British Journal of Educational Studies, ISSN 0007-1005
Vol. 53, No. 4, December 2005, pp 417–430

REVISITING A 90-YEAR-OLD DEBATE: THE ADVANTAGES OF THE MEAN DEVIATION

by Stephen Gorard, *University of York*

ABSTRACT: This paper discusses the reliance of numerical analysis on the concept of the standard deviation, and its close relative the variance. It suggests that the original reasons why the standard deviation concept has permeated traditional statistics are no longer clearly valid, if they ever were. The absolute mean deviation, it is argued here, has many advantages over the standard deviation. It is more efficient as an estimate of a population parameter in the real-life situation where the data contain tiny errors, or do not form a completely perfect normal distribution. It is easier to use, and more tolerant of extreme values, in the majority of real-life situations where population parameters are not required. It is easier for new researchers to learn about and understand, and also closely linked to a number of arithmetic techniques already used in the sociology of education and elsewhere. We could continue to use the standard deviation instead, as we do presently, because so much of the rest of traditional statistics is based upon it (effect sizes, and the F-test, for example). However, we should weigh the convenience of this solution for some against the possibility of creating a much simpler and more widespread form of numeric analysis for many.

Keywords: variance, measuring variation, political arithmetic, mean deviation, standard deviation, social construction of statistics

1. Introduction

This paper discusses the reliance of numerical analysis on the concept of the standard deviation, and its close relative the variance. Such a consideration suggests several potentially important points. First, it acts as a reminder that even such a basic concept as 'standard deviation', with an apparently impeccable mathematical pedigree, is socially constructed and a product of history (Porter, 1986). Second, therefore, there can be equally plausible alternatives of which this paper outlines one – the mean absolute deviation. Third, we may be able to create from this a simpler introductory kind of statistics that

is perfectly useable for many research purposes, and that will be far less intimidating for new researchers to learn (Gorard, 2003a). We could reassure these new researchers that, although traditional statistical theory is often useful, the mere act of using numbers in research analyses does not mean that they have to accept or even know about that particular theory.

2. What is a Standard Deviation?

The 'standard deviation' is a measure of 'dispersion' or 'spread'. It is used as a common summary of the range of scores associated with a measure of central tendency – the mean-average. It is obtained by summing the squared values of the deviation of each observation from the mean, dividing by the total number of observations,[1] and then taking the positive square root of the result.[2] For example, take the separate measurements:

13, 6, 12, 10, 11, 9, 10, 8, 12, 9.

Their sum is 100, and their mean is therefore 10. Their deviations from the mean are:

3, –4, 2, 0, 1, –1, 0, -2, 2, –1.

To obtain the standard deviation we first square these deviations to eliminate the negative values, leading to:

9, 16, 4, 0, 1, 1, 0, 4, 4, 1.

The sum of these squared deviations is 40, and the average of these (dividing by the number of measurements) is 4. This is defined as the 'variance' of the original numbers, and the 'standard deviation' is its positive square root, or 2. Taking the square root returns us to a value of the same order of magnitude as our original readings. So a traditional analysis would show that these ten numbers have a mean of 10 and a standard deviation of 2. The latter gives us an indication of how dispersed the original figures are, and so how representative the mean is. The main reason that the standard deviation (SD) was created like this was because the squaring eliminates all negative deviations, in a way that makes the result easier to work with algebraically than simply ignoring the sign of negative deviations (see below).

3. What is a Mean Deviation?

There are several alternatives to the standard deviation (SD) as a summary of dispersion. These include the range, the quartiles, and

the inter-quartile range. The most direct alternative for SD as a measure of dispersion, however, is the absolute mean deviation (MD). This is simply the average of the absolute differences between each score and the overall mean. Take the separate measurements:

13, 6, 12, 10, 11, 9, 10, 8, 12, 9.

Their sum is 100, and their mean is therefore 10. Their deviations from the mean are:

3, –4, 2, 0, 1, –1, 0, –2, 2, –1.

To obtain the mean deviation we first ignore the minus signs in these deviations to eliminate the negative values, leading to:

3, 4, 2, 0, 1, 1, 0, 2, 2, 1.

These figures now represent the distance between each observation and the mean, regardless of the direction of the difference. Their sum is 16, and the average of these (dividing by the number of measurements) is 1.6. This is the mean deviation, and it is easier for new researchers to understand than SD, being simply the average of the deviations – the amount by which, on average, any figure differs from the overall mean.[3] It has a clear meaning, which the standard deviation of 2 does not.[4] Why, then, is the standard deviation in common use and the mean deviation largely ignored?

4. Why do we use the Standard Deviation?

As early as 1914, Eddington pointed out that 'in calculating the mean error of a series of observations it is preferable to use the simple mean residual irrespective of sign [i.e. MD] rather than the mean square residual [i.e. SD]' (Eddington, 1914, p. 147). He had found, in practice, that the 'mean deviation' worked better with empirical data than SD, even though 'this is contrary to the advice of most text-books; but it can be shown to be true' (p. 147). He also subsequently claimed that the majority of astronomers had found the same.

Fisher (1920) countered Eddington's empirical evidence with a mathematical argument that SD was more efficient than MD *under ideal circumstances*, and many commentators now accept that Fisher provided a complete defence of the use of SD (e.g. Aldrich, 1997; MacKenzie, 1981). Fisher had proposed that the quality of any statistic could be judged in terms of three characteristics. The statistic, and the population parameter that it represents, should be 'consistent' (i.e. calculated in the same way for both sample and population).

The statistic should be 'sufficient' in the sense of summarising all of the relevant information to be gleaned from the sample about the population parameter. In addition, the statistic should be 'efficient' in the sense of having the smallest probable error as an estimate of the population parameter. Both SD and MD meet the first two criteria (to the same extent). According to Fisher, it was in meeting the last criterion that SD proves superior. When drawing repeated large samples from a normally distributed population, the standard deviation[5] of their individual mean deviations is 14 per cent higher than the standard deviation of their individual standard deviations (Stigler, 1973). Thus, the SD of such a sample is a more consistent estimate of the SD for a population, and is considered better than its plausible alternatives as a way of estimating the standard deviation in a population using measurements from a sample (Hinton, 1995, p. 50). That is the main reason why SD has subsequently been preferred, and why much of subsequent statistical theory is based on it.

One further concern has been that the absolute value symbols necessary to create a formula for the absolute mean deviation are quite difficult to manipulate algebraically (http://infinity.sequoias.cc.ca.us/faculty/woodbury/Stats/Tutorial/Disp_Var_Pop.htm). This makes the development of sophisticated forms of analysis more complicated than when using the standard deviation (http://mathworld.wolfram.com/MeanDeviation. html). So we now have a complex form of statistics based on SD (and its square – the variance) because SD is more efficient than MD under ideal circumstances, and because it is easier to manipulate algebraically. Of course, SD has now become a tradition, and much of the rest of the theory of statistical analysis rests on it (the definition of distributions, the calculation of effect sizes, analyses of variance, least squares regression, and so on). For example, SD is both based on and part of the definition of the widespread Gaussian or 'normal' distribution. This has the benefit that it enables commentators to state quite precisely the proportion of the distribution lying within each standard deviation from the mean. Therefore, much of the expertise of statisticians rests on the basis of using the standard deviation, and this expertise is what they pass on to novices.

5. Why might we use the Mean Deviation?

On the other hand, it is possible to argue that the mean deviation is preferable and that, since Fisher, we have taken a wrong turn in our analytic history. The mean deviation is actually more efficient than the standard deviation in the realistic situation where some of the

measurements are in error, more efficient for distributions other than perfect normal, closely related to a number of other useful analytical techniques, and easier to understand. I discuss each of these in turn.

Error Propagation

The standard deviation, by squaring the values concerned, gives us a distorted view of the amount of dispersion in our figures. The act of squaring makes each unit of distance from the mean exponentially (rather than additively) greater, and the act of square-rooting the sum of squares does not completely eliminate this bias. That is why, in the example above, the standard deviation (2) is greater than the mean deviation (1.6), as SD emphasises the larger deviations. Figure 1 shows a scatterplot of the matching mean and standard deviations for 255 sets of random numbers. Two things are noteworthy. SD is always greater than MD, but there is more than one possible SD for any MD value, and vice versa. Therefore, the two statistics are not measuring precisely the same thing. Their Pearson correlation over any large number of trials (such as the 255 pictured here) is just under 0.95, traditionally meaning that around 90 per cent of their variation is common. If this is sufficient to claim that they are measuring the same thing, then either could be used, but in the absence of other evidence the mean deviation should be preferred because it is simpler, and easier for newcomers to understand (see above). If, on the other hand, they are not measuring the same thing then the most important question is not which is the more reliable but which is measuring what we actually want to measure?

The apparent superiority of SD is not as clearly settled as is usually portrayed in texts (see above). For example, the subsequent work of Tukey (1960) and others suggests that Eddington had been right, and Fisher unrealistic in at least one respect. Fisher's calculations of the relative efficiency of SD and MD depend on there being no errors *at all* in the observations. But for normal distributions with small contaminations in the data, 'the relative advantage of the sample standard deviation over the mean deviation which holds in the uncontaminated situation is dramatically reversed' (Barnett and Lewis, 1978, p. 159). An error element as small as 0.2 per cent (i.e. two error points in 1,000 observations) completely reverses the advantage of SD over MD (Huber, 1981). So MD is actually more efficient in all life-like situations where small errors *will* occur in observation and measurement (being over twice as efficient as SD when the error element is 5 per cent, for example). 'In practice we should certainly prefer d_n [i.e. MD] to s_n [i.e. SD]' (Huber, 1981, p. 3).

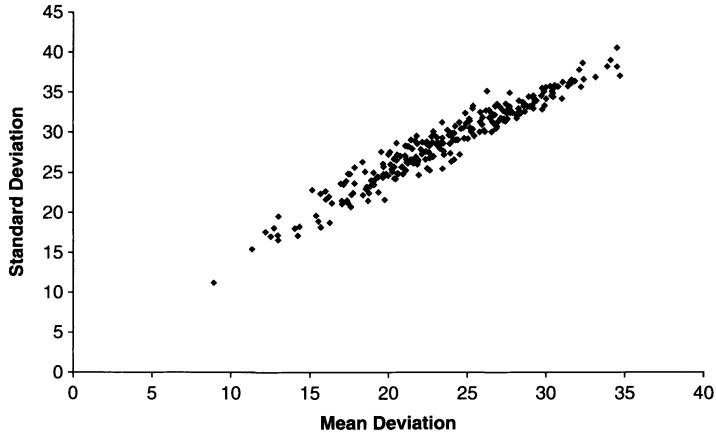

Figure 1. Comparison of mean and standard deviation for sets of random numbers *Note*: this example was generated over 255 trials using sets of 10 random numbers between 0 and 100. The scatter effect and the overall curvilinear relationship, common to all such examples, are due to the sums of squares involved in computing SD.

The assumptions underlying statistical inference are only mathematical conveniences, and are usually defended in practice by a further assumption, presented without an explicit argument, that minor initial errors should lead to only minor errors in conclusions. This is clearly not the case (see Gorard, 2003b). 'Some of the most common statistical procedures (in particular those optimized for an underlying normal distribution) are excessively sensitive to seemingly minor deviations from the assumptions' (Huber, 1981, p. 1). The difference between Fisher and Eddington is related to the difference between mathematics and science. The first is concerned with the Platonic world of perfect distributions and ideal measurements. Perhaps agriculture, where Fisher worked and where vegetative reproduction of cases is possible, is one of the fields that most closely approximates this world. The second is concerned with the Aristotelian world of empirical research. Astronomy, where Eddington worked and where the potential errors in calculated distances are substantial, highlights the importance of tracking the propagation of measurement errors. The imperfect measurements that we use in social research are more like those from the largely non-experimental astronomy than those from agriculture.

Another important, but too often overlooked, assumption underlying the purported superiority of SD is that it involves working with samples selected randomly from a fixed population (because this is

how its efficiency is calculated). However, there is a range of analytical situations where this is not so, such as when working with population figures, or with a non-probability sample, or even a probability sample with considerable non-response. In all of these situations it is perfectly proper to calculate the variation in the figures involved, but without attempting to estimate a population SD. Therefore, in what are perhaps the majority of situations faced by practising social scientists, the supposed advantage of SD in terms of efficiency is irrelevant.

Some statisticians may attempt to evade this conclusion by arguing that the data they work, even where these are for populations, are actually taken from an infinitely large super-population. For example, Camilli (1996) reports a debate between several prominent commentators in which it is argued that statisticians are not really interested in generalising from a sample to a specified population but to an idealised super-population spanning space and time. It is claimed that 'social statisticians are pretty much forced to adopt the notion of a "superpopulation" when attempting to generalise the results of an analysis' (p. 7). But it is clear that if such a super-population is involved then its variance is infinite (Fama, 1963), in which case the purported greater efficiency of SD, based on estimating the population variance, is impossible to establish. An analyst cannot, therefore, use a super-population *and* argue the efficiency of the standard deviation at the same time.

Distribution-free

In addition to an unrealistic assumption about error-free measurements, Fisher's logic also depends upon an ideal normal distribution for the data. What happens if the data are not perfectly normally distributed, or not normally distributed at all?

Fisher himself pointed out that MD is better for use with distributions other than the normal/Gaussian distribution (Stigler, 1973). This can be illustrated for uniform distributions through the use of repeated simulations. However, we first have to consider what appears to be a tautology in claims that the standard deviation of a sample is a more stable estimate of the standard deviation of the population than the mean deviation is (e.g. Hinton, 1995). We should not be comparing SD for a sample versus SD for a population with MD for a sample versus SD for a population. MD for a sample should be compared to the MD for a population, and Figure 1 shows why this is necessary – each value for MD can be associated with more than one SD and vice versa, giving what is merely an illusion

of unreliability for MD when compared with SD. The exact inverse conclusion to that drawn by Fisher could be argued by comparing the SD for a sample and the MD of a population, with the MD for a sample and the MD of a population. In this case, MD would be the better estimate of the population parameter, but the comparison would be as unfair as the one Fisher made.

Repeated simulations show that the efficiency of MD is at least as good as SD for non-normal distributions. For example, I created 1,000 samples (with replacement) of 10 random numbers each between 0 and 19, from the population of 20 integers between 0 and 19. The mean of the population is 9.5, the mean deviation is 5, and the standard deviation is 5.77. The 1,000 sample SDs varied from 2.72 to 7.07, and the sample MDs varied from 2.30 to 6.48. The standard deviation of the 1,000 estimated standard deviations around their true mean of 5.77 was just over 1.025.[6] The standard deviation of the 1,000 estimated mean deviations around their true mean of 5 was just under 1.020. These values and their direction of difference are relatively stable over repeated simulations with further sets of 1,000 samples. Readers may like to try this for themselves. This is an illustration that, for uniform distributions of the kind involving random numbers, the efficiency of the mean deviation is at least as good as that of the standard deviation.

The normal distribution, like the notion of measurement without error, is a mathematical artifice. In practice, scientists will be dealing with observations that merely resemble or approximate such an ideal. But strict normality was a basic assumption of Fisher's proof of the efficiency of SD. What Eddington had realised was that small deviations from normality, such as always occur in practice, have a considerable impact on *ideal* statistical procedures (Hampel, 1997). In general, our observed distributions tend to be longer-tailed, having more extreme scores, than would be expected under ideal assumptions. Because we then square the deviations from average to produce SD, but not MD, such longer-tailed distributions tend to 'explode' the variation in SD (Huber, 1981). The act of squaring makes each unit of distance from the mean exponentially (rather than additively) greater, and the act of square-rooting the sum of squares does not completely eliminate this bias. In practice, of course, this fact is often obscured by the widespread deletion of 'outliers' (Barnett and Lewis, 1978). In fact, our use of SD rather than MD forms part of the pressure on analysts to ignore any extreme values.

The distortion caused by squaring deviations is part of a culture in which advice is routinely given to students to remove or ignore valid

measurements with large deviations because these unduly influence the final results. This is done regardless of their importance as data, and it means that we no longer allow our prior assumptions about distributions to be disturbed merely by the fact that they are not matched by the evidence. Good science should treasure results that show an interesting gulf between theoretical analysis and actual observations, but we have a long and ignoble history of too often ignoring results that threaten our fundamental tenets (Moss, 2001). Extreme scores are important occurrences in a variety of natural and social phenomena, including city growth, income distribution, earthquakes, traffic jams, solar flares, and avalanches. We cannot simply dismiss them as exogenous to our models. If we take them seriously, as a few commentators have, then we find that many approximately normal distributions show consistent departures from normality. In particular, they have more of the extreme scores than expected. Statistical techniques based on the standard deviation give misleading answers in these cases, and so 'concepts of variability, such as ... the absolute mean deviation ... are more appropriate measures of variability for these distributions' (Fama, 1963, p. 491).

Related Techniques

Another advantage in using MD lies in its links and similarities to a range of other simple analytical techniques, a few of which are described here. In 1997, Gorard proposed the use of the 'segregation index' (S) for summarising the unevenness in the distribution of individuals between organisational units, such as the clustering of children from families in poverty in specific schools (Gorard and Fitz, 1997).[7] The index is related to several of the more established indices such as the dissimilarity index and the isolation index. However, S has two major advantages over both of these. It is strongly composition-invariant, meaning that it is affected only by unevenness of distribution and not at all by scaled changes in the magnitude of the figures involved (Gorard and Taylor, 2002). Perhaps even more importantly, S has an easy to comprehend meaning. It represents the proportion of the disadvantaged group (children in poverty) who would have to exchange organisational units (schools) for there to be a completely even distribution of disadvantaged individuals. Other indices, especially those like the Gini coefficient that involve the squaring of deviations, lead to no such easily interpreted value. The similarity between S and MD is striking. MD is what you would devise in adapting S to work with real numbers rather than the frequencies of categories. MD is also, like S, more tolerant of problems

within the data, and has an easier to understand meaning than its potential rivals.

In social research there is a need to ensure, when we examine differences over time, place or other category, that the figures we use are proportionate (Gorard, 1999). Otherwise misleading conclusions can be drawn. One easy way of doing this is to look at differences between figures in proportion to the figures themselves. For example, when comparing the number of boys and girls who obtain a particular examination grade, we can subtract the score for boys from that of girls and then divide by the overall score (Gorard *et al.*, 2001). If we call the score for boys b and for girls g, then the 'achievement gap' can be defined as $(g-b)/(g+b)$. This is very closely related to a range of other scores and indices, including the segregation index (see above and Taylor *et al.*, 2000). However, such approaches give results that are difficult to interpret unless they are used with ratio values having an absolute zero, such as examination scores. When used with a value, such as the Financial Times (FT) index of share prices, which does not have a clear zero, it is better to consider differences in terms of their usual range of variation. If we divide the difference between any two figures by the past variation in the figures then we automatically deal with the issue of proportionate scale as well.

This approach, of creating 'effect' sizes, is growing in popularity as a way of assessing the substantive importance of differences between scores, as opposed to assessing the less useful 'significance' of differences (Gorard, 2006). The standard method is to divide the difference between two means by their standard deviation(s). Or, put another way, before subtraction the two scores are each standardised through division by their standard deviation(s).[8] We could, instead, use the mean deviation(s) as the denominator. Imagine, for example, that one of the means to be compared is based on two observations (x,y). Their sum is $(x+y)$, their mean is $(x+y)/2$, and their mean deviation is $(|\, x-(x+y)/2\,|+|\, y-(x+y)/2\,|)/2$. The standardised mean, or mean divided by the mean deviation, would be:

$$\frac{(x+y)/2}{(|\, x-(x+y)/2\,|+|\, y-(x+y)/2\,|)/2}$$

Since both the numerator and denominator are divided by two these can be cancelled, leading to :

$$\frac{(x+y)}{(|\, x-(x+y)/2\,|+|\, y-(x+y)/2\,|)}$$

If both x and y are the same then there is no variation, and the mean deviation will be zero. If x is greater than the mean then $(x+y)/2$ is subtracted from x but y is subtracted from $(x+y)/2$ in the denominator. This leads to the result:

$$\frac{(x+y)}{(x-y)}$$

If x is smaller than the mean then x is subtracted from $(x+y)/2$ but $(x+y)/2$ is subtracted from y in the denominator. This leads to the result:

$$\frac{(x+y)}{(y-x)}$$

For example, if the two values involved are actually 13 and 27, then their standardised score is $20/7$.[9] Therefore, for the two value example, an effect size based on mean deviations is the difference between the reciprocals of two 'achievement gaps' (see above).

Similarities such as these mean that there is a possibility of unifying traditional statistical approaches, based on probability theory, with simpler arithmetic approaches, of the kind described in this section. This, in turn, would allow new researchers to learn simpler techniques for routine use with numbers that may also permit simple combination with other forms of data (Gorard with Taylor, 2004).

Simplicity

In an earlier era of computation it seemed easier to find the square root of one figure rather than take the absolute values for a series of figures. This is no longer so, because the calculations are done by computer. The standard deviation now has several potential disadvantages compared to its plausible alternatives, and the key problem it has for new researchers is that it has no obvious intuitive meaning. The act of squaring before summing and then taking the square root after dividing means that the resulting figure appears strange. Indeed, it *is* strange, and its importance for subsequent numerical analysis usually has to be taken on trust. Students are simply taught that they should use it, but in social science most opt not to use numeric analysis at all anyway (Murtonen and Lehtinen, 2003). Given that both SD and MD do the same job, MD's relative simplicity of meaning is perhaps the most important reason for henceforth using and teaching the mean deviation before, or even instead of, the more complex and less meaningful standard deviation. Most researchers wishing to provide a summary statistic of the dispersion

in their findings generally do not want to manipulate *anything*, whether algebraically or otherwise. For these, and for most consumers of research evidence, the algebraic advantages of SD are irrelevant, and using the mean deviation is more 'democratic'.

6. Conclusion

Even one of the most elementary things taught on a statistics course, the standard deviation, is more complex than it need be, and is considered here as an example of how convenience for mathematical manipulation often over-rides pragmatism in research methods. In those rare situations in which we obtain full response from a random sample with no measurement error, and we wish to estimate the dispersion in a perfect Gaussian population from the dispersion in our sample, then the standard deviation has been shown to be a more stable indicator of its equivalent in the population than the mean deviation has. Note that we can only calculate this via simulation, since in real-life research we would not know the actual population figure, else we would not be trying to estimate it via a sample. In essence, the claim made for the standard deviation is that we can compute a number (SD) from our observations that has a relatively consistent relationship with a number computed in the same way from the population figures. This claim, in itself, is of no great value. Reliability alone does not make that number of any valid use. For example, if the computation led to a *constant* whatever figures were used then there would be a perfectly consistent relationship between the parameters for the sample and population. But to what end? Surely the key issue is not how stable the statistic is but whether it encapsulates what we want it to. Similarly, we should not use an inappropriate statistic simply because it makes complex algebra easier.

Of course, much of the rest of traditional statistics is now based on the standard deviation, but it is important to realise that it need not be. In fact, we seem to have placed our 'reliance in practice on isolated pieces of mathematical theory proved under unwarranted assumptions, [rather] than on empirical facts and observations' (Hampel, 1997, p. 9). One result has been the creation since 1920 of methods for descriptive statistics that are more complex and less democratic than they need be. The lack of quantitative work and skill in social science is usually portrayed via a deficit model, and more researchers are exhorted to enhance their capacity to conduct such work. One of the key barriers, however, could be deficits caused by the unnecessary complexity of the methods themselves rather than their potential users. The standard deviation is one such example.

7. Notes

[1] Sometimes the sum of the squared deviations is divided by one less than the number of observations. This arbitrary adjustment is intended to make allowance for any sampling error.

[2] $SD = \sqrt{(\Sigma(|M - Oi|)^2/N)}$, where M is the mean, Oi is observation i, and N is the number of observations.

[3] $MD = \Sigma(|M - Oi|)/N$, where M is the mean, Oi is observation i, and N is the number of observations.

[4] Traditionally, the mean deviation has sometimes been adjusted by the use of Bessel's formula, such that MD is multiplied by the square root of Pi/2 (MacKenzie, 1981). However, this loses the advantage of its easy interpretation.

[5] SD is used here and subsequently for comparison, because that is what Fisher used.

[6] Of course, it is essential at this point to calculate the variation in relation to the value we are trying to estimate (the population figure) and not merely calculate the internal variability of the estimates.

[7] The segregation index is calculated as $0.5 * \Sigma (|Ai/A - Ti/T|)$, where Ai is the number of disadvantaged individuals in unit i, Ti is the number of all individuals in unit i, A is the number of all disadvantaged individuals in all units, and T is the number of all individuals in all units.

[8] There is a dispute over precisely which version of the standard deviation is used here. In theory it is SD for the population, but we will not usually have population figures, and if we did then we would not need this analysis. In practice, it is SD for either or both of the groups in the comparison.

[9] Of course, it does not matter which value is x and which is y. If we do the calculation with 27 and 13 instead then the result is the same.

8. References

ALDRICH, J. (1997) R. A. Fisher and the making of maximum likelihood 1912–1922, *Statistical Science*, 12 (3), 162–176.

BARNETT, V. and LEWIS, T. (1978) *Outliers in Statistical Data* (Chichester, John Wiley and Sons).

CAMILLI, G. (1996) Standard errors in educational assessment: a policy analysis perspective, *Education Policy Analysis Archives*, 4 (4).

EDDINGTON, A. (1914) *Stellar movements and the structure of the universe* (London, Macmillan).

FAMA, E. (1963) Mandelbrot and the stable Paretian hypothesis, *Journal of Business*, 420–429.

FISHER, R. (1920) A mathematical examination of the methods of determining the accuracy of observation by the mean error and the mean square error, *Monthly Notes of the Royal Astronomical Society*, 80, 758–770.

GORARD, S. (1999) Keeping a sense of proportion: the 'politician's error' in analysing school outcomes, *British Journal of Educational Studies*, 47 (3), 235–246.

GORARD, S. (2003a) Understanding probabilities and re-considering traditional research methods training, *Sociological Research Online*, 8 (1), 12 pages.

GORARD, S. (2003b) *Quantitative Methods in Social Science: the Role of Numbers Made Easy* (London, Continuum).

GORARD, S. (2006) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27 (1).

GORARD, S. and FITZ, J. (1997) The missing impact of marketisation, *School Leadership and Management*, 17 (3), 437–438.
GORARD, S. and TAYLOR, C. (2002) What is segregation? A comparison of measures in terms of strong and weak compositional invariance, *Sociology*, 36 (4), 875–895.
GORARD, S., with TAYLOR, C. (2004) *Combining Methods in Educational and Social Research* (London, Open University Press).
GORARD, S., REES, G. and SALISBURY, J. (2001) The differential attainment of boys and girls at school: investigating the patterns and their determinants, *British Educational Research Journal*, 27 (2), 125–139.
HAMPEL, F. (1997) *Is Statistics too Difficult?* Research Report 81, Seminar fur Statistik, Eidgenossiche Technische Hochschule, Switzerland.
HINTON, P. (1995) *Statistics Explained* (London, Routledge).
HUBER, P. (1981) *Robust Statistics* (New York, John Wiley and Sons).
MacKENZIE, D. (1981) *Statistics in Britain 1865–1930* (Edinburgh, Edinburgh University Press).
MOSS, S. (2001) *Competition in Intermediated Markets: Statistical Signatures and Critical Densities* Report 01-79, Centre for Policy Modelling, Manchester Metropolitan University.
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.
PORTER, T. (1986) *The Rise of Statistical Thinking* (Princeton, Princeton University Press).
STIGLER, S. (1973) Studies in the history of probability and statistics XXXII: Laplace, Fisher and the discovery of the concept of sufficiency, *Biometrika*, 60 (3), 439–445.
TAYLOR, C., GORARD, S. and FITZ, J. (2000) A re-examination of segregation indices in terms of compositional invariance, *Social Research Update* , 30, 1–4.
TUKEY, J. (1960) A survey of sampling from contaminated distributions. In I. OLKIN, S. GHURYE, W. HOEFFDING, W. MADOW and H. MANN (Eds) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (Stanford, Stanford University Press).

Correspondence
Professor Stephen Gorard
Department of Educational Studies
University of York
YO10 5DD
E-mail: sg25@york.ac.uk