Improving the Power of Outlier Detection through Reasonably Choosing Ratio Composition

Cheng Bangwen[1], Yang Hongjin[1], Shi Linfen[1], Wang Yali[1], He Rong[2] and Xu Ji[1]
[1]School of Management, Huazhong University of Science and Technology, Wuhan, China
[2]School of Management, Wuhan Textile University, Wuhan, China
Corresponding author: Cheng Bangwen, e-mail: Chengbw@hust.edu.cn

Abstract

In socio-economic statistics, the outlier detection based on ratio edits is widely used. The paper discusses the power of ratio edits in detecting outliers and the way to improve the power by reasonably choosing ratio composition. The study shows the power of ratio edits is mainly determined by the variation coefficients of ratio variables and the power can be improved if an appropriate denominator is chosen which will make the variation coefficient of the ratio variable be small. Firstly, the correlation coefficient between the numerator and the denominator should be as large as possible. Secondly, the variation coefficient of the denominator should be less than the one of the numerator, but the difference between them should not be too large. Besides, the critical conditions with which we can judge if the detection power is improved and the improved degree are given, which is useful for judging if the composition of ratio variables is reasonable.

Keywords: quality of statistical data; ratio edit; power of outlier detection; Tailor approximation of the variance of the ratio variable

1. Introduction

In socio-economic statistics, ratio edits are widely used and play an important role when detecting the abnormal points in survey data from basic statistical units. The ratio edit approach is to compare the ratio of two highly correlated data items with the upper and lower bounds of preassigned tolerance interval and find out the outliers in survey data. Ratios outside the tolerance are considered outliers, and one or both of the items in a ratio outlier may be wrong and need to be flagged by analyst review.

Socio-economic survey data are multidimensional, therefore, a statistical item can be tested in many ratio edits by selecting different denominators. Different ratio edits have different power of detecting outliers. One always hopes a ratio edit has high power, that is, as many abnormal data as possible can be detected under a high confidence level. Obviously, studying how to reasonably choose the ratio edit to improve the detection power is quite significant. Besides, it is impossible and unnecessary to test survey data in all possible ratio edits and we need to choose those important and efficient ones from many ratio edits.

At present, the study on ratio edits is focused on the design of ratio statistics, the determination of the upper and lower bounds of tolerances and the comparison of different ratio edit approaches (Thompson et al 1999, 2007; Zhang, 2010), while there is little research on how to reasonably choose the ratio composition. Cheng et al (2013) studied this issue and gave the conditions that the correlation coefficient of the two variables composing ratios should satisfy under the circumstance that two variables

obey joint lognormal distribution. Under this assumption, the ratio variables are normally distributed after the logarithmic transformation, which is a special case. Generally, socio-economic data are highly right skewed. Although after a power transformation, they can be turned into a normal distribution (Box et al, 1964), but only marginally correlated. At this time, the conditions given by the paper (Cheng et al 2013) are not applicable. It is the aim of this paper to discuss the power of outlier detection by ratio edits under more common situations that the two variables of ratios have marginal correlated normal distribution.

2. Power of outlier detection by ratio edits

Let $R \quad X/Y$ be a ratio variable with sample $r_1 \quad x_1/y_1,\cdots,r_n \quad x_n/y_n$. For X with sample $x_1,\cdots,x_n$, it is maybe difficult to judge and find out which ones have large statistical deviation from their real value only by their own values because those individuals near the center of probability distribution may not be necessarily detected as outliers even if they show severe statistical deviation. In addition, some data are detected as outliers due to their large numerical values, and not due to their large deviation. But it is easier to make the judgments through the ratio $r_1 \quad x_1/y_1,\cdots,r_n \quad x_n/y_n$ because the data with large deviation usually make their ratios outliers.

Ratio edits belong to the method of univariate test. In the test, null hypothesis is $H_0: r_i \quad F \ (i \quad 1,\cdots,n)$, where $\boldsymbol{F}$ is the target population distribution. An alternative hypothesis is $\overline{H}: K \quad 1$ ratios are not from F, which are called contaminants. Ratios outside the tolerance (R_L, R_U) are considered outliers, and the interval (R_L, R_U) is determined based on statistical hypothesis and preassigned significance level . Let

$$out\ (R_L, R_U; F, \quad) \quad (\quad ,R_L] \quad [R_U, \quad) \tag{1}$$

then

$$P(R \quad out(R_L, R_U; F, \quad)|H_0) \tag{2}$$

Significance level denotes the probability that the sample data are from the target population but judged as outliers, namely the probability that accurate ratios are misjudged as wrong ones.

Under preassigned , we always expect high power of detecting outliers. According to David (1981), the power can be expressed by

$$P_1 \quad P(R \quad out(R_L, R_U; \quad ,F)|\overline{H}) \tag{3}$$

which is called power function. This is the probability under $\overline{H}$ that the outliers are detected as discordant. Obviously, the higher P_1, the stronger the ability of detecting outliers is. Because contaminants are not always manifested as outliers, in order to distinguish whether the outliers are contaminants, the following two indicators can be also used to measure the power.

$$P_3 \quad P(k \text{ contaminants are outliers}, R \quad out(R_L, R_U; \quad ,F)|\overline{H}) \tag{4}$$

This is the probability that k contaminants are the outliers and are detected as discordant.

$$P_5 \quad P(R \quad out(R_L, R_U; \quad ,F)|\overline{H}, k \text{ contaminants are outliers}) \tag{5}$$

This is the probability that, when k contaminants are outliers, they are detected as discordant.

For a 'good' test, we require P_1, P_3, P_5 to be high, and $P_1 \quad P_3$ to be low under fixed (usually 0.05), obviously, according to (3), (4) and (5), the interval (R_L, R_U) should

be small.

Therefore, when the condition (11) is satisfied, the ratio R can be approximately expressed by (6). The condition (12) will be explained later. The variable Y is chosen by us, so the conditions (10), (11) and (12) are generally achievable.

From (6), the second-order approximation of R is

$$R_2 \quad {}_{x}/{}_{y}\ 1 \quad (x \quad y) \quad (x \quad y)y \ . \tag{13}$$

The mean of R_2

$$_{R_2} \quad {}_{x}(1 \quad {}_{x\ y} \quad {}^{2}_{y})/{}_{y} \ . \tag{14}$$

When (12) holds, and ${}_{x/y} \quad 1$, ${}_{R_2}$ is minimum. Obviously

$$_{x}(1- {}^{2}_{y})/{}_{y} \quad {}_{R_2} \quad {}_{x}(1 \quad {}^{2}_{y})/{}_{y} \tag{15}$$

According to (11), we obtain

$$_{R_2} \quad {}_{x}/{}_{y} \tag{16}$$

Using (16), the variation coefficient ${}^{2}_{R_2}$ of R_2 can be expressed as

$$^{2}_{R_2} \quad E\, x \quad y \quad (x \quad y) \ y \quad {}_{x/y\ x\ y} \quad {}^{2}_{y}{}^{2} \tag{17}$$

Let

$$^{2}_{R_2}(1) \quad E(x \quad y)^2 \tag{18}$$

$$^{2}_{R_2}(2) \quad E\,(xy \quad {}_{x/y\ x\ y}) \quad (y^2 \quad {}^{2}_{y})^{2} \tag{19}$$

It is obvious

$$_{R_2}(1) \quad {}_{R_2}(2)^{2} \quad {}^{2}_{R_2} \quad {}_{R_2}(1) \quad {}_{R_2}(2)^{2}$$

Then, if ${}_{R_2}(1) \quad {}_{R_2}(2)$, we have

$$_{R_2} \quad {}_{R_2}(1) \quad k \quad {}_{R_2}(2), \quad 1 \quad k \quad 1 \ . \tag{20}$$

From (18), (19)

$$^{2}_{R_2}(1) \quad ({}^{2}_{x} \quad 2 \quad {}_{xy\ x\ y} \quad {}^{2}_{y}); \tag{21}$$

$$^{2}_{R_2}(2) \quad E\,(xy \quad {}_{x/y\ x\ y})^2 \quad 2(xy \quad {}_{x/y\ x\ y})(y^2 \quad {}^{2}_{y}) \quad (y^2 \quad {}^{2}_{y})^2 \ . \tag{22}$$

For normal random variable y, we have

$$E(y^2 \quad {}^{2}_{y})^2 \quad 2 \ {}^{4}_{y} \tag{23}$$

Based on (23)

$$E(xy \quad {}_{x/y\ x\ y})^2 \quad E(x^2 \quad {}^{2}_{x})(y^2 \quad {}^{2}_{y}) \quad (1 \quad {}^{2}_{x/y}) \ {}^{2}_{x} \ {}^{2}_{y}$$
$$_{x^2/y^2}\sqrt{E(x^2 \quad {}^{2}_{x})^2 E(y^2 \quad {}^{2}_{y})^2} \quad (1 \quad {}^{2}_{x/y}) \ {}^{2}_{x} \ {}^{2}_{y}$$
$$(1 \quad 2 \quad {}_{x^2/y^2} \quad {}^{2}_{x/y}) \ {}^{2}_{x} \ {}^{2}_{y} \tag{24}$$

where ${}_{x^2/y^2}$ is the correlation coefficient between x^2 and y^2. Based on (23), (24)

$$E(xy \quad {}_{x/y\ x\ y})(y^2 \quad {}^{2}_{y}) \quad {}_{xy/y^2}\sqrt{E(xy \quad {}_{x/y\ x\ y})^2 E(y^2 \quad {}^{2}_{y})^2}$$
$$_{xy/y^2}\sqrt{2(1 \quad 2 \quad {}_{x^2/y^2} \quad {}^{2}_{x/y})} \ {}_{x} \ {}^{3}_{y} \tag{25}$$

where ${}_{xy/y^2}$ is the correlation coefficient between xy and y^2. Substituting (23), (24) and (25) into (17), we obtain

$$^{2}_{R_2}(2) \quad (1 \quad 2 \quad {}_{x^2/y^2} \quad {}^{2}_{x/y}) \ {}^{2}_{x} \ {}^{2}_{y} \quad 2 \quad {}_{xy/y^2}\sqrt{2(1 \quad 2 \quad {}_{x^2/y^2} \quad {}^{2}_{x/y})} \ {}_{x} \ {}^{3}_{y} \quad 2 \ {}^{4}_{y} \tag{26}$$

Now we analyze the correlation coefficients ${}_{x/y}$, ${}_{x^2/y^2}$ and ${}_{xy/y^2}$. Obviously, if x and y are independent in their joint distribution, then ${}_{x/y} \quad {}_{x^2/y^2} \quad {}_{xy/y^2} \quad 0$; and if ${}_{x/y} \quad 1$, then $P(x \quad y) \quad 1$, therefore ${}_{x^2/y^2} \quad {}_{xy/y^2} \quad 1$. Usually in ratio edits, is

numerator and denominator of ratio variables are normally distributed. For the case of non-normal distribution, we should first take the power transformation to make them to be normal, and then analyze and choose an appropriate ratio variable.

5. Main conclusions

This paper studies the power of outlier detection by ratio edits and the way that the composition of ratio variables is reasonably chosen to improve the power. First, the analysis on the power function shows the power of ratio edits is mainly determined by the variation coefficient of the ratio variable and the power can be improved if an appropriate denominator is chosen which will make the variation coefficient of the ratio be low. Based on the method of error propagation, the paper analyzes the variation coefficient of ratio variables which is expressed in terms of the variation coefficient of the ratio numerator and denominator, and their correlation coefficient by the second order Tailor approximation. The conditions have been obtained that the power can be improved through reasonably choosing the ratio composition. Firstly, the correlation coefficient between the numerator and the denominator should be as large as possible. Secondly, the variation coefficient of the denominator should be lower than the one of the numerator, but the difference between them should not be too large. According to the analysis, the critical conditions with which we can judge if the detection power is improved and the improved degree are given, which is useful for judging if the composition of ratio variables is reasonable.

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. (2013) A Method for Detecting Outliers in Survey Data by Ratio Tests. Proceedings 59th ISI World Statistics Congress, Hong Kong.

David, H. A. Order Statistics. 2nd edn. Wiley, New York. 1981.

Díaz-Francés, E.,Rubio, F. J. On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. Published online: 24 January 2012. © Springer-Verlag 2012.

Pham-Gia, T. et al., (2006) Density of the Ratio of Two Normal Random Variables and Applications. Communications in Statistics—Theory and Methods, 35: 1569-1591, 2006.

Thompson, K. J. and Sigman, R. S.(1999) Statistical methods for developing ratio edit tolerances for economic data. Journal of Official Statistics, 15:517-535.

Thompson, K. J. (2007) Investigation of macro editing techniques for outlier detection in survey data [J]. In Proceedings of the Third International Conference on Establishment Surveys. Survey Methods for Businesses, Farms, and Institutions: 1186 -1193.

Tukey, JW. Exploratory data analysis [M]. New York: Addison-Wesley, 1977.

Zhang Lanju et al. (2010) Tolerance Limits for a Ratio of Normal Random Variables. Journal of Biopharmaceutical Statistics, 20: 172–184.