How to determine f

Document 3 in a series of 3

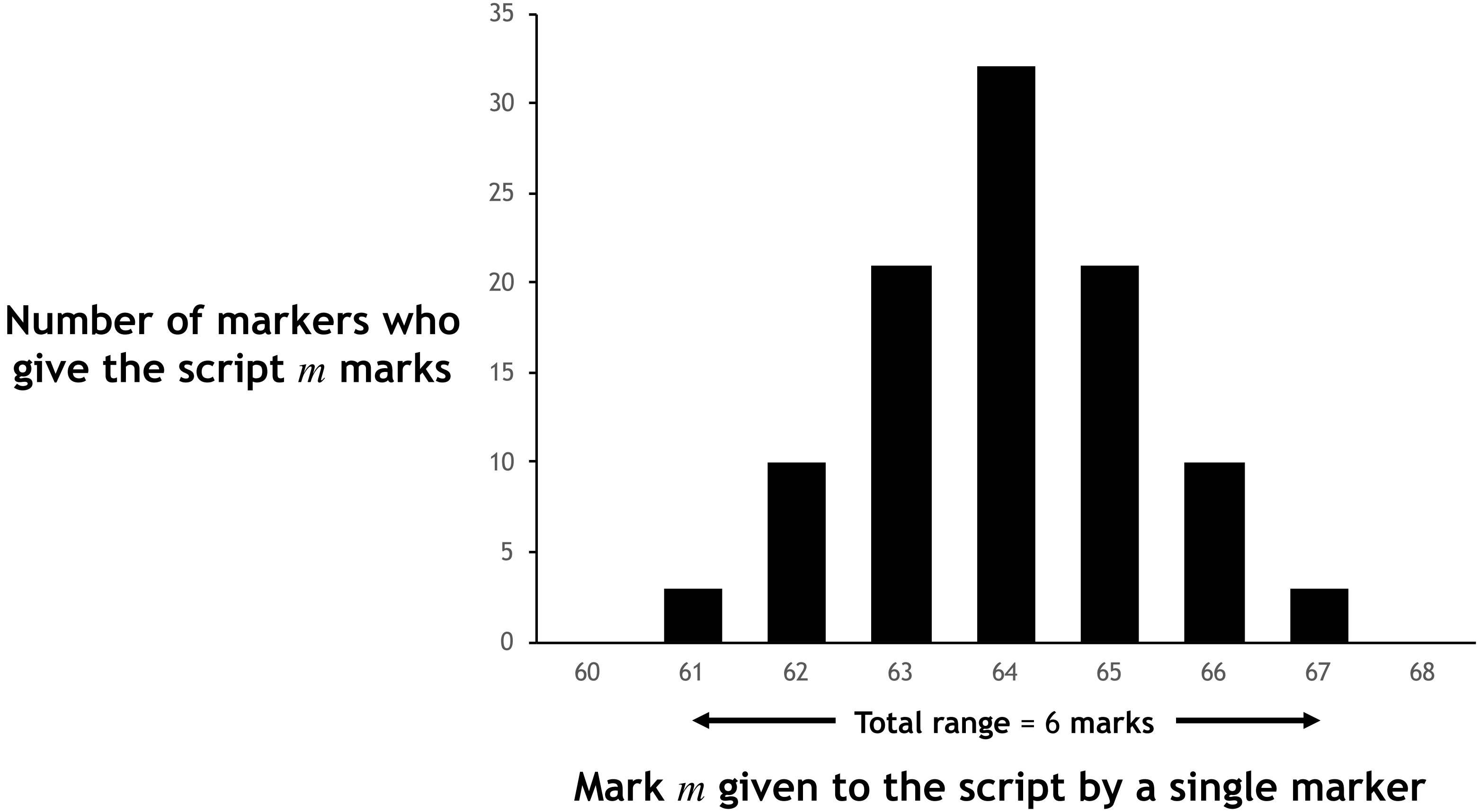
Dennis Sherwood, July 2017

As discussed in the accompanying documents *The great grading scandal* and *How to make grading fair*, the grading of examinations can become much more fair if the grades are based not on the mark m, as originally given, but on an 'adjusted mark' $m + f$, where f is a number, specific to each examination, and applied to the marks of all candidates.

The significance of f is that it is a measure of the intrinsic variability in marking: a variability attributable to the reality that different markers can give (slightly) different marks to the same submission, whilst remaining compliant with the examination marking scheme.

This document describes four different possible methods as to how f might be determined: a policy choice needs to be made to agree the actual method to be used.

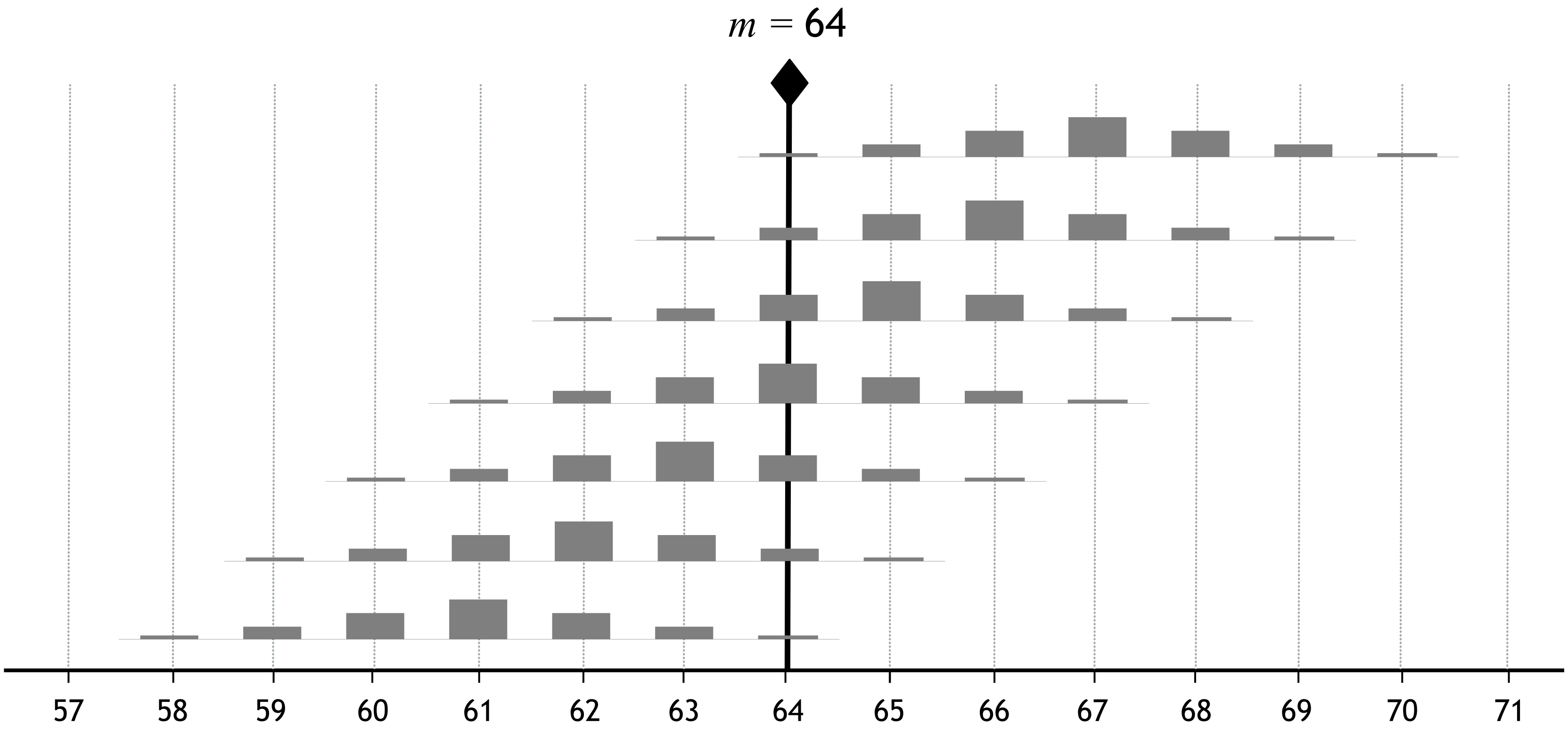
The variability of marking - the ‘panel distribution’



If a panel of, say, 100 markers were each to mark the same submission, they will not all give that submission the same mark m: rather, the marks might be distributed as illustrated here. Most markers give $m = 64$, whilst all the marks are between $m = 61$ and $m = 67$, a total range of 6 marks. The shape of this distribution is known as the ‘panel distribution’. In practice, it is not possible for any particular submission to be marked by a panel - any submission is given a single mark m by a single marker, and this mark must be a sample from the corresponding panel distribution.

The distribution shown here is hypothetical, and for illustrative purposes; it also happens to be symmetrical, as will be the case for some real distributions, whilst other real distributions will be skewed. The analysis presented in this document will use this symmetrical distribution; the analysis for a skewed distribution is very similar: full details are available from the author who may be contacted at dennis@silverbulletmachine.com.

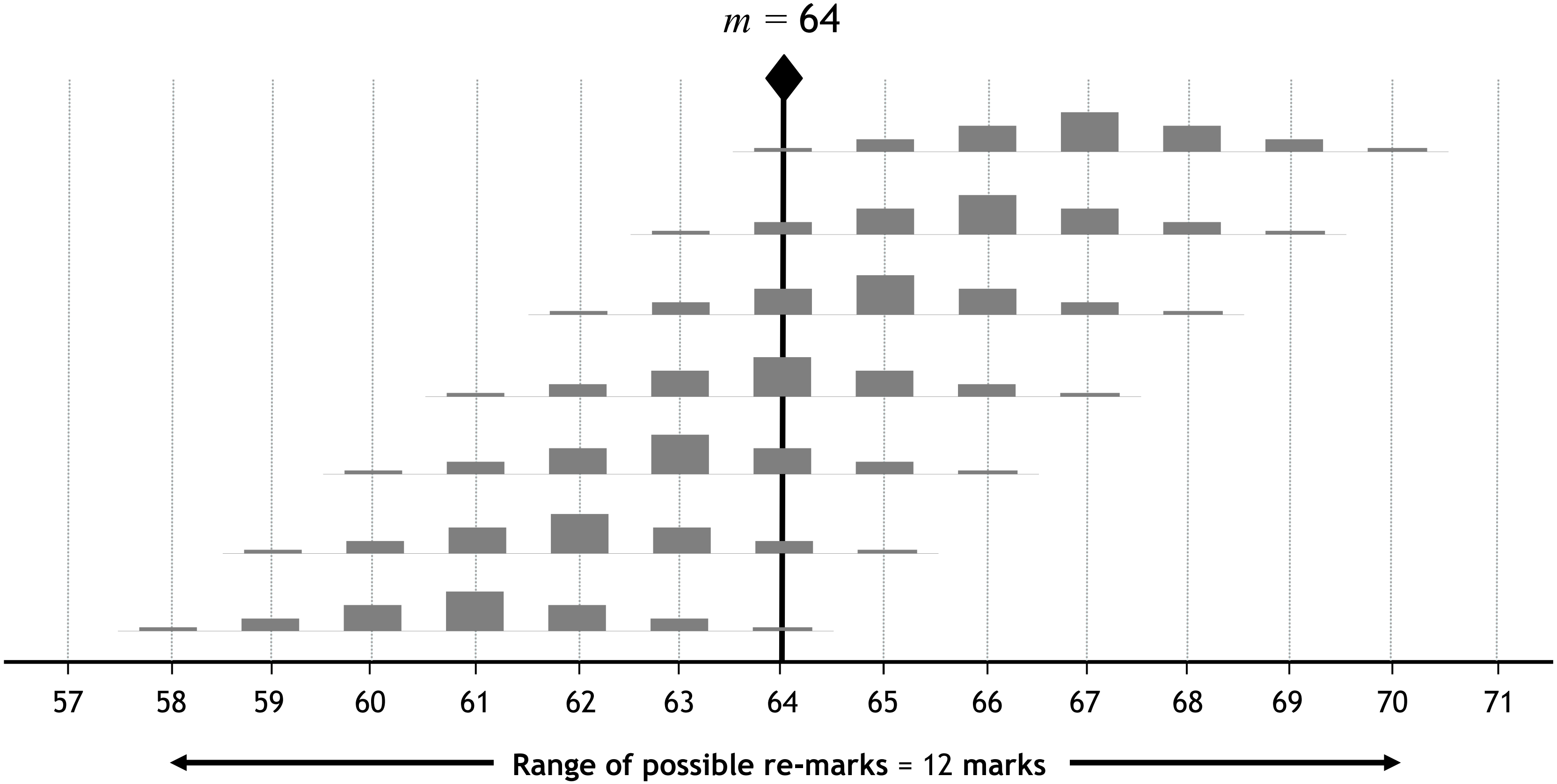
A consequence of the variability of marking



Suppose that a particular submission is given a mark $m = 64$ by a single marker. This mark must be a sample drawn from the panel distribution that would be generated if 100 markers were to mark that script.

In practice, it is impossible for every submission to be marked by a panel: each submission is given a single mark by a single marker. But although the specific panel distribution for this particular submission is unknown, it must be one of the seven possibilities represented here, in which each 'squashed' distribution has the same shape as the panel distribution shown on page 2.

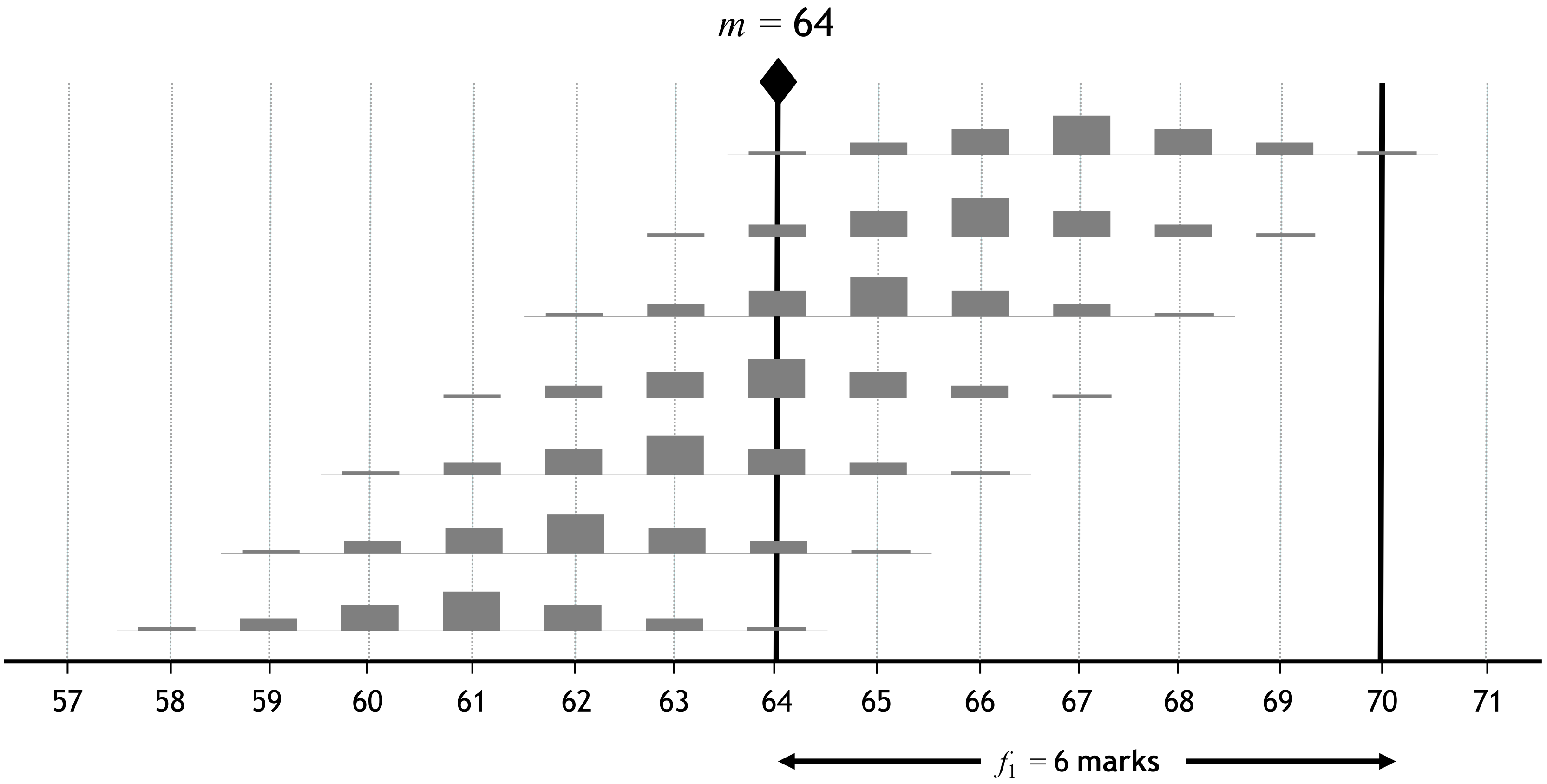
What happens when a script is re-marked



Suppose that a particular submission is given a mark $m = 64$ by a single marker, and that, perhaps as a result of an appeal, the submission is given a re-mark m^* by a different marker. Both the original mark m, and the re-mark m^*, must be a member of the same panel distribution, but in practice, there is no knowledge as to which of the seven possible panel distributions that particular panel distribution might be.

This implies that, in principle, for an original mark $m = 64$, the re-mark might be as low as $m^* = 58$ (if the original mark $m = 64$ is a member of the panel distribution centred on 61, as at the bottom) or as high as $m^* = 70$ (if the original mark $m = 64$ is a member of the panel distribution centred on 67, as at the top). The range of possible re-marks is therefore 12 marks, this being twice the range of the underlying panel distribution, as shown on page 2.

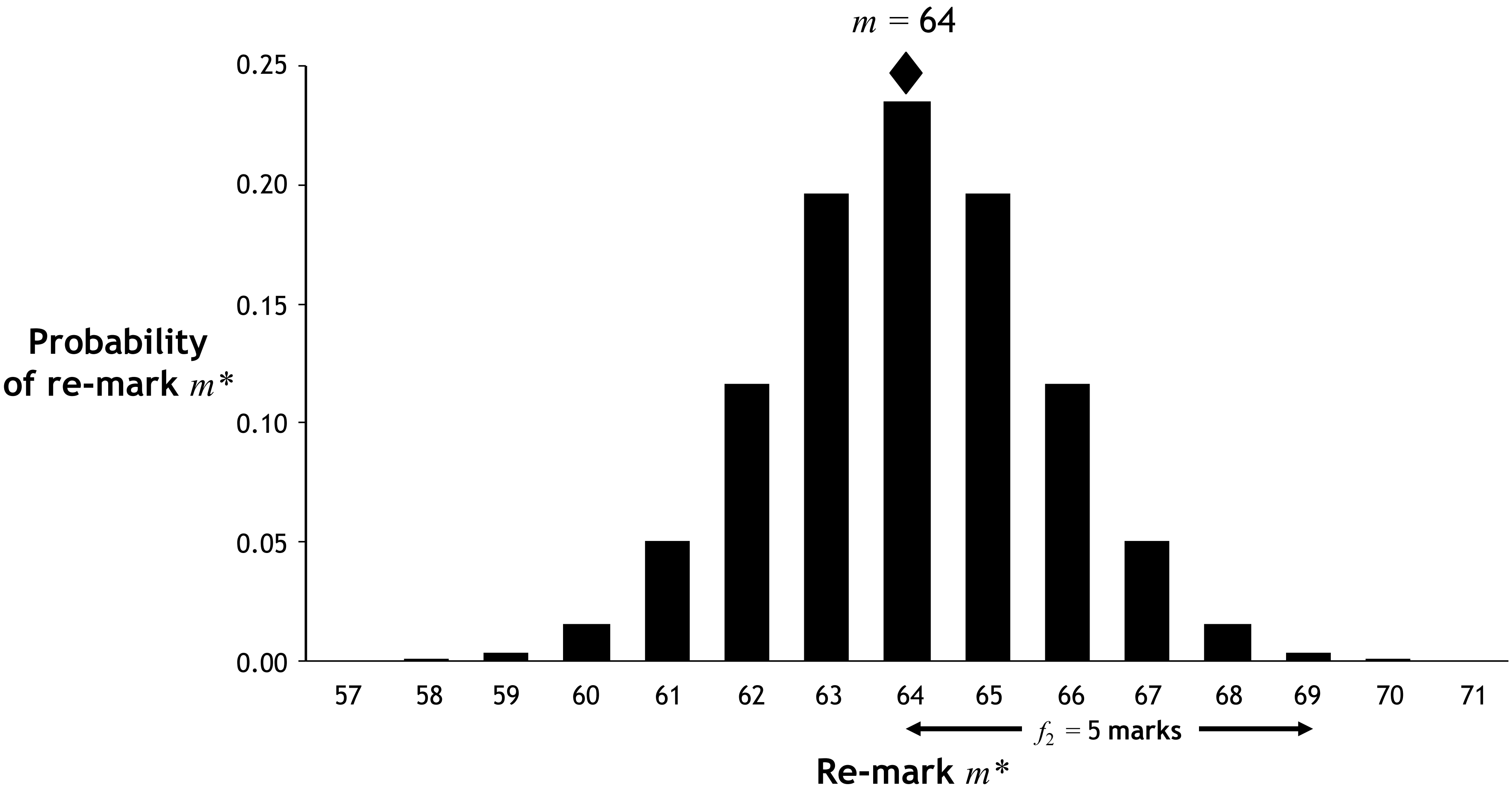
Making grading robust under appeal - one measure of f



Suppose that a particular submission is given a mark $m = 64$ by a single marker, and that, perhaps as a result of an appeal, the submission is given a re-mark $m*$ by a different marker. If both the original mark m and the re-mark $m*$ are within the same grade width, then the awarded grade is the same. But if m and $m*$ are on different sides of a grade boundary, and if the policy is that the re-mark $m*$ takes precedence, then the grade is changed. This is what happens currently, and is intrinsically unfair: the candidate's grade depends, in effect, as to which of the two marks m and $m*$ happens to be given first. Furthermore, which grade is right? They can't both be right, resulting in the problem of grade misallocation, as discussed in *The Great Grading Scandal*.

If, however, the grade is determined not by the original mark m, but by the 'adjusted mark' $m + f$, where f is the distance to the maximum possible re-mark (in this case, shown as $f_1 = 6$ marks, the same as the total width of the panel distribution shown on page 2), then the grade as awarded in robust under appeal: any re-mark $m*$ must be equal to, or less than, $m + f_1$.

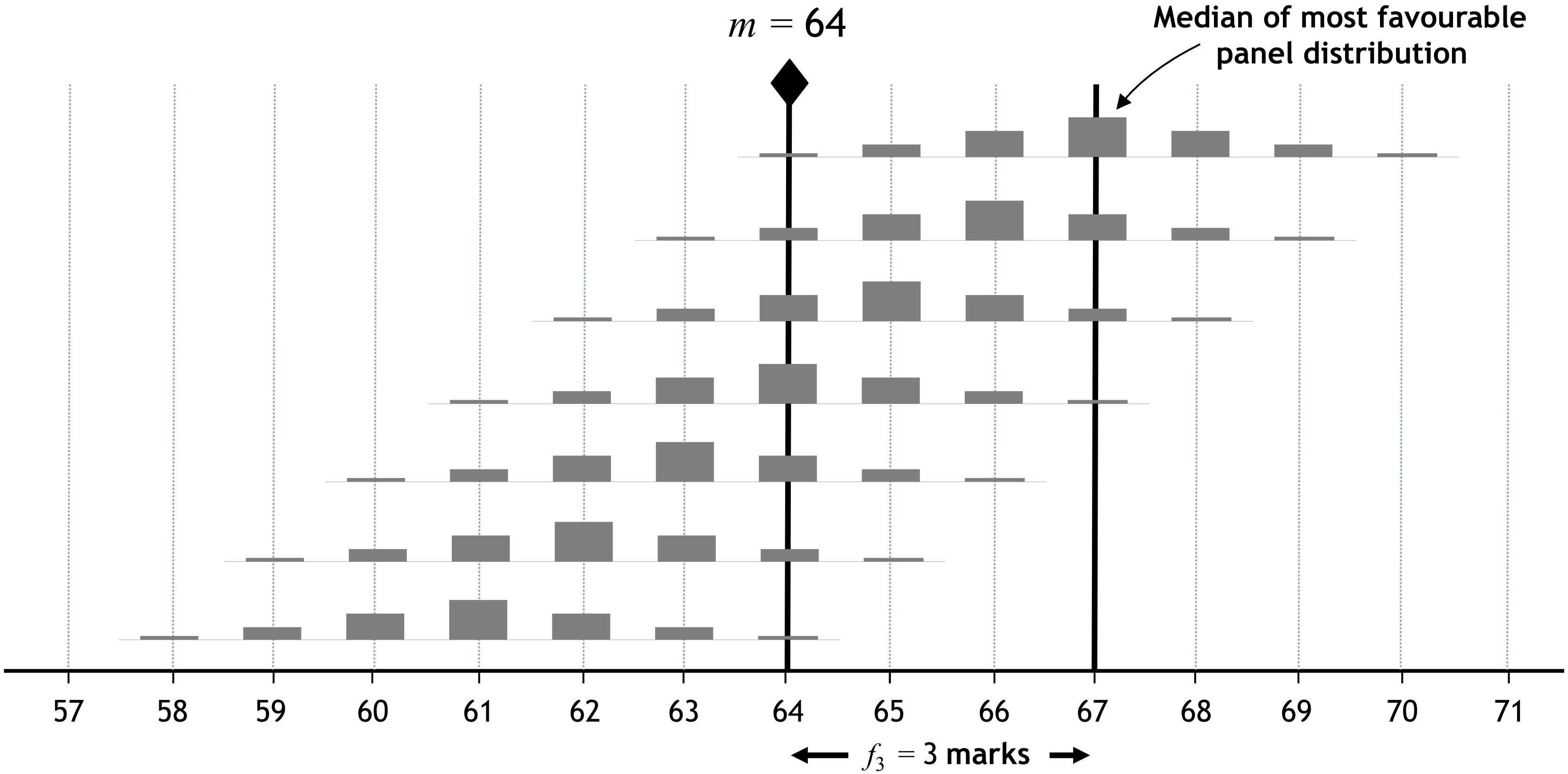
Making grading robust under appeal - a second measure of f



For a submission given an original mark $m = 64$, and for which the panel distribution shown on page 2 is valid, although all re-marks m^* from $m^* = 58$ to $m^* = 70$ are possible, they are not equally probable. The probabilities of each re-mark m^* are shown here, from which it can be seen that a re-mark $m^* = m = 64$ is the most likely, and re-marks of $m^* = 58$ and $m^* = 70$ are highly unlikely. A second possible definition of f takes these probabilities into account, suggesting, in this case, that $f_2 = 5$ marks, somewhat less than the value of $f_1 = 6$ marks as shown on page 5.

Mathematically, the shape shown here is known as the 'self-convolution' of the panel distribution shown on page 2 (or the 'autocorrelation' if the panel distribution is asymmetrical).

A third, statistical, measure of f



A third possible method of determining f is by reference to the number of marks between the given mark m and the median of the most favourable panel distribution containing the given mark. In the case shown, $m =$ 64, the median of the most favourable panel distribution is 67, and so $f_3 = 3$ marks.

For a symmetrical panel distribution, f_3 in one-half of the total width; for an asymmetrical panel distribution, f_3 is equal to the number of marks between the median and the panel distribution's lowest value.

This method of determining f is statistically rigorous, and is based on the policy that the grade for any submission is most fairly based on the median of the corresponding actual panel distribution. There is, however, a potential problem as regards appeals, for a fair re-mark could be greater than the 'adjusted mark' $m + f_3 = 67$, as used to determine the submission's grade.

A fourth, mathematical, measure of f

A fourth possible method of determining f is by reference to the 'tolerances' used by all the examination boards, whereby a senior examiner will review the marks given by a particular marker marks to verify that all marks are within 'tolerance'. So, for example, a board might have the policy that an item with a maximum of 3 marks has a tolerance of 0; an item of 4 to 10 marks, a tolerance of 1 mark; an item of 11 to 17 marks, a tolerance of 2 marks; and an item of 18 to 25 marks, a tolerance of 3 marks.

Consider, then, an examination structured as

5 questions, each of 10 marks, and each with a tolerance of 1 mark
2 questions, each of 15 marks, and each with a tolerance of 2 marks
1 question, of 20 marks, and with a tolerance of 3 marks

Suppose further that, for these eight questions, a candidate receives marks of 7/10, 4/10, 6/10, 3/10, 8/10, 9/15, 13/15 and 14/20, giving a total of 64/100. According to standard statistical theory, the variability of the total can be estimated from the tolerances of each question as

$$(\text{Variability of total})^2 = 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 2^2 + 2^2 + 3^2 = 22$$

The variability of the total is therefore estimated as 4.7, implying that the candidate's total mark is 64 ± 4.7.

If a similar exercise is carried out for a number of submissions, then the average of the resulting total variabilities can serve as a fourth method of estimating f, so defining f_4 as, say, $f_4 = 5$ marks.

Which method is best?

This document has described four possible methods of determining f - there are undoubtedly others too, and these should be identified. Having done so, a debate then should be held so as to agree which method is the most suitable for operational use.

Of the four methods presented here, however, my preference* is for the first method - to determine f as the total width of the panel distribution, as described on page 5. Why so?

- This method makes grading robust under appeal, which is both important, and also very easy to explain and to understand: by assigning grades in this way, it is extremely unlikely that a submission, given an original mark m and subsequently re-marked m^*, will result in a re-grade.
- This confirmation of the original grade will build confidence in the examination system, and will eliminate the unfairness associated with the current system.
- It also implies that there are almost no 'disadvantaged' candidates at any grade boundary.
- It is easy to put into practice, for it requires no statistics, other than the identification of the width of the underlying panel distribution, which can readily be done from a sample of submissions.

Method 2, which also has the objective of making grading robust under appeal, is statistically more sophisticated than method 1, and perhaps more valid. My view, however, is that the statistical complexity of 'convolutions' and 'autocorrelations' makes this method much harder to explain, and vastly harder to understand, for relatively little additional benefit.

Method 3, based on the distribution of medians of the relevant panel distributions, will appeal to experts and statisticians, for the argument that the fair grade for a submission should be based on the median of the actual panel distribution has considerable validity. This method, however, is complex; furthermore, although grades might be based on the 'adjusted mark' $m + f_3$ (see page 7), a *different* rule has to be used for fair appeals, which makes the application of this method even more complex.

Method 4, to me, is not suitable: it is based on mathematical calculations and therefore somewhat separated from reality, whereas methods 1, 2 and 3 are all based on empirical data, as expressed by the panel distribution.

* There are two important objections to this method. The first is that it is the most generous, in that it implies more 'lucky' candidates than the other method. The second is that it implies that the most fair way of determining a submission's grade is based on the high-end mark of the panel distribution ($m = 67$ in the diagram on page 2), and not the median ($m = 64$ in the diagram on page 2). This is controversial, and merits discussion. Experts will certainly, and validly, assert that the median is 'right'. My personal opinion, however, is that it does not matter which specific mark within the panel distribution is used for determining the grade, for all members of the panel are suitably qualified. What is important is that *the same* mark is used for every submission. In that context, the ease of explanation and understanding is to me very important, so my preference in method 1. But that's just my opinion - we need to have the debate.

Silver Bullet

The Silver Bullet Machine Manufacturing Company Limited

Building ultimate competitive advantage

Barnsdale Grange, The Avenue, Exton, Rutland LE15 8AH
E-mail: dennis@silverbulletmachine.com
Website: www.silverbulletmachine.com
Mobile and messages: 07715-047947
Telephone: 01572-813690

Idea generation, evaluation and development

Strategy development and scenario planning

Training and knowledge transfer

Business and market modelling

Conferences

Building high-performing teams

Making innovation happen