Journal of Modern Accounting and Auditing, ISSN 1548-6583
May 2013, Vol. 9, No. 5, 711-718

Analysis of Salaries and Some Non-traditional Measures of Location*

Milan Terek, Nguyen Dinh He
University of Economics in Bratislava, Bratislava, Slovakia

The paper deals with an analysis of how to use certain measures of location in analysis of salaries. One of the traditional measures of location, the mean should 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 a very biased idea about the location of the distribution. In these cases, using different measures of location could be useful. Trimmed mean is described. The trimmed mean refers to a situation where a certain proportion of the largest and smallest observations are removed and the remaining observations are averaged. The construction of some measures of location is based on the analysis of outliers. Outliers are characterized. Then the possibilities of the detection of outliers are analyzed. Computing of one-step M-estimator and modified one-step M-estimator of location is described. A comparison of the trimmed means and M-estimators of location is presented. Finally, the paper focuses on the application of the trimmed mean and M-estimators of location in analysis of salaries. The analysis of salaries of employers of the big Slovak companies in second half of the year 2009 is realized. The data from the census are used in the analysis. The median, 20% trimmed mean and the characteristics, based on the one-step M-estimator of location and modified one step M-estimator, are calculated.

Keywords: trimmed mean, detecting outliers, one-step M-estimator, modified one-step M-estimator, analysis of salaries

Introduction

One of the traditional indicators of the standard of living of population is average monthly salary (Muchová, 2011). The distribution of salaries is obviously skewed and outliers are present. Then, the interpretation power of this indicator is very small (Halley, 2004). It will be shown that the using of some non-traditional measures of location could be interesting.

One of these measures of location 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 average 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 description of the trimmed mean and M-estimators and on their application in the analysis of salaries.

* This paper was elaborated with the support of the grant agency VEGA in the framework of the project Number 1/0761/12.
Milan Terek, professor, Department of Statistics, University of Economics in Bratislava. Email: milan.terek@euba.sk.
Nguyen Dinh He, assistant professor, Department of Statistics, University of Economics in Bratislava.

Methods

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% of the smallest observations were removed.

A fundamental issue is deciding how much to trim. When addressing a variety of practical goals, 20% trimming often offers considerable advantages 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 the squared errors:

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

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 Equation (1), this leads to using the sample mean. But if we measure how close *c* is to the *n* values using the sum of the 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 setting $a = 2$ 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 that the probability curve is symmetric, if we choose *c* so that it satisfies:

$$\Psi(X_1 - c) + \Psi(X_2 - c) + ... + \Psi(X_n - c) = 0 \tag{2}$$

Measures of location based on Equation (2) are called M-estimators.

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

Outliers. In almost every series of observations, some are found, which differs so much from the others as to indicate some abnormal sources of errors 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 called 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).

Sometimes, outliers are defined simply as unusually large or small values (Terek, 2008).

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 situations in which an outlier can indicate, for example, some unexpectedly useful industrial treatments. 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 losing 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.

We would like to note that the outlier problem is an unavoidable one. It is not enough to say that one should not consider dealing with outlying observations. In fact, the persons who have to deal with data and take decisions are forced to make judgments about outliers—whether or not to include them, whether to make allowances for them on some compromise bases, and so on.

The detection of outliers requires assessing the integrity of a set of data. We need the techniques for assessing the integrity of a set of data with respect to an assumed model. We require the methods for assessing, rejecting, making allowances for, or minimizing the influence of outlying observations.

The outlier problem progresses in the following way. We examine the data set. Suppose, we decide that outliers exist. Then we must ask: How should we react to the outliers, and what principles and methods can be used to support rejecting them, adjusting their values, or leaving them unaltered, prior to processing the main mass of data? The answers depend on the form of the population, the used techniques depend mainly on the postulated model for the population.

Detecting outliers. The first strategy is based on the sample mean and sample standard deviation. If we suppose normal distribution of the population, it is obvious to consider as outlier a value which is more than 2.24 standard deviation from the mean $\frac{|x-\mu|}{\sigma} > 2.24$. If we suppose the normal distribution of the population with the mean μ and standard deviation σ, the probability of declaring a value an outlier using this equation is 0.025.

Generally, μ and σ are unknown, 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:

$$\frac{|x-\bar{x}|}{s} > 2.24 \tag{3}$$

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 (3) (Wilcox, 2003).

A rule for detecting outliers that is not itself affected by outliers is needed. One robust method of outliers detection will be described.

The method based on median absolute deviation (MAD). Firstly, a measure of dispersion called as *MAD*

will be described. To compute it, we will first compute the value $x_{1/2}$ of the sample median $X_{1/2}$, then we will compute the absolute values of the differences: $\left|x_i - x_{1/2}\right|$ for $i = 1, 2,... n$.

Generally, *MAD* does not estimate σ, but it can be shown that when sampling from a normal distribution, $MADN = \frac{MAD}{0.6745}$ estimates σ as well (Wilcox, 2003, p. 73).

The robust decision rule for the detection of outliers is as follows.

The value *x* is declared to be an outlier if:

$$\frac{\left|x - x_{1/2}\right|}{MADN} > 2.24 \tag{4}$$

One-step M-estimator. Let n_1 be the number of observations X_i, for which $\frac{\left(X_i - X_{1/2}\right)}{MADN} < -K$ and let n_2 be the number of observations such that $\frac{\left(X_i - X_{1/2}\right)}{MADN} > K$, where typically $K = 1.28$ is used. The one-step M-estimator of location (based on Huber' s Ψ) is:

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

where $X_{(i)}$ is *i*-th order statistic[1].

The calculation of the value of M-estimator requires the determination of outliers using the method based on *MAD*, except that Equation (4) is replaced by:

$$\frac{\left|x - x_{1/2}\right|}{MADN} > K \tag{6}$$

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 the one-step M-estimator is used:

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

where $K = 2.24$ is used to determine n_1 and n_2.

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

The Comparison of Trimmed Means and M-estimators

What are the fundamental differences among trimmed means, one-step M-estimator, and modified one-step M-estimator? Each of them represents a different approach to measuring location. Trimmed means discard a fixed proportion of small and large observations. The M-estimators empirically determine how many observations are to be trimmed. They also include the possibility of different amounts of trimming in the tails as well as no trimming at all.

It is also possible to make inferences based on these measures. For example, it is possible to compute the confidence intervals for trimmed mean and for the measures based on M-estimators.

Results and Discussion

Measures of Location and Analysis of Salaries

The possibilities of application of the different measures of location will be illustrated on the analysis of the gross monthly salaries of 956,844 employees in Slovak republic in the second half year 2009. Table 1 presents the values of some descriptive measures, computed with the aid of the software Statgraphics Centurion, according to Dinh He (2011).

Table 1

Descriptive Measures of the Gross Monthly Salaries

Count	956,844
Average	812.106
Median	659.333
Variance	642,891
Standard deviation	801.805
Coeff. of variation	98.73%
MAD	202.198
Minimum	6.21167
Maximum	99,144.7
Range	99,138.48
Lower quartile	482.283
Upper quartile	911.056
Skewness	21,590.8

Average monthly salary is 812,106 EUR. It is highly different from the median which is 659.333 EUR. We can point at another interesting value of the measures as well. The lowest salary is, for example, 6.21167 EUR. It is probably the salary of the employee working only part time, or it is an error. The highest salary 99,144.7 EUR is also interesting[2]. It could be for example untypical salary of a top manager. These untypically low and high values were taken into account in the computing of the average. Standard deviation is 801.805 EUR, and the coefficient of variation is 98.73 %. The distribution of salaries is highly skewed on the right—the value of the skewness is 21,590.8. It seems that the mean is not the best measure of the "typical" salary of an employee in the set.

Trimmed mean and M-estimators. The 20% trimmed mean was computed. The results of the analysis according to Dinh He (2011) are presented in Table 2. The value of the 20% trimmed mean is 676.719 EUR

[2] Really, it is the average gross monthly salary of an employee in the second half year 2009.

which is only a little different from the median. The median remains the same—it is, as before, 659.333 EUR. The standard deviation decreased to 147.306 EUR and the coefficient of variation to 21.77%. The range of salaries decreased to 551.043 EUR. The skewness also decreased, and now it is negative (-103.8).

Table 2

Descriptive Measures—20% Trimming

Count	574,107
Median	659.333
20% trimmed mean	676.719
Standard deviation	147.306
Coeff. of variation	21.77%
MAD	115.468
Minimum	446.377
Maximum	997.42
Range	551.043
Lower quartile	553.023
Upper quartile	788.068
Skewness	-103.8

It can be seen that the trimming of 20% of data from each side of the distribution offers less values of measures of variability and less skewness. In the light of these results, it seems that 20% trimmed mean, 676.719 EUR, could be a good measure of the "typical" monthly salary of an employee in the analyzed period.

Then the values of M-estimators were computed. The values used in the computing of one-step M-estimator are in Table 3 (Dinh He, 2011).

Table 3

Values Used in Computing of M-estimator

Median	659.333
MAD	202.198
MADN	299.774
K	1.28
n_1	24,504
n_2	170,475
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	472,189,611.9

From the left tail of the distribution, 24,504 values were discarded, from the right tail -170,475 values. After substitution from Table 3 to Equation (5), we get:

$$\hat{\mu}_{os} = \frac{K(MADN)(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = \frac{1.28299774(170{,}475 - 24{,}504) + 472{,}189{,}611.9}{956{,}844 - 170{,}475 - 24{,}504} \approx 693.299$$

We got another measure for the describing of the "typical" monthly salary of an employee in the analyzed period. It can be seen that the value 693.299 EUR is only slightly different from 20% triemmed mean and median, but it is highly different from the mean which is equal to 812.106 EUR.

It is clear that the value of the mean is highly influenced by the small number of untypically high salaries. It is the reason of its loosing of the "interpretation power" as "typical value".

Now the value of the modiefied M-estimator will be computed. The values used in the computing are in Table 4 (Dinh He, 2011).

Table 4

Values Used in Computing of Modified M-estimator

MAD	202.198
MADN	299.775
Median	659.333
n_1	0
n_2	90,379
K	2.24
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	570,441,925

After substitution from Table 4 to Equation (7), we get:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = \frac{570{,}441{,}925}{956{,}844 - 90{,}379} \approx 658.355$$

Again, there is one measure for a good characterization of the "typical" monthly salary of an employee in the analyzed period. The value 658.355 EUR is only slightly different from 20% trimmed mean, M-estimator, and median, but highly different from the mean.

Conclusions

The value of 20% trimmed mean is 676.719 EUR, value of M-estimator is 693.299 EUR, value of modified M-estimator is 658.355 EUR, and the value of median is 659.333 EUR. The values of measures are only slightly different and they oscillate around the value of median. Each of these values certainly better characterizes typical monthly salary of an employee in the analyzed period as mean, equaling to 812.106 EUR, which is evidently highly influenced by a small number unusually high salaries.

According to our opinion, in the analysis of salaries area like in other areas is useful to compute and interpret except basic measures like mean, median, modus, quartiles, or other quantiles, also some non-traditional measures of location (Terek, 2010). They can enrich the overall view of the situation. It is also certainly useful to compute the measures of dispersion, skewness, or other measures which can offer another view of the overall situation as well (Terek & Dinh He, 2011).

Trimmed mean and M-estimators in the given analysis are efficient tools for obtaining more real and true views of the typical monthly salary of an employee in the given set.

References

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. New York, NY: John Wiley and Sons.

Dinh He, N (2011). Analysis of selected socioeconomic indicators in respect of outlying data (Ph.D. theses, Bratislava University of Economics).

Halley, R. M. (2004). Measures of central tendency, location, and dispersion in salary survey research. *Compensation and Benefits Review, 36*(5), 39-52.

Muchová, E. (2011). The labor market and competitiveness in the European area. Proceedings of the contributions from the *International Scientific Conference*—Background and Challenges for Social Policy at the Impending Decades, Bratislava.

Terek, M. (2008). The analysis of outlying data. *Forum Statisticum Slovacum, 6*, 152-157.

Terek, M. (2010). Analysis of outlying data and certain characteristics of location. Proceedings of the contributions from the *International Scientific Conference of the Knowledge Economy and its Reflection in Economic Theory and Economic Practice* (pp. 371-384), Bratislava.

Terek, M., & Dinh He, N. (2011). The possibility of using some of the non-traditional characteristics in the analysis of wages. *Economic Outlook, 1*, 74-91.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. USA: Academic Press.