OUTLIERS AND SOME NON-TRADITIONAL MEASURES OF LOCATION IN ANALYSIS OF WAGES

Milan Terek, Prof., PhD
Matúš Tibenský, Mgr.
University of Economics in Bratislava, Department of Statistics, Slovakia

Abstract

The paper deals with an analysis of how to use certain measures of location in analysis of wages. One of traditional measures of location – the mean should to offer typical value of variable, representing all its values by the best way. Sometimes the mean is located in the tail of the distribution and gives very biased idea about the location of the distribution. The removing of outliers, if any, or using of different measures of location could be useful in these cases. Outliers are characterized and some robust methods of their detecting are described in the paper. Then the trimmed mean and M-estimators are characterized. Computing of one-step M-estimator and modified one-step M-estimator of location is described. The possibilities of using these tools are illustrated on the analysis of the gross yearly wages of employers of one Slovak firm in the year 2013.

Keywords: Detecting outliers, trimmed mean, one-step M-estimator, modified one-step M-estimator, analysis of wages

Introduction

The distribution of wages is obviously skewed and outliers are present. Then, the interpretation power of the mean is very small[43] It will be shown that the removing outliers or alternatively, using of some non-traditional measures of location could be interesting.

Outliers are defined as unusually large or small values. Sometimes introduction of outliers into the investigations can lead to the loss of interpretation power of results, so the methods of their detection are applied. Sometimes the using of some non-traditional measures of location is appropriate. One from these measures is trimmed mean. Trimmed mean refers to a situation where a certain proportion of the largest and smallest values are removed and from the rest, the mean is calculated. M-estimators provide another class of measures of location that have practical value. Their construction requires the detection of outliers. The paper focuses on the analysis of outliers, description of the trimmed mean and M-estimators and on their application in the analysis of wages.

Outliers

In almost every series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions and the introduction of which into the investigations can only serve to perplex or mislead the inquirer (Barnett, Lewis, 1994). Such observations are call outliers. We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the reminder of that set of data (Barnett, Lewis, 1994).

[43]More in details in Halley, 2004; Terek, 2008; Terek, Nguyen Dinh He, 2011.

What characterizes the outlier is its impact on the observer – not only it will appear extreme but it will seem, in some sense, surprisingly extreme.

Outlying observations are not necessarily bad or erroneous. There are the situations in which an outlier can indicate for example some unexpectedly useful industrial treatment. Frequently, outliers are very useful in the fraud recognition. In the situations like these, it may not be necessary to adopt either of the extremes: of rejection (with a risk of the loss genuine information) or inclusion (with the risk of contamination). Sometimes the using of the robust methods of inference which employ all the data but minimize the influence of any outliers is useful. The detection of outliers requires the assessing the integrity of a set of data.

Detecting outliers

The first strategy is based on the sample mean and sample standard deviation. If the normal distribution of the population is supposed, it is obvious to consider as outlier a value which is more than 2,24 standard deviations $\sigma$from the mean μ:

$$\frac{|x-\mu|}{\sigma} > 2{,}24$$

Generally μ and σ are not known but they can be estimated from the data using the value of the sample mean $\bar{x}$ and of the sample standard deviation s. Then the following decision rule can be formulated:

The value x is declared to be an outlier if

$$\frac{|x-\bar{x}|}{s} > 2{,}24 \qquad (1)$$

where $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$

The described method can lead to the problem known as masking. Outliers inflate both the sample mean and the sample standard deviation, which in turn can mask their presence when using equation (1) (Wilcox, 2003).

The rule for detecting outliers that is not itself affected by outliers is needed. Two robust methods of outliers detection will be describe.

The method based on quartile range

This method is based on quartile range R_Q :

$$R_Q = x_{\frac{3}{4}} - x_{\frac{1}{4}}$$

where $x_{\frac{3}{4}}$ is the third quartile,

$x_{\frac{1}{4}}$ – the first quartile

The value is outlier if:

- is greater or equal as ($x_{\frac{3}{4}}$ + 1,5 R_Q),
- is less or equal as ($x_{\frac{1}{4}}$ – 1,5 R_Q).

The method based on MAD

Firstly a measure of dispersion called median absolute deviation – MAD will be described. To compute it, first compute the value $x_{1/2}$ of the sample median $X_{1/2}$, then compute the absolute values of the differences:

$\left|x_i - x_{1/2}\right|$ for $i = 1, 2, ...n$

Generally, MAD does not estimates σ, but it can be shown that when sampling from a normal distribution,

$$\text{MADN} = \frac{\text{MAD}}{0{,}6745}$$

estimates σas well (Wilcox, 2003).

The robust decision rule for the detection of outliers is following:
The value x is declared to be an outlier if

$$\frac{\left|x - x_{1/2}\right|}{\text{MADN}} > 2{,}24 \qquad (2)$$

A trimmed mean

The value of the trimmed mean is calculated from the data, from which a certain proportion of the largest and smallest observations are removed and the remaining observations are averaged. For example 10% trimmed mean is calculated from the data from which 10% of the largest and 10% and smallest observations were removed.

A fundamental issue is deciding how much to trim. When addressing a variety of practical goals, 20% trimming often offers a considerable advantage over not trimming and the median (Wilcox, 2003).

M-estimators

M-estimators are from another class of measures of location. For example, if for any n values $X_1, X_2, ... X_n$ we want to choose c so that it minimizes the sum of squared errors,

$$\sum_{i=1}^{n}\left(X_i - c\right)^2 \qquad (3)$$

It can be shown that it must be the case that $\sum_{i=1}^{n}\left(X_i - c\right) = 0$. From this last equation $c = \overline{X}$.

So, when we choose a measure of location based on minimizing the sum of the squared errors given by (3), this leads to using the sample mean. But if we measure how close c is to the n values using the sum of absolute differences, the sample median minimizes this sum (Wilcox, 2003).

Generally, there are infinitely many ways of measuring closeness that lead to reasonable measures of location. For example, if we measure the closeness by $\sum_{i=1}^{n}\left|X_i - c\right|^a$, then setting $a = 1$ leads to the median, and $a = 2$ leads to the mean.

Let us have any function Ψ having the property: $\Psi(-x) = -\Psi(x)$, we get a reasonable measure of location, provided the probability curve is symmetric, if we choose c so that is satisfies

$$\Psi(X_1 - c) + \Psi(X_2 - c) + ... + \Psi(X_n - c) = 0 \qquad (4)$$

Measures of location based on (4) are called M-estimators.

The calculation of the M-estimators requires the detection of outliers.

One-step M-estimator

Let n_1 be the number of observations X_i, for which

$$\frac{(X_i - X_{1/2})}{\text{MADN}} < -K$$

and let n_2 be the number of observations such that

$$\frac{(X_i - X_{1/2})}{\text{MADN}} > K$$

where typically K = 1,28 is used. The one-step M-estimator of location (based on Huber` s Ψ) is

$$\hat{\mu}_{os} = \frac{K(\text{MADN})(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} \qquad (5)$$

where $X_{(i)}$ is i-th order statistic[44].

The calculation of the value of M-estimator requires the determination of outliers using the method in 2.1.2, except that (2) is replaced by

$$\frac{|X - X_{1/2}|}{\text{MADN}} > K \qquad (6)$$

Next, remove the values flagged as outliers and average the values that remain. For technical reasons, the one-step M-estimator makes an adjustment based on MADN, a measure of scale plus the number of outliers above and below the median (Wilcox, 2003).

A modified one-step M-estimator

Sometimes a simple modification of one-step M-estimator is used:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} \qquad (7)$$

Here, K = 2,24 is used to determine n_1 and n_2 (Wilcox, 2003).

Analysis of wages

The possibilities of application of the described tools will be illustrated on the analysis of the gross yearly wages of 251 employees of the firm Medirex in Slovak republic in the year 2013. Table 1 presents the values of some descriptive measures, computed with aid of the software MS Excel 2007 (Tibenský, 2014).

Tab. 1 Descriptive measures of the gross yearly wage

Count	251
Average	9873,68
Median	9114,43
Standard deviation	5438,019
Minimum	921,36
Maximum	55303,78
Range	54382,42

[44]Order statistic is determined by its ranking in a non-decreasing arrangement of random variables.

The frequency distribution is in the table 2 (Tibenský, 2014).

Tab. 2 Frequency distribution of the gross yearly wage

Gross yearly wage	Frequency
- 5 000	15
5 000 - 10 000	144
10 000 - 15 000	78
15 000 - 20 000	7
20 000 - 25 000	4
25 000 - 30 000	1
30 000 -	2

In figure 1 is the histogram for the gross yearly wages.

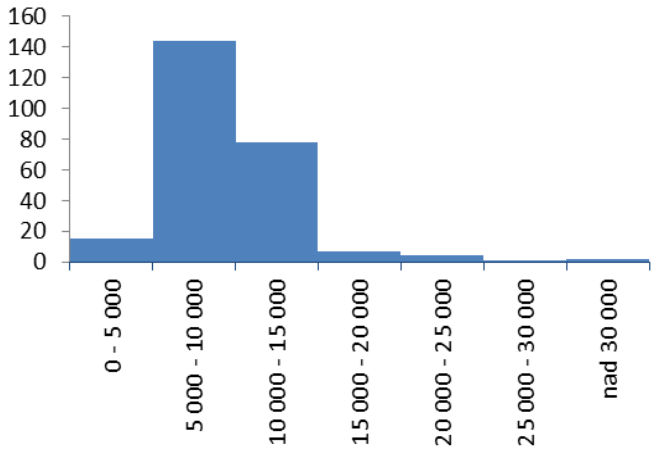

Figure 1 Histogram for the gross yearly wages

It is evident from the histogram, that the distribution of wages is highly skewed on the right. It seems that the mean is not the best measure of the „typical“ wage of employee in the population.

Detecting and removing outliers

The outliers will be detected with aid of the method based on quartile range. 22 outliers were detected by this method. Table 3 presents the values of some descriptive measures after removing these 22 outliers (Tibenský, 2014).

Tab. 3 Descriptive measures of the gross yearly wage after removing of outliers

Count	229
Average	9397,36
Median	9111,95
Standard deviation	2298,947925
Minimum	3398,17
Maximum	15783,95
Range	12385,78

In the table 3 can be seen that the range and standard deviation decreased. The change of median is only small, the change of the mean is important. The distance between median and mean is much smaller. Now, the mean better characterizes the typical yearly wage.

M-estimators in analysis of wages

Firstly, the detection of outliers with aid of the method based on MAD is necessary. The values MAD = 1606,7 and MADN = MAD/0,6745 = 2382,061 were calculated. Then, detecting outliers based on condition (6) was applied and the values of M-estimators were

computed. The values used in the computing of one-step M-estimator are in table 4 (Tibenský, 2014).

Tab. 4 Values used in computing ofone-step M-estimator

Count	251
MADN	2382,061
K	1,28
n_1	23
n_2	44

$$\hat{\mu}_{os} = \frac{K(\text{MADN})(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = \frac{1{,}28 \cdot 2382{,}061(44 - 23) + 1\,643\,129{,}64}{251 - 23 - 44} \approx 9\,278{,}04$$

The values used in the computing of modified M-estimator are in table 5 (Tibenský, 2014).

Tab. 5 Values used in computing ofmodified one-step M-estimator

Count	251
MADN	2382,061
K	2,24
n_1	13
n_2	15

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = \frac{2\,089\,446{,}25}{251 - 13 - 15} \approx 9\,369{,}71$$

Then the 20% trimmed mean was computed. All results are in table 6.

Tab. 6 Descriptive Measures

Mean	9873,68
Median	9114,43
20% Trimmed mean	9 375,36
$\hat{\mu}_{os}$	9 278,04
$\hat{\mu}_{mom}$	9 369,71

Conclusion

It can be seen that the detecting and removing outliers were very useful in the analysis of wages. The mean and median became much closer, the variability of the data was much less as before as well. Then the traditional measures of location – mean and median can better characterize the data. In our opinion this is the first possibility for improving the analysis of wages and making the results more consistent, giving better idea about the typical wage in the population.

The second possibility is the calculation of some non-traditional measures of location. The value of 20% trimmed mean is 9 375,36 EUR, value of one-step M-estimator is 9 278,04 EUR, value of modified one-step M-estimator is 9 369,71 EUR and value of median is 9114,43 EUR in the analysis. The values of measures are only slightly different. Each of these certainly better characterizes typical yearly wage of an employee in analyzed period as mean, equal to 9873,68 EUR, which is evidently highly influenced by a small number unusually high wages.

In our opinion, analysis of outliers followed by calculating of traditional measures of location or alternatively the calculating of some non-traditional measures of location like, trimmed means and M-estimators are efficient tools for obtaining more real and true view to the typical wage.

This paper was elaborated with the support of the grant agency VEGA in the framework of the project no. 1/0761/12.

References:

Barnett, V., Lewis, T.: Outliers in Statistical Data. New York: Wiley and Sons, 1994.
Halley, R. M.: Measures of Central Tendency, Location, and Dispersion in Wage Survey Research. In: Compensation and Benefits 2004/36, 39 (2004) p. 39 - 52.
Terek, M.: Analýza odľahlých údajov. In: Forum Statisticum Slovacum 6 (2008) p. 152 – 157.
Terek,M., Nguyen Dinh He:Možnosti využitia niektorých netradičných charakteristík v analýze miezd. In: Ekonomické rozhľady 1 (2011) p. 74 - 91.
Tibenský, M.:Chakteristiky polohy založené na analýze odľahlých údajov. Ing. theses, Bratislava: University of Economics, 2014.
Wilcox, R. R.: Applying Contemporary Statistical Techniques. USA: Academic Press 2003.