Mark Scaling in Physics

In any system of assessment where all candidates do not necessarily take the same modules, like that in the Physics department, measures ought to be taken to ensure that, as far as is possible, equal assessment standards apply across all modules. Similarly measures should be taken to ensure that standards remain similar over the years so that a mark of 60% in year 2005 in a particular module means much the same as a mark of 60% in 2015. Internal and external checking of exam papers does this to some extent. However, in addition, module marks are often scaled in some way. Indeed, as you probably know, the exam boards do this at A-level (by shifting the grade boundaries up or down) based on knowledge acquired from candidates GCSE performance.

In the Physics department we do not scale marks in year 1 as their overall contribution to final classification is small, and the degree of optionality is limited. All other examination marks are scaled using the procedure described below. Assessed work marks are moderated, but not scaled since assessed work often tests different skills to examinations and there is therefore no obvious data set to scale with reference to.

There are two stages to the process we use, the first of which deals largely with comparability from year to year (and typically has very little effect). The second deals with module to module fairness. In this second stage we take all candidates based in physics taking a particular module, and scale their marks with reference to their average performance in all other physics modules. Students from other departments taking a physics module will have their marks scaled, but their performance will not be used to determine how the scaling will occur.

The scaling procedure determines what the average for a module should be. We then remap the dataset (your marks) to obtain this average in what we believe to be the fairest way (it requires iteration on a computer). Rank order is of course maintained, whilst very high, or very low, marks will not be changed much (0 and 100 are fixed points). This approach has been extensively reviewed by a number of external examiners, who have all expressed their approval.

The Procedure

Examiners of all modules in years 2, 3 and 4 are given a guide average mark based on the performance of their cohort of students on examinations the previous year and are expected to return an average mark within 8% of this figure.

All examination papers taken by 2nd, 3rd and 4th year students are automatically scaled. The first stage of the procedure for a given year of study is to determine the global average (ignoring marks of less than 10%) of all physics-based students on physics examination papers. This value, A_p, is compared with the target figure of 62% for years 2 and year 3 and 67% for year 4 and the quantity $A_p' = \frac{A_p + 62(\text{or } 67)}{2}$ is determined. A_p' becomes the new global target average and the appropriate global scaling factor is determined. This is expected to be (and always has been) very close to unity. A scaling factor for each module is then determined by finding the overall average on all other physics papers, for the cohort taking that particular module (again ignoring marks of 10% or less), and scaling the module to that average using the procedure described below.

We denote the original marks by X_i , $i = 1$ to N, where N is the number of candidates, and the adjusted marks $Y_i = y(X_i)$. It is convenient to take these in rank order with respect to X so $X_1 \leq X_2 \leq \ldots X_N$, and to extend the notation such that $X_0 = 0$ and $X_{N+1} = 100$. We require that the mapping $y(x)$ has the properties $y(0) = 0$, $y(100) = 100$ and $\sum_i Y_i = N\overline{Y} = S$, where S is a fixed known value determined by the required average mark after scaling. Subject to these constraints, we extremise a "fairness" functional F:

$$F[y] = \int_0^{100} dx\, f\left(\frac{dy}{dx}\right)$$

which is chosen to minimise the distortion of incremental marks governed by $\frac{dy}{dx}$. We require $f(s)$ to be a convex function, and ideally f should be such that negative values of $\frac{dy}{dx}$ cannot arise. Using the method of Lagrange multipliers, we extremise $F + \lambda \sum_i Y_i$ leading to the equation

$$\frac{d}{dx}\left(f'\left(\frac{dy}{dx}\right)\right) = \lambda \sum_i \delta(x - X_i),$$

where the Lagrange multiplier λ has to be adjusted to force the desired average mark subject to the boundary conditions $y(0) = 0$ and $y(100) = 100$. The general solution satisfies

$$f'\left(\frac{dy}{dx}\right) = A + \lambda\, i, \quad X_i < x < X_{i+1},$$

where λ and A are determined from the overall average and the boundary conditions on $y(x)$.

Various choices for $f(s)$ are possible. The one we have chosen is $f(s) = s \ln s$ (the Gibbs measure), which leads to $\frac{dy}{dx} = B \exp(\lambda\, i)$, $X_i < x < X_{i+1}$,where B and λ are determined by the constraint and the boundary conditions on $y(x)$. It requires iteration but converges rapidly. Using this measure means that all scaling is reversible and negative slopes are not possible. It keeps the changes to marks slowly varying, and, on tests on previous years' datasets, we have found that it leads to minimal distortions in the mark increments.

To avoid any issues concerning pass marks for modules, no raw mark, that was above 40%, will be scaled to below 40%, though the reverse will be permitted.

Discussion

Inevitably you may ask yourself whether the above approach is fair. We believe that it is fairer than just using raw examination marks. Firstly, setting an appropriate target average mark for an exam means that the examination paper has to be designed to achieve this average, and hence should be a fair assessment of the module. The first stage of scaling, where we compare the global average of the marks on physics papers gained by physics students in a given year of study, results in almost no change to marks (typically a couple of tenths of a percent). The second stage of scaling has more of an effect and it deals with the following issue. Suppose that there is an optional year 3 module which is taken largely by students who generally get first class marks. The exam paper is sat and marked and there is an average mark of say 65%, which is well inside the target range. However it is a lower mark than the cohort of students taking it generally obtains, either because the exam paper was actually rather tough, or perhaps because the marking was tough. The effect of our procedure is that the marks on this paper will be normalised upwards so that the average matches that which this group of students obtains on all their other papers, which might have been 75%.

October 2015