

Method Article**An absolute deviation approach to assessing correlation****Abstract**

This paper describes two possible alternatives to the more traditional Pearson's R correlation coefficient, based on using the mean absolute deviation, rather than the standard deviation, as a measure of dispersion. Pearson's R is well-established and has many advantages. However, these newer variants also have several advantages, including greater simplicity and ease of computation, and greater tolerance of underlying assumptions (such as the need for linearity). Their introduction forms part of an attempt to simplify the use of numeric analysis, to make it more 'everyday', for the benefit of both analysts and consumers of evidence.

Introduction

In the UK, there is a long-standing concern about the quality and utility of social science research. There is a further long-standing concern about a shortage of young people studying science and maths once these become optional, leading to an apparent shortage of professionals in STEM areas. These twin concerns have been combined in a solution for at least part of the purported problem – a national Quantitative Methods Initiative for social science (e.g. <http://www.esrc.ac.uk/research/skills-training-development/qmi/index.aspx>). The initiative is funded by a range of partners including the Higher Education Funding Council for England, the British Academy, the Nuffield Foundation, and the Economic and Social Research Council. The focus is on the development of new methods courses, teaching resources, new training opportunities, work placements, scholarships, and the embedding of 'quantitative' evidence in substantive courses for all social scientists. The model underlying all of this activity and expenditure is that 'quantitative' work in social science is good, especially in economics and psychology, but that it is not always well taught, and there is just not enough of it outside those two areas.

Much of this diagnosis may be correct, although it is possible that the problems of social science are as much to with lack of design, genuine curiosity and research integrity as to do with lack of number. Many of the solutions also may be necessary but they will probably not be sufficient. What is being taught is currently too often wrong (the meaning of significance tests, for example), and even more often needlessly complex (the use of multi-level modelling with population data, for example). It is these confusing approaches more than anything that are driving researchers away from the relatively simple use of measurements and frequencies in their work (Copper and Shore 2008, Hampel 1997, Murtonen and Lehtinen 2003, Watts 1991). An alternative approach is to simplify what it means to use numbers in research, to stress the solid links between the logic of analysis for any kind of data, and then try to improve uptake through the pedagogy and persuasion of something like the Quantitative Methods Initiative. This paper is one step in that approach.

The paper introduces two possible alternatives to the more traditional Pearson's R correlation coefficient, both based on using the mean absolute deviation, rather than the standard deviation, as a measure of dispersion. These variants have several advantages, including

greater simplicity and ease of computation, and greater tolerance of underlying assumptions (such as the need for linearity). They form part of an attempt to simplify the use of numeric analysis, to make it more ‘everyday’, for the benefit of both analysts and consumers of evidence. The approach includes replacing the standard deviation (and its needless squaring and square rooting) with the absolute deviation where possible, introducing a robust absolute deviation effect size, and absolute deviation regression models, and of course the removal of significance testing and all of its components from everyday analysis (Gorard 2010, 2014a).

Correlation is a measure of the strength of a relationship between two variables, where each value in one variable has a corresponding or paired value in the other variable. An example might be list of test scores in maths and reading for the same children. Correlation is the key idea underlying more complex modelling including forms of regression and data reduction. It is usually based on a scatterplot showing a linear or near linear relationship between the two variables, and the correlation ‘coefficient’ is an estimate of how well the pairs of values in both variables ‘fit’ a straight line (Gorard 2003). In this type of correlation, the values must be real numbers (and there are other techniques for categorical data). The traditional, and generally excellent, correlation coefficient is Pearson’s R.

Pearson’s R correlation coefficient

Correlation generally starts with covariance. ‘Covariance’ is an estimate of the scale of inter-relatedness of two variables. For two variables x and y, we can find the deviation of each value from its mean, multiply the results for each pair of values, and sum these products for all values, before dividing by the number of values (1). This covariance, or average co-deviation, is large where each value in a pair differs from its mean in the same direction consistently.

$$\frac{n}{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]/n} \quad (\text{Covariance})$$

The covariance is standardised and adjusted for the scale of measurement by dividing it by the product of the two standard deviations for the two variables. This produces the correlation coefficient - R.

$$\frac{n}{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]/n} \quad (\text{Pearson’s R})$$

$$\frac{n}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2/n] \cdot [\sum_{i=1}^n (y_i - \bar{y})^2/n]}}$$

The numerator is the covariance between the two sets of numbers x (from 1 to n) and y (from 1 to n). It is the sum of: each numbers’ deviation from the mean of one set multiplied by the equivalent numbers’ deviation from the mean of the other set, all divided by the number of numbers in both sets. The denominator is the product of the standard deviations of the x and y sets of numbers individually. Although usually invisible to analysts using software like SPSS this is quite a complex formula, especially when the definition of the means of the x and y sets of numbers is included in full. It is not an easy formula to explain to general under-

graduates in social science, for example. The squaring and square-rooting involved in calculating the standard deviation is not intuitive, and has the effect of over-emphasising larger deviations from the mean. There are simpler alternatives.

Using the mean absolute deviation

Originally the standard deviation was devised as a way of eliminating the signs for each deviation. By definition, the sum of all of the deviations from the mean for any set of numbers will be zero. Squaring them before summing makes them all positive, and so the sum of all deviations would only be zero if every number in the set was equal to the mean (i.e. there were no deviations).

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / n} \quad (\text{Standard deviation})$$

A simpler alternative to the standard deviation, as a measure of dispersion, is the mean of the absolute deviations from the mean (Gorard 2005). Here the sum is of the absolute values (irrespective of sign) of each deviation, and because it is divided by the number of values, it is the mean absolute deviation. It gives the same substantive result as the standard deviation.

$$\sum_{i=1}^n (x_i - \bar{x}) / n \quad (\text{Mean absolute deviation})$$

It is the average by which each value deviates from the mean in each set of numbers. This everyday meaning is lost in the more complex calculation of the standard deviation. And that is part of the problem when introducing new researchers to a measure of dispersion. Ease of understanding is only one reason for increasingly preferring the absolute deviation over the standard deviation. The absolute deviation is also more robust when dealing with real-life data not necessarily following a normal distribution.

There is anyway a widespread ambiguity in the use and reporting of standard deviations. The formula makes it clear that the sum of squares is divided by n and then the square root is taken. A square root for a real positive number must always consist of two values (one positive and one negative). In practice, however, the positive square root is often the only one quoted, and propagated into future calculations. Statistical software such as SPSS only ever displays the positive result as the standard deviation. This will either cause errors, since in some calculations half of the possible ensuing outcomes will be ignored, or the formula should really be adjusted to portray that it is the absolute value of the square root that is being used. A widespread example is an effect size such as Cohen's d , where a difference between means is standardised through division by their standard deviation(s). The result should therefore really be both positive and negative. This would mean that it is not possible to tell even via the convention of putting an intervention group before the control group whether the outcome is in favour of the intervention or not. Of course, it is possible to tell by other means (such as inspecting the data), but the summary effect size itself would be neutral on the matter.

Treating a standard deviation as an absolute value to overcome such confusion makes sense. But the main reason why the standard deviation was devised was to evade the use of absolute

values, which are inconvenient in many ways for algebraic manipulation. In practice, the use of absolute values with standard deviations means that any advantage they may have had over absolute deviations disappears. When teaching new analysts how to work with numbers, convenience for existing ‘experts’ is not a factor. Therefore, a simpler form of analysis based on absolute deviations could be presented instead. This is the idea examined further in this paper.

An absolute deviation correlation coefficient

A simpler alternative for a correlation coefficient would be to replace the standard deviations in the denominator with the mean absolute deviation of each set of numbers. This would be:

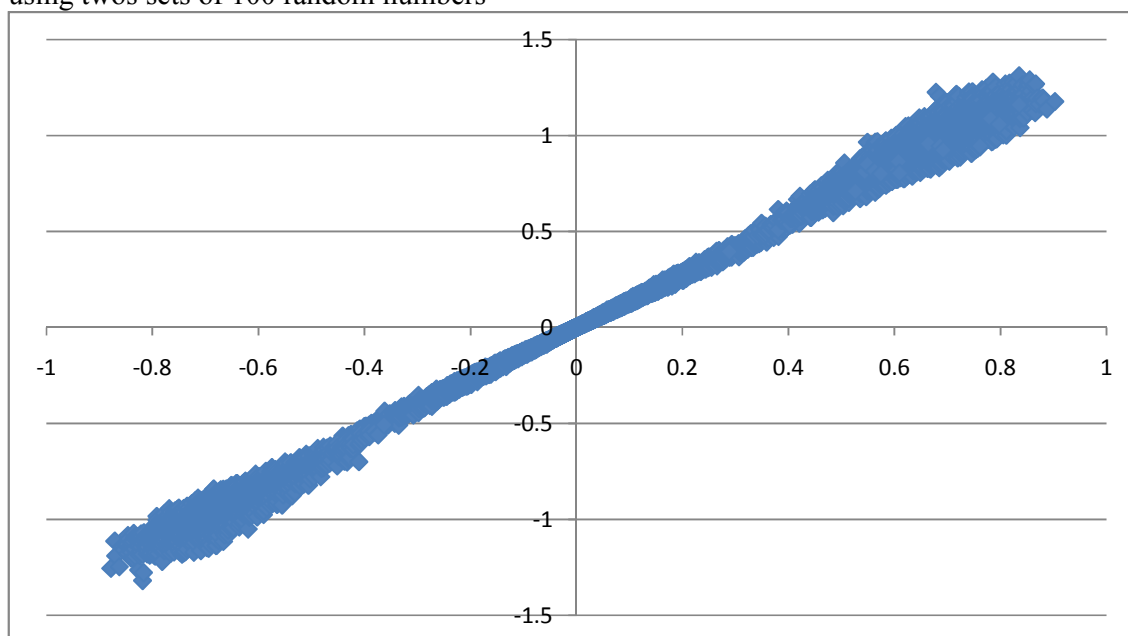
$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/n}{\sum_{i=1}^n |x_i - \bar{x}|/n \cdot \sum_{i=1}^n |y_i - \bar{y}|/n} \quad \text{(new correlation coefficient 1)}$$

Because the number of values in each set of numbers appears in both the numerator and the denominator, it can be cancelled in a way that is not possible for the coefficient based on standard deviations. This would lead to:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n |x_i - \bar{x}| \cdot \sum_{i=1}^n |y_i - \bar{y}|} \quad \text{(new correlation coefficient 1)}$$

This approach does not involve squaring the deviations of each value from the mean, summing the results and then finding the square root. The resulting formula is much simpler than for Pearson’s R – as it also does not require the extra brackets to deal with n. It is the sum of the products of each pair of deviations divided by the product of the sums of each pair of deviations. As a measure of dispersion, the mean absolute deviation is robust and efficient. Unlike the standard deviation, it does not artificially and unnecessarily inflate the larger deviations by squaring them, and then incompletely square-rooting the results. In almost all other practical circumstances it does not matter whether the standard or absolute deviation is used. For example, creating 30,000 pairs of samples, each of 100 random numbers, allows the correlation between each pair to be assessed. Doing this with both Pearson’s R and the absolute deviation correlation (RA1) yields two sets of 30,000 examples (Figure 1). This shows that there is a clear and straight-line equivalence between the two measures of correlation. Using R, their inter-‘correlation’ is reported as +1.0.

Figure 1 – Scatterplot of values for R (x axis) and RA1 (y axis), each based on 30,000 trials using two sets of 100 random numbers



A standard deviation is always larger than an absolute deviation for the same set of figures (because it inflates the figures by squaring). This means, in turn, that the absolute deviation correlation coefficient (created by dividing the covariance by the absolute deviations) is always larger than R for the same two sets of figures. RA1 is usually greater than R by a factor of around 1.3, because the absolute deviation is usually smaller than the standard deviation by a factor of about 0.77. Some commentators have stated that there is a simple conversion from SD to M|D|, through multiplying by $\sqrt{2/\pi}$, or around 0.8. In fact this is not so. Each SD can have more than one M|D|, again because of the vagaries caused by squaring the deviations, summing and then square-rooting. And this is why the graph in Figure 1 appears to diverge very somewhat from a straight line towards the extremes. Nevertheless, either coefficient will work as well as the other in practice, and the 1.3 or 0.77 conversion will usually work well enough to see what the other coefficient would have been. Before comparing the strengths and weaknesses of each approach more fully, it is important to see if the coefficient can be simplified even further.

An additive coefficient

In some respects it seems obvious to create a correlation coefficient that matches Pearson's R but with the standard deviations replaced by absolute deviations. However, it is also true that the absolute deviation is clearly additive where the standard deviation is multiplicative. It might be possible therefore to create a correlation coefficient in which the deviations for each pair of numbers are summed, and are divided by the sum of the sums of the deviation for each number separately. Where both numbers in a pair deviate in the same direction they will add together, and when in the opposite direction they will tend cancel out. The sum of these combined deviations for both pairs together can be assessed as a proportion of the sum of the absolute deviations for each set of numbers alone.

n

$$\frac{\sum_{i=1}^n |(x_i - \bar{x}) + (y_i - \bar{y})|/n}{\sum_{i=1}^n |x_i - \bar{x}|/n + \sum_{i=1}^n |y_i - \bar{y}|/n}$$

(new correlation coefficient 2)

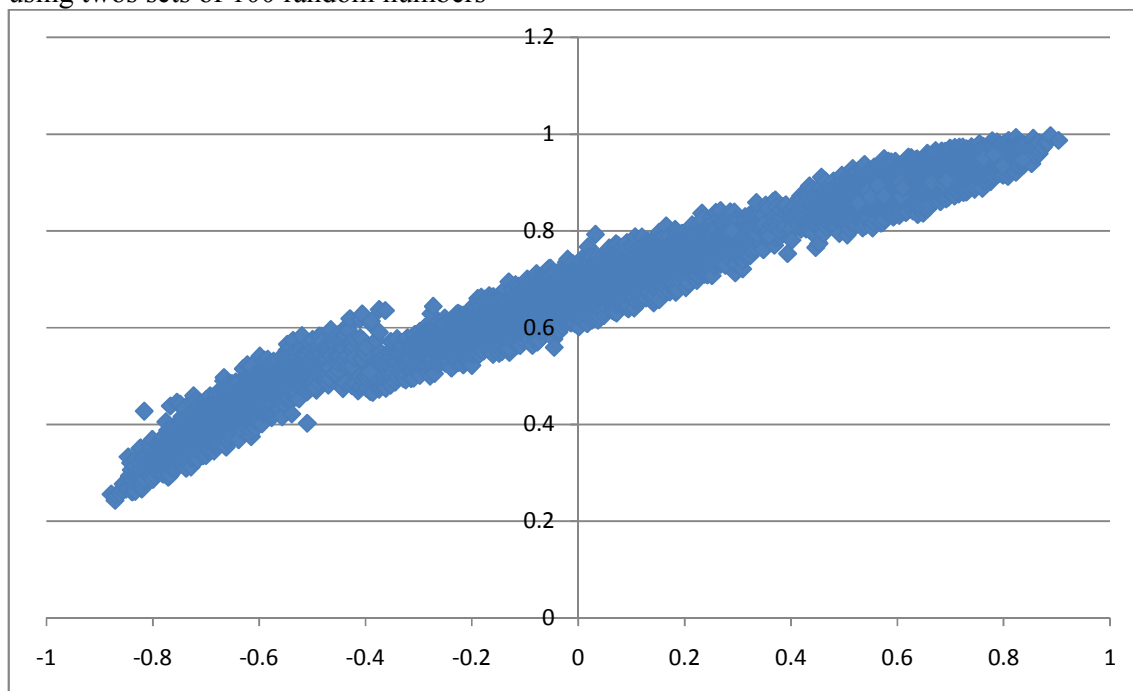
And as with new correlation coefficient 1 the value of n can be cancelled in numerator and denominator. This will yield:

$$\frac{\sum_{i=1}^n |(x_i - \bar{x}) + (y_i - \bar{y})|}{\sum_{i=1}^n |x_i - \bar{x}| + \sum_{i=1}^n |y_i - \bar{y}|}$$

(new correlation coefficient 2)

This is simpler again, and as with RA1 the results correlate very highly with those for Pearson's R. Here the results are illustrated for 30,000 pairs of samples, each of 100 random numbers. Running this simulation with both Pearson's R and the absolute deviation correlation coefficient (RA2) yields two sets of 30,000 examples (Figure 2). This shows that there is a clear and straight-line equivalence between the two measures of correlation. Using R, their 'correlation' is reported as over +0.99. The use of absolute values in both top and bottom of the formula means that RA2 is always positive (or zero). Zero for RA2 is equivalent to -1 for R. Zero for R is equivalent to 0.667 for RA2 (where the regression line crosses the y axis in Figure 2). When R is 1, RA2 is also 1. What this might mean is discussed in the next section.

Figure 2 – Scatterplot of values for R (x axis) and RA2 (y axis), each based on 30,000 trials using two sets of 100 random numbers



Comparing the three approaches

What this brief paper has shown is that it is possible to create alternatives to Pearson's R that are simpler and based on the absolute mean deviation. In practice, and as long as all three coefficients are understood, which is used in practice would make little difference (and many more trials have been conducted than are reported in the Figures here). They all give the same, or very nearly the same, substantive result in the contexts presented here. In many respects also they all have similar characteristics.

Pearson's R has the advantages of being built into software, its current use in other contexts, and its familiarity for existing users, especially concerning its limits, scale and importance in any context. The latter is important given that significance testing and its derivatives are inappropriate in most real-life contexts, and do not provide an accurate probability of the 'significance' of a correlation even when all assumptions are met. What is needed instead is judgement based on the scale of the correlation (the effect size), the number of cases, and the quality of data. Familiarity helps such judgement. It has the advantage over RA2 of having a zero that means there is no linear covariance. It is symmetrical, as long as it is remembered that this is based on using the absolute values of the standard deviations instead of the two standard deviations themselves. Perhaps most importantly, it is clearly and conveniently bounded by -1 and +1.

RA1 is very similar to R as would be expected where the covariance is divided by two closely related estimates of dispersion. Clearly, when x and y are unrelated the covariance or denominator for both coefficients is zero, and thus both coefficients are zero. However, RA1 is simpler to explain to new researchers and slightly easier to calculate. Based on the mean absolute deviation, it will be more robust in the majority of real-life situations where the data contains errors and/or is not perfectly normally-distributed. It is less distorted by larger deviations, and so does not promote the deletion of inconvenient data. It also appears more tolerant of non-linearity. As a scatterplot diverges from a linear crossplot of x and y but the two sets of data retain a clear relationship, such as when one is a power of the other, R decreases rapidly. In these circumstances, RA1 does not decrease. The real-life correlation of x with x+1 is no stronger than the correlation of x with x^2 , for example. R merely makes it appear to be so, because x is not clearly *linearly* related with x^2 . This stable and robust characteristic of RA1 is one reason why the factor of 1.3 (above) is only a guide for when x and y are linearly related. Using RA1 early in an investigation might help pick up indications of a more complex relationship than a simple linear one between x and y.

RA2 is different to both R and RA2 in some respects. Based on addition not multiplication, and on absolute values in both numerator and denominator, it does not have a negative and a positive form. In this respect it is no different to traditional effect sizes calculated properly (see above). It might therefore be necessary to make a cursory examination of the data or look at a crossplot in order to be certain of the direction of any relationship. The major unique advantage of RA2 is its conceptual and arithmetic simplicity. Like RA1 it is more robust, less distorted by large deviations, and more tolerant of non-linearity than R.

Where the results from each approach differ somewhat, as they do at extremes in Figure 1 and throughout Figure 2, it not immediately clear which result is preferable. To the extent that the different coefficients yield the same substantive results (as in measuring a

temperature in either Centigrade or Fahrenheit) then scientifically it does not matter which one is used. Where they differ, given that the distortion caused by squaring is already known, there is an argument that the absolute deviation coefficients might more accurately portray what was being estimated.

Clearly more work would be needed than presented in this introductory paper to make either RA1 or RA2 something that could be used safely in practice, and offered as a genuinely simpler alternative to R. Analysts would also need more practice in assessing the meaning or a new coefficient value. It is not as easy, for example, as using the mean absolute deviation to create a new effect size (Gorard 2013). But the work so far shows that in principle the mean absolute deviation can be used instead of the standard deviation in a variety of settings, probably throughout descriptive statistics (Gorard 2014b). The objective is to make working with numbers as simple as possible for reluctant social scientists. At present, in the UK at least, considerable resources are being spent on trying to widen the use of ‘quantitative’ methods. Almost all of the effort lies in persuasion and pedagogy. These might be necessary, but will not be sufficient in themselves. This paper is part of an attempt to point in a different way – correcting the logical errors and simplifying what it is deemed necessary to know, before then improving the teaching of the corrected and simpler methods.

References

- Cooper L. and Shore F. (2008) Students' Misconceptions in Interpreting Center and Variability of Data Represented via Histograms and Stem-and-leaf Plots, *Journal of Statistics Education*, 16, 2, 13 pages
- Gorard, S. (2003) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum
- Gorard, S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation, *The British Journal of Educational Studies*, 53, 4, 417-430
- Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77
- Gorard, S. (2013) The possible advantages of the mean absolute deviation ‘effect’ size, *Social Research Update*, 65, Winter 2013, pp.1-4, <http://sru.soc.surrey.ac.uk/SRU64.pdf>
- Gorard, S. (2014) Introducing the mean absolute deviation ‘effect’ size, *International Journal Research and Methods in Education*, (forthcoming)
- Gorard, S. (2014) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10
- Hampel, F. (1997) *Is statistics too difficult?*, Research Report 81, Seminar fur Statistik, Eidgenossische Technische Hochschule, Switzerland
- Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences
- Murtonen, M. and Lehtinen, E. (2003) Difficulties experienced by education and sociology students in quantitative methods courses, *Studies in Higher Education*, 28, 2, 171-185
- Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291