Statistics Based on Adjusted Metered Water Supply Manual (2001)

Jane Zou

February 11, 2024

A) Introduction

Surcharges and connection fees are determined based on the strength data for suspended solids (SS) and chemical oxygen demand (COD), sourced from district monitoring events and self-monitoring reports. This data is provided in two sets through split samples, with averages calculated from both sets. Occasionally, analytical results may substantially differ from the typical strength data observed, prompting the need to identify such results as outliers and subsequently reject them. Statistical hypothesis testing is employed to apply outlier identification techniques, determining whether an observation significantly deviates from the norm. These tests ascertain whether an observation should be retained in the dataset as typical or classified as an outlier based on sufficient evidence.

Before concluding that data outliers retained in the dataset inevitably signify a change, it is crucial to consider various factors that could have led to their presence. These outliers could stem from a malfunctioning instrument, inaccuracies in data transcription, unnoticed laboratory errors, contaminated sampling, incorrect instrument readings, mistakes in experimental or analytical procedures, intentional concealment, inconsistent sampling methods, and so on. Determining whether a monitoring result might be attributed to an unnatural, catastrophic event such as a spill requires careful examination. Whether an observation should be included or excluded in a particular situation depends on an assessment of how much variability the dataset can accommodate. If an outlier can be linked to one of the error-producing processes mentioned above, it may be safely removed from consideration. Therefore, the primary task is to utilize lab notebooks to meticulously track the experimental setup and methodology behind the aberrant observation. If a logical reason for exclusion is identified, the questionable value should be removed, subsequently facilitating the data analysis process.

Instead of relying solely on traditional methods, several alternative procedures have been devised to identify outliers. Many of these protocols operate under the assumption that a sufficiently large and representative sample set, combined with an underlying normal distribution, enables the identification and removal of suspicious data points. In cases where the sample set is limited, outliers must exhibit significant deviation from the rest of the population to be identified as such.

For the Districts' purposes, two procedures will be outlined to test for outliers:

- Dixon's Outlier Test: This test is designed for screening outlier concentrations in datasets containing fewer than 25 samples, assuming the data (with outliers removed) originates from a normal (Gaussian) distribution.
- Rosner's Generalized Extreme Studentized Deviate Test: This method is suitable for datasets with a minimum of 25 samples. It permits up to 10 potential outliers in a dataset,

under the assumption that the data (with outliers removed) follows a normal (Gaussian) distribution.

Transformed data refers to the dataset that has undergone a normalization process, typically through methods like Box-Cox or log transformations. This normalization aims to make the distribution of the data more closely resemble a normal (Gaussian) distribution, which is a prerequisite for the application of Rosner's test.

If the data does not follow a normal distribution, it may be necessary to apply transformation techniques to normalize the data before conducting outlier tests.

B) Shapiro-Wilk Normality Test

The Shapiro-Wilk Test is a method for evaluating normality in datasets, especially beneficial for small to medium-sized samples. It involves comparing observed data order statistics with those anticipated from a normal distribution. Lower values of the W statistic suggest deviations from normality. The test evaluates the null hypothesis that a sample is drawn from a population with a normal distribution.

1) Formula

The test statistic W for the Shapiro-Wilk Test is computed using the following formula:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- $x_{(i)}$ denotes the i-th smallest number in the sample.
- $\bar{x}$ denotes the sample mean.
- a_i are constants derived from the means, variances, and covariances of the order statistics based on the properties of the normal distribution and are calculated to yield the best fit (in a least-squares sense) to the expected values of order statistics from a normal distribution.

This statistic is employed to test the null hypothesis that the data were drawn from a normal distribution. A low value of W indicates that the null hypothesis can be rejected when $p < 0.05$, implying that the data are not normally distributed due to significant deviation from the expected values of a normal distribution.

2) Implementation

The computation of the constants a_i in the Shapiro-Wilk Test involves a complex process typically carried out using statistical software. These constants are determined based on the sample size and the expected values of the order statistics for a standard normal distribution. This

calculation accounts for the unique characteristics of the sample data and ensures accurate assessment of the test statistic W in evaluating the normality of the dataset. Such computations can also be executed using Microsoft Excel, as demonstrated in the tool developed by Kristopher McGinnis of LACSD.

C) Dixon's Outlier Test for Samples Less than Size 25

To identify outliers using Dixon's Outlier Test, the set of observations is initially arranged based on their magnitudes. Then, utilizing a formula that varies according to the sample size, the ratio is computed by dividing the difference between an extreme value and one of its closest neighbor values within the range of sample values.

$$\frac{Gap}{Range}$$

- Gap represents the absolute difference between the outlier in question and the nearest number in the dataset.
- Range represents the absolute difference between the dataset's maximum and minimum values.

1) Formula

In 1950, Dixon introduced the following ratio statistic to identify outliers:

$$r_{j,i-1} = max\{\frac{x_n - x_{n-j}}{x_n - x_i}, \frac{x_{1+j} - x_1}{x_{n-i} - x_1}\}$$

The specific calculation of the gap depends on whether one is searching for outliers at the high end, low end, or both ends of the dataset. The standard version of the test utilizes different sample size ranges to generate a range of critical values.

$$r_{10} \text{ for } 3 \leq n \leq 7,$$
$$r_{11} \text{ for } 8 \leq n \leq 10,$$
$$r_{21} \text{ for } 11 \leq n \leq 13,$$
$$r_{22} \text{ for } 14 \leq n \leq 30.$$

In Dixon's Outlier Test, the number of suspected outliers at the high end is denoted by the first subscript on r, while the number of suspected outliers at the low end is indicated by the second subscript. When using r_{10}, Dixon's test is referred to as the Q-Test.

The test's null hypothesis, which claims that there are no outliers, is contrasted with the alternative hypothesis, which claims that at least one observation is an outlier. An extreme value is defined as an outlier at either the 5% or 1% significance threshold when it surpasses a tabulated number. The 5% criteria threshold, or an α value of 0.05, is usually applied for Districts' purposes. In situations where the sample size is small, the stricter 1% criteria threshold, which corresponds to an α value of 0.01, may be used. By contrasting the computed ratio with

tabulated comparison values, the criteria level denotes the significance threshold at which extreme results are categorized as outliers.

The formula for the ratio varies based on the sample size and whether the alleged outlier has the highest or lowest value. During each test iteration, the greatest or lowest outlier value is identified, and the remaining values are retained for subsequent iterations. Iterations continue until no data points are identified as outliers. The calculated ratio is then compared with tabulated comparison values, considering the significance level and the number of samples that have not yet been confirmed as outliers.

Spreadsheet programs like Microsoft Excel make it easy to use Dixon's Outlier Test. This technique provides a useful tool for outlier discovery in research projects by utilizing Excel's arithmetic computation and statistical analysis features.

2) Example

Consider the concentration values of Benzo(a)pyrene: 2.77, 2.80, 2.90, 2.92, 3.45, 3.95, 4.44, 4.61, 5.21, and 7.46. To determine if 7.46 is an outlier using Dixon's test:

$$r_{11} = \frac{(X_{10} - X_9)}{(X_{10} - X_2)} = \frac{7.46 - 5.22}{7.46 - 2.80} = 0.48$$

As $r_{11} = 0.48$ exceeds the critical value of 0.477 for $N = 10$ at the 5% significance level (from the provided table), 7.46 is considered an outlier.

CRITICAL VALUES AND CRITERIA FOR TESTING FOR EXTREME VALUES

N \ α	.30	.20	.10	.05	.02	.01	.005	Criterion
3	.684	.781	.886	.941	.976	.988	.994	
4	.471	.560	.679	.765	.846	.889	.926	
5	.373	.451	.557	.642	.729	.780	.821	$r_{10} = \frac{x_N - x_{N-1}}{x_N - x_1}$
6	.318	.386	.482	.560	.644	.698	.740	
7	.281	.344	.434	.507	.586	.637	.680	
8	.318	.385	.479	.554	.631	.683	.725	
9	.288	.352	.441	.512	.587	.635	.677	$r_{11} = \frac{x_N - x_{N-1}}{x_N - x_2}$
10	.265	.325	.409	.477	.551	.597	.639	
11	.391	.442	.517	.576	.638	.679	.713	
12	.370	.419	.490	.546	.605	.642	.675	$r_{21} = \frac{x_N - x_{N-2}}{x_N - x_2}$
13	.351	.399	.467	.521	.578	.615	.649	
14	.370	.421	.492	.546	.602	.641	.674	
15	.353	.402	.472	.525	.579	.616	.647	
16	.338	.386	.454	.507	.559	.595	.624	
17	.325	.373	.438	.490	.542	.577	.605	
18	.314	.361	.424	.475	.527	.561	.589	
19	.304	.350	.412	.462	.514	.547	.575	
20	.295	.340	.401	.450	.502	.535	.562	$r_{22} = \frac{x_N - x_{N-2}}{x_N - x_3}$
21	.287	.331	.391	.440	.491	.524	.551	
22	.280	.323	.382	.430	.481	.514	.541	
23	.274	.316	.374	.421	.472	.505	.532	
24	.268	.310	.367	.413	.464	.497	.524	
25	.262	.304	.360	.406	.457	.489	.516	

Table 1: Dixon, W. J. (1953). Processing data for outliers. Biometrics, 9(1), 74-89.

D) Rosner's Generalized Extreme Studentized Deviate Test

Rosner's Test is designed for detecting up to 10 outliers in datasets with 25 or more samples, provided that the data (or transformed data) follows a normal (Gaussian) distribution. Transformed data refers to a dataset that has undergone normalization, often through methods like Box-Cox or log transformations, to make its distribution more akin to a normal distribution. Transformed data improves the validity and reliability of Rosner's Test by guaranteeing that the data satisfy the assumption of normality, allowing for more precise outlier detection. If the sample size is less than 25, Rosner's Test cannot be applied because the distribution of residuals may not approximate a normal distribution well enough for the test to be reliable. This procedure's ability to spot outliers that could be hidden by other outliers is one of its advantages. Additionally, Rosner's Test can detect suspiciously large or suspiciously small data points.

1) Implementation

To implement Rosner's Test, an upper limit k must be specified on the number of potential outliers present. The procedure involves repetitively removing the data point (either large or small) farthest from the mean and recalculating the test statistic after each deletion. A table, typically provided, is used to evaluate the test. For sample sizes between 50 and 500, linear interpolation may be utilized to obtain critical values not explicitly listed in the tables.

2) Formula

Certain notation is necessary to illustrate Rosner's Test. Let ${x_m}^{(i)}$ and ${s_m}^{(i)}$ represent the sample mean and standard deviation, respectively, of the $n - i$ observations that remain after the removal of the i most extreme observations. For instance, ${x_m}^{(1)}$ and ${s_m}^{(1)}$ denote the sample mean and standard deviation for the $n - 1$ data remaining after the most outlying datum from ${x_m}^{(0)}$ has been eliminated. Furthermore, $x^{(i)}$ refers to the most outlying observation, i.e., the data point farthest from the mean ${x_m}^{(i)}$, remaining in the dataset after the removal of i more extreme data points (either large or small).

$$R_{i+1} = \frac{|x^{(i)} - {x_m}^{(i)}|}{s^{(i)}} \quad (1)$$

$$\lambda_{i+1} = \text{tabled critical value for comparison with } R_{i+1} \quad (2)$$

- R_{i+1} is the test statistic for deciding whether the $i + 1$ most extreme values in the complete data set are outliers from a normal distribution.

3) Example

Consider the following dataset comprising the logarithms of $n = 55$ total suspended particulate (TSP) air data, collected every sixth day at a monitoring site. The logarithms are arranged in ascending order. Our aim is to investigate the presence of outliers in this dataset, assuming it follows a lognormal distribution. We adopt a significance level of 5%.

Given our prior decision to investigate for three outliers (setting $k = 3$), our task is to test whether such outliers exist within the dataset.

2.56	3.58	3.83	4.06	4.32
3.18	3.58	3.91	4.08	4.33
3.33	3.64	3.91	4.09	4.33
3.40	3.69	3.97	4.17	4.44
3.43	3.69	3.99	4.17	4.47
3.43	3.71	4.03	4.23	4.48
3.43	3.74	4.04	4.26	4.48
3.50	3.76	4.04	4.26	4.62
3.50	3.76	4.04	4.29	4.68
3.50	3.81	4.04	4.32	5.16

Values of $y_m^{(i)}$, $s_y^{(i)}$, and $R_{y,i+1}$ for $i = 0, 1,$ and 2 were computed from the data and presented in the table below.

i	$n - i$	$y_m^{(i)}$	$s_y^{(i)}$	$y^{(i)}$	$R_{y,i+1}$	λ_{i+1}
0	55	3.94	0.444	2.56	3.11	3.165
1	54	3.96	0.406	5.16	2.96	3.155
2	53	3.94	0.374	4.68	1.98	3.150

- $y_m^{(0)}$ and $s_y^{(0)}$ denote the mean and standard deviation for the entire dataset.
- $y_m^{(1)}$ and $s_y^{(1)}$ denote the mean and standard deviation after removing 2.56.
- $y_m^{(2)}$ and $s_y^{(2)}$ denote the mean and standard deviation after removing both 2.56 and 5.16.

The most extreme datum, $y^{(i)}$, at each stage is also indicated. The critical values in the last column were derived through linear interpolation for the 5% significance level for $n = 50$ and $n = 60$.

- $R_{y,3} = 1.98$ is less than $\lambda_3 = 3.150$, we fail to reject 4.68.
- $R_{y,2} = 2.96$ is less than $\lambda_2 = 3.155$, we fail to reject 5.16.
- $R_{y,1} = 3.11$ is less than $\lambda_1 = 3.165$, we fail to reject 2.56.

We conclude that there are no outliers within the assumed lognormal distribution.

The calculation of the mean, standard deviations, and *R* values for the various iterations can be efficiently conducted using the data analysis tool in Excel, programmed by Kristopher McGinnis of LACSD.

Approximate Critical Values for Rosner's Outlier Procedure

n	i+1	5%	n	i+1	5%
25	1	2.82	31	1	2.92
	2	2.80		2	2.91
	3	2.78		3	2.89
	4	2.76		4	2.88
	5	2.73		5	2.86
	10	2.59		10	2.76
26	1	2.84	32	1	2.94
	2	2.82		2	2.92
	3	2.80		3	2.91
	4	2.78		4	2.89
	5	2.76		5	2.88
	10	2.62		10	2.78
27	1	2.86	33	1	2.95
	2	2.84		2	2.94
	3	2.82		3	2.92
	4	2.80		4	2.91
	5	2.78		5	2.89
	10	2.65		10	2.80
28	1	2.88	34	1	2.97
	2	2.86		2	2.95
	3	2.84		3	2.94
	4	2.82		4	2.92
	5	2.80		5	2.91
	10	2.68		10	2.82
29	1	2.89	35	1	2.98
	2	2.88		2	2.97
	3	2.86		3	2.95
	4	2.84		4	2.94
	5	2.82		5	2.92
	10	2.71		10	2.84
30	1	2.91	36	1	2.99
	2	2.89		2	2.98
	3	2.88		3	2.97
	4	2.86		4	2.95
	5	2.84		5	2.94
	10	2.73		10	2.86

Table 2: Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2), 165-172.

Approximate Critical Values for Rosner's Outlier Procedure

n	i+1	5%	n	i+1	5%
37	1	3.00	44	1	3.08
	2	2.99		2	3.07
	3	2.98		3	3.06
	4	2.97		4	3.05
	5	2.95		5	3.04
	10	2.88		10	2.98
38	1	3.01	45	1	3.09
	2	3.00		2	3.08
	3	2.99		3	3.07
	4	2.98		4	3.06
	5	2.98		5	3.05
	10	2.89		10	2.99
39	1	3.03	46	1	3.09
	2	3.01		2	3.09
	3	3.00		3	3.08
	4	2.99		4	3.07
	5	2.98		5	3.06
	10	2.91		10	3.00
40	1	3.04	47	1	3.10
	2	3.03		2	3.09
	3	3.01		3	3.09
	4	3.00		4	3.08
	5	2.99		5	3.07
	10	2.92		10	3.01
41	1	3.05	48	1	3.11
	2	3.04		2	3.10
	3	3.03		3	3.09
	4	3.01		4	3.09
	5	3.00		5	3.08
	10	2.94		10	3.03
42	1	3.06	49	1	3.12
	2	3.05		2	3.11
	3	3.04		3	3.10
	4	3.03		4	3.09
	5	3.01		5	3.09
	10	2.95		10	3.04
43	1	3.07	50	1	3.13
	2	3.06		2	3.12
	3	3.05		3	3.11
	4	3.04		4	3.10
	5	3.03		5	3.09
	10	2.97		10	3.05

Table 3: Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2), 165-172.

Approximate Critical Values for Rosner's Outlier Procedure

n	i+1	5%	n	i+1	5%
60	1	3.20	200	1	3.61
	2	3.19		2	3.60
	3	3.19		3	3.60
	4	3.18		4	3.60
	5	3.17		5	3.60
	10	3.14		10	3.59
70	1	3.26	250	1	3.67
	2	3.25		5	3.67
	3	3.25		10	3.66
	4	3.24			
	5	3.24	300	1	3.72
	10	3.21		5	3.72
				10	3.72
80	1	3.31			
	2	3.30	350	1	3.77
	3	3.30		5	3.76
	4	3.29		10	3.76
	5	3.29			
	10	3.26	400	1	3.80
				5	3.80
90	1	3.35		10	3.80
	2	3.34			
	3	3.34	450	1	3.84
	4	3.34		5	3.83
	5	3.33		10	3.83
	10	3.31			
			500	1	3.86
100	1	3.38		5	3.86
	2	3.38		10	3.86
	3	3.38			
	4	3.37	750	1-10	3.95
	5	3.37	1000	1-10	4.02
	10	3.35	2000	1-10	4.20
			3000	1-10	4.29
150	1	3.52	4000	1-10	4.36
	2	3.51	5000	1-10	4.41
	3	3.51			
	4	3.51			
	5	3.51			
	10	3.50			

Table 4: Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2), 165-172.

E) Rosner's Test Formulas

The R programming language's Environmental Statistics package contains functions that have been incorporated into the research approach that is being discussed below.

In 1983, Rosner introduced the generalized Extreme Studentized Deviate (ESD) test for detecting one or more outliers in a univariate dataset that approximately follows a normal distribution. This test requires only an upper bound for the suspected number of outliers. It conducts separate tests for each potential number of outliers up to the specified upper bound, denoted as r. The test is particularly accurate for sample sizes of 25 or more.

Let $x_1, x_2, \cdots, x_n$ denote the n observations. We assume that $n - k$ of these observations originate from the same normal (Gaussian) distribution, while the k most extreme observations may or may not be outliers from a different distribution.

Let $x_1^*, x_2^*, \cdots, x_n^*$ denote the $n - i$ observations left after omitting the i most extreme observations, where $i = 0, 1, \cdots, k - 1$. Furthermore, let $\bar{x}^{(i)}$ and $s^{(i)}$ represent the mean and standard deviation, respectively, of the $n - i$ observations remaining after removing the i most extreme observations. Thus, $\bar{x}^{(0)}$ and $s^{(0)}$ denote the mean and standard deviation for the full sample, and in general,

$$\bar{x}^{(i)} = \frac{1}{n-i} \sum_{j=1}^{n-i} x_j^*$$

$$s^{(i)} = \sqrt{\frac{1}{n-i-1} \sum_{j=1}^{n-i} (x_j^* - \bar{x}^{(i)})^2}$$

For a specified value of i, the most extreme observation $x^{(i)}$ is the one that has the greatest distance from the mean for that particular dataset, i.e.,

$$x^{(i)} = max_{j=1,2,\cdots,n-i} |x_j^* - \bar{x}^{(i)}|$$

Thus, an extreme observation may be either the smallest or the largest one in that dataset.

Rosner's Test is founded on the k statistics $R_1, R_2, \cdots, R_k$, which denote the extreme Studentized deviates calculated from successively reduced samples of size $n, n - 1, \cdots, n - k + 1$:

$$R_{i+1} = \frac{|x^{(i)} - \bar{x}^{(i)}|}{s^{(i)}}$$

Critical values for R_{i+1} are denoted λ_{i+1} and are computed as:

$$\lambda_{i+1} = \frac{t_{p,n-i-2}(n-i-1)}{\sqrt{(n-i-2+t_{p,n-i-2})(n-i)}}$$

- $t_{p,v}$ denotes the p-th quantile of Student's t-distribution with v degrees of freedom

$$p = 1 - \frac{\alpha}{2(n-i)}$$

- α denotes the Type I error level

The algorithm for determining the number of outliers proceeds as follows:

1) Compare R_k with λ_k. If $R_k > \lambda_k$, conclude that the k most extreme values are outliers.

2) If $R_k \leq \lambda_k$, compare R_{k-1} with λ_{k-1}. If $R_{k-1} > \lambda_{k-1}$, conclude the $k - 1$ most extreme values are outliers.

3) Continue this process iteratively until a certain number of outliers have been identified or until Rosner's Test finds no outliers at all.

Based on a study utilizing $N = 1,000$ simulations, Rosner (1983) presented Table 1, which displays the estimated true Type I error rate of declaring at least one outlier when none actually exists. This analysis spans various sample sizes n ranging from 10 to 100 and the declared maximum number of outliers k ranging from 1 to 10. Rosner (1983) concluded that for an assumed Type I error level of 0.05, as long as $n \geq 25$, the estimated α levels closely approximate 0.05. Similar results were observed assuming a Type I error level of 0.01.

F) References

Dixon, W. J. "Analysis of Extreme Values." *The Annals of Mathematical Statistics*, vol. 21, no. 4, 1950, pp. 488–506. *JSTOR*, http://www.jstor.org/stable/2236602.

Dixon, W. J. "Processing Data for Outliers." *Biometrics*, vol. 9, no. 1, 1953, pp. 74–89. *JSTOR*, https://doi.org/10.2307/3001634.

Rosner, Bernard. "Percentage Points for a Generalized ESD Many-Outlier Procedure." *Technometrics*, vol. 25, no. 2, 1983, pp. 165–72. *JSTOR*, https://doi.org/10.2307/1268549.

Royston, P. Approximating the Shapiro-Wilk W-test for non-normality. *Stat Comput* **2**, 117–119 (1992). https://doi.org/10.1007/BF01891203

Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, vol. 52, no. 3/4, 1965, pp. 591–611. *JSTOR*, https://doi.org/10.2307/2333709.