A Method for Detecting Outliers in Survey Data by Ratio Tests

Cheng Bangwen, Yang Hongjin, Shi Linfen, Wang Yali, Xu Ji
School of Management, Huazhong University of Science and Technology, Wuhan, China
Corresponding author: Yang Hongjin, e-mail: Chengbw@hust.edu.cn

Abstract

A new method based on ratio tests is proposed. This paper discusses the basic principle and basis of the method, explores the mechanism, condition, and the way to improve its detection effectiveness. It has been shown that when a certain condition is satisfied the ratio indicator will have lower coefficient of variation so that outliers can more easily be detected. Reasonably selecting ratio indicators is the key to effectively detect outliers. The proposed method is also suitable for data with probability distribution unknown. Empirical study shows that the method is simple, practical, easy to use, and has application value.

Keywords: Quality of statistical data, probability distribution test, coefficient of variation, lognormal distribution, boxplot

1. Introduction

In socio-economic statistics, detecting abnormal data in original survey data is an important measure to improve the quality of statistical data. Much work (Barnett et al. 1994; Hawkins 1980; Victoria et al. 2004) has been done on detecting outliers in univariate and multivariate data, which provide a theoretical foundation and methods for outlier detection. In the socio-economic statistical work, the survey items most frequently encountered are "scale indicators" reflecting the size or scale of research objects, such as yield, output value, fund, personnel, and assets. For this kind of indicators, outliers in original survey data from samples can be identified by univariate methods such as lognormal distribution test (Cheng et al. 2000, 2001), normal distribution test at Cox-Box transformation scale (Box 1964; Hosseini 1998), and boxplots (Tukey 1977). But the power of the methods in detecting outliers is limited (Wang et al. 2009; Xu et al. 2011). Two possible errors commonly occur during the process of searching for outliers. The first error is that some of data, although with larger statistical deviation from its real value, may not be detected as outliers because they are close to the center of probability distribution of data. Also, some data detected as outliers are due to their large scale and not real statistical bias. The main reason for this situation is that the methods cannot exclude the influence of the scale of research objects on data. To overcome this defect, the methods of multivariate outlier detection should be employed. Multivariate methods already got wide uses in many areas, however, in detecting outliers in original data of socio-economic statistics there are still many difficulties, especially outlier detection has become increasingly difficult (Rocke et al. 1996) with the increase in the dimension of the data.

Outlier detection is an independent work step in data collection and processing in socio-economic statistics. The method to be used should be simple, practical, without too much mathematical calculation, not taking up too much time, should be able to check out the mistake place of outliers. To meet this request, a new method based on ratio tests is proposed. This paper discusses the basic principle and basis of the method, explores the mechanism, condition, and the way to improve its detection effectiveness. An empirical study based on the original data of government research

institute survey of China has also been taken.

2. Ratio and its Probability Distribution

A ratio refers to the relative magnitudes of two scale indicators reflecting the scale of research objects, such as the ratio of output value to employees of enterprises (per capita output), the ratio of household income to expenditure (income-expenditure ratio), the ratio of published papers to expenditures of research institutes (input-output ratio). Using different scale indicators can form a variety of ratios. The meaning of the ratio is extremely broad, which can be used to express various concepts.

Let $x_1,\ldots,x_n$ be m-dimensional scale indicators which generally are continuous and positive in value, Q stands for the population composed of research objects. Clearly $x_1,\ldots,x_n$ are random variables defined over Q. Based on theoretical and empirical analyses, Cheng (2003) pointed out that in socio-economic systems, probability distributions of scale indicators $x_1,\ldots,x_n$ are usually approximately joint lognormal, i.e. $\ln x_1,\ldots,\ln x_n$ are normally joint distributed. Let x and y be two scale indicators from $x_1,\ldots,x_n$ and then the ratio $z=x/y$ is also a random variable over Q. Because $\ln z=\ln x-\ln y$ is a linear combination of $\ln x_1,\ldots,\ln x_n$, according to the property of joint normal distribution, $\ln z$ is also normally distributed (Mardia 1979). The Empirical study in section 6 will demonstrate that the conclusion on ratio's lognormal distribution is correct in certain conditions and certain ranges.

The conclusion on lognormal distribution of the ratio of two scale indicators is very important. In the following, unless a special statement is made, it is always assumed that the conclusion is correct.

3. Ratio Statistical Test

According to the lognormal distribution property of ratio indicators, statistical tests may be used to identify outliers in ratio data. An outlier in ratio data, with value out of a normal range, is likely to be caused by the statistical bias of the numerator or denominator of the ratio indicator; in the other hand, a large statistical bias of a scale indicator in an individual often makes its ratio value more abnormal so that it could be detected. Ratio tests are based on the attributes or characteristics of research objects and exclude the interference of individual scale on tests such that can make the tests point at abnormal data more accurately.

It is desirable that statistical tests has good power in outlier detection i.e. with high confidence can detect outliers as many as possible. This problem is directly related to the concentricity of data probability distributions. The more concentrated a distribution is the more easily detected outliers are. The concentricity of probability distributions can be measured by a coefficient of variation which is a dimensionless quantity and can be employed to make a comparison between different probability distributions. In the following, based on analysis of coefficients of variation the mechanism and conditions to improve effectiveness of outlier detection by ratio tests will be revealed.

Let x denote a scale indicator of research objects with lognormal distribution. Its mean and variance (Laha et al. 1979）

$$E(x)=\exp(\mu_x+\frac{1}{2}\sigma_x^2),\quad D(x)=\exp(2\mu_x+\sigma_x^2)\cdot(\exp\cdot\sigma_x^2-1)$$

$$E(\ln x)=\mu_x,\qquad D(\ln x)=\sigma_x^2.$$

Therefore the coefficient of variation of random variable x

$$CV_x=\frac{\sqrt{D(x)}}{E(x)}=\sqrt{\exp\cdot\sigma_x^2-1} \tag{1}$$

Obviously, the coefficient of variation only relates to its $D(\ln x)$.

Assume y is another scale indicator of same research objects, which is also log normally distributed, a ratio indicator $z=x/y$, then

$$\sigma_z^2=D(\ln x-\ln y)=D(\ln x)+D(\ln y)-2Cov(\ln x,\ln y)$$

$$= \sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x \cdot \sigma_y \quad (2)$$

where $\sigma_z^2 = D(\ln z)$, $D(\ln y) = \sigma_y^2$, ρ_{xy} is the correlation coefficient between ln*x* and ln*y*. Obviously, the following relationships are hold

$$(\sigma_x - \sigma_y)^2 \le \sigma_\eta^2 \le (\sigma_x + \sigma_y)^2 \quad (3)$$

Under $\sigma_x > \sigma_y$

$$\text{if } \frac{\sigma_y}{2\sigma_x} > \rho_{xy} \quad \text{then } \sigma_y < \sigma_x < \sigma_\eta,\ CV_y < CV_x < CV_\eta \quad (4)$$

$$\text{if } \frac{\sigma_x}{2\sigma_y} \ge \rho_{xy} \ge \frac{\sigma_y}{2\sigma_x}, \quad \text{then } \sigma_y \le \sigma_\eta \le \sigma_x, CV_y \le CV_\eta \le CV_x \quad (5)$$

$$\text{if } \rho_{xy} > \frac{\sigma_x}{2\sigma_y}, \quad \text{then } \sigma_\eta < \sigma_y,\ CV_\eta \le CV_y \le CV_x \quad (6)$$

Under $\sigma_x < \sigma_y$

$$\text{if } \frac{\sigma_x}{2\sigma_y} > \rho_{xy}, \quad \text{then } \sigma_x < \sigma_y < \sigma_\eta,\ CV_x < CV_y < CV_\eta \quad (7)$$

$$\text{if } \frac{\sigma_y}{2\sigma_x} \ge \rho_{xy} \ge \frac{\sigma_x}{2\sigma_y}, \quad \text{then } \sigma_x \le \sigma_\eta \le \sigma_y, CV_x \le CV_\eta \le CV_y \quad (8)$$

$$\text{if } \rho_{xy} > \frac{\sigma_y}{2\sigma_x}, \quad \text{then } \sigma_\eta < \sigma_x, CV_\eta \le CV_x \le CV_y \quad (9)$$

The above results show that:

a. A ratio indicator may have lower coefficient of variation so that outliers can more easily be detected. However, it needs to meet certain condition. Otherwise the coefficient of variation of the ratio indicator is highest compared with that of the denominator as well as numerator scale indicators so that the ratio test result will become worse.

b. When condition (5) or (8) is satisfied, the coefficient of variation of the ratio indicator is middle, and so is the detection effectiveness of the ratio test. In this case, the ratio test can improve the effectiveness of outlier detection compared with that of the scale indicator of the denominator or numerator with high coefficient of variation.

c. When condition (6) or (9) is satisfied, the coefficient of variation of the ratio indicator is smallest so that the capacity of the ratio test on outlier detection is strongest. In this case, the role of ratio tests to improve detection effectiveness has got most effective exhibition.

4. Outlier detection

Based on above analysis, ratio tests on lognormal distribution could be used to detect outliers in original survey data of samples. Main steps are as follows:

a. Design and selection of ratio indicators. This is the key to effectively detecting outliers. First of all, ratio indicators should have a clear meaning, and concepts should be clear. On this basis, quantitative analyses on previous or current data should be employed to check whether ratio indicators meet condition (5), (6), (8), (9). It is necessary for a selected ratio indicator to meet at least one condition; otherwise the effectiveness of outlier detection will become worse. It is worth to design more candidate ratio indicators, from which ratio indicators for detecting outliers will be screened out by comparing.

b. Outlier Detection. After logarithmic transformation to the data of the ratio indicators screened in step 1, a univariate normal test method selected from the existing variety of methods is used to detect outliers. In accordance with statistical systems, for ease of data quality management, the outlier detection may be made in a sub-population such as provinces separately. In computer data processing systems, the outlier detection should be set as an independent module such that the operators can

selectively use it.

c. Verification and correction of outliers. Once outliers are detected, they should be individually examined and dealt with in different ways depending on their specific situations. For ratio's outliers with large value in a certain scale indicator, it should be verified and corrected by the surveyed units; and for ratio's outliers with small value of scale indicators, taking into account the cost factor, it may not be verified and corrected because they will only have small influence on summary data.

5. Ratio Probability Distribution unknown

In the case that the probability distribution of ratio indicator *z* is unknown, according to Chebyshev inequality

$$P(|\ln z - E(\ln z)| \geq k\sqrt{D(\ln z)}) < \frac{1}{k^2}. \quad (10)$$

where *k* is any positive integer. Because ln*z*-*E*(ln*z*) is a dimensionless quantity, it is obvious that D(ln*z*) can also be used to compare the effectiveness of detecting outlier. So the conclusion on the mechanism and condition of ratio tests to improve detection effectiveness is still applicable in the case of probability distributions unknown.

Under the case of probability distributions unknown, boxplot approach can be used to identify outliers for ln*z* (Tukey, 1977). Without prior assumption that data subject to specific distributions, without any restrictive requirements on data, boxplot approach has certain advantages in detecting outliers. A logarithmic transformation can make ratio distribution close to symmetric shape so that it could be conducive to detecting outliers.

6. Empirical Study

To understand the situation of China's scientific and technological (S&T) activities, since 1986 the Ministry of Science and Technology has been conducting the S&T survey in more than 4,000 government research institutes annually. Beginning from 2009, the ratio method is employed to detect outliers in original data.

Table 1 presents 5 scale indicators and 4 ratio indicators which are designed to detect outliers in the original survey data. The results in Table 1 show that for all indicators the significance levels of K-S test are less than 0.05 so that the assumptions of lognormal distribution cannot be rejected. Table 1 also shows that the effectiveness of outlier detection of ratio indicator tests will be much better than that of scale indicator tests. For example, the coefficient of variation of fixed assets is high up 8.46, while that of the ratio of fixed assets to employee is 2.18. Clearly, using the ratio indicator (ratio of fixed assets to employee) to detect outliers in data of fixed assets is much better than using scale indicator (fixed assets) in the effectiveness of detection. Also, for other 4 ratio indicators, their coefficients of variation are lower than that of not only numerators but also denominator respectively, so that their role of ratio tests has got efficient play in detecting outliers. By ratio tests, 433outliers were found, 351 research institutes were asked to check and correct the outlier data, and finally 298 outliers had been verified and corrected by the surveyed research institute.

Table 1. Main statistics of survey data of government research institute (2010)

	Scale and ratio indicator	Significance Level (K-S test)	Standard deviation $(D(\ln x))^{0.5}$	Coefficient of variation	Correlation coefficient ($\ln x$, $\ln y$)
1	fixed assets	0.017	2.06967	8.4555	
2	employee	0.014	1.21489	1.8372	
3	S&T income	0.031	1.80948	5.0421	
4	S&T expenditure	0.047	1.7529	4.5386	
5	S&T personnel	0.037	1.1666	1.7029	
6	Ratio of fixed assets to employee	0.000	1.3236	2.1830	0.797

7	Ratio of S&T income to S&T personnel	0.000	1.0866	1.5022	0.975
8	Ratio of S&T expenditure to S&T personnel	0.080	0.9326	1.1774	0.865
9	Ratio of S&T income to S&T expenditure	0.000	0.6986	0.7932	0.923

7. Conclusions

Detecting abnormal data in original survey data is an important and difficult work in statistical surveys. But in practice, it mainly relies on the checks based on logical balanced relationship among the data, and lacks effective quantitative methods to detect the non-logical balance abnormal data. The method by ratio tests provides a quantitative method for solving this problem. Starting from essential attributes or characteristics of research objects, the method uses distribution tests or boxplots of ratio data to detect outliers, which can eliminate the influence of individual scale on outlier detection with the help of ratio indicators of two scale indicators. It has been shown that when a certain condition is satisfied, a ratio indicator will have a lower coefficient of variation so that outliers can more easily be detected, and when the conditions are not satisfied the ratio test result will become worse. Reasonably selecting ratio indicators is the key to effectively detect outliers. The proposed method is also suitable for data with probability distribution unknown. Empirical study shows that the method is simple, practical, easy to use, and has application value.

References

Barnett, V. and Lewis, T. (1994) Outliers in Statistical Data, 3rd edn. John Wiley & Sons.

Box, G. E. P. and Cox, D. R. (1964) "An analysis of transformation," Journal of the Royal Statistical Society, Series B, 26, 211-252.

Cheng Bangwen et al. (2000) "Sci-tech scale index conforming to the law of logarithmic normal distribution," Science of Science and Management of S&T, 9, 9-11.

Cheng Bangwen et al. (2001) "A model and method for evaluating the quality of statistical data and identifying outliers from the data," System Engineering, 3, 85-89.

Cheng Bangwen et al. (2003) "The model and method for checking quality of multidimensional statistics and identifying outliers from the data," Mathematics in Practice and Theory, 4, 4-7.

Hawkins, DM. (1980) Identification of outliers, New York, chapman and hall.

Hosseini, M et al. (1998) "Identification of outlying height and weight data in the Iranian National Health Survey 1990-92," Journal of Applied Statistics, 5, 601-612.

Laha, R. G. and Rohatgi, V. K. (1979) Probability Theory, John Wiley & Sons, Inc.

Mardia, K. V et al. (1979) Multivariate analysis, Academic Press Inc, New York.

Rocke, David M. and Woodruff, David L. (1996) "Identification of outliers in multivariate data," Journal of the American Statistical Association, 435, 1047-1061.

Tukey, JW. (1977) Exploratory data analysis, Addison-Wesley, New York, NY.

Victoria, J. Hodge and Jim Astin, (2004) "A survey of outlier detection methodologies," Artificial Intelligence Review, 22, 85-126.

Wang Hua and Jing Yongjin. (2009) "Statistical data accuracy assessment: methods classification and applicability," Statistical Research, 1, 32-39.

Xu Dilong and Ye Shaobo, (2011) "Reviews on assessment method quality of statistics," Statistics & Information Forum, 7, 3-13.