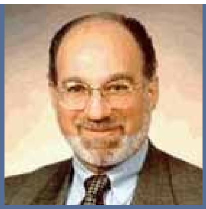
The Bizcomps Database, the Size Effect Phenomenon and a Necessary and Sufficient Sample Size: A Response to Toby Tatum

Mark G. Filler, CPA/ABV, CBA, AM, CVA



Once again you have published two thought provoking articles by Toby Tatum that inspire and require an equally considered response. His article in the 2012 First Quarter issue entitled *Revisiting the Size Effect Phenomenon Among Small Businesses* followed by *Some Observations on Statistical Analysis and Sample Size* published in the 2012 Second Quarter issue contain two ideas that could mislead your readers if they are unaware of certain statistical concepts that further explore and explain the ideas Mr. Tatum puts forth.

The first idea that needs attention is the one contained in the first article's title. In that article, it appears from Figure 4 that there is an upward trend in the Selling Price to Seller's Discretionary Earnings (SP/SDE) ratio relative to firm size as measured by annual revenue, indicating that as a firm gets larger so does its SP/SDE ratio. However, this appearance is illusory, not only for the Bizcomps transaction database as a whole, but also for a selected subset of that database, Standard Industrial Classification (SIC) Code Number 782 – Lawn Maintenance. The reason for the illusory upward trend is the notion of variance, or variation or dispersion about the mean, and how it affects confidence intervals about the mean. For the Bizcomps transaction database, after removing 606 transactions that either had a negative SDE or neither revenue, SDE or sales price, the remaining 12,767 transactions had a coefficient of variation of 171.5% derived by dividing the SP/SDE standard deviation of 4.11 by the average SP/SDE of 2.39. The individual 16 size categories had coefficients of variation ranging from a minimum of 50.9% to a maximum of 270.7%. For the 262 observations in the lawn maintenance SIC Code No. the coefficient of variation was 49.9%, and the minimum and maximum for the 10 size deciles I created were 32.1% and 83.1%, respectively. These coefficients are quite large as my experience has been that after I have removed outliers in excess of 2.5 standard deviations from a sample Bizcomps database, the coefficient of variation is typically about 25%.

With standard deviations this large, the supposed differences in SP/SDE ratios could be obtained merely by chance. That is, because they are so wide, the confidence intervals for each size category, which are a function of standard deviation and sample size, will overlap all the other size categories, signifying that no statistically significant difference exists among the average SP/SDE ratios. For example, when a political poll indicates that candidate X has a 48% chance of winning with a ±4% margin of error, the lower bound of the 95% confidence interval is 44%. If candidate Y has 42% chance of winning with a ±3% margin of error, the upper bound of that 95% confidence interval is 45%. Since the two confidence intervals overlap we can say that there is no statistical difference between the two candidates' chances of winning – it is a statistical dead heat.

Performing a pairwise mean difference test on the 16 size categories in the complete Bizcomps database and on the 10 size categories for the lawn maintenance subset indicates that is exactly the case, except for the largest size categories, for both databases. That is, for all but the largest size categories, there is no statistically significant difference among all the other size categories of the SP/SDE ratio – they are essentially of equal value. This result is not obviated simply because Mr. Tatum correctly calculated his average SP/SDE ratios using the weighted harmonic mean, and the paired mean difference test uses arithmetic means and standard deviations. The obvious variation in the data is not avoided by the use of a different measure of central tendency such as the weighted harmonic mean.

With this information in hand, it is not necessary to follow Mr. Tatum's suggestion that you select the size category that fits your subject company and pick the appropriate SP/SDE ratio associated with that size category. Since most SIC Codes have a lot fewer transactions than that of lawn maintenance, you cannot afford to give up 9/10ths or 15/16ths of your transactions. One suggestion would be to delete the largest size category and then use the weighted harmonic mean of all the remaining transactions in your selected database to come up with your SP/SDE ratio, because, as we have seen, there is no real size effect in the Bizcomps database. Another suggestion would be to remove the outliers from the database, i.e., those transactions that have multiples greater than 2.5 or 3 times the database average, as companies that sell for extreme multiples are dissimilar and irrelevant to your subject company. This procedure should restore the largest size category to a comparable position vis-a-vis the other size categories.

The second idea offered in the articles that needs further elucidation is that of the need to use a sample size of approximately 30. Mr. Tatum quotes a well-known textbook to support this assertion, but he doesn't tell us either the context of the quote, nor its meaning. The quote concerns itself with the Central Limit Theorem, which says that if the size of repeated samples taken from a population is large enough, the distribution of the averages of those samples will be approximately normal, regardless of the distribution of the population. For example, the distribution of the roll of one die is uniform, as each of the 6 sides has an equal chance of occurring each time we roll the die. But if we were to roll the die 30 times, and then compute the average of those 30 rolls, and then repeat the process 100 times, the sampling distribution of the 100 averages would be approximately normal. This is true even though the distribution of each of the 30 rolls, the samples themselves, is uniform, and remains so regardless of sample size. It's the sample averages that are approximately normal.

The average of the 100 sample averages will be equal to the population average, and the standard deviation of the 100 sample averages will be equal to the population standard deviation divided by the square root of the sample size. What is most amazing is that each sample average, if the sample size is large enough, will approximate the sampling distribution average, and the sample standard error, computed by dividing the sample standard deviation by the square root of the sample size will approximate the standard deviation of the sampling distribution. What this means is that we have no need to draw more than one adequate sample to compute confidence intervals about a mean or do tests of hypothesis.

An obvious question is: how large should the sample size be in order for the Central Limit Theorem to hold? A rule of thumb in place for years is that the Central Limit Theorem will hold when the sample size is greater than 30. However, one should not apply the rule blindly. If the population is heavily skewed, the sampling distribution of the average will still be skewed even if the sample size is greater than 30. On the other hand, if the population is symmetric, the Central Limit Theorem holds for sample sizes less than 30. Before we take a look and see if the sample size of 30 rule of thumb applies to either the total Bizcomps database or to the lawn maintenance SIC Code subset, we need to review Mr. Tatum's later article in which he defines the rule of thumb as a natural consequence of increasing sample size.

In the later article Mr. Tatum shows us in Figure 1 that the critical value of t stabilizes around the number 2 as the sample size approaches 30. He also demonstrates in Figure 3 that the absolute size of the errors about the mean begin to stabilize when the sample size reaches 30. But these results are both a mere coincidence and a consequence of the Central Limit Theorem and the Law of Large Numbers – as the sample size and the number of trials increase we expect to see the sample average equate to the population average, something that happens rather rapidly as sample size increases. But equality with the population average is not the test. Rather, the test is whether or not the sampling distribution is normal or, at a minimum, near-bell shaped. Now we can explore that notion and the rule of thumb as it pertains to the Bizcomps database.

Let's begin with the Bizcomps transaction database as a whole. But first we need to create some metrics to see if the sampling distribution of the mean is approximately normal so as to conform to the Central Limit Theorem. Since confidence intervals and hypothesis tests are robust as to non-normality, we do not need to attain perfectly normal distributions as measured by skewness and kurtosis values of zero. Therefore, if skewness and kurtosis lie between -.5 and +.5, we will assume the distribution is approximately normal. If skewness and kurtosis lie between -1 and +1 we will assume a near-bell-shaped distribution, the minimum requirement needed for confidence intervals and hypothesis testing. Any distribution with skewness and kurtosis beyond -1 and +1 will be rejected as non-conforming to the requirement of normality. Returning to the 12,767 transactions in the complete Bizcomps database referenced above, we note that this population's SP/SDE ratio is highly kurtic and skewed, with metrics of kurtosis and skewness of 643 and 21, respectively. With so much skewness and kurtosis, a sample size of 200 fails to produce even a near-bell-shaped sampling distribution of the mean with 1,000 samples, never mind a sample size of 30. Not until the sample size reached 300 did skewness and kurtosis become less than 1.

I next removed 73 transactions to bring the SP/SDE ratio down to less than 10 which reduced the total count to 12,694 transactions. Kurtosis and skewness remained very high at 19.1 and 3.4, respectively. It took a sample size of 84 to get 1,000 samples to attain approximate normality, while a sample size of 15 produced a near-bell-shaped distribution for 1,000 samples.

Finally, I removed another 259 transactions to bring the SP/SDE ratio down to less than 6.45 which further reduced the total count to 12,435 transactions. Kurtosis and skewness remained high at 2.0 and 1.2, respectively. In this case it took only a sample size of 8 to get 1,000 sample averages to attain approximate normality.

Turning now to the lawn maintenance SIC Code subset, I began with 266 transactions but had to remove 4 transactions that had no SDE. The 262 remaining transactions had kurtosis and skewness values of 7.5 and 2.1, respectively, indicating that the database is neither approximately normal nor near-bell-shaped. It took a sample size of 83 to get 1,000 sample averages to be approximately normal and a sample size of 54 to get near-bell-shaped results.

I then removed 9 transactions that were more than 3 standard deviations from the mean. This brought the total count down to 253 transactions with kurtosis and skewness values of .12 and .57, respectively, With kurtosis and skewness values this low we should not be surprised when a sample size of 5 produces approximately normal results for 1,000 sample averages.

So much for the rule of thumb that says one needs a sample size of at least 30 for the Central Limit Theorem to hold. Such a count is neither necessary nor sufficient. Like most things in life, it all depends – especially on how symmetric the population is. If you clean up your transaction database sample by removing the outliers and obvious errors, thereby getting your kurtosis and skewness values to be between -1 and +1, you can obtain good results with sample sizes of less than 10, as Ray Miles has been preaching for years.

Mark G. Filler, CPA/ABV, CBA, AM, CVA is President of and heads Filler & Associates' valuation and litigation and support practice. He was recently chair of the editorial board of The Value Examiner *and coauthor of NACVA's quarterly marketing newsletter* Insights on Valuation. *Filler has published various articles on valuation and commercial damages, and is the coauthor of* A Quantitative Approach to Commercial Damages: Applying Statistics to the Measurement of Lost Profits *published by John Wiley & Sons.*