Detecting statistical outliers in psychophysical data

Pete R. Jones[1,2]

[1]*Institute of Ophthalmology, University College London (UCL), UK, EC1V 9EL*
[2]*NIHR 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 experiment[1], 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.

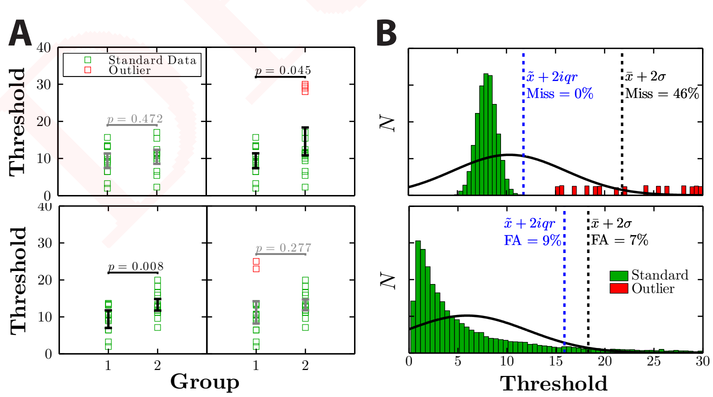

Fig 1. Examples of (***A***) how the presence outliers can qualitatively affect the overall pattern of results, and (***B***) common errors made by existing methods of outlier identification heuristics. P-values pertain to the results of between-subject t-tests. See body text for details.

2. General approaches and outstanding questions

One way to militate against outliers is to only ever use non-parametric statistics (i.e., which have a high breakdown point[2], and so tend to be robust against extreme values). In reality though, this approach often proves impractical, since non-parametric methods are less powerful, less well understood, and less widely available than their parametric counterparts. Alternatively, some experimenters identify and remove outliers ‘manually’, using some unspecified process of ‘inspection’. This approach is not without merit. However, when used in isolation, manual inspection is susceptible to bias and human error, and it precludes rigorous replication or review. Finally then, statistical outliers can be identified numerically. If the underlying sampling distribution is known, then it is trivial to set a cutoff based on the likelihood of observing a given data point. However, when the sampling distribution is unknown, researchers are often compelled to use numerical heuristics, such as “was the data point more than N standard deviations from the mean?”. Currently, however, a plethora of such heuristics exist. It is unclear which method works best, and at present unscrupulous individuals are free to pick-and-choose whichever yields the outcome they expect/desire. The goal of this work was therefore *(i)* to describe what methods are currently available for identifying statistical outliers (in datasets generated from an unknown sampling distribution), and *(ii)* to use simulations to assess how well each method performs in a typical psychophysical context.

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

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

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

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

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

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

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

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

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

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

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

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

$$|x_i| > (\tilde{x} + \lambda iqr)\,. \qquad \textbf{(Eq 4)}$$

This is a non-parametric analog of the *SD* rule: substituting median and iqr for mean and standard deviation. Unlike *SD*, the key statistics are relatively robust. Thus, the breakdown points for $\tilde{x}$ and iqr are 50% and 25% (respectively), meaning that outliers can constitute up to 25% of the data before the statistics start to be distorted[3]. However, like *SD*, the *IQR* method only considers absolute deviation from the center. It is therefore insensitive to any asymmetry in the sampling distribution (Fig 1B, *bottom*).

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

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

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

Tukey x_i=outlier if it lies more than λ times the iqr from the 25$^{\text{th}}$/75$^{\text{th}}$ percentile:

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

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

MAD$_n$ x_i=outlier if it lies farther from the median than λ times the median absolute distance [MAD] of every point from the median:

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

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

S$_n$ x_i=outlier if the median distance of x_i from all other points, is greater than λ times the median distance of every point from every other point:

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

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

4. Comparison of techniques using simulated psychophysical observers

To assess the eight methods described in *Section 3*, we applied each to random samples of data prelabeled as either 'good' or 'bad'. However, rather than simply specifying arbitrary sampling distributions for each of these categories, we generated data by simulating a typical two-alternative forced-choice [2AFC] experiment in which a 2-down 1-up transformed staircase[4] was applied to N simulated observers. Each observer consisted essentially of a randomly generated psychometric function, and made stochastic, trial-by-trial responses based on the current stimulus level and a random sample of additive internal noise (i.e., the variance of which was determined by the slope of their psychometric function). Trial-by-trial response data were then processed and analyzed as if from human participants, leading, for example, to the sampling-distributions of 70.7% thresholds shown in Fig 2 (*bottom right*).

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

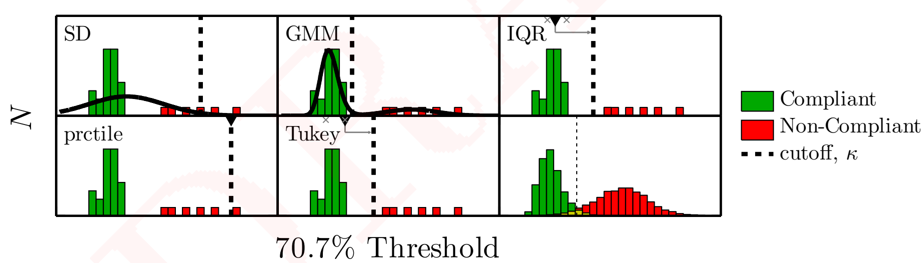

Fig 2. Simulation methods. Random sample of thresholds were generated, of which $X\%$ came from 'non-compliant' simulated observers (here: N=32, $X\%$=19). Each of eight methods were then used to identify which observations were generated by 'non-compliant' observers (i.e., likely statistical outliers). 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.

Results and Discussion

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

As expected, the *SD* rule proved poor. When λ=3, it was excessively conservative – seldom exhibiting false alarms, but missing the great majority of outliers, particularly as the number of outliers increased. Lowering the criterion to λ=2 yielded more reasonable results. However, *SD* still exhibited a lower hit rate than most other methods, and also exhibited a high false alarm rate when there were few/no outliers. The modified *GMM* and *rSD* rules exhibited increased robustness and accuracy, respectively. However, compared to non-parametric methods, they were generally only more sensitive than the *prctile* method, which was only accurate when the predefined exclusion rate matched the true number of outliers exactly.

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

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

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

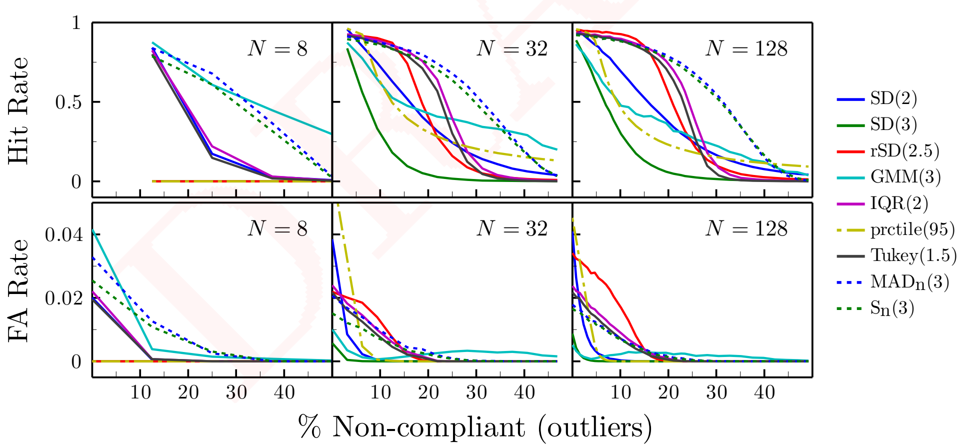

Fig 3. Simulation results. The eight classifiers described in *Section 3* were used to distinguish between random samples of 'compliant' and 'non-compliant' simulated observers (see Fig 2). Numbers in parentheses indicate the criterion level, λ, used by each classifier.

5. Summary and concluding remarks

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

Limitations of the present study

The present findings are predicated on finite simulations of a single experimental paradigm, and so cannot be guaranteed to generalize. Anecdotally, the same overall pattern of results remained unchanged when key parameters were varied (e.g., properties of the observers and/or of the experimental paradigm). However, there exist an infinite number of possible circumstances, and some experimental paradigms — particularly those involving advanced adaptive procedures — are capable of producing quite complex (e.g., bimodal) sampling distributions. With this in mind, the code in *Supplemental Material* also provides support for a variety of paradigms (transformed/weighted staircases, Constant Stimuli, and various more advanced procedures, implemented via the Palamedes toolbox[5]). Readers are encouraged to simulate their own experimental configurations, to assess how each method performs.

On the ethics of excluding statistical outliers

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

Acknowledgments

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

References

[1] Osborne, J. W. & Overbay, A. The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation* **9**, 1–12 (2004).

[2] Huber, P. J. *International Encyclopedia of Statistical Science*, chap. Robust Statistics, 1248–1251 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

[3] Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *Journal of the American Statistical association* **88**, 1273–1283 (1993).

[4] Levitt, H. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America* **49**, 467–477 (1971).

[5] Kingdom, F. & Prins, N. *Psychophysics: a practical introduction* (Elsevier Academic Press, 2010), 1 edn.