upa
home

Featured UPA Links:

Job Bank

Member Benefits: Discounts and Special Offers:

Event Discounts

THE UPA VOICE

Why a Completion Rate is Better with a Confidence Interval

By Jeff Sauro

Jeff Sauro is a Six Sigma trained Statistician at Oracle in Denver, CO. Before Oracle, Jeff was a Human Factors Engineer at PeopleSoft, Intuit and General Electric. Jeff has presented and published on the topic of usability metrics at CHI, UPA and HFES conferences and maintains the website **measuringusability.com**. He received bachelors degrees from Syracuse University and a Masters from Stanford University.

You just ran an 8 participant usability test and watched as 7 out of 8 users completed a task to provide an 87.5% completion rate. You need to summarize your findings in a test report, and you’ve heard that you should include confidence intervals. But what exactly do confidence intervals do that the completion rate cannot?

Confidence intervals have been lauded by the APA (American Psychological Association) as the preferred technique when presenting data for any size sample, and this has been echoed in the usability literature as well **[3][5]**. They are so strongly recommended because they combine information on “location and precision and can often be directly used to infer significance levels” **[1]** More on location and precision will be discussed below.

From the usability test described above, we don’t know what proportion of all users actually will complete the task (for that we’d need to test all users, which could be thousands). The observed proportion tells us that perhaps 87.5 percent of all users will be successful, and that’s certainly more than we knew before running the test. But the chance that this estimate is absolutely, as opposed to approximately, correct is very close to zero, especially when sample sizes are small.

There are some simple techniques that consistently improve the accuracy of small-sample estimates In this case, it’s to add one success and one failure to the observed proportion [2]. This provides a slightly better estimate of (7+1)/ (8+2) = 80 percent (better because in the long run, this estimate will be closer to the true population value). Even this adjustment only slightly improves the estimate In fact, no matter what adjustment we make to a point estimate, it will almost always be wrong. Here’s why. If I asked you how many seconds it will take you to get to work tomorrow, you’d probably come up with a reasonable guess based on your past experience, say 1250 seconds (not quite 21 min). But what are the chances you’d be accurate to the second? Not very likely. Now if I asked you to tell me a range of time that it would take you to get to work, you might say, between 1200 to 1500 seconds (20 to 25min) and you’d probably be right on many days.

It’s much easier and more accurate to report a likely range of values than hazarding a guess. Yet, when we report just a completion rate in a usability report, it’s tantamount to stating the number of seconds it will take to get to work. The formal way of providing a likely range on an unknown population parameter (the mean usually) is to use a confidence interval.

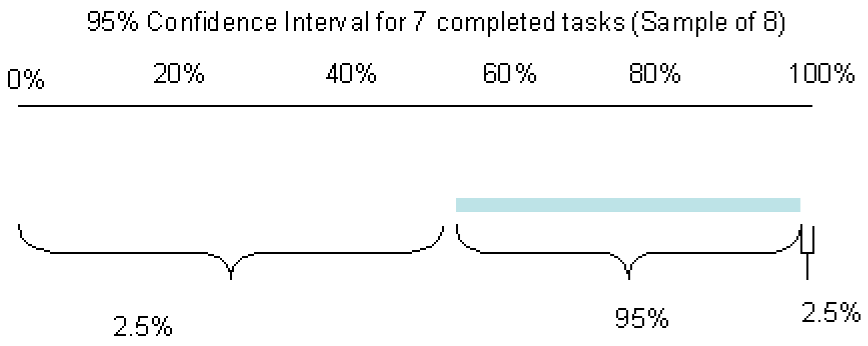
Calculating a Confidence Interval on Small Sample Completion Rates

To calculate a confidence interval on this data (small-sample proportion) there is also an adjustment you need to make for the smaller sample size: add two successes and two failures to the observed completion rate (when sample sizes are less than about 100)**[5]**. So instead of 7/8 you have 9/12. From this proportion you build a confidence interval through the following steps (or use the online calculator available at **[4]**):

1. Add two successes and two failures to the observed proportion 7/8 = 9/12
2. Find the Standard Deviation:
   a. Multiple the completion rate times 1 minus the completion rate (use the adjusted completion rate). 9/12 = (.75) x (.25) = .1875. This is the Variance.
   b. Take the square root of the Variance to get the Standard Deviation: SQRT (.1875) = .433
3. Find the Standard Deviation of the Mean: This is the Standard Deviation divided by the Square Root of the sample size: SQRT(12) = 3.464, so SEM = .433/3.464 = .125. This figur is also called the Standard Error of the Mean.
4. Find the Margin of Error: Multiply your Standard Error times your desired level of confidence Let's use the common 95% Confidence Level which is just about 2 standard errors: 1.96. Multiple the Standard Error x 1.96 = .125 x 1.96 = .245
5. Calculate the Low and High end of the Confidence Intervals: Proportion +/- the Margin of Error = .75 +/- .245 = 50.5% to 99.8%

So the likely range of what the total proportion of users to complete the task is between 50% and 99.8%. The blue shaded rectangle in Figure 1 indicates this likely range.

Figure 1: 95% Confidence Interval for 7 completed tasks (Sample of 8)



With 7 out of 8 users you can make the statistically significant statement that: "The chances are less than 5% that the true population completion rate is less than 50% or above 99.8%." The completion rate will fall somewhere in the blue range of Figure 1 and more likely in the middle of the blue range than on the edges. In fact, the most likely point given this data is the first adjustment described above, which was 80%.

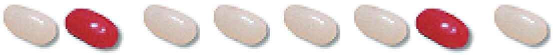
Perhaps you can see what all the rage is about at the APA. Just looking at Figure 1 you can see the location (centered around 75%) and the precision (in this case, not too precise since the range of probable values is quite large). Precision, not just location is an important component in making decisions about a real life event, in this case the number of users that will likely complete the task. You might see this range of almost 50 percentage points and think it not terribly informative given the probable range of values. While it is wide, after watching just 8 users you've narrowed the probable values from 100 percentage points to about 50!

That's what the statistics say, but how does that work? How can we make such fine-delineated statements with such a small sample? After all, if we just use these statistics on faith, how much better is this approach that using our professional intuition?

How it Works

Imagine there's a jar full of red and white jelly beans. You don't know what proportion are red and white, only that there are 239 jelly beans in the jar. Why 239? Because it's a weird number, just like the number of users in the population you'll probably be sampling from. Now you need to make the best guess as to the proportion of whites. Let's say those red jelly beans are task failures and the white ones are task successes.

Instead of taking wild guesses, you're allowed to take out just 8 jelly beans. Instead of imagining this scenario, I actually tried this with my wife. I went out and bought several bags of jelly beans, separated out the colors and put a combination of two colors in one jar and covered it (another reason there are 239 is that's all I could fit into the biggest jar I had). I asked her to draw 8 beans from the jar without looking. This is what she drew on her first try:



That's 6 whites and 2 reds, or the equivalent of watching 6 completed tasks and 2 failures. So, one guess is that there are 70% whites and 30% reds (using the adjustment of adding one success and one failure). So following the steps described above, I'll build a confidence interval around the sample:

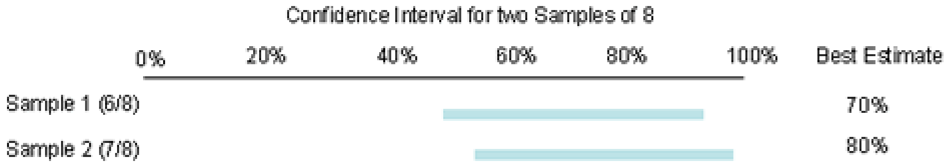
1. Adjust the observed proportion 6/8 to 8/12 = .6667
2. Standard Deviation of the Sample is: .6667*.3333 = .2222 (Variance), SQRT (.2222) = .4714
3. Standard Error of the Mean = .4714/ 3.364 = .1361
4. Margin of Error: Standard Error of the Mean times the critical value for a one-sided 95% confidence level. .1361*1.96 = .2667
5. Low & High ends of the CI = .6667+/- .2667 = 40% to 93.3%

With this sample of 8, my wife and I can be 95% confident that the jar contains between 40% and 93.3% white jelly beans. Since it didn't cost me anything to run another sample and I had more time (just a bit more because my wife was wondering why I was asking her to do this), I had her put the 8 jelly beans back into the jar, I shook it up, and I had her draw another 8. This time she drew one red and 7 white.

Computing the confidence interval on this sample provided the estimate as in the usability test above, or a 95% Confidence Interval of between 50% and 99%.

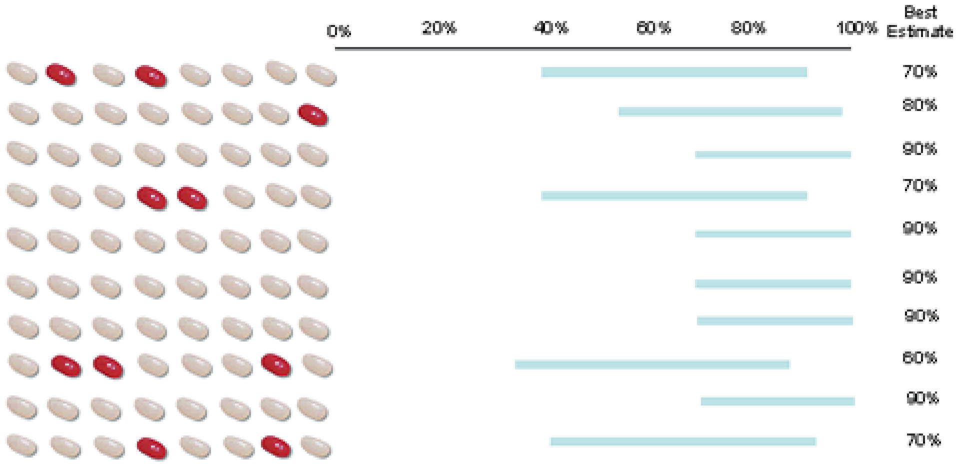
So I now have two estimates as to what the proportion of jelly beans are (70% and 80%) and two confidence intervals that most likely contain the actual proportion as shown in Figure 2 (40% to 93.3% and 50% to 99.8%).

Figure 2:



Just to make the most of this jar of beans, I then had my wife take 8 more samples, each time replacing the jelly beans before the next draw. The final tally is shown in Figure 3:

Figure 3: Confidence Intervals & Best Estimate for 10 Samples of 8



As it happens, there are 39 red jelly beans and 200 white jelly beans or 83.68% of the jar contains white jelly beans. I know because I put them in there. The best guess proportion and sample proportion were wrong every time (as expected). The confidence interval contained the true proportion in all 10 samples. We'd expect on average 95 out of 100 samples to contain the true proportion so it's not surprising to see that all 10 did.

This jelly bean exercise should summarize the strategy of using confidence intervals and completion rate estimates from small samples: the best estimate of the true proportion will almost always be wrong. The properly constructed confidence interval will almost always contain the true proportion but will force you to consider several likely values.

Were those Results Typical?

Finally, when you test with 8 users there's a tendency to think that you just happened to get a grou of good users, that the next crop could very well contain only 1 success and 7 failures. Of course thi is possible, but if you draw 7 completions and 1 failure from a population, the chance you will also draw 7 failures and one completion from the same population is very small. This would be similar to drawing seven reds from the jelly bean jar above. It's possible, but the chance this will happen is only .00206% or about twice in 100,000 draws.

Small samples are a fact of life for the Usability Practitioner, so it's no wonder we have strong opinions on what to report and what to conclude from our studies. On one hand, it's tempting to dismiss the results from small samples concluding that you cannot have statistically significant findings, or that figures cannot be accurately projected onto the larger population of users. On the other hand, it's also tempting to be overly confident in our tests and conclude that the proportion of observed will be nearly identical in the total population. The best approach lies somewhere in between: you can make statistically significant conclusions with small samples; however, extending the results to a larger population requires showing the boundaries of your findings through confidence intervals.

References

1. American Psychological Association (2001) Publication manual of the American Psychological Association (5th ed) Washington, DC: Author.

2. Lewis, J.R. & Sauro, J. (2006) "**When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates**" (PDF) in *Journal of Usability Studies* Issue 3, Vol. 1, May 2006, pp. 136-150.

3. 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.

4. Measuring Usability: Confidence Interval Around a Completion Rate Calculator **http://www.measuringusability.com/wald.htm**

5. Sauro, J & Lewis, J R (2005) "**Estimating Completion Rates from Small Samples using Binomial Confidence Intervals: Comparisons and Recommendations**" (PDF) in Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES 2005) Orlando, FL

Contact the Vo

Usability Professionals' Association
140 N. Bloomingdale Road
Bloomingdale, IL 60108-1017
Tel: +1.630.980.4997
Fax: +1.630.351.8490
UPA: **office@upassoc.org**