Statistics Based on Adjusted Metered Water Supply Manual (2001)

Jane Zou

February 11, 2024

Table of Contents

3

Introduction

Importance of accurate water quality data for determining surcharges and connection fees

Data Analysis Overview

- Surcharges and connection fees determined by SS and COD strength data.
- Data from district monitoring events, split samples.
- Outlier detection via statistical hypothesis testing.

Factors Affecting Outliers and Examination Process

- Factors causing outliers: instrument issues, errors in transcription or sampling.
- Careful examination needed to distinguish natural variation from abnormal events.
- Lab notebooks track setup for logical exclusion.

Alternative Outlier Detection Methods

- Alternative outlier detection relies on large, representative samples.
- Dixon's Outlier Test for < 25 samples assuming normal distribution.
- Rosner's Generalized Extreme Studentized Deviate Test for ≥ 25 samples assuming normal distribution post-outlier removal.

7

Shapiro-Wilk Test

Formula and implementation details for computing the test statistic

Normality Assessment

- Method for evaluating normality, particularly useful for small to medium-sized samples.
- Compares observed data order statistics with those expected from a normal distribution.
- Lower W statistic values indicate deviations from normality.

Formula

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

- $x_{(i)}$: i-th smallest number in the sample.
- $\bar{x}$: sample mean.
- a_i : constants derived from means, variances, and covariances of order statistics.

Implementation

- Computation of constants a_i involves a complex process, often done using statistical software.
  - These constants depend on sample size and expected values of order statistics for a standard normal distribution.
- Can also be implemented in Microsoft Excel using tools like those developed by Kristopher McGinnis of LACSD.

11

Dixon’s Outlier Test

Calculation of the ratio statistic and determination of outliers

Identification Process and Formula

- Observations sorted by magnitude.
- Ratio (Gap / Range) computed using sample size-dependent formula.
  - Gap: absolute difference between outlier and nearest value.
  - Range: difference between data set's maximum and minimum values.
- $$r_{j,i-1} = max\{\frac{x_n - x_{n-j}}{x_n - x_i}, \frac{x_{1+j} - x_1}{x_{n-i} - x_1}\}$$

Implementation and Example

- Different sample size ranges have specific critical values.
- Excel facilitates computation and analysis.

Example: 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.

- $r_{11} = \frac{(X_{10} - X_9)}{(X_{10} - X_2)} = \frac{7.46 - 5.22}{7.46 - 2.80} = 0.48$
- As $r_{11} = 0.48$ exceeds the critical value of 0.477 for $N = 10$ at the 5% significance level, 7.46 is considered an outlier.

14

Rosner’s Test

Calculation of extreme Studentized deviates and comparison with critical values

Background

- Designed for datasets with 25+ samples, assuming normal distribution or transformed data.
- Transformed data enhances reliability by normalizing distributions.
- Detects up to 10 outliers; robust against hidden outliers.

Implementation

- Specify upper limit (k) for potential outliers.
- Remove extreme data points iteratively; recalculate test statistic.
- Utilize provided table or linear interpolation for critical values.

Formula and Example

- $R_{i+1} = \frac{|x^{(i)} - x_m^{(i)}|}{s^{(i)}}$
- λ_{i+1} : tabled critical value for comparison with R_{i+1}
- R_{i+1} : test statistic identifies outliers from normal distribution

Example: dataset of log(TSP) air data ($n = 55$) arranged in ascending order. Detect 3 outliers ($k = 3$) with a 5% significance level.

- Computed values $y_m^{(i)}$, $s_y^{(i)}$, and $R_{y,i+1}$
- Conclusion: no outliers within assumed lognormal distribution

18

Rosner’s Test Formulas

Algorithm for determining the number of outliers based on the calculated test statistics

Algorithm

- Conducts separate tests for potential outliers up to the specified upper bound, denoted as r.
- Assumptions: n - k observations from the same normal distribution, while k most extreme may be outliers.
- Utilizes extreme observation statistics R_1 to R_k for outlier detection.

Formula

- Calculation of extreme observation $x^{(i)}$ and standard deviation $s^{(i)}$.
- Critical values computed using the p-th quantile of Student's t-distribution with v degrees of freedom.
- Algorithm iteratively compares R_k with k and R_{k-1} with $k-1$ to identify outliers.
- Utilizes R programming language's Environmental Statistics package functions.

Study Findings

- Rosner's analysis (1983) using 1,000 simulations presents Type I error rates for various sample sizes (n) and declared maximum outliers (k).
- Concluded that for Type I error level of 0.05, α levels approximate 0.05 if $n \geq 25$.
- Rosner's Generalized ESD Test provides robust outlier detection for large datasets with normal distribution approximations, aiding in accurate data analysis and interpretation.

22

References

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.