BINOMIAL CONFIDENCE INTERVALS FOR SMALL SAMPLE USABILITY STUDIES

James R. Lewis
IBM Human Factors Group
P. O. Box 1328
Boca Raton, FL 33429-1328
Tel: +1 (407) 443-1066
Fax: +1 (407) 443-2778
E-mail: JIMLEWIS@US.IBM.COM

usability studies, sample size, usability measurement

Efficiency is an important consideration in the design of industrial usability studies. One way to reduce the cost of a usability study is to reduce its sample size. Small samples are not always appropriate, but in this paper I will describe a way to use binomial confidence intervals to determine rapidly if a usability defect rate exceeds a criterion. In these situations, relatively small samples are often adequate to meet the goals of a usability evaluator. I will also discuss the risks of using small samples in these situations.

INTRODUCTION

To be as cost-effective as possible, industrial usability studies must be efficient (Lewis, 1994; Lewis, 1991; Virzi, 1992). A study conducted with a small sample is less costly than one with a large sample. Small samples are not always appropriate, but in this paper I describe a way to use binomial confidence intervals to determine rapidly if a usability defect rate exceeds a criterion. In these situations, relatively small samples are often adequate to meet the goals of a usability evaluator. I also discuss the risks of using small samples in these situations.

A problem that occurs during a usability study may be indicative of a defect in the design of the system (Norman, 1983). In usability studies, a usability defect rate for a specific problem is the number of participants who experience the problem divided by the total number of participants. Usability defect rates can be proportions or percentages. 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, often with a small sample. The report of a binomial confidence interval usually takes the form of:

Advances in Applied Ergonomics (ed., Ahmet F. Ozok & Gavriel Salvendy)
West Lafayette, IN: USA Publishing

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 and 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. The smallest sample size he evaluated was 25. I compared exact 90-, 95-, and 99-percent binomial confidence intervals with the estimated intervals for sample sizes of two, five, and ten. 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. (Appendix A contains the results of this evaluation. Appendix B contains the BASIC code for a program to calculate the approximation.)

"Ideally, we prefer a short interval with a high degree of confidence." (Walpole, 1976, p. 123) The factors that affect the width of a confidence interval are very similar to those that affect sample size estimates (Kraemer and Thiemann, 1987; Walpole, 1976). All other factors being equal, a confidence interval from a large sample will be shorter than one from a small sample. A confidence interval with a high degree of confidence will be wider than one with a lower degree of confidence. Unlike confidence intervals that use the normal distribution, binomial confidence intervals will usually not be symmetrical, especially if the observed percentage is close to 0 or 100 percent. Table 1 illustrates the effects of these factors.

TABLE 1. Effect of Varying Sample Size and Degree of Confidence on Interval Width

		Confidence Interval			
Sample Size	Degree of Confidence	Lower Limit	Obs. Value	Upper Limit	Interval Width
8	90	40	75	96	56
80	90	66	75	83	17
8	95	35	75	97	62
80	95	64	75	84	20

For a 95-percent binomial confidence interval, it is 95-percent likely that the interval contains the true percentage of defects. If the true percentage is high, then the lower limit of a 95-percent binomial confidence interval will be high, even with a small sample. If the lower limit of the confidence interval is unacceptable, then it is legitimate to conclude that the defect rate is unacceptable, regardless of the sample size. The following examples illustrate this technique.

EXAMPLE 1: ERRATA SHEET EFFECTIVENESS

King, Lee and Lewis (1990) studied the effectiveness of an errata sheet placed on top of the documentation ship group of a product. The primary dependent variable was whether a participant who unpacked the system would use the errata sheet. We designed the errata sheet to attract the user's attention, and placed large (24-point) print at the top of the page that stated "DO THIS FIRST!" Six out of eight participants ignored the errata sheet. We defined ignoring the errata sheet as a usability defect, so the observed defect rate was 75 percent. A 95-percent binomial confidence interval for this defect rate ranged from 35-percent to 97-percent failures. Even with this small sample, we could reliably predict that the minimum failure rate for this type of errata sheet would be 35-percent. We concluded that the use of an errata sheet in this situation will generally be an unacceptable strategy.

EXAMPLE 2: EVALUATION OF GRAPHIC SYMBOLS FOR *PHONE* AND *LINE*

Lewis and Pallo (1991) studied the effectiveness of graphic symbols for phone and phone line connection for attaching telephony equipment to a computer. Computer-naive participants attached a telephone to a computer with only the graphic symbols to guide the installation. Nine of eleven installations (82 percent) were incorrect. The 95-percent confidence interval for this percentage ranged from 48 percent to 98 percent -- unacceptably high. We were 95-percent confident that, unless the product developers provided additional information to users, the failure rate for installation would be at least 48 percent. After we provided a wordless setup sheet, 10 of 10 participants correctly installed the telephony equipment. (The 95-percent confidence interval for the second phase of the study ranged from 0- to 31-percent incorrect installations.)

DISCUSSION

For studies that will measure the percentage of usability defects, I recommend a strategy similar to that for studies in which a mean is compared to a criterion (Lewis, 1991). Study a small sample of participants and record the percentage of usability defects (such as incorrect installations or failures to complete tasks). Calculate the binomial confidence interval. Report the observed defect percentage and the lower limit of the binomial confidence interval to the product developer. Compare the lower limit of the confidence interval to the maximum acceptable defect rate (the criterion). If there is no criterion, ask the product developer if the lower limit is an acceptable defect rate because it is 95-percent likely that the true defect rate will be at least as high as the lower limit of the interval. When the defect rate is high, this can be a very convincing argument to redesign the product or system.

This method can rapidly demonstrate with a small sample that a usability defect is unacceptably high if the criterion is low and the true defect rate is high. Although the confidence interval will be wide (62 percentage points in the errata sheet example and 50 percentage points in the graphic symbols example), the lower limit of the interval may be clearly unacceptable. When the true defect rate is low or the criterion is low, this procedure may not work without a large sample size. The decision to continue sampling or to stop the study should be determined by a reasonable business case that balances the cost of continued data collection against the potential cost of allowing defects to go uncorrected.

This procedure cannot be used with a small sample to prove that a success rate is acceptably high. With small samples, even if the observed defect percentage is 0 or close to 0 percent, the interval will be wide, so it will include defect percentages that are unacceptable (as in the graphic symbols example). Therefore, it is relatively easy to prove (requires a small sample) that a product is unacceptable, but it is difficult to prove (requires a large sample) that a product is acceptable.

REFERENCES

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

KING, L., LEE, R., and LEWIS, J. R. (1990). Errata sheet effectiveness: A risk assessment (IBM Tech. Report 54.567). (IBM Corp., Boca Raton, FL).

KRAEMER, H. C. and THIEMANN, S. (1987). How many subjects? Statistical power analysis in research. (Sage, Newbury Park, CA).

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

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

LEWIS, J. R. and PALLO, S. (1991). Evaluation of graphic symbols for phone and line (IBM Tech. Report 54.572). (IBM Corp., Boca Raton, FL).

NORMAN, D. (1983). Design rules based on analyses of human error. Communications of the ACM, vol. 4, 254-258.

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

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

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

APPENDIX A. EVALUATION OF THE PAULSON-TAKEUCHI APPROXIMATION FOR SMALL SAMPLE SIZES

I evaluated the Paulson-Takeuchi approximation for 90-, 95-, and 99-percent binomial confidence intervals and for sample sizes of two, five, and ten. Appendix Table 1 (on the following page) contains the results. N is the sample size, x is the number of defects, PP is the observed percentage of failures and CC is the confidence level. I used the procedure described in Steele and Torrie (1960) to calculate the exact interval, and I used the program from Lewis (1991) to calculate the approximate interval. For PPs that exceed 50 percent, 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 - 100 (100-78, 100-00). The table shows that the approximate procedure is very accurate, with approximate end points within two percentage points of the exact end points in all cases. Also, in all cases the approximate interval contains the exact interval. Therefore, the approximate interval errs only slightly and always conservatively, even for these small samples.

App. 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

APPENDIX B. PROGRAM FOR APPROXIMATE BINOMIAL CONFIDENCE INTERVALS

```
10 ' Approximate binomial confidence limits:  Level 2.0, 8/29/90, J. R. Lewis
30 ' This BASIC program uses the Paulson-Takeuchi approximation described in
40 ' Fujino, Y.  (1980).  Approximate binomial confidence limits.
50 ' Biometrika, Volume 67, Page 679, to calculate approximate 2-sided
60 ' 90-, 95-, and 99-percent binomial confidence limits.
70 '
80 Z(1)=1.645      ' Z value for 90-percent confidence limits.
90 Z(2)=1.96       ' Z value for 95-percent confidence limits.
100 Z(3)=2.575     ' Z value for 99-percent confidence limits.
110 CLS
120 PRINT "Approximate 2-sided binomial confidence intervals (90%, 95%, 99%)"
130 PRINT
140 PRINT "Enter the observed number of occurrences (x) and the number of"
150 PRINT "opportunities for occurrence (n, the sample size) separated "
160 PRINT "by a comma.  The results displayed are proportions.  "
170 PRINT
180 PRINT "(Move the decimal place over two positions (i.e., multiply by 100)
190 PRINT "to convert to percentages.)  "
200 PRINT
210 PRINT
220 INPUT "Enter x,n: ",X,N:PRINT:PRINT
230 FOR CNT=1 TO 3
240 LET U=Z(CNT)
250 XPRIME=X:IF XPRIME=N THEN P2=1:IF P2 = 1 THEN GOTO 280
260 GOSUB 470
270 P2=PTPROB
280 ' Get lower limit pprime by replacing x by n-x and
290 ' carry out calculation as before ; then pprime = 1 - ptprob.
300 IF X=0 THEN PPRIME=0:IF PPRIME=0 THEN GOTO 340
310 XPRIME=N-X
320 GOSUB 470
330 PPRIME=1-PTPROB
340 PRINT
350 IF CNT=1 THEN PRINT "90% confidence interval: ";
360 IF CNT=2 THEN PRINT "95% confidence interval: ";
370 IF CNT=3 THEN PRINT "99% confidence interval: ";
380 PRINT USING "#.###";PPRIME;:PRINT " - ";:PRINT USING "#.###";X/N;
390 PRINT " - ";:PRINT USING "#.###";P2
400 NEXT CNT
410 PRINT:PRINT:PRINT "(Use Print Screen if hardcopy is required.)"
420 PRINT:PRINT:INPUT "Do you want to do more calculations? (Y/N)",A$
430 IF LEFT$(A$,1)="y" THEN LET A$="Y"
440 IF LEFT$(A$,1)="Y" THEN 110
450 SYSTEM
460 ' Binomial confidence interval subroutine
470 A=1/(9*(XPRIME+1)):APRM=1-A
480 B=1/(9*(N-XPRIME)):BPRM=1-B
490 F=((APRM*BPRM+U*SQR(APRM^2*B+A*BPRM^2-A*B*U^2))/(BPRM^2-B*U^2))^3
500 PTPROB=(XPRIME+1)*F/((N-XPRIME)+(XPRIME+1)*F)
510 RETURN
```