ROBUST MEASURES OF LOCATION IN WATER-QUALITY DATA

Kyung-Sub Kim[1], Bomchul Kim[2], and Jin-Hong Kim[3]

[1] Dept. of Env. Ngineering, Hankyong National University, Kyonggi-Do, Korea

[2] Dept. of Env. Science, Kangwon National Univrsity, Kangwon-Do, Korea

[3] Department of Civil Engineering, Chung-Ang University, Kyonggi-Do, Korea

Abstract: The mean is generally used as a point estimator in water-quality data. Unfortunately, the nonnormal and skewed distributions of data hinder the direct application of the mean, which is inappropriate statistics in this case. The use of robust statistics such as L, M, and R-estimators are recommended and become more efficient. The median (L-estimator), the biweight (M-estimator), and the Hodges-Lehmann method (R-estimator) are briefly introduced and applied in this paper. From the actual data analyses, it is known that the median does not guarantee robustness for a small number of data sets, and robust measures of location or the arithmetic mean without outliers are highly recommended if the distribution has tails or outliers. Care must be taken to measure the location because water quality level within a water body can change depending on the selected point estimator.

Key Words: statistics, point estimator, biweight, R-estimator, outlier

1. INTRODUCTION

The water quality level of streams and reservoirs can be inferred from the point estimator of measured data. The arithmetic mean is generally employed as the location statistic and is a good measure of central tendency when the data set has a normal distribution. Researchers, however, have found that distributions of actual environmental data do not follow the normal distribution and sometimes have straggling tails (Reckhow and Chapra, 1983). In such cases, statistics such as the mean and standard deviation of the data having a nonnormal and skewed distribution may be inappropriate measures. Since many situations exist within the water quality data where we are unable to determine the shape of the true distribution, robust and nonparametric statistics can be quite useful (Mosteller and Tukey, 1977; Reckhow *et al.* 1990). Robust statistics are also less susceptible to the influence of outliers (Sprent, 1998). Barnett and Lewis (1994) list 48 tests for outliers in a normal distribution.

The mean, median, trimmed means and Windsorized means are categorized as L-esti-

L-estimators in which linear functions of the order statistics are involved. The trimmed mean is used to handle a long tailed-distribution. The observations are arranged in ascending order, and the top and the bottom t % are rejected. An alternative of trimming approach is Windsorization, where extreme observations are rounded to the value of the nearest remaining observation, thus reducing their influence (Sprent, 1998). M-estimators, which are similar to the optimal maximum likelihood estimators, use the distance of a data point from the center of the distribution (Hoaglin et al., 1983). These estimators have a mechanism for reducing the effect of outliers. One estimator employed is the biweight. Another type of robust estimator is the R-estimators, which is based on ranked data (Hettmansperger, 1984; Staudte and Sheather, 1990).

In this paper, the several methods of robust measures of location, such as the biweight and the Hodges-Lehmann's R-estimator, are introduced and applied to real environmental data consisting mainly of BOD. Analysis of each estimator was conducted to help the researcher understand the statistical problem.

2. ROBUST MEASURES OF LOCATION

Several robust measures of location are briefly described with an introduction to robust estimation of outliers.

2.1 Outlier

The extreme data points can be identified by the several methods (Barnett and Lewis, 1994). Many of them are not good for the detection of

Table 1. Robust Estimation of BOD (mg/L) at Kyungan Stream 5

mean	median	biweight	R-estimator	mean (w/o outlier)	R-estimator (w/o outlier)
6.63	5.00	5.50	5.65	5.70	5.35

Table 2. Weights of Final Iteration in Biweight

BOD	weight
1.8	0.870
3.0	0.939
3.2	0.949
3.9	0.975
4.2	0.983
4.2	0.983
5.8	0.999
6.8	0.983
6.8	0.983
9.8	0.826
13.2	0.501
16.8	0.138

more than one outlier and tend to miss some others. Sprent (1998) recommends the use of the following equation, which is effective in the identification of all outliers in the tails

$$\frac{|x_0 - \text{med}(x_i)|}{\text{med}\left[|x_i - \text{med}(x_i)|\right]} > 5 \quad (1)$$

where med(x_i) is the median of all observations in the sample and x_0 is the value of outlier. The denominator is the median absolute deviation (MAD). The value of 5 is chosen to pick up observations more than 3 standard deviations from the mean.

2.2 Median

The median is a middle point of data set when the data are arranged in order or magnitude. Generally it is used when distributions are skewed or a small number of measurements are given.

2.3 Biweight

M-estimators minimize functions of the deviations of the observations from the estimate. The estimating equation is given by

$$\sum \Psi(x_i, t) = 0 \quad (2)$$

where Ψ is the derivative form of the objective function in each estimator and t is the estimate.

The biweight, which is one of M-estimators and strongly recommended by several authors (Mosteller and Tukey, 1977; Reckhow and Chapra, 1983), iteratively determines location based on a weighting of the data points according to their distance from the center of the distribution. The weights are calculated as

$$w(u_i) = \begin{cases} (1-u_i^2)^2 & |u_i| \leq 1 \\ 0 & \text{elsewhere} \end{cases} \quad (3)$$

with

$$u_i = \frac{x_i - \hat{x}}{cs} \quad (4)$$

where $\hat{x}$ is the estimate of location from a previous iteration, c is a constant usually between 6 and 9 (often 6), and s is a measure of spread given by $\frac{1}{2}I$ where I is the interquartile range. The interquartile range is the difference between the ascending data point at the 75% and 25% level. The weight on a particular data point becomes smaller if that point is found to deviate more from the last iteratively calculated value. The extreme data points in which the absolute value of u is greater than 1 are eliminated from the calculation. When c is equal to 6 in Eq.(4), the difference between the estimate and the data point is three times greater than the interquartile range. The new location is estimated by the weighted approach

$$\hat{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (5)$$

If the new estimation value of location is the same as the previous calculation value, more iteration is not needed.

2.4 R-estimator

When the data sets are ranked from low to high, the R- estimator is then obtained by the following equation

$$\text{med}\left|\beta x_i + (1-\beta)x_j\right| \quad (6)$$

where β is the weighting factor and $x_i < x_j$. The Hodges-Lehmann estimator, which has a strong robustness and efficiency properties, uses β = 0.5. This value performs well for a variety of distributions. Other values of β (0.9 or 2) are used for medium to wide tailed distributions.

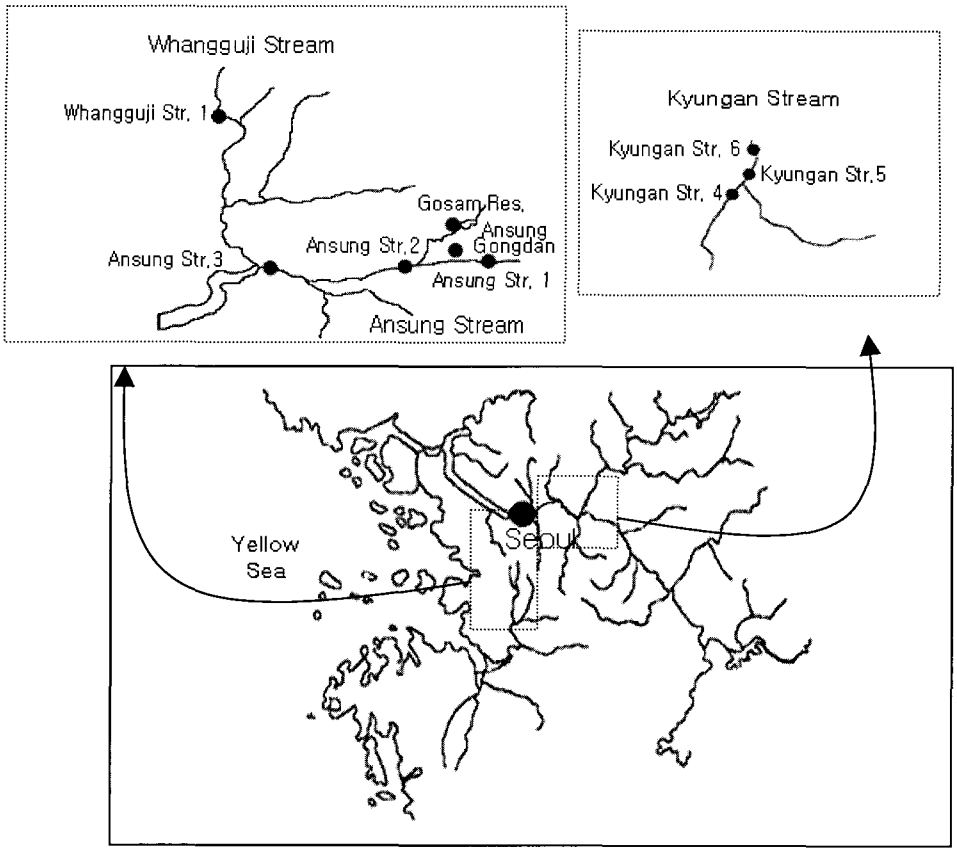

Fig. 1. Location of Sampling Sites

3. APPLICATION OF ROBUST ESTIMATION

An open data set from the Ministry of Environment in Korea is used for the application of robust estimation. Since the monthly based value is applied for only one year, each data set consists of 11 or 12 measurements. One data set from Gosam Reservoir, however, has a total of 33 data points since the data of three different measuring sites are gathered. The total number of data sets is 10 (Table 3). Fig.1 shows the location of sampling sites.

3.1 BOD of Kyungan Stream 5

Kyungan Stream 5 is selected as a representative sampling site for the detailing description of robust measures of location. A Stem and Leaf diagram of BOD taken in 2000 is shown in Fig. 2.

Left and right column of Fig.2 represents one digit number and decimal point value, respectively. From the plot we know that the distribution has a straggling tail and a big difference exists between the arithmetic mean (6.63 mg/L) and the median (5.0 mg/L). The highest value of 16.8 is identified as an outlier from Eq.(1). Thus, robust measures of location are recommended.

The results of each method are summarized in Table 1 including or excluding an outlier. Table 2 shows the weights of the final calculation step in the biweight. The weights were obtained after 5 iteration beginning with the mean. It is known

0	
1	8
2	
3	0 2 9
4	2 2
5	8
6	8 8
7	
8	
9	8
10	
11	
12	
13	2
14	
15	
16	8
17	
18	

Fig. 2. Stem and Leaf Diagram of BOD (mg/L) at Kyungan Stream 5

that each observation has a weight and that the weights of the data point deviated from the center of the distribution, such as 13.2 and 16.8 mg/L, are relatively small. Even though an outlier is detected, the weight of an outlier is assigned in this approach because u of Eq.(4) is less than 1. The robust measure of the biweight is easily obtained from Eq.(5). When using the R-estimator, the total number of the pairs originating from the 12 measuring points becomes 66 After ranking the data, the median is the averaged value of 33rd and 34th data points.

Fig. 3 is a representation of the differences among the robust estimators including the mean. The mean is overestimated and the median is underestimated when compared with the other robust measures. Generally, the biweight, the mean without an outlier, and R-estimator with or without an outlier produces similar results. It is noted that the water quality level of each estimator can be different. The mean shows a level Ⅳ water quality standard, but the estimators based on the biweight and R-estimator show a level Ⅲ. Since the reasonable point estimates by robust statistics have a property of resistance and stretched-tail efficiency, the water quality

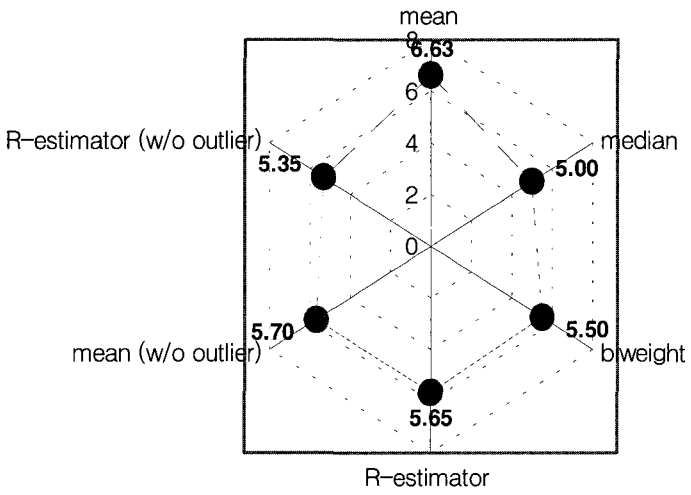

Fig. 3. Diagram of the Robust Estimators at Kyungan Stream 5

Table 3. Robust Estimation of Sampling Sites (mg/L)

No.	sampling site	mean	median	biweight	R-estimator	mean (w/o outlier)	R-estimator (w/o outlier)	variable
1	Ansung Str. 1	5.16	5.00	4.37	4.60	4.36	4.40	BOD
2	Ansung Str. 2	6.74	7.00	6.74	6.55	-	-	BOD
3	Ansung Str. 3	9.56	7.90	9.46	9.65	-	-	BOD
4	Ansung Gongdan	19.95	19.05	19.23	19.80	-	-	BOD
5		29.15	20.15	19.25	20.78	21.16	20.05	SS
6	Kyungan Str. 4	8.55	6.75	7.56	7.75	7.31	7.15	BOD
7	Kyungan Str. 5	6.63	5.00	5.50	5.65	5.70	5.35	BOD
8	Kyungan Str. 6	4.29	3.95	4.28	4.30	-	-	BOD
9	Whangguji Str. 1	52.69	54.20	55.16	53.95	-	-	BOD
10	Gosam Res.	0.041	0.036	0.033	0.037	0.038	0.036	T-P

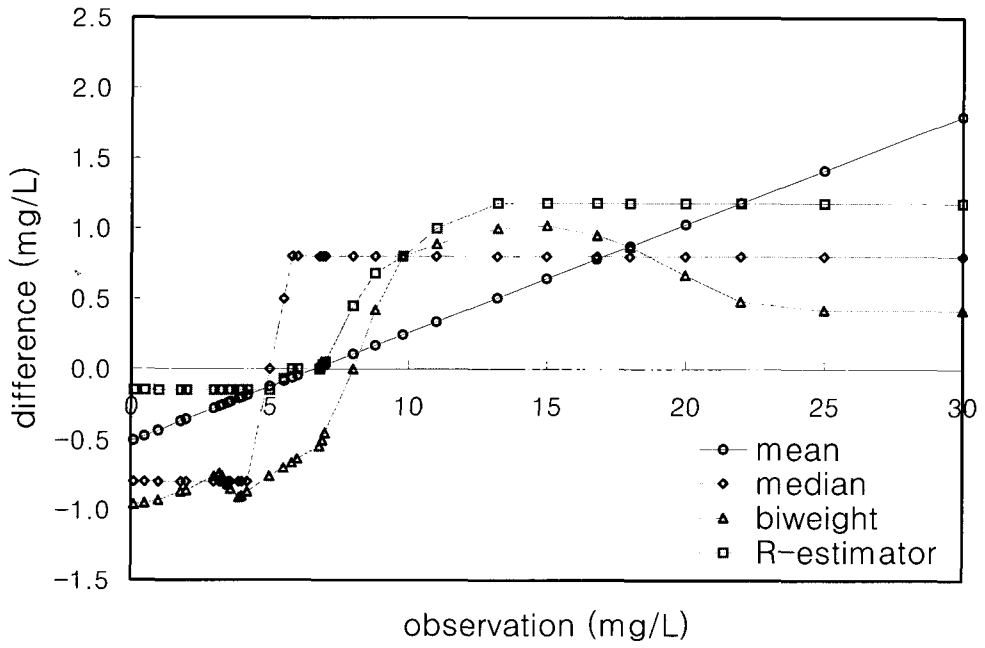

Fig. 4. Influence Curves at Kyungan Stream 5

level of Kyungan Stream 5 can be inferred as a level Ⅲ.

The influence curves, which are made by adding one observation value in the sample, is drawn in Fig. 4 where we plot the value of an adding observation versus the difference between the original and new estimate. It is known that the best estimator for the case of adding one large value is the biweight because of the small difference of Y axis. The maximum influence of the biweight occurs around 15 mg/L. The graphs follow the general shape of influence curves.

3.2 Other Data Set

Robust measures of 9 other sampling sites are given in Table 3. Water quality levels of Ansung Gongdan (SS) and Kyungan Stream 4 is changed according to the applied point estimator. The mean and the other robust estimators are very close if outliers are discarded. If the data set has outliers or the distribution of data has tails, the arithmetic mean is not a good point estimate and is generally higher than any other estimates except for Whangguji Stream 1. This shows that the outliers are located in a higher tail. The mean after discarding outliers and the use of robust measuring techniques such as the biweight and R-estimator are highly recommended for this case. The median which comes from a small number of data points causes some deviation from other estimates whether outliers exist or not.

4. CONCLUSIONS

The mean is generally used as a point estimator in water-quality data, but the nonnormal distribution with long tails or outliers make it difficult for direct application of the mean. Robust measures of location are briefly introduced and applied to a real field data set in this paper.

Several features are of note :

(1) Outliers

It is important to know if the data sets contain outliers. If outliers exist, robust point estimation or the arithmetic mean excluding outliers is recommended. R-estimators with and without outliers do not vary greatly.

(2) Robust Estimation

The median of wide spread or a small number of data set does not guarantee robustness in statistics. The biweight and R-estimator gives us reasonable results in this case. Even though the biweight requires more calculation effort, it is highly recommended.

(3) Water Quality Standard

Water quality levels of a water body can be different according to the applied point estimator. Thus, care must be taken to measure the location. Reasonable conclusions should be reached after reviewing the estimates of several robust statistics.

REFERENCES

Barnett, V., and Lewis, T. (1994). *Outliers in statistical data.* 3rd. edn. John Wiley & Sons, Inc., New York, NY., pp. 216-250.

Hettmansperger, T.P. (1984). *Statistical inference based on ranks.* John Wiley & Sons, Inc., New York, NY., pp. 12-17.

Hoaglin, D.C, Mosteller, F., and Tukey, J.W. (1983). *Understanding robust and exploratory data analysis.* John Wiley & Sons, Inc., New York, NY., pp. 339-347.

Mosteller, F., and Tukey, J.W. (1977). *Data analysis and regression; A second course*

in Statistics. Addison-Wesley, MA., pp. 203-219.

Reckhow, K.H., and Chapra, S.C. (1983). *Engineering approaches for lake management, Vol. 1 : Data analysis and empirical modeling*. Butterworth, Woburn, MA., pp. 85-96.

Reckhow, K.H., Clements, J.T., and Dodd, R.C. (1990). "Statistical evaluation of mechanistic water-quality models." *Journal of Environmental Engineering*, ASCE, Vol. 116, No. 2, pp. 250-268.

Sprent, P. (1998). *Data driven statistical methods*. Chapman & Hall, London., pp. 57-75.

Staudte, R.G., and Sheather, S.J. (1990). *Robust estimation and testing*. John Wiley & Sons, Inc., New York, NY., pp. 119-121.

Kyung-Sub Kim, Associate Professor, Dept. of Envir. Engineering, Hankyong National University

(E-mail : kskim@hnu.hankyong.ac.kr)