CADMUS

Memorandum

To: Erik Helm; EPA

From: Richard Krop, Yitzhak Henry; Cadmus

Subject: Comparison of Estimate of Cost of Lead Service Line Replacements using 7th DWINSA and CMD Smith Data

Date: September 12, 2024

The average cost of lead service line replacements in the Lead and Copper Rule Improvements (LCRI) Economic Analysis (EA) is based on a sample of systems collected in the 7th Drinking Water Infrastructure Needs Survey and Assessment (DWINSA). The sample is a large probability sample selected randomly from the inventory of public water systems (PWS). Not every system provided information about the cost of lead service lines, but the EPA can use the sample to estimate the design-based mean and standard error of the cost of service line replacements.

CDM Smith collected cost information from a sample independent of the DWINSA sample. Unlike the 7th DWINSA, the CDM Smith sample is not a probability sample. It does not include sampling weights or other information about the sample. It also includes the number of lead service lines associated with the cost of each project for only some observations, not all.

Because of the lack of information about the CMD Smith sample, it is difficult to conduct formal tests of differences between the means of the two samples. One approach is to ignore the sampling information and the number of lines from the 7th DWINSA and conduct a difference-in-means test between the two samples, assuming unequal variances. This treats the 7th DWINSA data like the CDM Smith data: it does not weight the estimate by the number of service lines and removes sample weights and other sampling information. Table 1 shows the average cost of service line replacements, the standard error of the mean, and the 95 percent confidence interval based on data from each dataset. It also shows the difference between the two estimates, the standard error of the difference, and 95 percent confidence interval of the difference. The standard error of the difference is given by $(s_D^2/n_D + s_C^2/n_C)^{1/2}$, where $s_D^2/n_D$and s_C^2/n_C are the standard errors of the mean of the cost of service line replacements in the 7th DWINSA dataset and the CDM Smith dataset, respectively.

Table 1. Cost per Line Replaced in the 7th DWINSA and CDM Samples Unweighted Estimates (in 2020 Dollars)

Data Source	Mean Cost of LSL Replacement	Standard Error	[95% conf. interval]	
7th DWINSA	$7,419	$660	$6,051	$8,787
CDM Smith	$8,717	$755	$7,124	$10,131
Difference	$1,299	$1,003	-$734	$3,332

The mean cost in the CDM Smith sample is approximately $1,300 higher than the mean using the 7th DWINSA sample. The distribution of the data for each sample is shown in Figure 1. The ranges are similar, but the CDM Smith data are skewed slightly higher than the 7th DWINSA data.

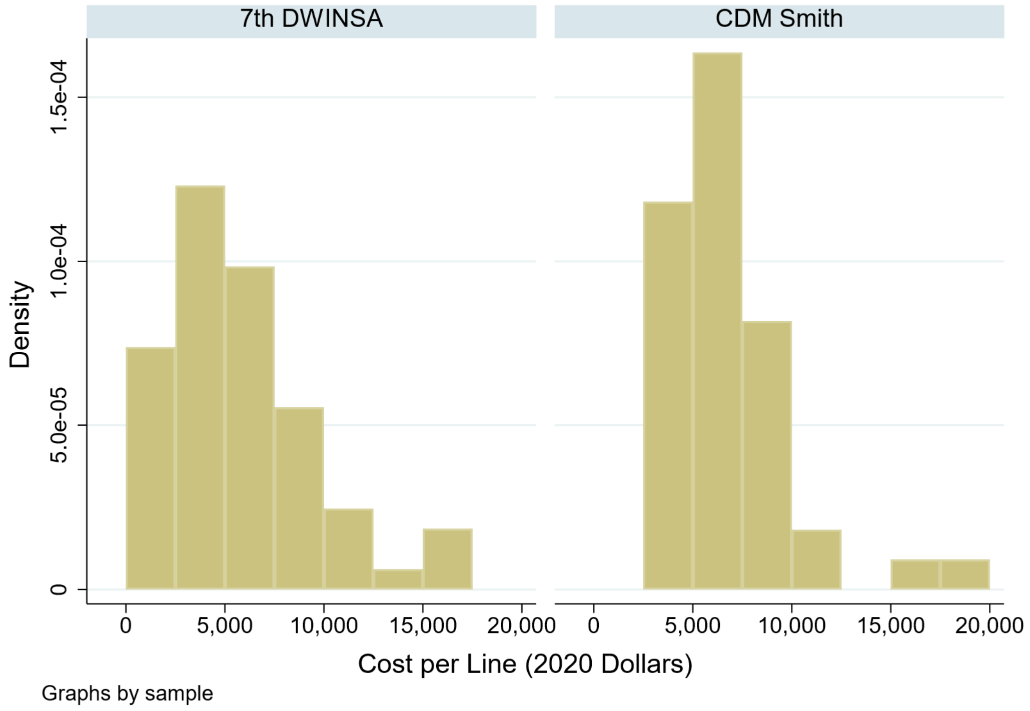

Figure 1. Distribution Cost per Service Line Replacement in the 7th DWINSA and CDM Smith Samples

To formally test the difference in the average cost per line, we conducted a two-sample t-test with unequal variances. The test statistic is a t statistic with 36 degrees of freedom. The difference in the means is not statistically significant at the five percent level:

t = 1.2952
Degrees of freedom = 36
critical value = 2.028
Probability of observing a t >1.2952 if the difference is 0 = 0.2034

Not accounting for the weighting of the DWINSA data may over- or under-estimate the difference in costs. Unweighted comparisons can distort the results, especially when there are disparities in how the samples were collected or how representative they are of the general population of public water systems. If we assume the CDM Smith is a simple random sample with equal weights, we can test the difference in weighted means. The CDM Smith data includes the number of lines replaced for a portion of the sample. For purposes of this analysis, we assume the number of lines for the other observations is equal to the average of the reported number of lines. Table 2 shows the weighted mean for both samples.

Table 2. Cost per Line Replaced in the 7th DWINSA and CDM Samples Weighted Estimates (in 2020 Dollars)

Data Source	Mean Cost of LSL Replacement	Linearized Standard Error	[95% conf. interval]	
7th DWINSA	$6,930	$40	$6,847	$7,013
CDM Smith	$8,305	$346	$7,588	$9,021

We use a Wald test to test the difference in the weighted means. The adjusted Wald test statistic is an F with 1 degrees of freedom in the numerator and 22 degrees of freedom in the denominator. The difference of approximately $1,375 is statistically significant at the 5 percent level.

F(1, 22) = 15.62
Critical value = 4.30
Probability of observing an F > 15.62 if the difference is 0 = 0.0007

The weighted mean from the 7th DWINSA reflects the probability-based sampling design, giving a more representative estimate of lead service line replacement costs. However, the CDM Smith sample lacks such sampling information, requiring the assumption that it is a simple random sample with equal weighting. This assumption may not be valid, as the lack of sampling weights and other information can lead to biased estimates that do not reflect the broader population. As a result, the difference detected by the weighted difference in means test may be skewed, and the statistical significance might not accurately represent the true differences in costs between the two samples. We do not recommend using this weighted analysis; we provide it to demonstrate how the weights can change the results of the analysis.