



City Research Online

City St George's, University of London

Citation: Jones, P. R. (2016). A note on detecting statistical outliers in psychophysical data (10.1101/074591). .

This is the draft version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/20361/>

Link to published version: <https://doi.org/10.1101/074591>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

Detecting statistical outliers in psychophysical data

Pete R. Jones^{1,2}

¹Institute of Ophthalmology, University College London (UCL), UK, EC1V 9EL

²NIHR Moorfields Biomedical Research Centre, London, UK, EC1V 2PD

p.r.jones@ucl.ac.uk

Abstract: This paper considers how best to identify statistical outliers when the underlying sampling distribution is unknown. Eight methods are described, and each is evaluated using Monte Carlo simulations of a typical psychophysical experiment. The best method is shown to be one based on a measure of absolute-deviation known as S_n . In particular, this method is shown to be more accurate than popular heuristics based on standard deviations from the mean, and more robust than non-parametric methods based on interquartile range.

PACS numbers: 43.66.Yw

1. The problem of outliers

A statistical outlier is an observation that diverges abnormally from the overall pattern of data. They are often generated by a process qualitatively distinct from the main body of data. For example, in psychophysics, spurious data can be caused by technical error, faulty transcription, or — perhaps most commonly — participants being unable or unwilling to perform the task in the manner intended (e.g., due to boredom, fatigue, poor instruction, or malingering). Whatever the cause, statistical outliers can profoundly affect the results of an experiment¹, making similar populations appear distinct (Fig 1A, *top panel*), or distinct populations appear similar (Fig 1A, *bottom panel*). For example, it is tempting to wonder how many ‘developmental’ differences between children and adults are due to a small subset of non-compliant children.

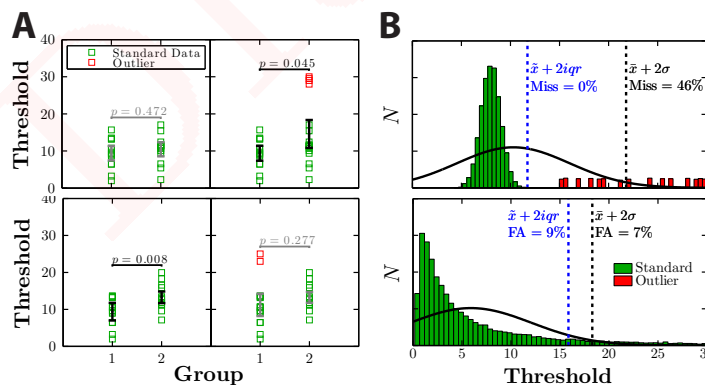


Fig 1. Examples of (A) how the presence outliers can qualitatively affect the overall pattern of results, and (B) common errors made by existing methods of outlier identification heuristics. P -values pertain to the results of between-subject t -tests. See body text for details.

18 2. General approaches and outstanding questions

19 One way to militate against outliers is to only ever use non-parametric statistics (i.e.,
 20 which have a high breakdown point², and so tend to be robust against extreme
 21 values). In reality though, this approach often proves impractical, since non-parametric
 22 methods are less powerful, less well understood, and less widely available than
 23 their parametric counterparts. Alternatively, some experimenters identify and remove
 24 outliers ‘manually’, using some unspecified process of ‘inspection’. This approach is
 25 not without merit. However, when used in isolation, manual inspection is susceptible
 26 to bias and human error, and it precludes rigorous replication or review. Finally then,
 27 statistical outliers can be identified numerically. If the underlying sampling distribution
 28 is known, then it is trivial to set a cutoff based on the likelihood of observing a given
 29 data point. However, when the sampling distribution is unknown, researchers are
 30 often compelled to use numerical heuristics, such as “was the data point more than N
 31 standard deviations from the mean?”. Currently, however, a plethora of such heuristics
 32 exist. It is unclear which method works best, and at present unscrupulous individuals
 33 are free to pick-and-choose whichever yields the outcome they expect/desire. The
 34 goal of this work was therefore (i) to describe what methods are currently available
 35 for identifying statistical outliers (in datasets generated from an unknown sampling
 36 distribution), and (ii) to use simulations to assess how well each method performs in
 37 a typical psychophysical context.

38 3. State-of-the-art methods for identifying statistical outliers

39 Here we describe eight methods for identifying statistical outliers. Five of these
 40 methods are also shown graphically in Fig 2.

41 ***SD*** x_i =outlier if it lies more than λ standard deviations, σ , from the mean, \bar{x} :

$$|x_i| > (\bar{x} + \lambda\sigma), \quad \text{(Eq 1)}$$

42 where λ is typically between 2 (liberal) and 3 (conservative). This is one of the most
 43 commonly used heuristics, but is theoretically flawed. Both the \bar{x} and σ terms are easily
 44 distorted by extreme values, meaning that more distant outliers may ‘mask’ lesser ones.
 45 This can lead to false negatives (identifying outliers as genuine data; Fig 1B, *top panel*).
 46 The method also assumes symmetry (i.e., attributes equal importance to positive and
 47 negative deviations from the center), whereas psychometric data are often skewed.
 48 This can lead to false positives (identifying genuine data as outliers; Fig 1B, *bottom*
 49 *panel*). Furthermore, while *SD* does not explicitly require normality, the $\pm\lambda\sigma$ bracket
 50 may include more or less data than expected if the data are not Gaussian distributed.
 51 For example, $\pm 2\sigma$ includes 95% of data when Gaussian distributed, but as little as 75%
 52 otherwise (*Chebyshev’s inequality*).

53 ***GMM*** x_i =outlier if it lies more than λ standard deviations from the mean of the
 54 *primary component of a Gaussian Mixture Model*:

$$|x_i| > (\bar{x}_1 + \lambda\sigma_1) \quad \text{where} \quad pdf(x) = \omega\Phi(x; \mu_1, \sigma_1) + (1 - \omega)\Phi(x; \mu_2, \sigma_2). \quad \text{(Eq 2)}$$

55 An obvious extension to *SD*: The two methods are identical, except that when fitting
 56 the parameters to the data, the *GMM* model also includes a secondary component
 57 designed to capture any outliers (see Fig 2). The secondary component is not used to
 58 identify outliers per se, but prevents extreme values from distorting the parameters
 59 of the primary component. In practice the fit of the secondary component must be
 60 constrained to prevent it ‘absorbing’ non-outlying points (see *Supplemental Material*).

61 ***rSD*** Same as *SD*, but applied recursively until no additional outliers are identified:

$$\begin{cases} |x_i^0| > (\bar{x}_0 + \lambda\sigma_0) \\ |x_i^n| > (\bar{x}_n + \lambda\sigma_n). \end{cases} \quad (\text{Eq 3})$$

62 This approach aims to solve the problem of masking by progressively peeling away
63 the most extreme outliers. However, like *SD*, it remains intolerant to non-Gaussian
64 distributions. In situations where samples are sparse/skewed, this approach therefore
65 risks aggressively rejecting large quantities of genuine data (see Fig 1B). Users typically
66 attempt to compensate for this by using a relatively high criterion level, and/or by
67 limiting the number of recursions (e.g., $\lambda \geq 3, n_{\max} = 3$).

68 ***IQR*** x_i =outlier if it lies more than λ times the interquartile range from the median:

$$|x_i| > (\tilde{x} + \lambda iqr). \quad (\text{Eq 4})$$

69 This is a non-parametric analog of the *SD* rule: substituting median and *iqr* for mean
70 and standard deviation. Unlike *SD*, the key statistics are relatively robust. Thus, the
71 breakdown points for \tilde{x} and *iqr* are 50% and 25% (respectively), meaning that outliers
72 can constitute up to 25% of the data before the statistics start to be distorted³.
73 However, like *SD*, the *IQR* method only considers absolute deviation from the center. It
74 is therefore insensitive to any asymmetry in the sampling distribution (Fig 1B, *bottom*).

75 ***prctile*** x_i =outlier if it lies above the λ^{th} percentile, or below the $(1 - \lambda)^{\text{th}}$:

$$x_i > P_\lambda \quad \text{or} \quad x_i < P_{1-\lambda}. \quad (\text{Eq 5})$$

76 This effectively ‘trims’ the data, rejecting the most extreme points, irrespective of their
77 values. Unlike *IQR*, this method is sensitive to asymmetry in the sampling distribution.
78 But it is otherwise crude in that it ignores any information contained in the spread of
79 the data points. The *prctile* method also begs the question in that the experimenter
80 must estimate, *a priori*, the number of outliers that will be observed. If λ is set
81 incorrectly, genuine data will be excluded, or outliers missed.

82 ***Tukey*** x_i =outlier if it lies more than λ times the *iqr* from the 25th/75th percentile:

$$x_i > (P_{75} + \lambda iqr) \quad \text{or} \quad x_i < (P_{25} - \lambda iqr). \quad (\text{Eq 6})$$

83 Popularized by John W. Tukey, this attempts to combine the best features of the *IQR* and
84 *prctile* method. The information contained in the spread of data, *iqr*, is combined with
85 the use of lower/upper quartile ‘fences’ that provide some sensitivity to asymmetry.

86 ***MAD_n*** x_i =outlier if it lies farther from the median than λ times the median absolute
87 distance [MAD] of every point from the median:

$$\left(\frac{|x_i - \tilde{x}|}{MAD_n} \right) > \lambda \quad \text{where} \quad MAD_n = 1.4826 \operatorname{med}_{i=1:n} |x_i - \operatorname{med}_{j=1:n} x_j|, \quad (\text{Eq 7})$$

88 where 1.4826 is simply a scaling factor, used for consistency with the standard
89 deviation over a Gaussian distribution (see Ref [3]). Unlike the non-parametric
90 methods described previously, this method uses MAD rather than *iqr* as the measure of
91 spread. This makes this method more robust, as the MAD statistic has the best possible
92 breakdown point (50%, versus 25% for *iqr*). However, as with *IQR*, *MAD_n* assumes
93 symmetry, only considering the absolute deviation of datapoints from the center.

94 ***S_n*** x_i =outlier if the median distance of x_i from all other points, is greater than λ
95 times the median distance of every point from every other point:

$$\left(\frac{\operatorname{med}_{j \neq i} |x_i - x_j|}{S_n} \right) > \lambda \quad \text{where} \quad S_n = 1.1926 c_n \operatorname{med}_{i=1:n} \left\{ \operatorname{med}_{j \neq i} |x_i - x_j| \right\}, \quad (\text{Eq 8})$$

96 where 1.1926 is again for consistency with the standard deviation, and c_n is a finite
 97 population correction parameter (see Ref [3]). Like MAD, the S_n term is maximally
 98 robust. However, this method differs from MAD_n in that S_n considers the typical
 99 distance between all data points, rather than measuring how far each point is from a
 100 central value. It therefore remains valid even if the sampling distribution is asymmetric.
 101 The historic difficulty with S_n is its long computational time [$O(n^2)$]. However, for
 102 psychophysical applications this is trivial given modern computing.

103 **4. Comparison of techniques using simulated psychophysical observers**

104 To assess the eight methods described in Section 3, we applied each to random
 105 samples of data pre-labeled as either ‘good’ or ‘bad’. However, rather than simply
 106 specifying arbitrary sampling distributions for each of these categories, we generated
 107 data by simulating a typical two-alternative forced-choice [2AFC] experiment in which
 108 a 2-down 1-up transformed staircase⁴ was applied to N simulated observers. Each
 109 observer consisted essentially of a randomly generated psychometric function, and
 110 made stochastic, trial-by-trial responses based on the current stimulus level and a
 111 random sample of additive internal noise (i.e., the variance of which was determined
 112 by the slope of their psychometric function). Trial-by-trial response data were then
 113 processed and analyzed as if from human participants, leading, for example, to the
 114 sampling-distributions of 70.7% thresholds shown in Fig 2 (bottom right).

115 Of the N observers, $X\%$ were ‘non-compliant’ (on average, their psychometric
 116 functions had a higher mean, standard deviation, and lapse-rate), and were thus
 117 likely to produce outlying data points (Fig 2, red bars). The remaining observers were
 118 ‘compliant’ (on average lower mean, standard deviation, and lapse-rate), and produced
 119 the distribution of ‘good’ data shown in green. Precise details of all test parameters can
 120 be found in the *Supplemental Material*, which contains the complete MATLAB code used
 121 to generate all of the data presented here. N took the values {8, 32, 128}, representing
 122 small, medium, and large sample sizes, while the number of non-compliant observers
 123 varied from 0 to 50% of N (e.g., {0, 1, ..., 16}, when when $N=32$). For each condition,
 124 2,000 independent simulations were run, for a total of 108K simulations.

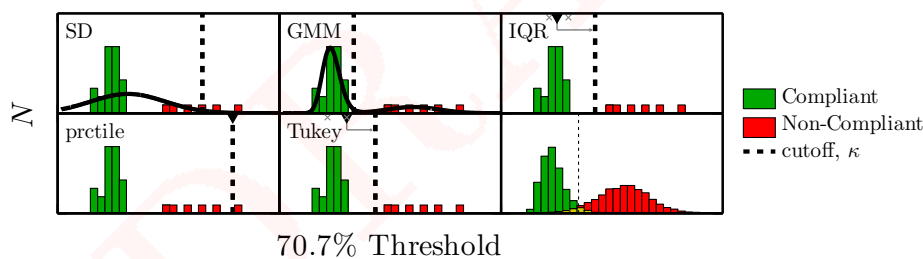


Fig 2. Simulation methods. Random sample of thresholds were generated, of which $X\%$ came from ‘non-compliant’ simulated observers (here: $N=32$, $X\%=19$). Each of eight methods were then used to identify which observations were generated by ‘non-compliant’ observers (i.e., likely statistical outliers). Only five methods are depicted here, as the other three (rSD , MAD_n and S_n) have no obvious graphical analog. The final panel shows the full sampling distributions over 20,000 trials, and the ideal unbiased classifier, for which: Hit rate = 0.97, False Alarm = 0.05.

125 **Results and Discussion**

126 The results are shown in Fig 3. We begin by considering only the case where $N=32$
 127 (Fig 3, middle column), before considering the effect of sample size.

128 As expected, the SD rule proved poor. When $\lambda=3$, it was excessively conservative –
 129 seldom exhibiting false alarms, but missing the great majority of outliers, particularly

130 as the number of outliers increased. Lowering the criterion to $\lambda=2$ yielded more
 131 reasonable results. However, *SD* still exhibited a lower hit rate than most other
 132 methods, and also exhibited a high false alarm rate when there were few/no outliers.
 133 The modified *GMM* and *rSD* rules exhibited increased robustness and accuracy,
 134 respectively. However, compared to non-parametric methods, they were generally only
 135 more sensitive than the *prctile* method, which was only accurate when the predefined
 136 exclusion rate matched the true number of outliers exactly.

137 The two *iqr*-based methods, *IQR* and *Tukey*, exhibited high sensitivity when the
 138 number of outliers was low ($\leq 20\%$). However, as expected, they exhibited a marked
 139 deterioration in hit rates when the number of outliers increased beyond 20% (i.e., in
 140 accordance with the 25% breakdown point for *iqr*).

141 The two median-absolute-deviation-based methods, MAD_n and S_n , were as sensitive
 142 as all other methods when outliers were few ($\leq 20\%$), and were more robust than
 143 the *iqr* methods – continuing to exhibit high hit rates and few false alarms even when
 144 faced with large numbers of outliers. Compared to each other, MAD_n and S_n performed
 145 similarly. However, the S_n statistic makes no assumption of symmetry, and so ought to
 146 be superior in situations where the sampling distribution is heavily skewed.

147 We turn now to how sample size affected performance. With large samples ($N=128$),
 148 the pattern was largely unchanged from the medium sample-size case ($N=32$), except
 149 that *rSD* exhibited a marked increase in false alarms, making it an unappealing option.
 150 With small samples ($N=8$), the *prctile* and *rSD* methods became uniformly inoperable,
 151 while most other methods were unable to identify more than a single outlier. The MAD_n
 152 and S_n methods, however, remained relatively robust, and generally performed well,
 153 though they did exhibit an elevated false alarm rate when there were few/no outliers.
 154 It may be that this could be rectified by increasing the criterion, λ , as a function of N ,
 155 however this was not investigated here. The *GMM* method also performed well overall
 156 in the small-sample condition. However, it did also exhibit the highest false alarm rate
 157 when there were no outliers, and was only more sensitive than MAD_n or S_n when the
 158 proportion of outliers was extremely high ($>33\%$).

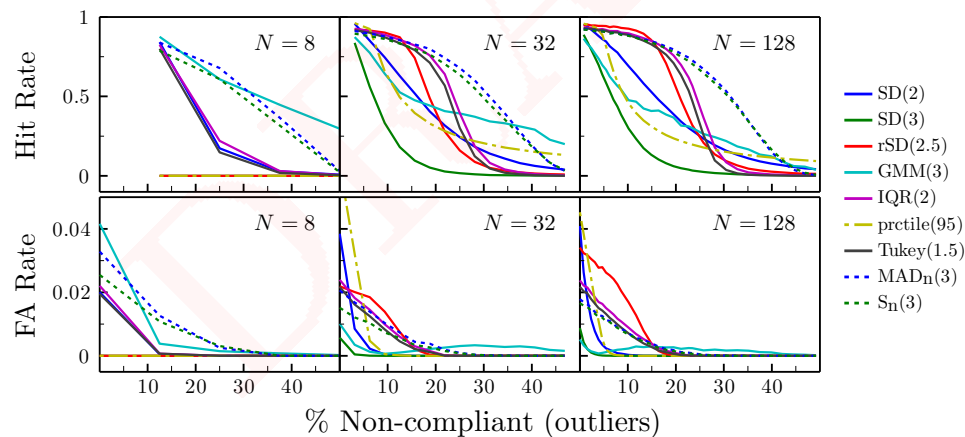


Fig 3. Simulation results. The eight classifiers described in Section 3 were used to distinguish between random samples of ‘compliant’ and ‘non-compliant’ simulated observers (see Fig 2). Numbers in parentheses indicate the criterion level, λ , used by each classifier.

159 5. Summary and concluding remarks

160 Of the eight methods considered, S_n proved the most sensitive and robust. Specific
161 situations were observed in which other heuristics performed as-well-as or even better
162 than S_n : for example, when the sample size was large (rSD), or when the proportion
163 of outliers was very low (IQR , $Tukey$) or very high (GMM). However, most methods
164 were less sensitive in than S_n in the majority circumstances, and failed precipitously
165 in some circumstances, making them unattractive alternatives. The related method
166 MAD_n also proved strong, and could be considered a viable alternative to S_n . However,
167 as discussed in *Section 3*, MAD_n assumes a symmetric sampling distribution, and so
168 would not be expected to perform as well if the sampling distribution was very heavily
169 skewed (e.g., when dealing with reaction time data). The popular SD metric proved
170 particularly poor in all circumstances, and should never be used. In short, S_n appears to
171 provide the best means of identifying statistical outliers when the underlying sampling
172 distribution is unknown. Its use may be particularly beneficial for researchers working
173 with small/irregular populations such as children, animals, or clinical cohorts. MATLAB
174 code for computing S_n is provided in the *Supplemental Material*.

175 *Limitations of the present study*

176 The present findings are predicated on finite simulations of a single experimental
177 paradigm, and so cannot be guaranteed to generalize. Anecdotally, the same overall
178 pattern of results remained unchanged when key parameters were varied (e.g.,
179 properties of the observers and/or of the experimental paradigm). However, there
180 exist an infinite number of possible circumstances, and some experimental paradigms
181 — particularly those involving advanced adaptive procedures — are capable of
182 producing quite complex (e.g., bimodal) sampling distributions. With this in mind,
183 the code in *Supplemental Material* also provides support for a variety of paradigms
184 (transformed/weighted staircases, Constant Stimuli, and various more advanced
185 procedures, implemented via the Palamedes toolbox⁵). Readers are encouraged to
186 simulate their own experimental configurations, to assess how each method performs.

187 *On the ethics of excluding statistical outliers*

188 Excluding outliers is often regarded as poor practice. As shown in *Section 1*, however,
189 the exclusion of outliers can sometimes be preferable to reporting misleading results.
190 Automated methods of statistical outlier identification should never be used blindly
191 though, and they are not a replacement for common sense. Where feasible, datapoints
192 identified as statistical outliers should only be excluded in the presence of independent
193 corroboration (e.g., experimenter observation). Furthermore, best practice dictates
194 that when outliers are excluded, they should continue to be shown graphically, and
195 all statistical analyses should be run twice: with and without outliers included.

196 **Acknowledgments**

197 This work was supported by the NIHR Biomedical Research Centre located at (both)
198 Moorfields Eye Hospital and the UCL Institute of Ophthalmology.

199 **References**

- 200 [1] Osborne, J. W. & Overbay, A. The power of outliers (and why researchers should always check for them).
201 *Practical assessment, research & evaluation* **9**, 1–12 (2004).
- 202 [2] Huber, P. J. *International Encyclopedia of Statistical Science*, chap. Robust Statistics, 1248–1251 (Springer
203 Berlin Heidelberg, Berlin, Heidelberg, 2011).
- 204 [3] Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *Journal of the American*
205 *Statistical association* **88**, 1273–1283 (1993).
- 206 [4] Levitt, H. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of*
207 *America* **49**, 467–477 (1971).
- 208 [5] Kingdom, F. & Prins, N. *Psychophysics: a practical introduction* (Elsevier Academic Press, 2010), 1 edn.