in the statistical literature (Agresti and Coull, 1998). The "add two successes and two failures" (or adding two to the numerator and 4 to the denominator) is derived from the critical value of the normal distribution for 95% intervals (1.96, which is approximately 2). Squaring this critical value provides the 4 for the denominator. For example, an observed completion rate of 80% with 10 users (8 successes and 2 failures) would be converted to 10 successes and 4 failures, and these values would then be used in the Wald formula.

Table 1 displays the four differing results for each of the interval methods for a sample of five users with four successes and one failure (80% completion rate).

Table 1: 95% confidence intervals by method for an 80% completion rate (4 successes, 1 failure)

CI Method	Low %	High %	CI Width
Exact	28.4	99.5	71.1
Score	37.6	96.4	58.8
Adj. Wald	36.5	98.3	61.8
Wald	44.9	100	55.1

As can be seen from Table 1, the different methods provide different end points and differing confidence interval widths. While one would like a narrower confidence interval (which provides less uncertainty), the interval should not be so narrow as to exclude more completion rates than expected from the stated or nominal rate – that is, a nominal 95% confidence interval should have a likelihood of 95% of containing the population parameter. The implication is clear, depending on which method the HF practitioner chooses, the boundaries presented with a completion rate can lead to different conclusions about the usability of an interface.

The Wald and Exact methods are by far the most popular ways of calculating confidence intervals. Depending on which method practitioners are using to calculate their intervals, they will either work with intervals that provide a false sense of precision (Wald method) or work with intervals that are consistently less precise than their nominal precision (Exact method). If the Adjusted Wald method can provide the best average coverage while still being relatively simple to compute (as suggested in the statistical literature, Agresti and Coull, 1998), it will provide the HF practitioner with the easiest and most precise way of computing binomial confidence intervals for small samples.

Method

One way to test the effectiveness of a confidence interval calculation is to take a sample many times from a larger data set and see how well the calculated confidence interval contained the actual completion rate of the data set. We took data from several tasks across five usability evaluations with completion rates between 20% and 97%. The usability analyses were performed on commercially available desktop and web-based software applications in the accounting industry. Each task had at least 49 participants, and we used these completion rates as the best estimate of the population completion rate.

Table 2: Percent coverage for nine task completion rates by confidence interval method and number of users. Expected width is 95.0. Values are derived from sampling 5, 10 or 15 completion rates (or hypothetical users) 10,000 times.

		Observed Task Completion Rate								
CI Method	Users	**20.4%**	**42.9%**	**61.2%**	**65.3%**	**77.6%**	**85.7%**	**91.8%**	**93.8%**	**97.8%**
Exact	5	99.5	98.74	99.11	99.73	99.34	98.55	99.78	99.88	100
	10	99.72	98.93	98.96	97.73	99.60	99.81	99.86	99.35	100
	15	97.73	99.02	99.68	99.81	98.88	99.70	100	100	100
Adjusted Wald	5	94.98	98.74	99.11	96.05	93.48	98.55	95.40	97.50	89.94
	10	98.23	98.93	96.54	97.73	96.89	97.46	97.50	99.35	100
	15	99.36	99.02	98.92	97.89	97.96	97.88	99.43	97.38	100
Score	5	94.98	93.50	91.47	96.05	93.48	98.55	95.40	97.50	89.94
	10	98.23	96.87	96.54	97.73	91.17	97.46	97.50	99.35	100
	15	97.73	99.02	97.70	97.89	97.96	97.88	99.43	97.38	100
Wald	5	69.35	84.93	85.70	84.84	73.10	53.75	35.93	28.30	10.06
	10	92.01	96.87	93.26	91.66	93.88	81.80	60.20	51.77	20.74
	15	88.11	96.46	97.70	94.82	92.04	92.87	77.61	67.15	30.53

Using a Monte Carlo simulation method written in Minitab, we took 10,000 unique random samples of 5, 10 and 15 completion rates to test each of the confidence interval methods (Wald, Exact, Score and Adjusted Wald). We then counted how many of the 10,000 completion rates fell outside the calculated intervals for each of the methods. For example, on one sample of 5 users from a dataset with a population completion rate of 65.3%, we observed one success and four failures (a 20% completion rate). The Exact method provided a 95% confidence interval from .5% to 71.6%, so it did contain the true population completion rate of 65.3%. The Score method provided intervals from 3.6% to 62.5%, so it did not contain the true rate. Since we calculated nominal 95% confidence intervals, we expect coverage of 95%. In other words, about 9,500 of the 10,000 intervals computed during a Monte Carlo simulation should contain the true value.

A Note on the Methodology

We could have chosen any hypothetical completion rates to test the confidence intervals (as is often the case in the statistical literature) but we used values from a known large sample usability study so as to focus our analysis on likely completion rates for commercially available software. While the HF practitioner usually doesn't know ahead of time what the population completion rate is, this exercise allowed us to work backwards to see how well the smaller samples predicted the known completion rates. We were in essence running 10,000 usability evaluations with small samples, calculating the confidence interval with the different methods, and seeing how many times the known completion rate was contained within the intervals. While a sample size of 49 may not seem large enough to test 10,000 combinations of completion rates, even this modest sample size contains about 2 million unique combinations of five users.

Results

Table 2 contains the results of Monte Carlo simulations for nine tasks with varying completion rates (e.g., 91.8%, 93.8%, etc.) for sample sizes of 5, 10 or 15. As expected, the Wald interval provided the worst coverage, only containing the actual proportion 10% of the time for the task with a 97.8% completion rate and 5 users. To find this value, start with the Wald method in the bottom left cell of Table 2. Next, find the intersection with the completion rate of 97.8% (the rightmost column). The first value in this cell (10.06) means that 10.06% of the calculated intervals contained the true values using the Wald method with a sample of 5 users (the second and third values are for 10 and 15 user samples respectively). For the Wald method to be a legitimate method to apply to these types of data, one would expect this value to be approximately 95%. Even at the less extreme completion rate of 85.7%, the Wald interval only contained the true value about half of the time (53.75%) – a far cry from the 95% many practitioners would have expected from a nominal 95% confidence interval calculation.

The Exact interval showed the expected conservative coverage with many of the nominally 95% confidence intervals capturing over 99% of the 10,000 completion rates (see especially the completion rates above 90% in Table 2). The Adjusted Wald and Score methods provided average coverage closest to the 95% nominal level, which confirms earlier recommendations in the statistical literature (Agresti and Coull, 1998). The

mean and standard deviation of the coverage for each of the methods appears in Table 3.

Table 3: Average coverage by confidence interval method (*n*= 27 for each cell). Expected mean is 95.00.

CI Method	Mean %	SD
Exact	99.39	0.64
Score	97.56	2.17
Adj. Wald	96.69	2.68
Wald	72.06	26.43

Discussion

The Monte Carlo simulations show that the Adjusted Wald method provides the coverage closest to 95%. An additional advantage of the Adjusted Wald method is its ease of calculation. Thus, HF practitioners should use the Adjusted Wald method to calculate confidence intervals for small sample completion rates. This can be accomplished by simply adding two successes and two failures to their observed sample, then computing a 95% confidence interval using the standard Wald method. If a practitioner needs a higher level of confidence than 95%, then he or she should substitute the appropriate *Z*-critical values for 2 and 4. For example, a 99% confidence interval would use the *Z*-critical value of 2.58. The confidence interval would then be calculated by adding 2.58 successes and 6.63 failures to the observed completion rate.

The Score method provided coverage better than the Exact and Wald methods but fell short of the Adjusted Wald method. Additionally, its drawback is its computational difficulty and its poor coverage for some values when the population completion rate is around 98% or 2%, regardless of sample size (Agresti and Coull, 1998). The only advantage in using the Score method is that it provides more precise endpoints when the ends of the intervals are close to 0 or 1. For some values (e.g. 9/10) the adjusted Wald's crude intervals go beyond 1 and a substitution of >.999 is used. For the Score method, however, the upper interval is calculated as a more precise .9975.

The Exact method was designed to guarantee **at least** 95% coverage, whereas approximate methods (such as the Adjusted Wald) provide an average coverage of 95% in the long run. HF practitioners should use the Exact method when they need to be sure they are calculating a 95% or greater interval – erring on the conservative side. For example, at the population completion rate of 97.8% both the Score and Adjusted Wald methods had actual coverage that fell to 89% (See Table 2 above). When the risk of this level of actual coverage is inappropriate for an application, then the Exact method provides the necessary precision.

The Wald method should be avoided if calculating confidence intervals for completion rates with sample sizes less than 100. Its coverage is too far from the nominal level to provide a reliable estimate of the population completion rate. As the sample size increases above 100, all four methods converge to similar intervals. A calculator for all four methods is available online at http://www.measuringusability.com/wald.htm.

When All Users Pass or Fail

With small sample sizes, it is a common occurrence that all users in the sample will complete a task (100% completion rate) or all will fail the task (0% completion rate). For these scenarios, it is often unpalatable to report 100% or 0%. After all, how likely is it that the true population parameter is as extreme as 100% or 0%? One alternative is to use the midpoint of the binomial confidence interval derived from the Adjusted Wald method as the point estimate (called the Wilson Point Estimator). For example, if 15 out of 15 users complete a task, the mid-point of the Adjusted Wald method provides a 94.01% completion rate. While this value may seem too far from the observed 100%, its attractiveness is that it is a function of the sample size—the greater the sample size, the closer this value will be to 100%. Whether this method provides a consistent advantage in improving the accuracy of point estimates is a topic for future research.

Conclusion

There is a strong need to continue to encourage HF practitioners to include confidence intervals when reporting estimates of completion rates. Because the Adjusted Wald method is just a slight modification to the widely-taught Wald method, it should be easy to teach with other basic statistics without overwhelming students.

Confidence intervals are a way to build a reasonable boundary to capture unknown population completion rates. For a 95% confidence interval, "reasonable boundary" means a 5% chance of not containing the population completion rate after repeated samples. "Reasonable boundary" is not a 1% chance and certainly not a 40% chance– the typical rates obtained when using the Exact or Wald methods to generate binomial confidence intervals. To use the Adjusted Wald interval, the HF practitioner can use their own software, a spreadsheet calculation, or the calculator at http://www.measuringusability.com/wald.htm, which also computes the Exact, Score and Wald intervals for comparison.

Acknowledgements

We'd like to thank Lynda Finn of Statistical Insight for providing the Monte Carlo macro in Minitab and assistance with interpreting the statistical literature. We'd also like to thank Erika Kindlund of Intuit for providing the large sample completion rates.

References

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

Clopper, C. J., and Pearson, E. (1934). The use of confidence intervals for fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404-413.

Landauer, T. K. (1997). Behavioral research methods in human-computer interaction. In M. Helander, T. K. Landauer, and P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (pp. 203-227). Amsterdam, Netherlands: North Holland.

Laplace, P. S. (1812). *Theorie analytique des probabilitites*. Paris, France: Courcier.

Lewis, J. R. (1996). Binomial confidence intervals for small sample usability studies. In G. Salvendy and A. Ozok (eds.), *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics -- ICAE '96* (pp. 732-737). Istanbul, Turkey: USA Publishing.

Sauro, J. (2004). Restoring confidence in usability results. From Measuring Usability, article retrieved Jan 2005 from http://www.measuringusability.com/conf_intervals.htm

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212.