京都大学防災研究所年報　第 50 号 B　平成 19 年 4 月
Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No. 50 B, 2007

Suitability of Sample Size for Identifying Distribution Function in Regional Frequency Analysis

Binaya Kumar MISHRA*, Yasuto TACHIKAWA* and Kaoru TAKARA

*Graduate school of Urban and Environment Engineering, Kyoto University

Synopsis

Estimation of environmental extremes using regional frequency analysis needs fitting of an appropriate frequency distribution function representing a homogeneous region. Best fit frequency distribution depends on the computed moment coefficients using either conventional method of moment or recently developed method of L-moment. These moment coefficients depend on the length of observed data record. This paper examines the deviation of moment coefficients with the length of record in terms of root mean square error (RMSE). It shows that percentage deviation is about 50% for sample size of 20 and decreases to near or below 20% for sample size of 50.

Keywords: Distribution function, L-moments, Sample size, Moment coefficients

1. Introduction

Reliable estimation of the hydro-metrological extremes for the given return period is of particular importance in planning and design of hydraulic structures. Frequency analysis relates the magnitude of extreme events to their frequency of occurrence through the use of probability distributions. It needs a large number of historical observed data at the place of interest. In practice, either there is no data or is of very short length. In such cases, regional frequency analysis can be an effective tool.

Estimation of frequency distribution function representing a homogeneous region is one of the important major steps in regional frequency analysis. Best fitting frequency distribution function representing the regional data can be determined from the plot of skewness versus corresponding kurtosis using either conventional method of moments or recently developed L-moment method against various frequency distribution function representing lines.

A diagram based on C_s and C_k, such as that in Figure 1, can be used to identify appropriate distributions in conventional methods. A diagram based on LC_s and LC_k, such as that in Figure 2, can be used similar to conventional moment ratio diagrams to identify appropriate distributions in the method of L-moment. The location of sample estimate with respect to the distribution gives an indication of the suitability of a distribution of the data. A suitable parent distribution is that which averages the scattered and around which the points spread consistently.

However, if the sample size is small, the bias in the values of higher moments may be large enough to give misleading results (Rao & Hamed, 2000). Misleading of best fitting distribution function takes place since moment ratios (coefficient of variance, skewness and kurtosis) varies largely with the length of data or sample size. Therefore, it is very much important to have the suitable length of data (sample size) for identifying best-fit distribution function in regional frequency analysis.

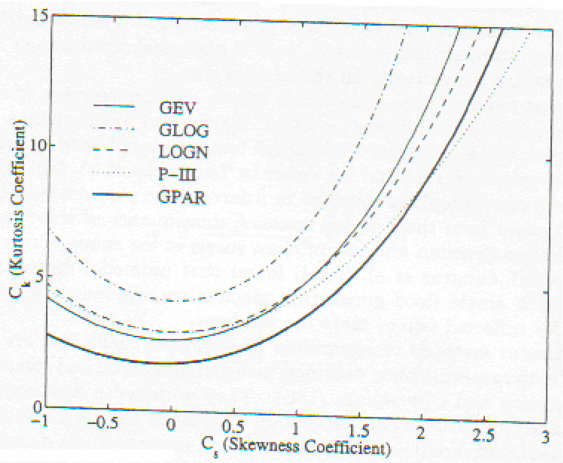

Fig. 1 C_s-C_k moment ratio diagram (Rao & Hamed, 2000)

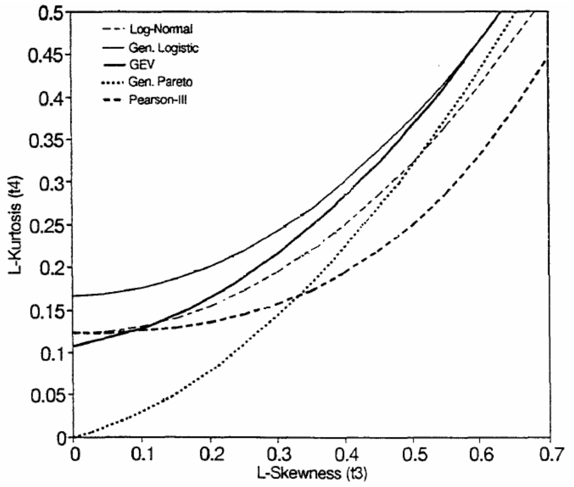

Fig. 2 LC_s-LC_k moment ratio diagram (Rao & Hamed, 1994)

In the present paper, it is intended to show the fluctuation pattern of moment coefficients (skewness and kurtosis) in terms of root mean square error (RMSE) with the change in sample size using conventional method of moments and recently developed method of L-moment. This fluctuation pattern will provide guidance for determining sample size in regional frequency analysis.

2. Methodology

Data screening of the collected data to check for independence, homogeneity, outliers *etc.*, calculation of moment coefficient for data set of different sample size and finally calculation of root mean square error (RMSE) are the major steps involved to carry out the present study. These steps have been discussed in the following sections.

2.1 Data screening

Annual maximum daily rainfall values of Hirakata, Gojyou, Kameoka and Katada rain-gauge stations lying in the Yodo River basin region, Japan have been used to carry out the job. All these data were made available from the data bank of Innovative Disaster Prevention Technology and Policy Research Lab, Disaster Prevention Research Institute, Kyoto University, Japan. At every considered rain-gauge station, sub-samples were formed from the whole sample considering continuous dataset of different sample size varying from 20 to 80 at an increment of 10. Test for independence and homogeneity (assumption that the whole set data come from the same distribution) of the data was performed as discussed by Mann-Whitney (1947) and checked at 5% significance level. Test for outlier (an observation that deviates largely from the bulk of the data) was performed as discussed by Grubbs (1969) and checked at 5% significance level.

2.2 Moment coefficients

The computation relationships for mostly used moment ratios namely coefficient of variance (C_v or LC_v), coefficient of skewness (C_s or LC_s) and coefficient of kurtosis (C_k or LC_k) have been discussed in following subsections.

2.2.1 Conventional moments

Moment coefficients, particularly coefficient of skewness and coefficient of kurtosis, are used to characterize regional probability distributions. If x_i is observed rainfall data at station i and n is the no. of rainfall records, these moment coefficients are calculated as:

Sample mean, $m_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance, $m_2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Standard deviation, σ = (variance)$^{1/2}$

Third moment, $m_3 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i - \bar{x})^3$

Fourth moment, $m_4 = \frac{n^2}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(x_i - \bar{x})^4$

Moment coefficients i.e. moment ratios are calculated as given below:

Coefficient of variation, C_v = σ/ mean $= m_2^{1/2} / \bar{x}$

Coefficient of skewness, $C_s = m_3/ \sigma^3$

Coefficient of Kurtosis, $C_k = m_4/ \sigma^4$

2.2.2 L-moments

L-moments (Hosking and Wallis, 1997) based method is a simpler and effective alternative as compared to conventional method to select the best-fit distribution function for a region. For the ordered sample $x_{j=1} \leq x_{j=2} \leq ---- \leq x_{j=n}$ at any station, the first four L-moments are calculated as:

$$l_1 = b_0,\ l_2 = 2b_1 - b_0$$

$$l_3 = 6b_2 - 6b_1 + b_0$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0$$

Where, $b_o = \frac{1}{n}\sum_{j=1}^{n} x_j$

$$b_1 = \frac{1}{n(n-1)}\sum_{j=2}^{n}(j-1)x_j$$

$$b_2 = \frac{1}{n(n-1)(n-2)}\sum_{j=3}^{n}(j-1)(j-2)x_j$$

$$b_3 = \frac{1}{n(n-1)(n-2)(n-3)}\sum_{j=4}^{n}(j-1)(j-2)(j-3)x_j$$

L-moment ratios are expressed as follows:

L-coefficient of variation, $LC_v = l_2/l_1$

L-coefficient of skewness, $LC_s = l_3/l_2$

L-coefficient of kurtosis, $LC_k = l_4/l_2$

2.3 Error calculation

Bias and root mean square error (RMSE) are common measures of the performance of an estimator (Hosking & Wallis, 1997). Bias is expressed as mean of the difference between parameter obtained from available sample and the parameter representing the population data set. But RMSE is expressed as mean of $((\theta_s-\theta_p)^2)^{1/2}$ where θ_s and θ_p represents sample parameter and population parameter respectively. Both the bias and RMSE of parameter have the same units of measurement as the parameter. It is convenient to express bias and RMSE as ratios with respect to the parameter itself and hence, we obtain dimensionless measures, the relative bias and relative RMSE. Contributions of negative and positive biases may cancel out to give a misleadingly small value of the bias. Hence, the present study has been dealt in terms of RMSE only as the average relative RMSE measures the overall deviation of sample quantiles from population quantiles.

3. Results and discussion

As discussed in previous section, data were first analyzed to check for independence, homogeneity and outliers. Data sets of all the considered stations were found independent and homogeneous at 5% significance level. For example, value of z, an indicator for independence and homogeneity test by Mann-Whitney (1947) method, was found as 1.94. This value is lesser than 1.96 which is the critical value for sample size greater than 20 at 5% significance level. Hence, the population data set was assumed as independent and homogeneous. The presence of outliers in the data causes difficulties when fitting a distribution to the data. Presence of even a single outlier, may be on upper or lower side, affects the moment coefficient much. For example, deviation in moment coefficient values in presence and absence of outlier at Hirakata station has been compared and shown in table 1. The fall in LC_s and LC_k was found as 23% and 30%; whereas in C_s and C_k was found as 64% and 69% after deleting the single outlier in upper side. The table 1 shows that conventional moment coefficients are much deviating with the presence of outlier as compared to L-moment coefficients. In other words, conventional moment coefficients are more sensitive to the presence of outliers as compared to method of L-moments.

Figures 3-8 shows variation pattern of moment coefficients (skewness and kurtosis) at Hirakata as an example. In general, similar variation pattern were obtained at all considered rain-gauge stations. Considering all the results, error in both conventional and L-moment coefficient of skewness and kurtosis is very large (points are deviated largely from one another) for sample size of 20. Error in moment coefficients is reduced largely to a sample size of 50 or more and hence, deviations in between points are smaller as shown in Figures 3, 4, 5 and 6.

As shown in Figures 7 & 8, points plotted for skewness versus kurtosis for dataset of smaller sample size deviated very much from each other as well as from the point of whole sample size and got concentrated as the sample size increased to 50. Deviation in moments coefficients for different sub-samples size from that of the whole sample (94 years) has been shown numerically in table 2 as an example at Hirakata in terms of RMSE.

Table 1 Comparison of moment coefficient due to presence of outliers at Hirakata, Japan

Moment Coeff.	Without outlier	With outlier	Error	% error
LC_s	0.158	0.194	0.370	23.332
LC_k	0.109	0.141	0.032	29.595
C_s	0.709	1.158	0.450	63.526
C_k	3.004	5.071	2.067	68.791

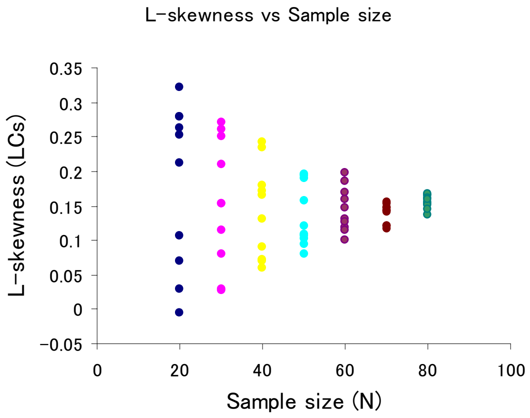

Fig. 3 L-moment skewness versus sample size at Hirakata, Japan

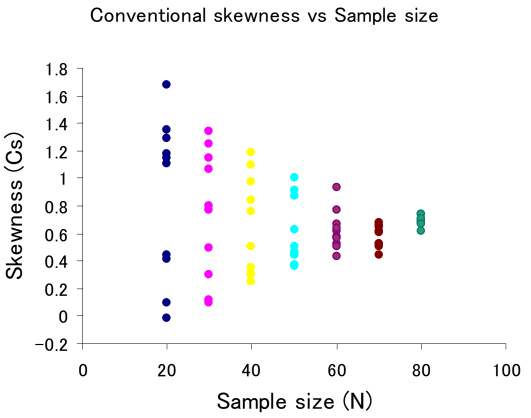

Fig. 4 Conventional moment skewness versus sample size at Hirakata, Japan

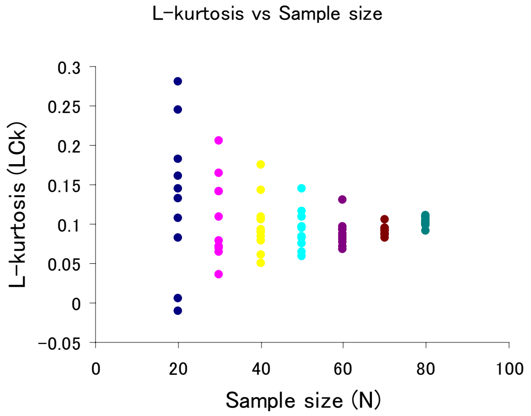

Fig. 5 L-moment kurtosis versus sample size at Hirakata, Japan

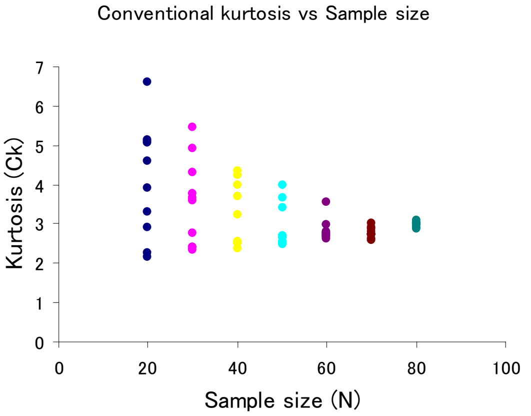

Fig. 6 Conventional moment kurtosis versus sample size at Hirakata, Japan

Table 2 Moment coefficient for different sample size dataset at Hirakata, Japan

Sample	L-moment		Conventional moment		L-moment		Conventional moment	
size	RMSE in LC_s	RMSE in LC_k	RMSE in C_s	RMSE in C_k	% error in LC_s	%ge error in LC_k	% error in C_s	% error in C_k
20	0.115	0.091	0.574	1.747	48.58	52.10	59.93	44.12
30	0.087	0.051	0.444	1.178	32.54	36.77	48.33	25.99
40	0.066	0.037	0.343	0.783	25.89	27.37	38.32	18.05
50	0.048	0.030	0.253	0.522	18.62	19.93	22.28	12.07
60	0.033	0.027	0.160	0.312	11.82	11.78	17.89	8.17
70	0.023	0.019	0.143	0.257	6.77	9.78	16.43	7.36
80	0.009	0.008	0.039	0.062	4.79	6.79	9.064	5.067

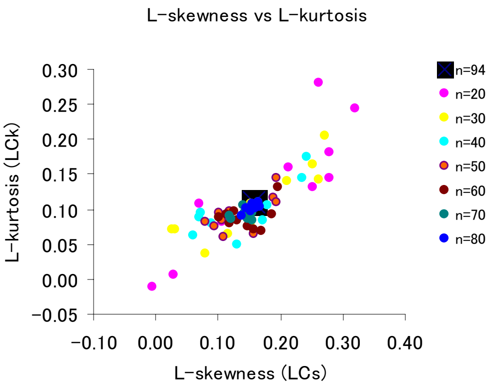

Fig. 7 LC_s-LC_k for different sample size at Hirakata, Japan

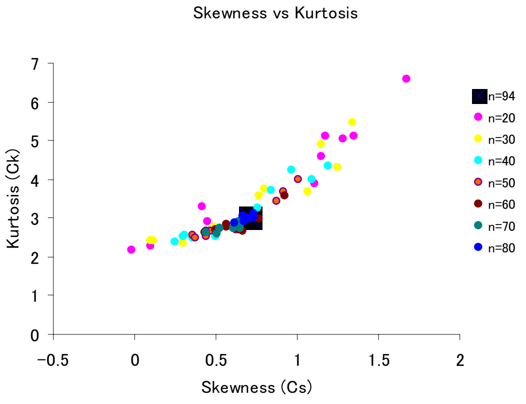

Fig. 8 C_s-C_k for different sample size at Hirakata, Japan

While calculating the value of root mean square error (RMSE) for each data set, population coefficients were assumed as that resulted from the whole data set which has length of more than 90 years. It shows that average % deviation is about 50% for sample size 20 and decreases near or below 20% as the sample increases to 50.

4. Conclusions

Findings of the paper are as follows:

1. Conventional moment coefficients are more sensitive to outlier observation as compared to L-moment coefficient.

2. Deviation pattern of moment coefficients (skewness and kurtosis) with respect sample size is similar for both methods showing there is not much comparative advantages of L-moments method over conventional in terms of sample size.

3. Identified frequency distribution function with the help of data set of sample size 40 or less may not be best-fit frequency distribution representing a region.

References

Grubbs, F.E. (1969): Procedures for detecting outlying observations in samples, Technometrics, Vol. 11 (1), pp. 1-21.

Hosking, J.R.M. and Wallis, J.R. (1997): Regional Frequency Analysis, Cambridge University Press, United Kingdom.

Mann, H.B. and Whitney, D.R. (1947): On a test of whether one of two random variables is stochastically larger than the other, The annuals of mathematical statistics, Vol. 18, pp. 50-60.

Rao, A. R. and Hamed, K.H. (1994): Frequency analysis of upper Cauvery Flood data by L-moments, Water resources management 8:183-201.

Rao, A. R. and Hamed, K.H. (2000): Flood Frequency Analysis, CRC Press LLC, Florida.

地域頻度解析における分布関数同定のためのサンプルサイズの適切性

Binaya Kumar MISHRA*・立川康人*・宝　馨

＊京都大学大学院工学研究科都市環境工学専攻

要　旨

地域頻度解析において極値水文量を推定するためには，同質と考えられる地域を代表する適切な頻度分布関数を決定する必要がある。頻度分布の母数推定には，積率法やL積率法が用いられ，推定された母数の値は同定に用いる時系列データの期間に依存する。本研究では，観測時系列データの長さによって現れる母数推定値の違いをRMSEを用いて分析する。その結果，20サンプルを用いた場合には50％近い母数推定値の違いが現れ，50サンプルとするとそれが約20％に減少することがわかった。

キーワード：分布関数，L 積率法，サンプルサイズ，モーメント係数