DETECTING OUTLIERS: USE ABSOLUTE DEVIATION AROUND THE MEDIAN BUT DO NOT USE STANDARD DEVIATION AROUND THE MEAN

JASPREET KAUR
ASSISTANT PROFESSOR
MATHEMATICS
B.Z.S.F.S Khalsa Girls College Morinda, Ropar, India

Abstract: A survey showed that according to researchers it still seems to be difficult to cope with outliers. Observe outliers by determining an interval spanning over the mean plus/minus three standard deviations remains a regular practice. Yet, since both the mean and the standard deviation are particularly sensitive to outliers, this method is difficult. We not only highlight the drawbacks of this method but also 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 software programming in SPSS and R software.

Keywords: MAD, Outliers, standard, JPSP.

1. INTRODUCTION

Moreover, **Simmons, Nelson, and Simonsohn (2011)** Displayed a new article that how outstanding results could easily turn out to be wrong effective (i.e., effects considered important although the null hypothesis is actually true) due to misuse of statistical tools. In their justification, they accurately pinpointed the importance of outliers. The aim of this paper is twofold: (1)) showing that most of the researchers do not use a very good method to detect outliers. (2) The way of dealing with the problem of outliers is outlining the median absolute deviation(MAD) method.

Outliers are not the latest review **(Orr, Sackett, & Dubois, 1991; Ratcliff, 1993; Rousseeuw & Croux, 1993)**. yet, we declare that scholars in the field of psychology still use unsuitable methods for no upright reason. Actually, we observe the methods used in two great psychology journals, namely the Journal of Personality and Social Psychology(JPSP) and Psychological Science(PSS) between2010 and 2012. Now We introduced the keywords “outlier” OR “outlying data” OR “maximum value” OR “nasty data” (in reference to **McClelland's** chapter on the subject) OR “extreme data” for searching this database. There were 127 related hits. We used a coded method to manage with outliers (see Fig. 1), whether the mean plus/minus a coefficient (2, 2.5 or 3) times the standard deviation or the interquartile method (Mostly used method to determine outliers, consider the example Rousseeuw& Croux, 1993), or another way (e. g. a method specifically established for reaction times by Ratcliff, 1993). No article mentioned used the Median Absolute Deviation described Downward.

This survey releases the lack of concern for the misconduct of outliers, even in recently published papers. Actually, in almost all cases researchers did not report the way used to handle outliers or excluded values over two or three standard deviations around the mean, which is a bad indicator.

In conclusions, we show a robust and easy to conduct method, for detecting outlying values in a univariate statistic the Median Absolute Deviation. Initially, that indicator was developed by statisticians but is comparatively hidden in psychology. In this paper, we present this method, building on the statistical literature, and consider its relevance to our field.

2. THE MEAN PLUS OR MINUS THREE STANDARD DEVIATIONS

Not-with-standing 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 involves extracting the values that occur only in 0.13% of all cases does not seem too conservative. Another writer (e.g. Miller, 1991) recommend being less demanding, and use 2.5 or even 2 standard deviations around the mean. This option clearly depends on the circumstances and on the point of view defended by the researcher.

Unfortunately, three problems can be analyzed when using the mean as the central tendency mark (Miller, 1991). Firstly, it considers 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.

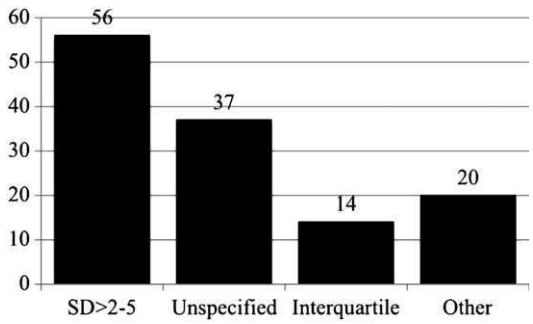

Fig 1: observe methods used to cope with outliers in JPSP and PS between 2010 and2012. Note: N = 127; SD > 2–5 = deviation from 2 to 5 SD around the mean; undefined = authors did not outline the method used to cope with outliers.

Correspondingly, 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, let suppose a small set of n= 8 observations with values 1, 3, 3, 6,8, 10, 10, and 1000. Clearly, one observation is an outlier (and we create it particularly salient for the argument). The mean is 130.13 and the not corrected standard deviation is 328.80. Therefore, using the basis of 3 standard deviations to be conservative, we could delete the values between−856.27 and 1116.52. The distribution is clearly not common (Kurtosis = 8.00; Skewness = 2.83), and the mean is inconsistent with the 7first 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.

3. An alternative: the median absolute deviation(MAD)

Absolute deviation from the median was again discovered and popularized by Hampel (1974) who allocate the idea to Carl FriedrichGauss (1777–1855). The median (M) is, like the mean, a measure of core tendency but offers the benefits of being very insensitive to the presence of outliers. The "breakdown point" is One indicator of this insensitivity (see, e.g., Donoho & Huber, 1983). The estimator's breakdown point is the maximum proportion of surveys that can be contaminated (i.e., set to infinity) without compelling the estimator to result in a false value (infinite or null in the case of an estimator of scale). For instance, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. On another hand, the median value remains the same. The median becomes instance only when more than 50% of the observations are Limitless. With a breakdown point of 0.5, the median is the location estimator that has the biggest breakdown point. Totally the same can be said about the Median Absolute Deviation as an estimator of scale (see the formula below for a definition). In addition, the MAD is totally resistant 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). To cit an example it is more robust than the classical interquartile range (see Rousseeuw &Croux, 1993), which has a breakdown point of 25% only.

Now, to calculate the median, observations have to be arranged in ascending order to identify the mean rank of the statistical series and to determine the value associated with that rank. Let us now assume 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). Therefore, the median 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. other precisely, the MAD is defined as follows (Huber, 1981): MAD=bMixi−Mjxj

where the xj is then original observations and Mi is the median of the series. Normally, b= 1.4826, a constant associated with the supposition of normality of the data, ignore the irregularity 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 regularity, 1/Q (0.75) = 1.4826 (Huber, 1981). This multiplication bybisa testing, as otherwise, the formula for the MAD would only estimate the scale up to a multiplicative constant.

Calculating the MAD implicit the next steps:(1) 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, and993; (2) when ranked, we obtain: 1, 1, 3, 3, 4, 4, 6, and 993; (3) and (4) the median equals 3.5 and will be multiplied by 1.4826 to find a MAD of 5.1891.

Then, we must define the unacceptable criterion of value. As for the mean and standard deviation, it is essential to define a level of decision: This remains required subjective points of the decision.

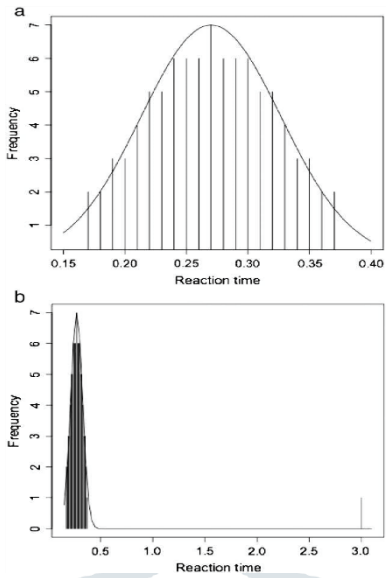

Fig 2. Outlier generating unsymmetrical. 1) Normal distribution, n = 91, mean = 0.27, median = 0.27, standard deviation = 0.06. 2) Asymmetry due to an outlier, n = 91, mean = 0.39, median = 0.27, standard deviation = 0.59

Steps 1)
DATASET ACTIVATE "Database's name".
Frequencies Variable =" first variable's name"
/STATISTICS=MEDIAN
/ORDER=ANALYSIS
Step 2)
COMPUTE" Computed second variable's name" =" First variable name"- "median"
EXECUTE
Step c)
FREQUENCIES VARIABLES=" Computed second variable's name"
/STATISTICS=MEDIAN
/ORDER=ANALYSIS

Fig 3. SPSS script for steps 1,2 and 3

Determined by the hardness of the researcher's basis, which should be explained and established by the investigator, Miller (1991) suggest the values of 3 (very conservative), 2.5 (relatively conservative) or even 2 (badly conservative). Let us use the identical limit as in the previous example and choose the threshold 3 for our example. The decision criterion becomes:

$$M-3*MA<xi<M+3*MAD$$

Or

$$\frac{|+3|}{MAD}$$

In our example, all values larger than 7 + (3*5.1891) = 22.57and all values smaller than 7 - (3*5.1891) =−8.57 can be erased. We show differently, we can abolish 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. Thus, we found (1000–7) / 5.19 = 191.36. We definitely see that this value strongly deviates from the threshold of 3 chosen previously.

Let us briefly examine the case of a fictional series in Fig. 2, which involve a huge number of observations. Fig. 2a shows a normal distribution and reports the mean, Standard deviation, and median. Fig. 2b shows a similar 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 standard deviation have extreme changed whereas the median remains the same.

Even if the dispersion was very small for sensitive reasons, we would have get an interval for identifying outliers of−0.57<xi<1.17by the method of the mean plus or minus three standard deviations and, by contrast, an interval of 0.09<xi<0.45 when using the method of the median plus or minus three times the MAD.

4. METHODOLOGY 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: (1) compute the median using the menu "Analysis" and the command "Frequency";(2) subtract this value from all observations in the statistical series using the command "Compute" in the menu "Transform"; (3) compute the median of the resulting new variable as in the first point ,and (4) multiply this value by 1.4826 (if we assume normality other data).Fig. 3shows the SPSS script for step 1 to 3 Step 4 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". Observe that this command assumes by default that b = 1.4826.

DISCUSSION

Our survey's results of two journals focus bad management of outliers, we represent that the way normally used ("The mean plus or minus three standard deviations" rule) is difficult and we state 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 the presence of outliers, and hence we strongly recommend the median plus or minus 2.5times 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 decent choice.
3. We motivate researchers to report information about outliers, namely: the number of outliers removed and them value (or at least the distance between outliers and the selected threshold).

Generally, we believe that faced with the issue presented by researchers' degrees of freedom (see Simmons et al., 2011), the inadequate knowledge of various outlier-detecting ways is not the prime challenge facing psychological science. Achieving a concert as to which method is most suitable and which individual threshold should be used (regardless of the method used) is of even greater importance. In other ways, the doubtful 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. According to outlier management, we expect that if such a concert can be achieved, our presentation of the MAD will have contributed to it.

References:

[1]. Cousineau D, Chartier S.2010.Outliers recognition and treatment: A review. International Journal of Psychological Research:3(1).58–67
[2]. Donoho D. L, Huber P. J.1983. In Bickel, Doksum, & Hodges (Eds.), The idea of breakdown point. California: Wadsworth
[3]. Hampel F. R. 1974. The influence curve and its role in robust estimation. Journal of the American Statistical Association:346(69).383–393
[4]. Howell D. C. 1998.Statistical technique in human sciences. New York: Wadsworth.
[5]. Huber, P. J. 1981. Robust statistics. New York: John Wiley.
[6]. McClelland, G. H. 2000. Nasty data: Unruly, ill-mannered observations can ruin your analysis.
[7]. H. T. Reis, & C. M. Judd (Eds.), instruction booklet of research methods in social and personality psychology. Cambridge: Cambridge University Press:393–411
[8]. Miller, J. 1991. Reaction time analysis with outlier exclusion: Bias varies with sample size. The Quarterly Journal of Experimental Psychology:43(4).907–912
[9]. 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
[10]. Ratcliff, R. 1993. Methods for conducting with reaction time outliers. Psychological Bulle-tin:114.510–532.
[11]. Rousseeuw, P. J., Croux, C. 1993. Substitute to the median absolute deviation. Journal of the American Statistical Association:424(88).1273–1283
[12]. Simmons, J. P., Nelson, L. D., Simonsohn, U. 2011. False positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. Psychological Science:22(11).1359–1366