GeoJournal 40.1–2: 209–212.

A robust-symmetric mean: A new way of mean calculation for environmental data

Zhang, Chaosheng; Zhang, Shen, Institute of Geography, Chinese Academy of Sciences, Beijing, P.R. China

Received 8 August 1995; accepted 31 December 1995

Abstract: Environmental data seldom follow normal distributions, so how to calculate their means is a very important problem. Commonly used methods for mean calculation, such as arithmetic mean, geometric mean, and median, were evaluated. Arithmetic means should only be used when datasets follow normal distributions. Geometric means are suitable for datasets which follow log-normal distributions. Medians are a kind of robust treatment. However, their 'efficiency' is very low. Based on the methods described, two new ideas are developed: 'robust' and 'symmetric', for calculating means. As far as the symmetric feature is concerned, Box-Cox power transformation is better than logarithmic transformation. Robust statistics and Box-Cox transformation are combined to produce the 'robust-symmetric mean'. As environmental data often follow log-normal or skewed distributions, this method is better than the previous ones and also is appropriate.

1. Introduction

One of the most important steps for environmental study is to obtain various environmental data. However, after obtaining the data, their analysis in an appropriate, scientific and statistical way often remains a problem. The scientific presentation of data is the basic requirement for environmental investigation, and is also the basis for statistical treatment. It is important to examine how to present the data in the appropriate scientific way.

There are various commonly used methods of mean calculation for a dataset, such as arithmetic mean, geometric mean, and median. However, all of these means have their own limitations. Arithmetic means should only be used for a dataset which follows a normal distribution, and geometric means, for a log-normally distributed dataset. Medians are in a way trimmed means, but they lose much information, and their efficiency is very low. We have studied these means, and propose a new way for mean calculation, which is called the 'robust-symmetric mean'.

2. Pretreatment of environmental data

To calculate the mean of an environmental dataset, it is important to detect outliers. Ordinarily, environmental investigations are carried out over wide areas, and it is possible that some sampling sites are located in polluted, or mineralized areas, and during the sample treatment procedure, artificial pollution might also be introduced. These factors could introduce outliers for an environmental dataset. The outliers have disadvantageous effects on statistical results. For aquatic environments, this situation is very serious. In order to let the statistical results reflect the common features of natural conditions, the outliers usually should be detected and rejected. However, if the objective of study is to seek 'outliers', such as mineral exploration, it is another case. There are many methods for the detection of outliers, such as Grubbs, t-test, and Dixon methods. Histograms for frequency distribution are also available for the detection of outliers. We would like to emphasize that the detection of outliers should not only be limited to statistical methods, and environmental analysis should also be involved. The outliers might not only be rejected, but also might be replaced by some kind of alternative values (Sanford et al. 1993).

The probability distribution characteristics of a dataset should also be studied, as many statistical techniques are based on particular distributions, and especially the assumption of normality. There are many methods for normal distribution tests, such as Kolmogorov-Smirnov (K-S), Shapiro-Wilk tests. Random variables often follow normal distributions. Environmental variables, however, rarely follow normal distributions, as they are regionalized variables. Many studies have shown that geochemical variables often follow a log-normal distribution (Krige 1951, 1960; Sichel 1952, 1966; Miesch and Riley 1961; Miesch 1967, 1976). For the application of statistical analysis, they should be transformed to satisfy the normality requirement. There are many techniques for normal transformation, such as logarithmic transformation (for log-normally distributed data), square root transformation (for data which follow a Poisson distribution), and Box-Cox transformation (for any distribution).

3. Commonly used means

After the treatment of outliers and transformation, statistical techniques are then applied. For the calculation of means, the arithmetic mean, the geometric mean, and the median are commonly used.

The arithmetic mean should only be applied to a dataset which follows a normal distribution, otherwise, large biases will be introduced. For example, for a dataset which follows a log-normal distribution, the arithmetic mean is larger than the geometric mean (such as in Table 1).

The geometric mean can be applied to a dataset which follows a logarithmic distribution. However, the 'logarithmic distribution' may be too artificial. For many cases, the distributions are 'pseudo-logarithmic' (Chapman 1976), and usually a dataset does not follow logarithmic distribution, which is called a skewed distribution. For these cases, logarithmic transformation is not optimal, so a geometric mean is not appropriate.

The main reason for the normality requirement of many statistical techniques is that the arithmetic mean calculation is often involved during the calculation procedures. The arithmetic mean should be appropriate. In fact, if the dataset distributes symmetrically (with a skewness of 0), the arithmetic mean is appropriate for the dataset. The Box-Cox power transformation should make a dataset distribute symmetrically (Jobsen 1991, pp. 68–74). In this case, the Box-Cox transformation is better than a logarithmic transformation. The main limitation of the Box-Cox transformation is the difficulty in carrying it out, which involves a relatively complicated calculation.

The calculation of the median is very simple. One just sorts the dataset, and the value in the middle is called median. If the number of the dataset is even, the median is the average of the two data points in the middle. This method is quite useful when data vary strongly, as the median is not sensitive to outliers. However, the method loses too much information of raw data, so it is not effective.

An alternative way for median calculation is the trimmed mean. One does not trim 50% of both high values and low values, but only some smaller percentage, such as 5%. This method could both avoid effects of potential outliers and yet retain most information in the raw data. It presents the idea of 'robustness', and 'trimming' is one way for robust treatment. However, the probability distribution of the trimmed data should also be studied.

4. A new way for mean calculation

All of the above commonly used methods for mean calculation have their own limitations. However, they

Table 1. Arithmetic mean, geometric mean, median, and robust-symmetric (R-S) mean of various element contents in raw sediments of the Yangtze River system, China

Element	Sample number	Arithmetic mean	Geometric mean	Median	R-S mean
Cu	248	26.7	22.7	21.6	21.5
Pb	249	23.7	21.5	21.8	21.4
Zn	251	80.1	73.9	72.8	73.6
Cd	243	0.215	0.153	0.140	0.148
Hg	250	0.044	0.032	0.034	0.034
Co	254	13.8	12.1	12.2	12.1
Ni	251	29.3	26.0	26.5	26.4
As	255	8.8	7.3	7.6	7.6
Cr	251	61.0	51.0	54.0	52.3
Mn	252	666	609	570	589
Fe	250	3.25	2.95	2.94	2.94

provide valuable information. For mean calculation, a dataset should be (1) symmetrically distributed and (2) robust. The 'symmetric' requirement should be completed by Box-Cox Transformation, and trimming is one way for getting 'robustness'.

The 'robust' requirement comes not only from the median and trimmed means, but also from the transformation methods. When a dataset follows a log-normal distribution, a geometric mean could be used (Bakr et al. 1978; Dagan 1979, 1981). However, the applicability of the geometric mean has been questioned by some authors (King 1988, 1989; Richardson 1990). The geometric mean is sensitive to small values, so it is not robust. Logarithmic transformation could reduce the weightings of high values, however, in the meantime, it increases the weightings of low values too much. Jensen (1991) proposed a method of j^{th} Winsorized mean to solve this problem. Here, we suggest a robust treatment combined with Box-Cox transformation to solve this problem.

Based on the above discussion, we propose a new method, the 'robust-symmetric mean', for mean calculation with the following procedure: (1) sort the raw dataset; (2) trim the dataset, such as 5% at both sides of low values and high values, so as to make the dataset more robust; (3) apply the Box-Cox transformation (Box and Cox 1962; Jobson 1991) to the trimmed data, in order to make the dataset distribute symmetrically, and to avoid the possibility of a pseudo-lognormal distribution (Chapman 1976); (4) calculate the arithmetic mean of the dataset which has been trimmed and transformed; and (5) retransform the arithmetic mean by the reverse process of a Box-Cox transformation.

5. Box-Cox transformation

The Box-Cox transformation is given by

$$\begin{aligned} y &= (x^{\lambda} - 1)/\lambda \quad && \lambda \neq 0; \\ &= \ln x && \lambda = 0;\ x > 0. \end{aligned}$$

For a given dataset ($x_1, x_2, \ldots, x_n$), the λ value should be estimated based on the assumption that the transformed values $y_1, y_2, \ldots, y_n$ are normally distributed (Jobson 1991, pp. 68–69). The maximum likelihood function for λ is:

$$L_m(\lambda) = -1/2n \ln \tilde{\sigma}_z^2$$

where

$$\tilde{\sigma}_z^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2/n,$$

$$z_i = x_i^{\lambda}/\tilde{x}_G^{\lambda} \text{ and } \tilde{x}_G = \text{the geometric mean of } x_1, \ldots, x_n = (x_1 x_2 \ldots x_n)^{1/n}$$

6. Case study: Heavy metal content in sediments of the Yangtze River system

During the seventh five-year plan of China (1986–1990), a massive investigation was carried out on the Yangtze River system. About 260 sediment samples were taken and detected for heavy metal contents (including As). We utilized these data for an example, showing the differences between commonly used means and the robust-symmetric mean (Table 1). Most of these datasets followed the log-normal distribution. The means were calculated after outliers were rejected. The robust-symmetric means were calculated based on both 5% of high values and low values being trimmed, and the Box-Cox transformation being done.

It is obvious that all the arithmetic means are much higher than the other means, showing that the arithmetic mean should be rejected for these data. The differences between arithmetic averages and robust-symmetric means are between 10.5% (for Fe) and 45.3% (for Cd), with an average value of 18.1%. The differences among the other means are relatively low, showing that means from the new method are not so different from commonly used methods. However, some differences do exist. The differences between geometric means and robust-symmetric means range from 0 (Co) to 5.9% (Hg), with the average value of 2.5%. The differences between medians and robust-symmetric means are within 0 (Hg, As, and Fe) to 5.4% (Cd), with the average value of 1.5%. Based on the above discussion, the robust-symmetric mean is recommended. Details about treatment of the data are available in Zhang et al. (1995).

7. Conclusions

The arithmetic mean should only be used when a dataset follows a normal distribution. Geometric mean calculation is acceptable for a dataset which follows a log-normal distribution. The median is one type of robust treatment. However, its efficiency is very low. These methods provide two important features for mean calculation: 'robustness' and 'symmetry'. As far as the symmetric feature is concerned, the Box-Cox power transformation is better than the logarithmic transformation. Robust statistics and Box-Cox transformation are combined to produce a 'robust-symmetric mean'.

The new method of 'robust-symmetric mean' was applied to calculation of the heavy metal content in raw sediments of the Yangtze River system, and satisfactory results were obtained. The differences between robust-symmetric means and arithmetic means are very large (with the average value of

18.1%), showing that arithmetic means should be rejected. The differences among the robust-symmetric means, geometric means and medians are small, with the largest value being less than 6%, showing that the new method is consistent with the commonly used methods. However, differences do exist. Based on the discussion for these methods, the robust-symmetric means are the most reasonable.

References

Bakr, A. A.; Gelhar, L. W.; Wutjahr, A. L.; MacMillan, J. R.: Stochastic analysis of spatial variability in subsurface flows 1: Comparison of one and three-dimensional flows. Water Resour. Res. 14, 263–271 (1978).

Box, G. E. P.; Cox, D. R.: An analysis of transformations. Journal of the Royal Statistical Society, Series B 26(2), 211–252 (1962).

Chapman, R. P.: Some consequences of applying lognormal theory to pseudolognormal distributions. Mathematical Geology 8(2), 209–214 (1976).

Dagan G.: Models of groundwater flow in statistically homogeneous porous formations. Water Resour. Res. 15, 47–63 (1979).

Dagan G.: Analysis of flow through heterogeneous random aquifers by the model of embedding matrix 1: Steady flow. Water Resour. Res. 17, 107–121 (1981).

Jensen J. L.: Use of geometric average for effective permeability estimation. Mathematic Geology 23(6), 833–840 (1991).

Jobson, J. D.: Applied Multivariate Data Analysis. Vol. I: Regressing and Experimental Design. Springer Verlag, New York 1991.

King, P. R.: Effective values in averaging. In: Edwards S.; King P. R. (eds.), Mathematics in Oil Production, pp. 217–234. Oxford University Press, Oxford 1988.

King, P. R.: The use of renormalization for calculation effective permeability. Transport in Porous Media 4, 37–58 (1989).

Krige, D. G.: A statistical approach to some basic mine valuation problem on the Witwatersrand. J. Chem. Metall. Mining Soc. S. Afr. 52, 119–139 (1951).

Krige, D. G.: On the departure of ore value distributions from lognormal models in South African gold mines. J. S. Afr. Inst. Mining Metall. 61, 231–244 (1960).

Miesch, A. T.: Methods of Computation for Estimating Geochemical Abundance. U.S. Geological Survey Open-file Report, 76–772 (1967). 1140 pp.

Miesch, A. T.: Geochemical survey of Missouri – Methods of sampling, laboratory analysis and statistical reduction of data. U.S. Geological Survey Professional Paper 954-A (1976). 39 pp.

Miesch, A. T.; Riley L. B.: Basic statistical methods used in geochemical investigations of Colorado Plateau uranium deposits. HIMMP Trans. (Mining) 220, 247–251 (1961).

Richardson J. G.: Letter to the Editor, J. Pet. Tech. 42, 1524 (1990).

Sanford, R. F.; Pierson, C. T.; Crovelli, R. A.: An objective replacement method for censored geochemical data. Mathematical Geology 25(1), 59–80 (1993).

Sichel, H. S.: New methods in the statistical evaluation of mine sampling data. London, Inst. Mining and Metall. Trans. 61, 261–288 (1952).

Sichel, H. S.: The estimation of means and associated confidence limits for small samples from lognormal populations. J. S. Afr. Inst. Mining and Metall., Symposium: Mathematical Statistics and Computer Application in Ore Valuation, 106–123 (1966).

Zhang, C. S.; Zhang, S.; Zhang, L. C.; Wang, L. J.: Background contents of heavy metals in sediments of the Yangtze River system and their calculation methods. Journal of Environmental Sciences 7(4), 422–429 (1995).