Average Task Times in Usability Tests: What to Report?

Jeff Sauro
Oracle Corporation
1 Technology Way, Denver, CO 80237
jeff@measuringusability.com

James R. Lewis
IBM Software Group
8051 Congress Ave, Suite 2227
Boca Raton, FL 33487
jimlewis@us.ibm.com

ABSTRACT
The distribution of task time data in usability studies is positively skewed. Practitioners who are aware of this positive skew tend to report the sample median. Monte Carlo simulations using data from 61 large-sample usability tasks showed that the sample median is a biased estimate of the population median. Using the geometric mean to estimate the center of the population will, on average, have 13% less error and 22% less bias than the sample median. Other estimates of the population center (trimmed, harmonic and Winsorized means) had worse performance than the sample median.

Author Keywords
Usability evaluation, task times, median, geometric mean, Monte Carlo simulations

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): User Interfaces–Evaluation/Methodology.

General Terms
Measurement, Human Factors, Experimentation

INTRODUCTION
In usability tests, task times are an often reported usability metric [7]. Small sample point estimates will be inaccurate and should be reported with confidence intervals [8]. When communicating usability test results, however, it is common to provide a typical or average value. Task times in usability tests tend to be positively skewed (longer right tails). This skew comes from users who take an unusually long time to complete the task, for example, users who have experienced more usability problems than other users. When data are roughly symmetric, the mean and median are the same, so either can serve as an unbiased statistic of central tendency. With a positive skew, however, the arithmetic mean becomes a relatively poor indicator of the center of a distribution due to the influence of the unusually long task times. Specifically, these long task times pull the mean to the right, so it tends to be greater than the middle task time.

Consistent with basic statistical advice [10], practitioners who are aware of this positive skew tend to report the median. The median has the advantage that it is not overly influenced by extreme values. Therefore, by definition, the median of a population will be the center value – what most practitioners are trying to get at when they report an "average." For example, the task times of 100, 101,102,103, and 104 have a mean and median of 102. Adding an additional task time of 200 skews the distribution, making the mean 118.33 and the median 102.5.

The strength of the median in resisting the influence of extreme values is also its weakness. The median doesn't use all the information available in a sample. For odd samples, the median is the central value; for even samples, it's the average of the two central values. Consequently, the medians of samples drawn from a continuous distribution are more variable than their means [2].

Furthermore, Cordes (1993) [3] demonstrated that the sample median was a biased estimate of the population median for usability test task times. Using two usability tasks and simulating large sample data with a Gamma Distribution tuned to simulate the typical skewness of task times, Cordes showed that small-sample (n = 5) medians overestimated the population median by as much as 10%. On the other hand, the sample mean was an unbiased estimator of the population mean in these skewed distributions. The Monte Carlo simulations showed the sample mean estimated the population mean to within 1% for all sample sizes. The conclusion was that the mean did a better job than the median at estimating their respective population values. In fact, it is a tenet of the central limit theorem that when repeatedly sampling from a population, the mean of the samples will be the population mean. The median does not have this property. The problem with the sample median overestimating the population median applies to many positively skewed distributions. This bias can lead to incorrect conclusions when comparing products or versions, especially if the sample sizes are unequal [5].

Although the sample mean may be a better estimate in the sense that it provides an unbiased estimate of the population mean, we know that due to the skewness of the distributions of usability task times, the population mean will be larger than the center value of the distribution (the population median). The "right" measure of central tendency depends on the research question, but for many usability situations, practitioners want to estimate (and perhaps compare) the

centers of the distributions. However, is there any method available to usability practitioners that can do a better job of estimating the population median than the sample mean or the median?

METHOD

In this analysis we assess a few of the more popular methods for the estimation of central tendency to determine which ones best estimate the center of the population of task time data in usability tests. Cordes (1993)[3] examined just the sample mean and median, but there are literally hundreds of ways to estimate the center of a distribution (see [1]). Such point estimates can take either a closed form (like the mean and median) or use an iterative algorithm to converge on an estimate.

Ease of computation is important for practitioners, so we excluded methods that require an iterative solution, focusing instead on the arithmetic mean, median, harmonic, geometric mean, two trimmed means and a Winsorized mean. We explain the less familiar point estimates below using, as an example, a sample of task times: 84, 85, 86, 103, 111, 122, 180, 183, 235, 278 seconds. The arithmetic mean for these times is 146.7 and the median is 116.5 (the mean of 111 and 122).

The Geometric Mean

To find the geometric mean, convert the raw times using a log-transformation, find the mean of the transformed data, then transform back to the original scale by exponentiating. Using the natural logarithm to transform the data, we get: 4.43, 4.44, 4.45, 4.63, 4.71, 4.80, 5.19, 5.21, 5.46, 5.63. The arithmetic mean of these values is 4.8965. Exponentiating the average back into the original scale, we get a geometric mean of 133.8 seconds. In Excel, use the LN function to transform to the natural log, and the EXP function to transform back.

Trim-Top Mean

The first of the trimmed means we investigated is what we called the trim-top mean, computed by dropping the longest 10% of times, then calculating the mean (see [6], for a recommendation of this method). We used a round-up strategy whereby at sample sizes between 3 and 15 we dropped the longest task time and between sample sizes of 16 and 25 we dropped the longest two task times. For the example sample times, drop the longest time of 278 and compute the arithmetic mean from the remaining times to get a trim-top mean of 132.1.

Trim-All Mean

The typical trimmed mean drops the top and bottom 10% of task times, then calculates the mean. We used a round-up strategy whereby at sample sizes between 4 and 15 we dropped the longest and shortest task times, and between sample sizes of 16 and 25 we dropped the two longest and shortest task times. For the example sample times, drop the 278 and 84 and compute the arithmetic mean from the remaining times to get a trim-top mean of 138.1.

Winsorized Mean

Winsorizing uses the same procedure as the trimmed means, but instead of dropping the extreme value(s) they are replaced with the less extreme adjacent values [9]. In the sample data, replace 278 with 235 and 84 with 85, then compute the arithmetic mean to get a Winsorized mean of 142.5.

Harmonic Mean

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. It is found by dividing the sample size by the sum of the reciprocal of each value. It is 123.21 for this data.

We used a large database of usability data gathered from an earlier analysis [7] to determine which sample average best predicts the population median using actual usability data. Tasks included in the analysis were those that contained at least 35 users who successfully completed the task. We used these tasks and their distributions of task times to estimate population medians. In total this provided 61 tasks from 7 distinct usability studies. There was a mix of attended and unattended usability tests. Figures 1 and 2 show data from two example tasks, illustrating the characteristic positive skew.
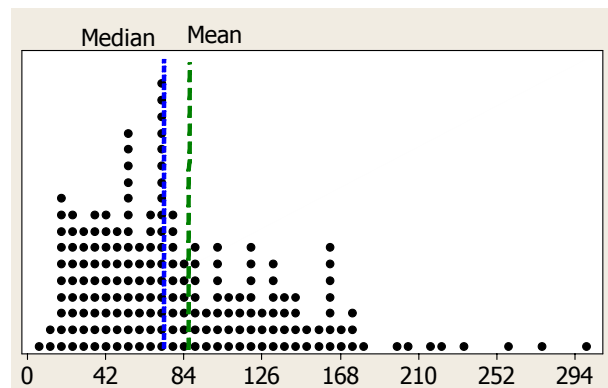

Figure 1. Sample task from an unattended usability test with 192 users who completed the task. The median is 71 and the arithmetic mean is 84.
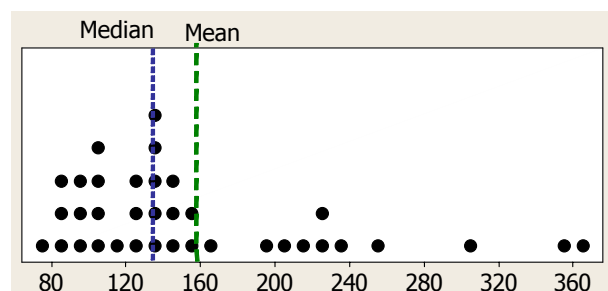

Figure 2. Sample task from a lab-based usability test with 36 users who completed the task. The median is 135 and the arithmetic mean is 158.

We used a Monte Carlo re-sampling procedure to estimate the median of the full sample for sample sizes between 2 and 25, taking 1000 random samples for each combination

of task and sample size, and for each, computing the various point estimates (mean, median, trim-top, trim-all, harmonic and Winsorized) and their accuracy as assessed by the dependent measures of normalized Root Mean Square Error and Bias.

To compute a Root Mean Square Error (RMS), subtract the median from the point estimate, square this difference, then take the square root (to eliminate negative values). Divide this result by the population median and multiply by 100 to get a percentage (normalized based on the magnitude of the population median). The overall RMS for each point estimate was the average of the RMS for each of the 61,000 values for each sample size between 2 and 25. The more accurate an estimate is, the lower will be its RMS, with a perfectly accurate estimate having an RMS of zero.

To compute Bias, subtract the population median from the point estimate and divide that difference by the population median and multiply by 100 to get a percentage (normalized based on the magnitude of the population median). If a point estimate exhibits little bias, we'd expect this value to be close to zero. Biased point estimates would have average errors noticeably above or below zero.

RESULTS

The results of the Monte Carlo simulation appear in the figures and tables below. Figure 3 and Table 1 show the average RMS error for each point estimate relative to the population median for sample sizes between 2 and 25.

The arithmetic, Winsorized, harmonic and trim-all means performed the poorest in predicting the population median. Table 1 shows the RMS error for select sample sizes from Figure 3. Table 2 shows the relative advantage or disadvantage in using an estimate other than the sample median to estimate the population median.
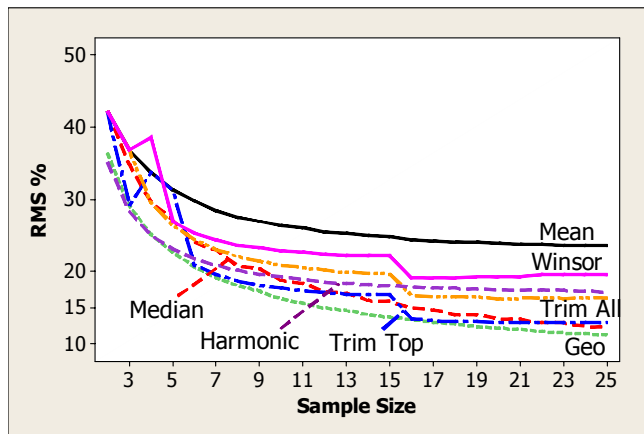

Figure 3: Percent Root Mean Square Error for the sample averages compared to the population median for sample sizes between 2 and 25 for all 61 usability tasks.

The geometric and trim-top means had, on average, 12.7% and 3% less error than the median. The trim-all, Winsorized, arithmetic and harmonic means had, on

average, 14.3%, 29.5%, 51.1% and 10.9% more error than the sample median in estimating the population median.

n	Mean	Med	Geo	Trim Top	Trim All	Winsor	Harm
5	31.3	27.1	22.6	31.3	26.4	26.9	23.1
10	26.3	18.7	16.1	17.6	20.8	22.8	19.2
15	24.8	15.8	13.6	16.7	19.6	22.1	18.0
20	23.8	13.4	12.1	12.8	16.2	19.2	17.4
25	23.5	12.2	11.1	12.8	16.3	19.6	17.1

Table 1: Average RMS error percentage relative to population median by point estimate from 1000 samples from 61 tasks for select sample sizes.

n	Mean	Geo	Trim Top	Trim All	Winsor	Harm
5	15.5	-16.6	15.5	-2.6	-0.7	-14.8
10	40.6	-13.9	-5.9	11.2	21.9	2.7
15	57.0	-13.9	5.7	24.1	39.9	13.9
20	77.6	-9.7	-4.5	20.9	43.3	29.9
25	92.6	-9.0	4.9	33.6	60.7	40..2
Avg	51.1	-12.7	-3.0	14.3	29.5	10.9

Table 2: Percent More/Less error relative to the sample median in estimating the population median for select sample sizes. Negative values show an estimate with less error than the sample median.

As expected, the average RMS error is rather high for sample sizes less than 5. At this sample size, even the best point estimate will, on average, be off by at least 22.6%.

At these low sample sizes many of the point estimates have RMS errors close to the arithmetic mean RMS error because trimming or Winsorizing two values is not possible for a sample of three. As the sample size increases, the amount of average RMS decreases and the point estimates diverge.

In examining the bias, Figure 4 and Tables 3 and 4 show the average bias by point estimate. In general, the point estimates have a positive bias, overestimating the population median (with the exception of the harmonic mean). The geometric mean has, on average, 22.7% less bias than the sample median. The trim-top, trim-all, Winsorized and arithmetic means have, on average, 81.4%, 298.4%, 412.9% and 584.5% more bias than the sample median in estimating the population median across sample sizes between 2 and 25.

Table 3 shows that the sample median has a bias of 7% and 2.4% for sample sizes of 10 and 20. For these same sample sizes, Cordes (1993)[3] found the median had an average bias of 7.6% and 1.9% for the two task times he examined – reassuringly similar to our results.
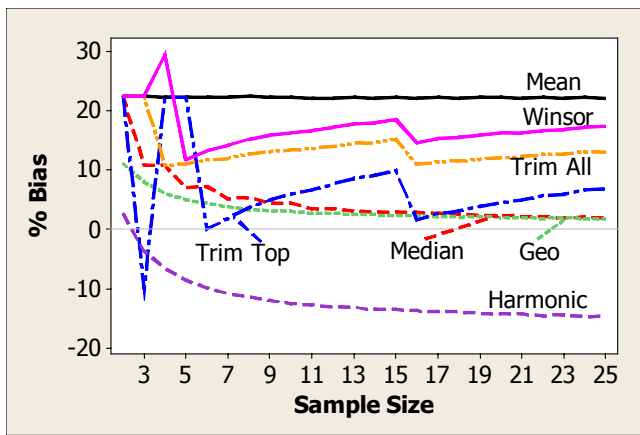
Figure 4: Percent Bias as a percent of the population median for the sample averages for sample sizes between 2 and 25 for all 61 usability tasks.

n	Mean	Med	Geo	Trim Top	Trim All	Winsor	Harm
5	22.2	7.0	5.0	22.2	11.0	11.8	-8.6
10	22.2	4.4	3.0	5.9	13.4	16.3	-12.6
15	22.3	2.9	2.3	9.9	15.2	18.5	-13.6
20	22.2	2.4	2.0	4.5	12.1	16.2	-14.3
25	22.1	1.8	1.8	6.9	13.1	17.3	-14.7

Table 3: Average Bias in estimating the population median by point estimate from 1000 samples from 61 tasks for select sample size. The harmonic mean consistently underestimates the population median.

n	Mean	Geo	Trim Top	Trim All	Winsor	Harm
5	217	-29	217	57	69	23
10	405	-32	34	205	271	186
15	669	-21	241	424	538	369
20	825	-17	88	404	575	496
25	1128	0	283	628	861	717
Avg	585	-23	81	298	413	312

Table 4: Percent more/less bias relative to the sample median in estimating the population median for select sample sizes. Negative values show an estimate with less bias than the sample median (i.e. the geometric mean).

DISCUSSION

For large sample sizes (n >25) the sample median does a good job of estimating the population median. For small sample sizes, these results suggest the sample median will not be the best estimate of the population median. As in [3] the sample median consistently overestimated the population median, especially for smaller samples. The arithmetic mean, trim-all mean, harmonic mean and Winsorized mean performed much more poorly than the sample median. The trim-top mean had less error than the sample median for some sample sizes, but more error for others. The geometric mean, with consistently less error and bias than the median or the trim-top mean, showed the best overall performance.

Furthermore, statisticians in the past have recommended the use of the logarithmic transformation for this type of skewed data to reduce nonnormality and heteroscedasticity and, consequently, to get better estimates of statistical significance when using analysis of variance [4]. Using and reporting the geometric mean is consistent with this practice. Free web-calculators are available to perform the transformation and calculations (e.g. http://www.measuringusability.com/time_intervals.php).

CONCLUSION

When providing an estimate of the average task time for small sample studies (n<25), the geometric mean is the best estimate of the center of the population (the median). The sample median will tend to over-estimate the population median by as much as 7% and will be less accurate than the geometric mean. The geometric mean is slightly more laborious to compute and less well-known to the current community of usability practitioners than the median, but is advantageous enough that we recommend its use in place of (or, at least, in addition to) the sample median when reporting the central tendency of task times.

REFERENCES

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Tukey, J. W., & Huber, P. J. *Robust estimates of location: Survey and advances*. Princeton, NJ: (1972). Princeton University Press.

2. Blalock, H. M. *Social statistics*. (1972). New York, NY: McGraw-Hill.

3. Cordes, R The effects of running fewer subjects on time-on-task measures. *International Journal of Human-Computer Interaction*, *5(4)*, (1993) 393-403.

4. Myers, J. L, Fundamentals of experimental design. . (1979) Boston, MA: Allyn and Bacon.

5. Miller, J. A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, (1988) 539-543.

6. Nielsen, J. *Quantitative Studies, How Many Users to Test*? (2006) Last viewed September 13, 2009, www.useit.com/alertbox/quantitative_testing.html.

7. Sauro, J. & Lewis J. R Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability. In *Proc CHI 2009* (pp. 1609-1618).

8. Sauro, J *Graph and Calculator for Confidence Intervals for Task Times* Last viewed Sept. 10, 2009, http://www.measuringusability.com/time_intervals.php

9. Tukey, J. W The future of data analysis. *Annals of Mathematical Statistics*, *33(1)*, (1962) pp 1-67.

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