Technical Report

IBM

Small-Sample Evaluation of the Paulson-Takeuchi Approximation to the Exact Binomial Confidence Interval

James R. Lewis

November 29, 1994

TR54.872

Unclassified

Boca Raton, Florida

Unclassified

Small-Sample Evaluation of the Paulson-Takeuchi Approximation to the Exact Binomial Confidence Interval

James R. Lewis

TR54.872

ABSTRACT

Evaluation of the Paulson-Takeuchi approximation to the exact binomial confidence interval showed that, even for very small samples, the approximation was very accurate. The estimate of the end points was always within two percentage points of the exact interval's end points, and the approximate interval always contained the exact interval. In other words, the approximate interval erred only slightly and conservatively. These results indicate that it is reasonable to use this approximation to estimate binomial confidence intervals for small-sample usability studies.

Introduction

To be as cost-effective as possible, industrial usability studies must be efficient (Lewis, 1991; Lewis, 1994; Virzi, 1990; Virzi, 1992). A study conducted with a small sample is less costly than one with a large sample. However, a large sample provides more reliable measurements than a small sample. Efficiency is the result of balancing these two considerations.

It is possible to use binomial confidence intervals to judge if a usability defect rate exceeds a criterion (Lewis, 1991). In these situations, relatively small samples are often adequate to meet the goals of a usability evaluator. The purpose of this report is to evaluate the small-sample accuracy of a technique that approximates binomial confidence intervals.

The statistical term for a study to estimate a defect rate is a binomial experiment, because a given problem either will or will not occur during the study. For example, a participant either will or will not install an option correctly. The point estimate of the defect rate is the observed percentage of failures. However, the likelihood is very small that the point estimate from a study is exactly the same as the true percentage of failures, especially if the sample size is small (Walpole, 1976). To compensate for this, it is possible to calculate interval estimates that have a known likelihood of containing the true percentage. These binomial confidence intervals can describe the percentage of usability defects effectively. The report of a binomial confidence interval usually takes the form of:

> The observed percentage of (a particular usability defect) was PP percent. The lower limit of the CC-percent binomial confidence interval was XX percent and the upper limit was YY percent, where:
>
> > The percentage PP is the observed percentage of failures.
> >
> > The value of CC is the likelihood (confidence) that the interval will contain the true defect rate.
> >
> > XX is the lower limit of the interval.
> >
> > YY is the upper limit of the interval.

Steele and Torrie (1960) described the technique to determine exact binomial confidence intervals. Fujino (1980) summarized and evaluated several techniques to approximate binomial confidence intervals. He concluded that the Paulson-Takeuchi approximation was the most complicated, but was also the most accurate. However, the smallest sample size he evaluated was 25, a larger sample than that used in many usability studies. Before using the Paulson-Takeuchi approximation for small-sample usability studies, it is important to evaluate its accuracy given small samples. The relative complexity of the Paulson-Takeuchi approximation is not an issue because it is easy to program. (See Lewis, 1991 for a listing of a BASIC program to calculate the Paulson-Takeuchi approximation.)

Evaluation

Method

For 90-, 95-, and 99-percent binomial confidence intervals for sample sizes of two, five, and ten, I calculated both the exact (Steel and Torrie, 1960) and approximate (Fujino, 1980) binomial confidence intervals for all possible values of the number of defects.

Results

Table 1 contains the results of the evaluation. N is the sample size, x is the number of defects, PP is the observed percentage of failures and CC is the confidence level. For PPs that exceed 50 percent (or x that exceed the midpoint of N), subtract PP from 100. For the confidence interval of (100-PP), subtract each interval end point from 100 to obtain the appropriate confidence interval. For example, if a sample size of two contains two defective units, then the 90-percent confidence interval for the defect rate is 22 to 100 percent (100-78, 100-00).

Table 1 shows that the approximation was accurate within two percentage points for the estimation of interval end points for these sample sizes. The approximate interval always contained the exact interval.

Discussion

The results show that the approximate procedure is very accurate, even with small sample sizes. In all cases the approximate interval contains the exact interval. Therefore, the approximate interval errs only slightly and conservatively, even for these small samples.

These results indicate that it is reasonable to use this approximation to estimate binomial confidence intervals for small-sample usability studies.

Table 1. Exact and Approximate Binomial Confidence Intervals for Small Samples

N	x	PP	CC	Exact Interval	Approximate Interval
2	0	0	90	00 - 78	00 - 78
			95	00 - 84	00 - 85
			99	00 - 93	00 - 94
	1	50	90	02 - 98	02 - 98
			95	01 - 99	01 - 99
			99	00 - 100	00 - 100
5	0	0	90	00 - 45	00 - 45
			95	00 - 52	00 - 52
			99	00 - 66	00 - 66
	1	20	90	01 - 65	01 - 66
			95	00 - 71	00 - 72
			99	00 - 81	00 - 82
	2	40	90	08 - 80	08 - 81
			95	05 - 85	05 - 86
			99	02 - 92	02 - 92
10	0	0	90	00 - 26	00 - 26
			95	00 - 31	00 - 31
			99	00 - 41	00 - 41
	1	10	90	01 - 38	00 - 39
			95	00 - 45	00 - 45
			99	00 - 54	00 - 55
	2	20	90	04 - 49	04 - 51
			95	03 - 56	02 - 56
			99	01 - 65	01 - 65
	3	30	90	09 - 59	09 - 61
			95	07 - 65	07 - 65
			99	04 - 74	03 - 74
	4	40	90	16 - 68	15 - 70
			95	12 - 74	12 - 74
			99	08 - 81	07 - 81
	5	50	90	24 - 76	22 - 78
			95	19 - 81	19 - 81
			99	13 - 87	13 - 87

References

Fujino, Y. (1980). Approximate binomial confidence limits. Biometrika, 67, 677-681.

Lewis, J. R. (1991). Legitimate use of small samples in usability studies: Three examples (Tech. Report 54.595). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. Human Factors, 36, 368-378.

Steele, R. G. D. and Torrie, J. H. (1960). Principles and procedures of statistics. New York, NY: McGraw-Hill.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. In Proceedings of the Human Factors Society 34th Annual Meeting (pp. 291-294). Orlando, FL: Human Factors Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? Human Factors, 34, 457-468.

Walpole, R. E. (1976). Elementary statistical concepts. New York, NY: Macmillan.