

Discarding the mean — An alternative way of analysing data.

N M Ridler¹

Division of Electrical Science, National Physical Laboratory, based at DRA Malvern.

Abstract

This paper presents a simple method for making quantitative estimates, *i.e.*, an average value, a measure of dispersion and a confidence interval, from measurement data. The method is based on arranging the data in ascending order of size and has been applied to analysis of measurement comparison results and for measurement uncertainty estimates. The method can out-perform current recommended practices, especially when the measurement data set contains a small number of erratic values.

1 Introduction

One of the fundamental tasks in measurement is estimating a value for the quantity being measured and determining the reliability of the estimate. The estimate, and its reliability, are often assessed by interpreting the dispersion of a series of scattered values. The scatter may have been produced by a single measurement system (*e.g.*, from multiple, repeated, measurements made during an uncertainty assessment) or from multiple systems (*e.g.*, during measurement comparison exercises).

Deriving a good estimate for a value, and its reliability, from a set of repeated measurements may, at first glance, seem to be a trivial task. The conventional method [1,2] is to calculate the arithmetic mean and standard deviation for the set of values and use these as the estimated value and a measure of the scatter, respectively. However, the mean and standard deviation can sometimes be of limited use, especially if the results being analysed exhibit erratic variation, *e.g.*, when one or two results are radically different from the majority. This situation sometimes occurs when analysing what should be similar data, from a variety of sources (inter-laboratory comparisons). It can also occur in repeat measurements made on a single system when one of the measurements is affected by an undetected gross error.

This paper presents an alternative method of analysing measurement data which relies on arranging the data in ascending order of size, termed *order statistics*. This alternative method produces estimates which are more resilient to the presence of a small number of erratic values in the data being analysed.

2 Order statistics

Arranging measurement data in ascending order of size enables a number of different statistics to be used during the data analysis. In this paper, just two are examined; the median and inter-quartile range (IQR), which can be used to provide an average value and a measure of dispersion, respectively. Having ordered the data, the lower quartile divides the order statistics so that 25% of the results are below its value, the median has 50% below and 50% above, and the upper quartile has 25% above its value. The IQR is the difference between the values for the upper and lower quartiles.

However, the ordered data cannot always be split exactly into four equal parts. This calls for more precise definitions for these terms, which are given in terms of a calculated

¹ Mr Ridler is employed by Assessment Services Ltd and works under contract to NPL.

'depth', *i.e.*, how far to go through the data to find the required value. For n observations, the depth for the median and quartiles are as follows;

$$\text{Depth of median} = \frac{n + 1}{2} ; \quad \text{Depth of quartiles} = \frac{\text{Depth of median} + 1}{2}$$

The depth for the lower quartile is found by entering the data from the low end (*i.e.*, starting with the lowest value) and the depth for the upper quartile is found by entering from the high end (starting with the highest value).

3 Confidence intervals

The IQR can be used to provide an estimate of the standard deviation, s . For a normal distribution, the standard deviation is approximately $\frac{3}{4}$ IQR. If we assume that the data being analysed conforms to a normal distribution (with an occasional unusual value) we can use this relationship to estimate the standard deviation and also a confidence interval for the estimated average (using Student's t).

$$\text{Confidence interval} = t \frac{s}{\sqrt{n}}$$

The confidence interval can be used, in conjunction with the median, to report the size of an uncertainty contribution to a measurement or when reporting results from measurement comparisons.

4 Example

To demonstrate finding the median, the IQR and estimating a confidence interval, some results obtained in a recent measurement comparison exercise, reported in [3], will be used. Eight participants supplied results, including measurements of the phase angle of the voltage reflection coefficient for a 3 dB attenuator at 1 GHz. For this particular measurement, seven of the participants' results were relatively tightly grouped whilst one participant's value was very different from the others.

The analysis of the data required finding an average value, based on the participant's results, and a measure of the dispersion of the results about this average. The average value and the measure of dispersion can be used by each participant to assess the performance of their own result.

The results from the seven participants showing good agreement will be analysed first, then the analysis will be repeated including the additional eighth value. This will enable the effect of the erratic value on the statistical summaries to be assessed.

4.1 Summarising the "good" results.

The seven results, arranged as order statistics, are given below.

$$+108.9^\circ, +112.9^\circ, +114.0^\circ, +115.2^\circ, +116.3^\circ, +118.2^\circ, +120.6^\circ$$

First, calculate the depths of the median and the quartiles:

$$\text{Depth of median} = \frac{7 + 1}{2} = 4 ; \quad \text{Depth of quartiles} = \frac{4 + 1}{2} = 2.5$$

The median is therefore the fourth value (from either end) and the quartiles are midway between the second and third values (lower quartile when entering the data from the low end and the upper quartile when entering from the high end).

$$\begin{aligned}\text{Median} &= +115.2^\circ \\ \text{Upper quartile} &= +117.25^\circ \\ \text{Lower quartile} &= +113.45^\circ \\ \text{IQR} &= 3.8^\circ \\ s &= 2.9^\circ \\ 95\% \text{ confidence interval} &= \pm 2.7^\circ\end{aligned}$$

Therefore, the summary statement for this data would be;

$$\text{phase angle} = +115.2^\circ \pm 2.7^\circ$$

This can be compared with a summary produced by the conventional method based on the arithmetic mean. Mean = +115.2°, $s = 3.8^\circ$, and 95% confidence interval = $\pm 3.5^\circ$.

$$\text{phase angle} = +115.2^\circ \pm 3.5^\circ$$

4.2 Including the erratic result.

The eight results, arranged as order statistics, are given below.

$$-58.3^\circ, +108.9^\circ, +112.9^\circ, +114.0^\circ, +115.2^\circ, +116.3^\circ, +118.2^\circ, +120.6^\circ$$

Re-calculate the depths of the median and the quartiles:

$$\text{Depth of median} = \frac{8 + 1}{2} = 4.5 \quad ; \quad \text{Depth of quartiles} = \frac{4.5 + 1}{2} = 2.75$$

The median value is therefore the midway between the fourth and fifth numbers and the quartile values are three-quarters of the way between the second and third numbers.

$$\begin{aligned}\text{Median} &= +114.6^\circ \\ \text{Upper quartile} &= +116.775^\circ \approx +116.8^\circ \\ \text{Lower quartile} &= +111.9^\circ \\ \text{IQR} &\approx 4.9^\circ \\ s &= 3.7^\circ \\ 95\% \text{ confidence interval} &= \pm 3.1^\circ\end{aligned}$$

Therefore, the summary statement for this data would be;

$$\text{phase angle} = +114.6^\circ \pm 3.1^\circ$$

Again, this can be compared with a summary produced by the conventional method based on the arithmetic mean. Mean = +93.5°, $s = 61.4^\circ$, and 95% confidence interval = $\pm 51.2^\circ$.

$$\text{phase angle} = +93.5^\circ \pm 51.2^\circ$$

5 Discussion

For the seven relatively well-behaved results the summary statistics derived from ordering the data were similar to those derived from the mean. However, this was not the case when the very different value was included in the analysis.

The median value remained relatively unaffected by the inclusion of the very different result (changing from $+115.2^\circ$ to $+114.6^\circ$). Similarly, the 95% confidence interval derived from the IQR remained relatively unaffected (changing from $\pm 2.7^\circ$ to $\pm 3.1^\circ$). By comparison, the summary statistics based on the mean vary considerably: the mean changing from $+115.2^\circ$ to $+93.5^\circ$ and the 95% confidence interval changing from $\pm 3.5^\circ$ to $\pm 51.2^\circ$.

This indicates that summary statistics based on the median and the IQR are more useful than the conventional statistics, based on the mean, when analysing a relatively small series of results containing one or two erratic values. Therefore, "discarding the mean" in favour of the median (or some other suitable statistic – see, for example, [4]) can sometimes be the best course of action when analysing measurement data.

6 Summary

This paper has presented an alternative method of analysing measurement data in terms of providing an average value, a measure of dispersion, and a confidence interval in the average value. The alternative statistics, based on the median and IQR, out-perform the conventional methods in that they continue to provide a useful average value and efficient confidence intervals in the presence of a small number of erratic values.

Traditionally, statistics based on ordering data have not been used extensively to analyse data. One possible reason for this is that manually ordering measurement results can be very time consuming, especially for large numbers of results. However, these statistics can now be calculated very easily using computer spreadsheet programs containing statistical toolkits which usually include medians and quartiles as analysis options.

7 Acknowledgements

The work described in this paper is part of the Electrical Science programme at NPL and is sponsored by the National Measurement System Policy Unit of the UK government's Department of Trade and Industry.

8 References

- [1] ISO 5725-2:1994(E), "Accuracy (trueness and precision) of measurement methods and results – Part 2: Basic methods for the determination of repeatability and reproducibility of a standard measurement method".
- [2] ISO, "Guide to the expression of uncertainty in measurement", First edition, 1993.
(BSI equivalent publication: PD 6461: Part 3: 1995.)
- [3] N M Ridler and J C Medley, "A comparison of complex scattering coefficient measurements in 50 ohm coaxial line to 26.5 GHz", NPL Report DES 138, June 1995.
- [4] Analytical Methods Committee, "Robust statistics – How not to reject outliers", Analyst, vol 114, December 1989, pp 1693-1702.