Confidence Intervals for the COD: Limitations and Solutions

Robert J. Gloudemans

Partner - Almy, Gloudemans, Jacobs & Denne

Abstract: The IAAO has adopted assessment uniformity standards as measured by the COD and many states and provinces have developed similar standards. However, the IAAO *Standard on Ratio Studies* (1999) states that "jurisdictions should not be mandated to reappraise unless the ratio study indicates failure to meet the standards presented in this section with an appropriate degree of statistical confidence" (page 36). This statement implies that one must calculate confidence intervals about the calculated COD or conduct related tests to determine whether required standard have been met. Unfortunately, unlike measures of assessment level (median, mean, weighted mean), there are no published formulas for calculating confidence intervals for the COD and the only method cited in the appraisal literature for doing so is the complex, computer-intensive "bootstrap" method. This paper evaluates the dilemma and proposes a simple, realistic test for determining whether COD standards have been achieved.

Background

There are three primary aspects of valuation performance: overall level, equity between property groups, and equity within property groups. The assessment industry has developed standards in all three areas and assessment agencies naturally desire to quantify performance in each area as accurately as possible. To this end, confidence intervals are used to measure the precision of computed assessment levels as measured by the median, mean, and weighted mean to determine whether one can assume with reasonable confidence that required measures have not met. Similarly, statistical tests can readily be applied to determine whether assessment levels for two or more property groups are reasonably similar.

The primary gauge of equity among individual properties *within* a property use class, neighborhood, or other group is the coefficient of dispersion (COD), which measures the average percentage deviation about the median ratio. IAAO and many state/provincial agencies have adopted standards for the COD and reappraisal contracts often call for the contractor to attain specified CODs. Of course, a computed COD, just like a computed measure of the assessment level, is only an indicator of true performance. The accuracy of the measure depends on sample size and distribution. Clearly, there is a need to determine with reasonable confidence whether assessment uniformity standards have been achieved. In fact, the 1999 IAAO *Standard on Ratio Studies* (page 36) states that jurisdictions "should not be required to reappraise unless the ratio study indicates failure to meet the standards presented in this section with an appropriate degree of statistical confidence". This language, not present in previous versions of the standard, makes it virtually imperative for oversight agencies to quantify whether required uniformity measures have been met.

Current Methodology and Limitations

If ratio data could be assumed to be normally distributed, a confidence interval for the standard deviation and hence the COD could be constructed. It is well known that in a normal distribution, a confidence interval for the population variance (σ^2) is given by the formula,

$$(n-1)s^2 / \chi^2_L \leq \sigma^2 \leq (n-1)s^2 / \chi^2_U$$

where n = sample size, s^2 is the sample variance, and χ^2_L and χ^2_U are the appropriate lower and upper critical values, respectively, of a chi-square distribution with n-1 degrees of freedom. One could then take the square root of the lower and upper confidence limits to extract the standard deviation, divide by the mean to obtain the coefficient of variation (COV), and multiply by .80 to obtain the corresponding confidence limits for the COD, since in a normal distribution the COD is approximately 80% of the COV[1]. Unfortunately, ratio data does not always approximate a normal distribution.

An alternative, nonparametric approach is to use a repeat sampling or "bootstrap" methodology. In this case, one draws a large number of samples with replacement of size *n* from the sample, calculates the COD for each draw, and determines the cut points (confidence limits) that correspond to the desired confidence level. For example, if a 95% confidence interval is desired, one could draw 1,000 samples (with replacement). The lower confidence limit would fall between the 25th and 26th smallest CODs and the upper limit would lie between the 975th and 976th largest CODs. Unfortunately, bootstrap confidence intervals are not part of software packages generally found in an assessor's office and require special (complex) programming[2].

The biggest problems with confidence intervals for the COD, however, go beyond the above limitations. As is well known, the precision of any statistic is a function of (1) sample size and (2) the distribution or dispersion of the data. As the dispersion of the data increases, confidence intervals widen. This is particularly problematic with measures of dispersion or uniformity, because dispersion is precisely what is being measured. Regardless of whether one assumes normality and uses the parametric approach or uses a bootstrap method, the resulting confidence limits for the COD will increase with the COD itself. The worse the COD, the wider the confidence interval. For example, given a certain sample size, a COD of 20 will have a much wider confidence interval (say 10 to 30) than a COD of 10 (say 5 to 15). Thus, poor dispersion masks or condones poor dispersion[3].

---

1 The author verified this by computing the COD and COV from a sample of 1,000,000 random draws from a normal distribution with mean of 1 and a standard deviation of 1. The COD and COV were 80.01 and 100.20, respectively.

2 The Kansas Department of Revenue has developed one such program.

3 Of course, poor dispersion also confounds the ability to reject poor assessment levels (a problem addressed further later in the paper).

Table 1 presents a hypothetical example of 30 residential ratios from a municipality that has not revalued in some time. The ratios range from 0.506 to 1.338, the median is .804, and the COD has deteriorated to 18.38. Can we conclude with 95% confidence that the true COD (for the population of residential values) is below 15?

Assuming normality[4], the 95% confidence limits for the population variance is calculated as:

$$(30\text{-}1)*.0382 /45.72 \leq \sigma^2 \leq (30\text{-}1)*.0382 /16.05$$

$$.0242 \leq \sigma^2 \leq .0690$$

where 45.72 and 16.05 are the critical chi-square values at the 95% confidence level for a sample of size 30 (and thus 29 degrees of freedom). The corresponding confidence limits for the population standard deviation (σ) and COD are then derived as follows:

Lower Limits	Upper Limits
σ = sqrt(.0242) = .1556	σ = sqrt(.0690) = .2627
COV = 100 x .1556/.8533 = 18.23	COV = .2627/.8533= 30.78
COD = 18.23 x .80 = 14.58	COD = 30.78 x .80 = 24.63

Notice that the confidence limits are not symmetric about the sample COD (18.38). This is typical and occurs because the upper end is skewed by ratios more distant from the center of the distribution (regardless of the distribution of the data).

Since the lower confidence limit for the COD is below 15, the null hypothesis that the true COD is 15 or less cannot be rejected with 95% confidence, despite the fact that the sample COD is 18.38 based on 30 sales. Similarly, a bootstrap algorithm (which does not require normality) run with 5,000 iterations also yielded a lower confidence limit of less than 15.0.

A better measurement tool is needed if IAAO standards for the COD are to have credibility and if tests of uniformity are to be commonly and easily conducted.

Suggested Approach

Rather than attempting to compute confidence intervals for the COD, consider an approach in which one tests the null hypothesis that the COD is not more than the value set in standards (which may be based on state/provincial requirements, professional guidelines, or in-house policy):

---

[4] Despite the ratios being skewed to the right, neither the powerful Kolmogorov-Smirnov or Shapiro-Wilk test were able to reject normality at the 90% confidence level.

$$H_o: COD \leq CODSTD$$

where CODSTD is the "standard" (required or target) COD. What is the maximum value of the *calculated* COD that can be accepted before H_o is rejected at the desired confidence level? This restates the problem in the form of a test and makes clear that what is really desired is to determine whether uniformity can be deemed *worse* than set forth in standards (thus a one-tailed test is appropriate). To make a straightforward mathematical solution possible, assume finally that the standard against which the calculated COD will be compared is a normal distribution with a COD equal to the standard COD. This does not imply that actual ratios must be normal; but only that the distribution (whatever it is) can not have a COD that significantly exceeds that of the benchmark distribution. Thus, the benchmark distribution is a normal distribution, where low and high ratios are approximately evenly balanced, with variance defined by the required or target COD. The fact that actual ratios may vary from a normal distribution will not result in rejection of Ho, but a significantly higher COD will.

Given this framework, the following test statistic is appropriate:

$$\chi^2 = \frac{(n-1) * COD^2}{CODSTD^2}$$

where CODSTD is again the required or target COD[5]. If the calculated chi-square value is less than or equal to the critical value (with n-1 degrees of freedom) at the specified confidence level, H_o (that uniformity complies with standards) is accepted; otherwise it is rejected.

To illustrate, consider the case of the hypothetical sample of 30 ratios in table 1. Assuming a 95% confidence level, the COD test is applied as follows:

$$\chi^2 = \frac{29 * 18.383^2}{15^2} = 43.56 .$$

The calculated chi-square value exceeds the critical one-tailed value of 42.56 and therefore H_o is rejected[6].

---

[5] Although chi-square tests traditionally compare standard deviations or variances, substituting the COD allows the actual COD to be used in testing. Further, as explained, in a normal distribution the COD is .80 times the COV (standard deviation divided by mean).

[6] The one-tailed t-value is appropriate because, as stated in H_o, we are specifically interested in whether the actual COD complies with the required COD.

Acceptable Limits for the COD

Acceptable upper limits for a calculated COD reflect three factors: (1) sample size, (2) the required or target COD, and (3) the specified confidence level. The above formula can be reworked to solve for the maximum acceptable COD based on the other three factors:

$$MAXCOD = CODSTD * sqrt[\chi^2 / (n-1)]$$

where MAXCOD is the maximum COD that can be accepted for the sample without concluding that the COD standard has not been met and χ^2 is the chi-square value for a one-tailed test with the appropriate degrees of freedom (n-1). For exposition purposes, we can rewrite the formula as follows:

$$MAXCOD = CODSTD * TF$$

where TF equals a "tolerance factor", computed as $sqrt[\chi^2 / (n-1)]$. The minimum value of TF, which would apply in the case of a sample of infinite size, is 1.00. The smaller the sample and the more demanding the confidence level for rejection of H_o, the higher TF is. For example, in a sample of only 10 sales and 95% confidence, TF = 2.11. For the previous example of 30 sales,

$$TF = sqrt(42.56 / 29) = 1.21$$

$$MAXCOD = 15 * 1.211 = 18.17$$

Since the actual COD (18.38) exceeds MAXCOD, the null hypothesis that the true COD is 15.0 or better can be rejected.

Table 2 contains values of TF and corresponding maximum allowable CODs for various sample sizes at the 95% confidence level based on COD standards of 10, 15, and 20. Notice that as sample sizes increase, TF factors move closer to 1.00 and maximum allowable CODs converge toward standards. Figure 1 graphs the relationship between TF and sample size.

Case Study: New York Towns

The COD test described above was applied to residential properties in 215 towns in the Northern region of New York State using sales from April, 1997 to March, 1999 (another 30 towns had fewer than 5 valid sales over the two year period and were excluded). Ratios that lay more than 1.5 interquartile ranges above the 75th percentile or below the 25th percentile (3.6% of cases in all) were removed as "outliers". After trimming, CODs ranged from 3.8 to 59.3, with an average of 13.8. Figure 2 shows a histogram of the CODs. Sixty-four towns had CODs above 15.0, of which 26 exceeded 20.0.

The null hypothesis that the COD was 20.0 or better could be rejected at the 95% confidence level in 10 of the 26 towns in which the COD exceeded 20.0. Figure 3 shows a plot of CODs by sample size for towns with approximately 100 sales or less, with markers indicating which CODs could be

rejected. The three towns with CODs of near 30 for which H_o could not be rejected all had only five sales.

Finally, the tests were rerun to determine how many of the 215 towns had CODs that exceeded a standard of 15.0. Thirty-four (15.8% of all 215 towns and 53.1% of those with CODs greater than 15) had CODs that could be rejected at the 95% confidence level. Figure 4 shows a plot of the results for towns with fewer than 200 sales.

Extensions to the Level of Assessment

The same problem in variability in ratios that makes it difficult to determine whether COD standards have been met also plagues the ability to determine whether assessment level standards have been achieved: the more variability in the ratios, the more tolerance confidence limits provide. A possible solution to this problem is to base accepted tolerance on the variability inherent in *acceptable* dispersion rather than observed dispersion. When the actual COD exceeds the required or target COD, tolerance limits could be based acceptable rather than actual dispersion. This would prevent poor dispersion from justifying poor assessment levels and would provide more equal tolerance standards to jurisdictions with good and poor performance.

Table 1
30 Ratios

Ratio	Ratio
0.530	0.814
0.577	0.828
0.614	0.846
0.655	0.863
0.688	0.889
0.701	0.915
0.717	0.939
0.730	0.953
0.744	0.977
0.746	1.028
0.755	1.069
0.769	1.115
0.775	1.200
0.787	1.242
0.794	1.338

Ratio Statistics	
Sales	30
Median	.80400
Mean	.85327
Std Dev	.19545
Variance	.03820
COV	22.91
Minimum	.53000
Maximum	1.3830
25^{th} Percentile	.72675
75^{th} Percentile	.95900
COD	18.38

Table 2
Tolerance Factors and Maximum Acceptable CODs

		COD Standard		
N	TF	10.00	15.00	20.00
5	1.540	15.40	23.10	30.80
6	1.488	14.88	22.32	29.76
7	1.449	14.49	21.73	28.97
8	1.418	14.18	21.26	28.35
9	1.392	13.92	20.88	27.85
10	1.371	13.71	20.57	27.42
11	1.353	13.53	20.30	27.06
12	1.337	13.37	20.06	26.75
13	1.324	13.24	19.86	26.47
14	1.312	13.12	19.67	26.23
15	1.301	13.01	19.51	26.01
16	1.291	12.91	19.36	25.82
17	1.282	12.82	19.23	25.64
18	1.274	12.74	19.11	25.48
19	1.266	12.66	19.00	25.33
20	1.260	12.60	18.89	25.19
21	1.253	12.53	18.80	25.06
22	1.247	12.47	18.71	24.95
23	1.242	12.42	18.63	24.84
24	1.237	12.37	18.55	24.73
25	1.232	12.32	18.48	24.64
26	1.227	12.27	18.41	24.54
27	1.223	12.23	18.34	24.46
28	1.219	12.19	18.28	24.38
29	1.215	12.15	18.23	24.30
30	1.211	12.11	18.17	24.23
35	1.196	11.96	17.93	23.91
40	1.183	11.83	17.74	23.66
45	1.172	11.72	17.59	23.45
50	1.164	11.64	17.45	23.27
60	1.149	11.49	17.24	22.99
70	1.138	11.38	17.07	22.76
80	1.129	11.29	16.94	22.59
90	1.122	11.22	16.83	22.44
100	1.116	11.16	16.73	22.31
200	1.082	10.82	16.23	21.64
300	1.067	10.67	16.00	21.34
400	1.058	10.58	15.87	21.16
500	1.052	10.52	15.78	21.04

Figure 1

Graph of TF with N

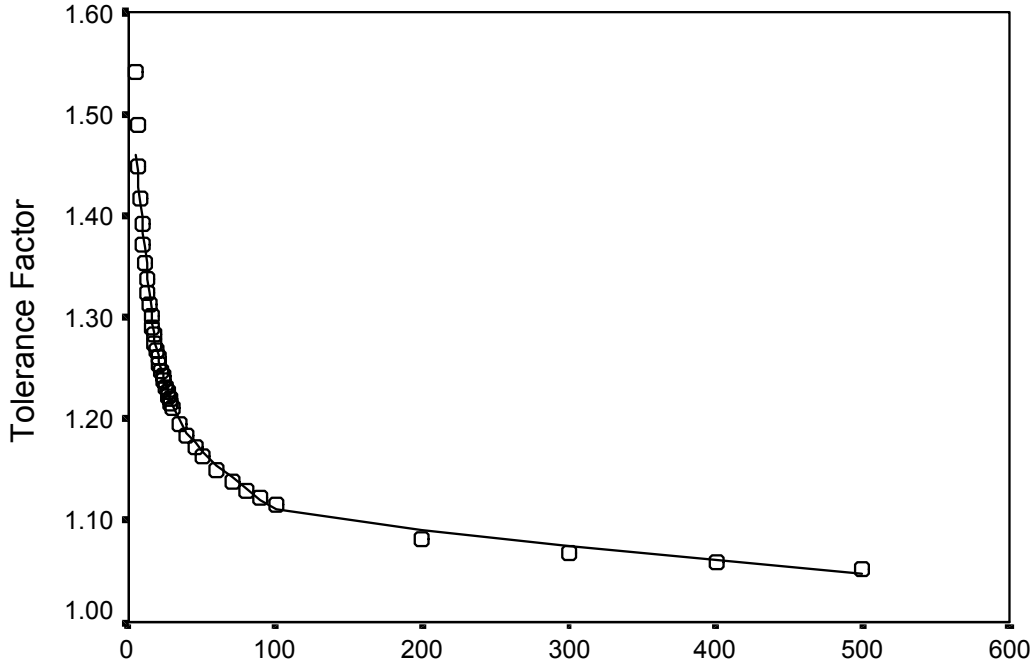

Figure 2

CODs of 215 NY Towns

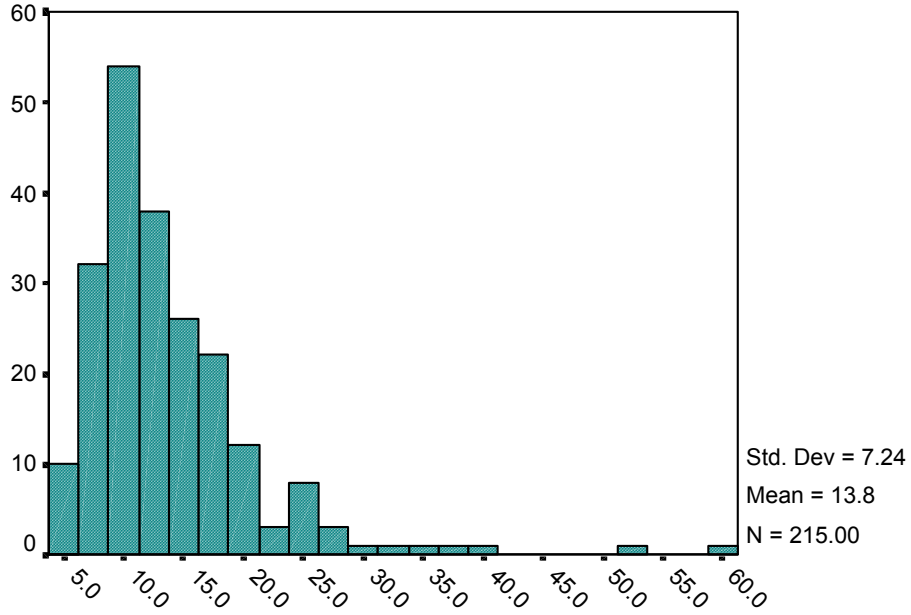

Figure 3

Compliance with Standard of 20

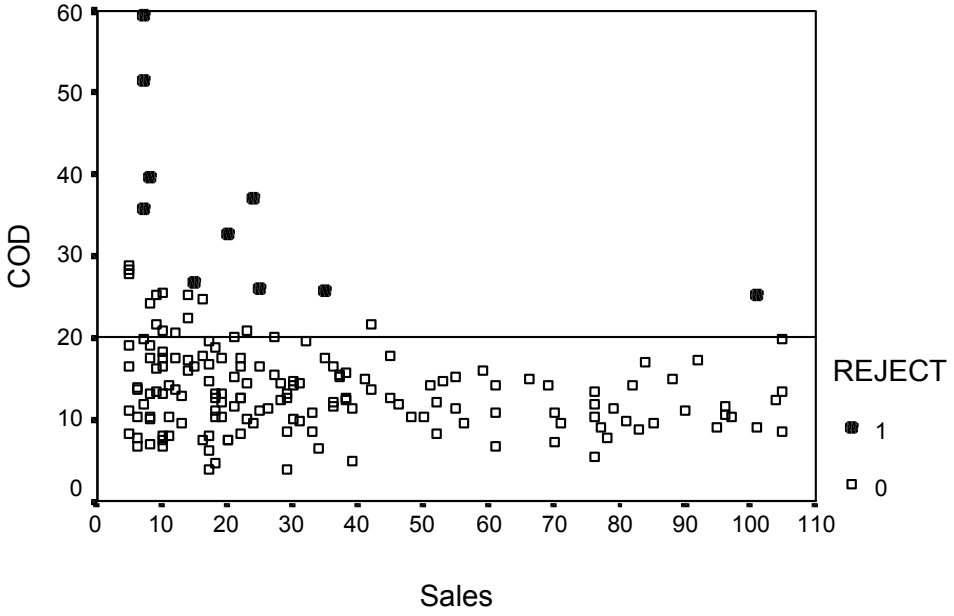

Figure 4

Compliance with Standard of 15