Journal of Experimental Social Psychology xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

FlashReport

Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median

Christophe Leys [a,*], Christophe Ley [b,1], Olivier Klein [a], Philippe Bernard [a,1], Laurent Licata [a]

[a] Université Libre de Bruxelles, Unité de Psychologie Sociale, Belgium
[b] Université Libre de Bruxelles, Département de Mathématique and ECARES, Belgium

ARTICLE INFO

Article history:
Received 24 November 2012
Revised 10 March 2013
Available online xxxx

Keywords:
Median absolute deviation
Outlier
MAD

ABSTRACT

A survey revealed that researchers still seem to encounter difficulties to cope with outliers. Detecting outliers by determining an interval spanning over the mean plus/minus three standard deviations remains a common practice. However, since both the mean and the standard deviation are particularly sensitive to outliers, this method is problematic. We highlight the disadvantages of this method and present the median absolute deviation, an alternative and more robust measure of dispersion that is easy to implement. We also explain the procedures for calculating this indicator in SPSS and R software.

In a recent article, Simmons, Nelson, and Simonsohn (2011) showed how, due to the misuse of statistical tools, significant results could easily turn out to be false positives (i.e., effects considered significant whereas the null hypothesis is actually true). In their argument, they accurately pinpointed the importance of outliers. The aim of this paper is twofold: (a) showing that many researchers use a very poor method to detect outliers; (b) outlining the Median Absolute Deviation (MAD) method as a way of dealing with the problem of outliers.

Outliers are not a new concern (Orr, Sackett, & Dubois, 1991; Ratcliff, 1993; Rousseeuw & Croux, 1993). However, we argue that scholars in the field of psychology still use inappropriate methods for no legitimate reason. Indeed, we surveyed the methods used in two major psychology journals, namely the *Journal of Personality and Social Psychology* (JPSP) and *Psychological Science* (PSS) between 2010 and 2012. We introduced the keywords "outlier" OR "outlying data" OR outliers OR "extreme value" OR "extreme values" OR "nasty data" (in reference to McClelland's chapter on the subject) OR "extreme data" for searching this database. There were 127 relevant hits. We then coded the method used to cope with outliers (see Fig. 1), either the mean plus/minus a coefficient (2, 2.5 or 3) times the standard deviation, or the interquartile method (a commonly used method to detect outliers, see for example Rousseeuw & Croux, 1993), or another method (e. g. a method specifically developed for reaction times by Ratcliff, 1993). No article mentioned used the Median Absolute Deviation described below.

This survey revealed the lack of concern for the mishandling of outliers, even in recently published papers. Indeed, in most cases researchers did not report the method used to handle outliers or excluded values over two or three standard deviations around the mean, which is a very poor indicator.

Facing these conclusions, we describe a robust and easy to conduct method, for detecting outlying values in univariate statistic the Median Absolute Deviation. This indicator was initially developed by statisticians but is relatively unknown in psychology. In this paper, we present this method, building on the statistical literature, and consider its relevance to our field.

The mean plus or minus three standard deviations

Notwithstanding the decision to remove, correct or leave an outlier (for a discussion on this topic see McClelland, 2000), it is necessary to be able to detect its presence. The method of the mean plus or minus three SD is based on the characteristics of a normal distribution for which 99.87% of the data appear within this range (Howell, 1998). Therefore, the decision that consists in removing the values that occur only in 0.13% of all cases does not seem too conservative. Other authors (e.g., Miller, 1991) suggest being less demanding, and use 2.5 or even 2 standard deviations around the mean. This choice obviously depends on the situation and on the perspective defended by the researcher.

Unfortunately, three problems can be identified when using the mean as the central tendency indicator (Miller, 1991). Firstly, it assumes that the distribution is normal (outliers included). Secondly, the mean and standard deviation are strongly impacted by outliers. Thirdly, as stated by Cousineau and Chartier (2010), this method is very unlikely to detect outliers in small samples.

* Corresponding author.
E-mail address: cleys@ulb.ac.be (C. Leys).

[1] Christophe Ley and Philippe Bernard thank the Fonds National de la Recherche Scientifique, Communauté Française de Belgique, for financial support via a Mandat de Chargé de Recherche FNRS.

0022-1031/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jesp.2013.03.013

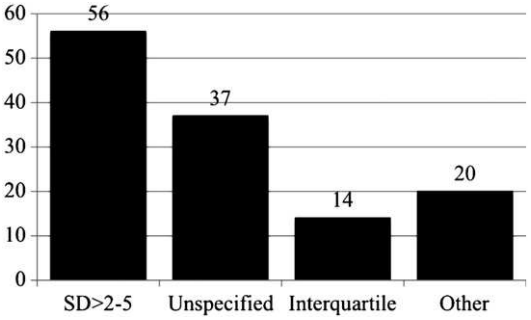

Fig. 1. Survey — methods used to cope with outliers in JPSP and PS between 2010 and 2012. Note: N = 127; SD > 2–5 = deviation from 2 to 5 SD around the mean; unspecified = authors did not report the method used to cope with outliers.

Accordingly, this indicator is fundamentally problematic: It is supposed to guide our outlier detection but, at the same time, the indicator itself is altered by the presence of outlying values. In order to appreciate this fact, consider a small set of $n = 8$ observations with values 1, 3, 3, 6, 8, 10, 10, and 1000. Obviously, one observation is an outlier (and we made it particularly salient for the argument). The mean is 130.13 and the uncorrected standard deviation is 328.80. Therefore, using the criterion of 3 standard deviations to be conservative, we could remove the values between −856.27 and 1116.52. The distribution is clearly not normal (Kurtosis = 8.00; Skewness = 2.83), and the mean is inconsistent with the 7 first values. Nevertheless, the value 1000 is not identified as an outlier, which clearly demonstrates the limitations of the mean plus/minus three standard deviations method.

An alternative: the median absolute deviation (MAD)

Absolute deviation from the median was (re-)discovered and popularized by Hampel (1974) who attributes the idea to Carl Friedrich Gauss (1777–1855). The median (M) is, like the mean, a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers. One indicator of this insensitivity is the "breakdown point" (see, e.g., Donoho & Huber, 1983). The estimator's breakdown point is the maximum proportion of observations that can be contaminated (i.e., set to infinity) without forcing the estimator to result in a false value (infinite or null in the case of an estimator of scale). For example, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. By contrast, the median value remains unchanged. The median becomes absurd only when more than 50% of the observations are infinite. With a breakdown point of 0.5, the median is the location estimator that has the highest breakdown point. Exactly the same can be said about the Median Absolute Deviation as an estimator of scale (see the formula below for a definition). Moreover, the MAD is totally immune to the sample size. These two properties have led Huber (1981) to describe the MAD as the "single most useful ancillary estimate of scale" (p. 107). It is for example more robust than the classical interquartile range (see Rousseeuw & Croux, 1993), which has a breakdown point of 25% only.

To calculate the median, observations have to be sorted in ascending order to identify the mean rank of the statistical series and to determine the value associated with that rank. Let us consider the previous statistical series: 1, 3, 3, 6, 8, 10, 10, and 1000. The average rank can be calculated as equal to (n + 1) / 2 (i.e., 4.5 in our example). The median is therefore between the fourth and the fifth value, that is, between six and eight (i.e., seven). Calculating the MAD is also straightforward, as it only involves finding the median of absolute deviations from the median. More precisely, the MAD is defined as follows (Huber, 1981):

$$MAD = b\ M_i\left(\left|x_i - M_j\left(x_j\right)\right|\right)$$

where the x_j is the n original observations and M_i is the median of the series. Usually, $b = 1.4826$, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers (Rousseeuw & Croux, 1993).

If another underlying distribution is assumed (which is seldom the case in the field of psychology), this value changes to $b = 1/Q(0.75)$, where $Q(0.75)$ is the 0.75 quantile of that underlying distribution. In case of normality, $1/Q(0.75) = 1.4826$ (Huber, 1981). This multiplication by b is crucial, as otherwise the formula for the MAD would only estimate the scale up to a multiplicative constant.

Concretely, calculating the MAD implies the following steps: (a) the series in which the median is subtracted of each observation becomes the series of absolute values of (1–7), (3–7), (3–7), (6–7), (8–7), (10–7), (10–7), and (1000–7), that is, 6, 4, 4, 1, 1, 3, 3, and 993; (b) when ranked, we obtain: 1, 1, 3, 3, 4, 4, 6, and 993; (c) and (d) the median equals 3.5 and will be multiplied by 1.4826 to find a MAD of 5.1891.

Then, we must define the rejection criterion of a value. As for the mean and standard deviation, it is necessary to define a level of decision: This remains the unavoidable subjective aspect of the decision.

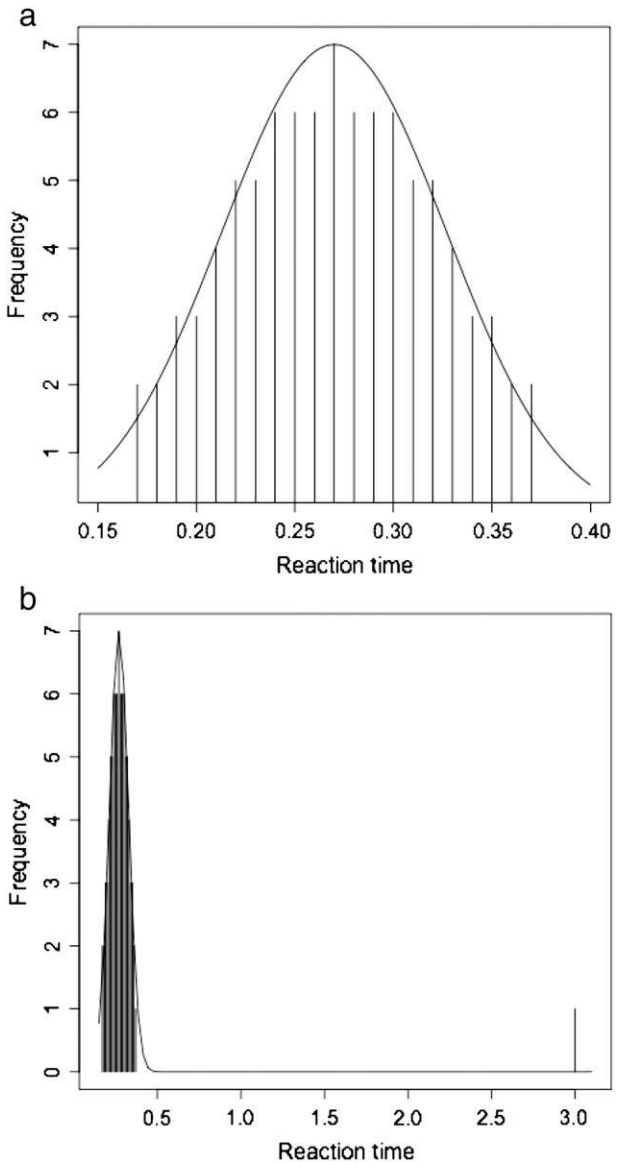

Fig. 2. Outlier generating asymmetry. a) Normal distribution, n = 91, mean = 0.27, median = 0.27, standard deviation = 0.06. b) Asymmetry due to an outlier, n = 91, mean = 0.39, median = 0.27, standard deviation = 0.59.

Please cite this article as: Leys, C., et al., Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology* (2013), http://dx.doi.org/10.1016/j.jesp.2013.03.013

```
Step a)
DATASET ACTIVATE "Database's name".
FREQUENCIES VARIABLES="first variable's name"
/STATISTICS=MEDIAN
/ORDER=ANALYSIS.
Step b)
COMPUTE "Computed second variable's name"="first variable's name"-"median".
EXECUTE
Step c)
FREQUENCIES VARIABLES="Computed second variable's name"
/STATISTICS=MEDIAN
/ORDER=ANALYSIS.
```

Fig. 3. SPSS script for steps a, b and c.

Depending on the stringency of the researcher's criteria, which should be defined and justified by the researcher, Miller (1991) proposes the values of 3 (very conservative), 2.5 (moderately conservative) or even 2 (poorly conservative). Let us use the same limit as in the previous example and choose the threshold 3 for our example. The decision criterion becomes:

$$M-3*MAD<x_i<M+3*MAD$$
$$\text{or}$$
$$\frac{x_i-M}{MAD}>|\pm 3|$$

In our example, all values greater than 7 + (3*5.1891) = 22.57 and all values smaller than 7 - (3*5.1891) = −8.57 can be removed. Stated differently, we can remove the observation "1000" of our series.

The second expression of our decision criterion leads to the same conclusion as the first but offers the advantage of indicating the distance of the value from the decision criterion, rather than proceeding by comparison with a specific value of the series. By doing so, we found (1000–7) / 5.19 = 191.36. We clearly see that this value strongly deviates from the threshold of 3 chosen previously.

Let us briefly consider the case of a fictional series in Fig. 2, which includes a larger number of observations. Fig. 2a shows a normal distribution and reports the mean, SD and median. Fig. 2b shows the same distribution but with one value (=0.37) changed into an outlier (=3). The same indicators are reported and we can see that the mean and SD have drastically changed whereas the median remains the same.

Even if the dispersion was very low for didactic reasons, we would have obtained an interval for detecting outliers of $-0.57 < x_i < 1.17$ by the method of the mean plus or minus three standard deviations and, by contrast, an interval of $0.09 < x_i < 0.45$ when using the method of the median plus or minus three times the MAD.

Procedure implemented in the statistical software SPSS and R

SPSS (statistical package for social sciences) is the software commonly used by many researchers in social sciences. The procedure for calculating the MAD is simple, we have to: (a) compute the median using the menu "Analysis" and the command "Frequency"; (b) subtract this value from all observations in the statistical series using the command "Compute" in the menu "Transform"; (c) compute the median of the resulting new variable as in the first point, and (d) multiply this value by 1.4826 (if we assume normality of the data). Fig. 3 shows the SPSS script for step a to c. Step d can be computed with any calculator.

The MAD can be easily calculated in the software R as well by utilizing the command "Mad" available in the package "Stats". Note that this command assumes by default that $b = 1.4826$.

Discussion

Given the results of our survey of two journals, emphasizing a poor management of outliers, we showed that the method conventionally used ("The mean plus or minus three standard deviations" rule) is problematic and we argued in favor of a robust alternative. We have finally explained that, whatever the method selected, the decision-making concerning the exclusion criteria of outliers (a deviation of 3, 2.5 or 2 units) is necessarily subjective. This leads us to three important recommendations:

1. In univariate statistics, the Median Absolute Deviation is the most robust dispersion/scale measure in presence of outliers, and hence we strongly recommend the median plus or minus 2.5 times the MAD method for outlier detection.
2. The threshold should be justified and the justification should clearly state that other concerns than cherry-picking degrees of freedom guided the selection. By default, we suggest a threshold of 2.5 as a reasonable choice.
3. We encourage researchers to report information about outliers, namely: the number of outliers removed and their value (or at least the distance between outliers and the selected threshold).

More generally, we believe that, faced with the pitfalls presented by researchers' degrees of freedom (see Simmons et al., 2011), the insufficient knowledge of various outlier-detecting methods is not the main challenge facing psychological science. Achieving a consensus as to which method is most appropriate and which subjective threshold should be used (regardless of the method used) is of even greater importance. Otherwise, the suspicion that researchers pick the method that yields the most promising results will remain in the air even when, as in most cases, it is unjustified. With respect to outlier management, we hope that if such a consensus can be achieved, our presentation of the MAD will have contributed to it.

References

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67.

Donoho, D. L., & Huber, P. J. (1983). In Bickel, Doksum, & Hodges (Eds.), *The notion of breakdown point*. California: Wadsworth.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393, http://dx.doi.org/10.1080/01621459.1974.10482962.

Howell, D. C. (1998). *Statistical methods in human sciences*. New York: Wadsworth.

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley.

McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 393–411). Cambridge: Cambridge University Press.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology*, *43*(4), 907–912, http://dx.doi.org/10.1080/14640749108400962.

Orr, J. M., Sackett, P. R., & Dubois, C. L. (1991). Outlier detection and treatment in I/O psychology: A survey of researchers' beliefs and an empirical illustration. *Personnel Psychology*, *44*(3), 473–486, http://dx.doi.org/10.1111/j.1744-6570.1991.tb02401.x.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283, http://dx.doi.org/10.1080/01621459.1993.10476408.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366, http://dx.doi.org/10.1177/0956797611417632.

Please cite this article as: Leys, C., et al., Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology* (2013), http://dx.doi.org/10.1016/j.jesp.2013.03.013