

In Praise of Medians: assessment, averages, aggregation, and adjustment

Stephen Bostock and Mike Brough, Keele University

Educational Developments 2.3, 22-24

Cuthbert (2001) considered the problem of detecting variation between markers and adjusting for it. We continue this assessment theme by considering how best to summarize student marks, and how to adjust and combine marks for different assessments. We recommend the use of medians and quartiles, and we are skeptical of aggregating marks by simple averaging.

The External Examiner in an Examiners' Meeting:

"Let me turn to the papers. I was rather alarmed by some of the marking disparities. On the semiology paper, the internal marker, Mr Odgers, covered a wide range of marks in his assessment, whereas the other internal marker, Dr Piercemüller, gave every one of the 42 candidates a mark of 57. I'd welcome your comments on that."

(Laurie Taylor, THES, July 6 2001)

This may be a familiar, if exaggerated, scene. Many academics recently sat in examiners' meetings considering lists of student names with columns of marks alongside them. We scanned averages (arithmetic means), and possibly standard deviations, for aberrant modules with unusually low or high means, or large standard deviations. Often we considered students' overall module marks generated from combining different types of assessments, and overall programme marks from combining module marks. At worst, final assessment processes degenerate into a statistical minefield where we struggle to make the numbers come out right according to our own view of student performance.

There are three issues worth considering: statistical measures of typical marks and their spread, the problem of aggregating marks, and methods of adjusting marks.

Averaging

Three common measures of the average (i.e. typical) level of marks are means, medians, and modes.

1. The arithmetic mean (or just 'the mean') of a list of marks is the sum of the marks divided by the number of marks.
2. The median is the middle mark when the list is ordered from lowest to highest (minimum to maximum). If there is an even number of marks the median is the number halfway between the two middle marks.
3. The mode is the most frequent value or, if the marks are grouped into, say, 0-10%, 11-20% ... classes, the modal class is the one with most members.

The counts of marks in all the classes form a frequency distribution. Usually shown as a histogram, this is informative about student performance (examples are below). But the mode is only reliable for large sets of marks and we do not consider it further.

Should we use means or medians? The median and mean will differ when the distribution of marks is skewed, or asymmetrical. In a symmetrical (e.g. normal) distribution the mean and median are equal, but real marks are often not symmetrical. Figure 1 shows a distribution of some real marks with a right skew, where the mean exceeds the median. Figure 2 shows an example of a left-skewed distribution of marks, where the mean is less than the median. The mean, which

includes a small part of each mark, is raised or lowered by the skew while the median is unaffected. We could check the distribution of marks for skew (although larger sample sizes would be needed to do this properly) and then only use means with symmetrically distributed marks, but why bother? The median makes fewer assumptions about the nature and distribution of marks.

Figure 1

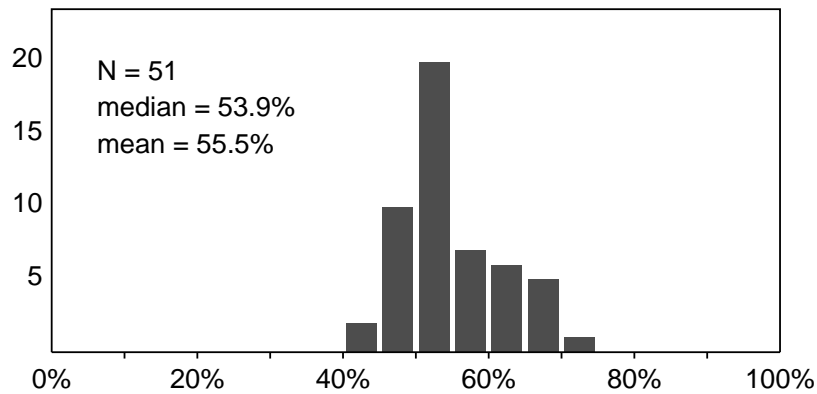
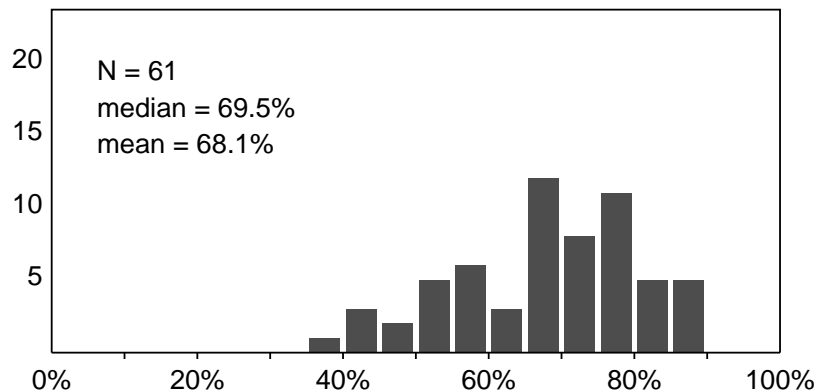


Figure 2



We conventionally allocate marks on a scale of 0 to 100%. This is an ordinal scale, not an interval scale. An interval scale, in addition, requires that every interval (1%) on the scale has the same significance. The interval between, say, 10% and 15% cannot be equated to the interval between 65% and 70%. For example, if a student has a mark on two examination questions of 40% and 50% we are happy to give them an overall mark of 45%. The difference between 40% and 45% is not much different from the difference between 45% and 50%. However, if a second student has marks of 10% and 80% we might again award an overall mark of 45% but this makes less sense. The second student clearly had a problem with one question but showed that she could sometimes have outstanding performance, while the first showed a consistent, but only basic, understanding.

Addition and taking means assume interval data and are therefore inappropriate for ordinal data, and if used will mislead. Medians work equally well for ordinal and interval data (and ratio data: see Siegel and Castellan, 1988, for a discussion of data types). So there are two reasons for using medians: data may be skewed and in principle ordinal marks should never be summarized by a mean.

Similar arguments can be made in favour of the use of quartiles as a measure of the spread of marks, rather than standard deviations. The first quartile is the value with the lowest quarter of marks below it when the marks have been ordered; the third quartile is the value with the highest quarter of marks above it. Therefore, the

distance between the first and the third quartile (the inter-quartile range) contains the middle half of the marks. This is an easy way of understanding the spread of marks around the median. To use standard deviation we need to know that one standard deviation on either side of the mean includes 68% of the marks, and in any case this is only true for a normal distribution, which we may not have.

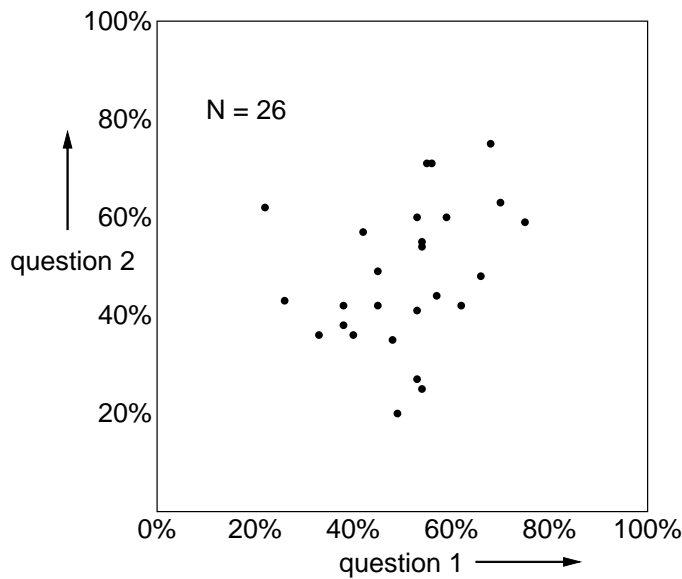
Aggregating

Examiners' meetings typically have numerous component marks for modules and programmes and need to provide a summary statistic. This necessarily throws away some information. Aggregation of marks always has this problem, whether it is combining marks for answers in an examination, combining marks for an examination and coursework, or combining marks for modules. How do we do that so that what is left is representative of the full information? There are no absolute rules - what is appropriate depends on the programme - but summing and averaging would only make sense if marks were on an interval scale. There are three simple alternatives, and many others possible.

1. It might be that the component skills in different modules should all require a *minimum competence*, as seems appropriate in a driving test or medical training. In this case we might apply a threshold (or gateway) algorithm: if the student has not scored 40% in all elements they cannot pass the course. Such a requirement is common in modular programmes and is widely used where health or safety depend on competence.
2. We might use the median of module (median) marks as an indicator of *typical performance*.
3. We might wish to assess *best performance* and use the maximum mark. For example, one of us once ran the mile in under 4.5 minutes (though not recently!) although his average over all races was more than 5.5 minutes. If he had ever run it in under 3.5 minutes he would be widely recognized as the best mile runner ever. It is the best performance that is of interest. Outside sports, creative arts might be more interested in the best performance than in the minimum or typical.

There is no global solution - certainly not the mean of module means. Different assessments measure different things. For example, Figure 3 shows the correlation between the marks students obtained for two essay questions in a closed examination, marked by the tutor. The correlation coefficient was only 0.32 (the coefficient of determination was only 10%) partly due to random variation in student performance and in assessment, and partly because the questions measure different skills. Simply taking the mean of the two marks not only loses information, it generally gives a poor summary.

Figure 3



The advantage of transcripts and student profiles is that these pitfalls are avoided, and a richer picture of performance is provided. If a summary mark is necessary, it should reflect the learning outcomes, which is an educational decision, not a statistical one. The use or combination of minimum, median and maximum marks (e.g. minimum competence, typical performance, best performance, as above) should be driven by educational criteria. For example, an algorithm to award a distinction is "if the median of module marks is above 60% and the dissertation is above 70%".

Adjusting

Why should we adjust a set of marks? There may be problems with the teaching or assessment of a module and we need to compensate for this before making a summary mark. In our putative examiners' meeting, we may note that one module has many high marks while another has many low ones. Their medians differ markedly from those of most modules. On enquiring into the assessment of the first module we find that it was a new module and the assessment methods were over-generous. On reviewing the second module, we find that there were problems with the teaching. A good summarizing algorithm will be less affected by an occasional unreliable mark but in both cases we may want to adjust the module marks before approving them, to compensate as best we can for errors in teaching or assessment.

Additional problems occur when students are allowed a choice of modules, or of assessments within modules, for example, a choice of questions in an examination. If a unit has suffered from errors in teaching or assessment we may want to adjust the marks so that the student transcript is more accurate even if no summary is to be made. Whether we should make an adjustment will depend upon the uses to be made of the marks: even if the driving instructor was at fault and taught the learner driver badly, or the examiner died of heart attack during the driving test, we would still not award a driving license if the driver were not competent (minimum competence). Ultimately, while we would not want to penalize students for our errors as teachers or assessors, a transcript must accurately reflect learning outcomes.

Nonetheless, an adjustment may be necessary so how should it be done? Adjusting marks should correct as far as possible for errors without simply giving all students the benefit of all doubt.

1. One approach maps module grades onto marks. So, while the norm may be that 'A' grades are obtained for 70% and above, in a particular module we may decide that they should be obtained for 80% and above, to compensate for an over-generous assessment instrument. The mapping of every grade can be adjusted individually – requiring decisions for each grade.
2. A second approach discards the marks from any module with differences in teaching or assessment before they are summarized. This requires care if there are many unreliable marks. The student profile will include some unreliable marks not included in the summary mark.
3. A third approach adjusts the percentage marks systematically. Here we recommend a transformation that leaves 0% and 100% unchanged and changes other marks in a linear way.

The first example of such a transformation (Figure 4) would be suitable for a module where typical students seem to have been marked too generously. Here a two-piece linear transformation is based on a median mark of 65% that is lowered to 55%. In a second example (Figure 5), students, whom we believe should have passed a module, have failing marks. The fail threshold of 30% is moved to 40% in a two-piece linear transformation. In a third example we believe students who should have received a mark of 70% have received 80% but do not want to affect students with the median mark or less. Figure 6 shows a three-piece linear transformation that achieves this. (example spreadsheets can be downloaded from <http://www.keele.ac.uk/depts/cs/Staff/Homes/Mdb3/papers/assess.htm>)

Figure 4

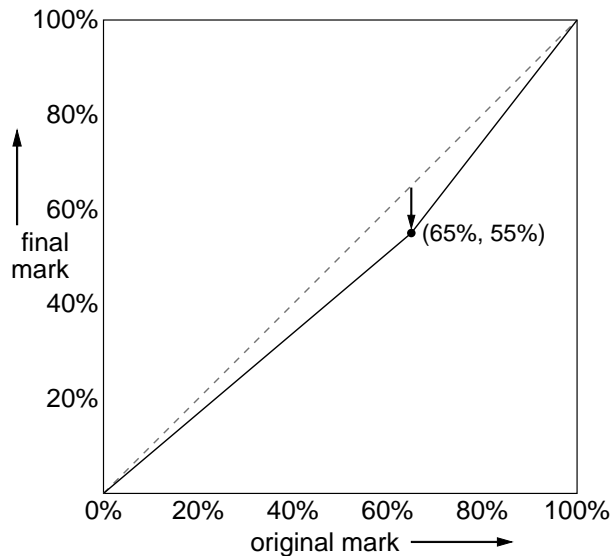


Figure 5

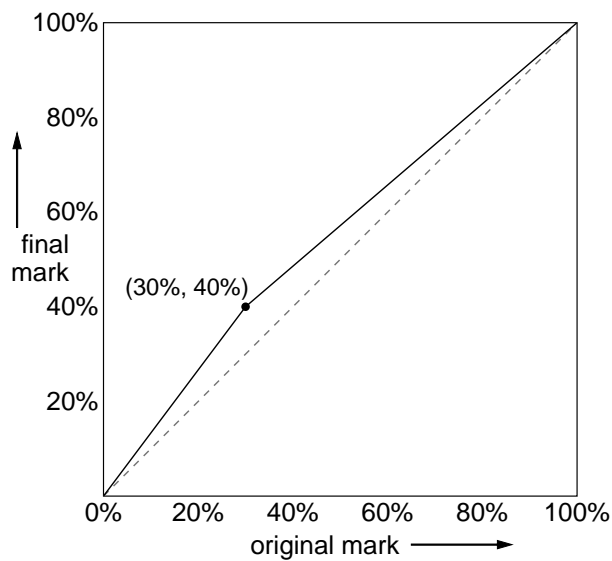
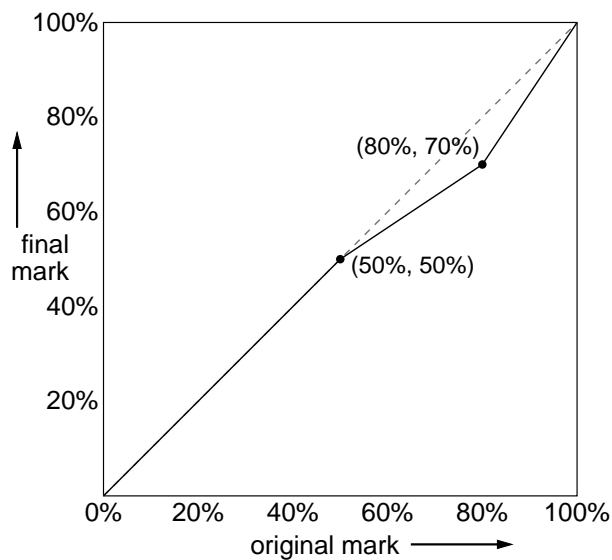


Figure 6



We stress that these adjustments should not be used in any blanket way to make all modules have similar medians or means. Any adjustment should be carried out in the context of the programme outcomes once the causes of unusual raw assessment marks are understood.

Cuthbert P. 2001 Large student groups: techniques for monitoring marking
Educational Developments 2.2 17-20

Siegel, S and Castellan, N.J. 1988 *Nonparametric Statistics for the Behavioral Sciences* New York: McGraw-Hill

Stephen Bostock is an Academic Staff Developer, and Mike Brough is Examinations Officer of the Computer Science Department, at the University of Keele. Both have been External Examiners.

Correspondence: s.j.bostock@keele.ac.uk