MATHEMATICAL THEORY OF STUDENT ASSESSMENT THROUGH GRADING

ADAM S. SIKORA

Abstract. We develop a mathematical theory of assigning grades at the conclusion of educational courses. Usually, such grades are derived from final scores (which can be assumed to range from 0% to 100%) and are assigned on the basis of various approaches and traditions, for example by fitting grades into a "bell curve", or by looking for gaps in between scores ("clustering" method), or by assigning grades on the basis of pre-determined thresholds, such as 90% for A, 80% for B, and so on. Are these methods rationally justified? Are they informative and fair to the students? We answer these questions by introducing a quantitative analysis of grading methods based on the information theory and statistics. Specifically, we introduce

(1) the "entropy" of a grading method which measures the amount of information carried by a student's grade, and
(2) the "inconsistency" of a grading method, which measures the influence of random factors on a student's grade.

We argue that the best grading methods are characterized by maximal entropy and minimal inconsistency. We show that the maximal entropy grading methods assign, on average all possible passing grades with equal frequencies e.g. frequency 1/10 for each grade in the 10 grade scale, A, A-, ..., D. On the basis of computer Monte Carlo computations, we show that a properly defined clustering grading method satisfies this property for classes up to around 50 students. However, the "quantile grading method" (defined in the paper) has the maximal entropy for classes of all sizes and is the most consistent among maximal entropy grading methods analyzed by us. We also show that the following popular grading methods are generally suboptimal:

(1) any overly lenient grading practice assigning A's and B's to majority of students.
(2) fitting grade distribution into a "bell curve", by assigning grades on the basis of the distance of a student's score from the class mean, measured in standard deviations. (We also show that most often fitting grades into a "bell curve" like that is difficult, if not impossible, from a practical point of view.)
(3) Assigning grades on the basis of the threshold 90% for A, 80% for B, and so on.

Finally, we discuss situations in which the teacher desires to assign grades with a pre-determined average. We show that among grading policies of that type, the highest entropy grading method has an exponential distribution. We discuss practical methods of assigning grades in that way.

1. **Introduction.** In the first part of the paper, we analyze distributions of final scores in educational courses on the basis of data available to us. We observe that although these scores are usually roughly normally distributed, various anomalies appear at the lower end of the distribution. We conclude that one obtains a much better model of score distributions by limiting oneself to passing scores and by using the truncated normal distribution approximations, cf. Sec. 3.

In the second, main part of the paper we develop a mathematical theory of grading policies at the conclusion of educational courses ("summative assessment"). We focus on the role of grading to provide an accurate assessment of students' mastery of contents in comparison to their peers.We do not discuss here formative assessments, which identify specific students' weaknesses to inform immediate instruction. Nor we discuss "pass-fail" assessments – that is assessments of sufficient mastery of a subject, with a pass-fail outcome – such as driver's license tests and professional bar exams.

Students' assessment in K-12 education, colleges, and graduate schools is traditionally achieved by computing students' percentile scores which are the sums or weighted sums of partial scores divided by the total number of points available. We will refer to these percentile scores, ranging between 0 and 100%, as scores. Afterwards these scores are usually translated into letter grades, by various "grading methods", which are the main focus of this paper. Such grading schemes often involve fitting grades into a "bell curve", or by looking for gaps in between scores ("clustering" method), or by assigning grades on the basis of pre-determined thresholds, such as 90% for A, 80% for B, etc. Are these methods rationally justified? Are they informative and fair to the students? We answer these questions by developing a quantitative analysis of grading methods based on the information theory and statistics. Specifically, we analyze different grading methods by evaluating their two major characteristics:

(1) the entropy of a grading method which measures the amount of information about students' scores encoded in grades, cf. Sec. 9.
(2) the "inconsistency" of a grading method, which measures the influence of random factors on a students grade. For example, a grade of a student randomly assigned to one of multiple sections of the same course should depend as little as possible on which section was she or he is assigned to. We measure this characteristics of a grading method by its "maximal inconsistency index" defined and analyzed in Sec. 11.

The desirability for a consistent grading method is self-evident as it follows from the basic principle of fairness.

We believe that the importance of high entropy value of a grading method is equally self-evident. A high entropy grading provides students with an accurate measurement of their knowledge and it informs their potential future educational institutions and employers about students' efforts, proficiency

and ability. This informative value is evident in many studies showing strong correlation between students' grades and their future educational outcomes and professional achievements. For example, there is a strong correlation between High School GPA and college GPA, [HF, Fig 3]. There is also a strong correlation between college GPA and the future income, [FHPR, Oe].

While we advocate for adoption of the best grading methods, we are aware that it takes a village to change a system. For example, if students are routinely assigned A grade in a certain educational setting then a single teacher adopting a different grading method with a much lower frequency of As may be able to provide much more informative assessment of her students to people familiar with her grading method, but a potentially misleading one to those who are unaware of it and ultimately hurtful to her students. One way of addressing this problem is by accompaning all grades by an explanation of the grading method. In more general terms though, this paper aims primarily at developing and promoting optimal grading methods awareness, leading to potential paradigm shift in grading policies and customs, rather than attempting to change the grading method of any teacher on an individual basis.

We have concluded already that desirable grading methods have high entropy and low inconsistency. We will arrive at the following further observations:

(1) Inconsistency differences between different methods are generally small in comparison with the entropy differences, especially for larger classes. Therefore, high entropy appears to be the most important characteristic of a desirable grading method. However, in general, high entropy methods often have also higher inconsistency.
(2) The highest entropy grading methods are those which assign different passing grades with equal frequencies. Among three different grading methods of this type analyzed in this paper, the most consistent one is the "quantile" grading method, cf. Sec. 7.
(3) Computer simulations suggest that the clustering method (defined in Sec. 8) has an optimal entropy for class sizes up to around 50 students. However, they also show that the clustering method is less consistent than the quantile grading method.
(4) Assigning grades on the basis of pre-determined thresholds can be an effective grading policy, although thresholds 90% for A, 80% for B, etc. are not effective.

Our computations of entropies and inconsistency indices of various grading policies assume truncated normal distributions of passing scores and are based on extensive Monte-Carlo computations in the computational statistics system "R". For reader's convenience, we enclose much of the code in this paper.

The results of this paper show that many popular grading methods are suboptimal from the point of view of their entropy value

(1) a grading with a little spread between the highest and lowest grades. In particular, an overly lenient grading assigning A and B's (from A to D scale) to majority of students.
(2) fitting grade distribution into a "bell curve", i.e. assigning grades on the basis of the distance of a student's numerical score from the mean, measured in standard deviations. (We also show that most often fitting grades into a "bell curve" like that is difficult, if not impossible, from a practical point of view.)

Since we have observed that the optimal (highest entropy) grading methods have on average an even distribution of grades, they yield grade average in the middle of grade scale. There are however at least two reasons why a teacher may desire to grade with a different grade average:

- the teacher may believe that due to a random factors his students are stronger or are weaker than average, in comparison to other comparable courses which he teaches concurrently or taught recently. For example, if the class scores are above average, it makes sense to assign above-average grades to students in that class. Such random variations in students' performance will be particularly likely in classes of small sizes.
- The teacher may wish to assign grades with a specific average following what is customary in the other similar courses, as discussed above.

In Section 13 we show that among grading policies with a pre-determined average, the highest entropy grading method has an exponential grade distribution. We discuss practical aspects of assigning grades in that way in Sec. 12.

2. **Acknowledgments.** We wish to thank Bernard Badzioch and Your Name for helpful discussions.

3. **Score Distributions.** Before delving into an in-depth analysis of grading methods, it is important to discuss final class score distributions first. For the several class scores analyzed by us, they are roughly normally distributed in general. There are however at least three limitations to this pattern occurring:

(1) two "peaks" may appear in score density (histogram) in inhomogeneous student groups (say, when students of two different college majors enroll in a course, or there is a large number of repeating students in a course).
(2) sometimes there is an unusually high number of students with scores very close to 100%, which may be indicative of overly lenient testing and scoring policies.

For these reasons, we will not assume any specific score distribution when it is not necessary. When it is, we will assume that the student group is fairly homogeneous and that the testing and scoring policies are adequate

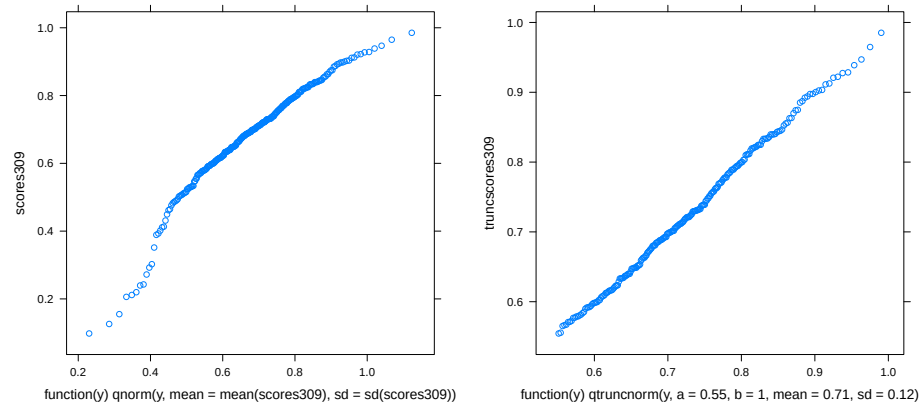

FIGURE 1. Left: Q-Q plot of scores of a class of 281 MTH309 students of Example 1 against the normal distribution. Right: Q-Q plot of the 234 scores above 0.55 against the normal distribution truncated to interval $[0.55, 1]$ with parameters $\mu = 0.71$, $\sigma = 0.12$ given by the maximum likelihood estimation method.

for the class. The scores available to us for such classes are roughly normally distributed, cf. Example 1 below. They exhibit however another limitation to the "normality" assumption:

(3) the "tails" of score distributions (the low scores) are often irregularly distributed. In our experience, it is usually a consequence of underperforming students stopping handing in assignments and preparing and taking tests when they give up in a course (say, after realizing that they have little chances of passing it with a satisfactory grade).

We are going to analyze this phenomenon using Q-Q plots which are a convenient graphic method of testing whether a data set follows a given distribution. Observed values are represented by points in XY-plane whose Y-coordinates are their values and the X-values are the corresponding quantiles of the conjectured distribution. If observed data fits a given distribution well, the scatter points are close to the diagonal line.

Example 1. *In recent Linear Algebra, MTH309, class of* 281 *students at University at Buffalo the score average was* $\mu_s = 67.7\%$ *and the standard deviation* $\sigma_s = 15.3\%$. *Figure 1(left) shows the Q-Q plot of the scores against the normal distribution* $\mathcal{N}(\mu_s, \sigma_s^2)$.

Addressing issue (3) above, let us consider now student scores above certain cut-off value m and let us try to fit them to normal distributions truncated to $[m, 1]$, denoted by $\mathcal{N}_{[m,1]}(\mu, \sigma^2)$, where μ and σ denote the mean and the standard deviation of the untruncated distribution. For the purpose

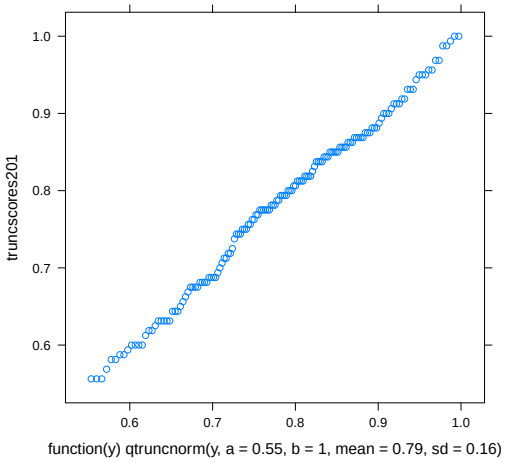

FIGURE 2. Q-Q plot of scores of 149 students above 0.55 in a class of 173 students enrolled in Math201 at UB versus the best fitting normal distribution truncated to interval $[0.55, 1]$.

of the examples in this paper, we will consider $m = 0.55$ which appears to be optimal for class scores data available to us.

Figure 1(right) shows the Q-Q plot of MTH309 scores above $m = 0.55$ against $\mathcal{N}_{[0.55,1]}(\mu, \sigma^2)$ for $\mu = 0.71$, $\sigma = 0.12$ (obtained by the maximum likelihood estimation method applied to MTH309 scores above m, cf. comments below). Clearly $\mathcal{N}_{[0.55,1]}(\mu, \sigma^2)$ gives a much better approximation of the distribution of the scores above m than the normal distribution with the class mean and standard deviation.

Other class score sets available to us, when truncated to $[0.55, 1]$, fit the truncated normal distributions equally well, especially for classes of large size. For example, Figure 2 shows the Q-Q plot for the truncated scores in MTH201 class of 173 students, taught recently at UB.

Obviously, truncated normal distribution does not model scores below m. These however will not be needed in our analysis of grading methods, as these scores usually result in the failing grade and this paper concerns passing grades only. See further comments in Sec 4.

We will call the scores above m the truncated scores or, simply, scores (since the scores below the cut-off value are to be ignored.)

We found the parameters μ, σ for the truncated normal distribution above by applying the best maximum likelihood estimation method. There is a number of software implementations of this method available. In the freely available program "R", [R], which we will use throughout this paper, an implementation of the maximum likelihood estimation for truncated normal distributions which we found most convenient is given by the command

`mle.tmvnorm` in the package `tmvtnorm`. Specifically, μ and σ can be obtained by the following code:

```
library(tmvtnorm);
ml<-mle.tmvnorm(cbind(scores),lower=m,upper=1, lower.bounds=c(m,0),
upper.bounds=c(1,Inf), method="L-BFGS-B");
c(coef(ml)[[1]],coef(ml)[[2]])
```

(Before an R package is loaded with `library()` it needs to be installed with `install.packages("...")`, where the triple dot is substituted with the package's name.)

For the purpose of examples in this paper we will assume $\mu = 0.75$ and $\sigma = 0.14$ which we find to be representative of values for the course scores available to us. The truncated normal distribution with these parameters has mean $\mu' = 0.76$ and standard deviation $\sigma' = 0.11$.

Finally, it may be useful to point out that maximum likelihood estimation methods rely on a sequence of numeric approximations which sometimes fail to converge. In particular, the above R code may result in error. In our Monte Carlo analysis, among samples of N scores generated by $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ the R computation above resulted in error around 0.5%-3% of the times, depending on the value of N. In practical terms, one may still compute the parameters μ, σ by a tiny modification (say by adding 0.001) of one of the scores.

4. **Grading methods.** The educational assessment literature divides grading methods into norm-referenced, criterion-referenced, and ipsative ones, cf. eg. [Gl]. Norm-referenced grades are based on a comparison of a student with his or her peers, unlike criterion-referenced grades which are assigned on the basis of pre-defined criteria independent of peer scores. The ipsative assessment measures students progress against his or her own past scores. Since ipsative grading is rare in K-12 and college education, it will not be discussed explicitly in this paper, even though some of our theory applies to that grading as well.

Since failing grades have very distinctive consequences for the students who receive them and often various circumstances have to be considered when assigning these grades, we will not discuss such decisions in this paper. Instead, we focus on assigning grades to passing students on the basis of their scores.

From now on we will assume for simplicity that the cut-off passing score value is precisely the value m of the last section. We will be assuming $m = 0.55$ in our examples.

The results of this paper hold for arbitrary grading scales. However, for the purpose of examples we will usually consider the scale: A, A-, B+,...., D+, D, F, which is typical in US, Canada, China, Kazakhstan, Pakistan, Saudi Arabia and some other countries. Note however that some institutions in these countries include A+ grade and/or exclude D+. Also graduate

schools in some countries (like US) often have shorter scales, A, A-,..., C+, C, F.

For practical purposes, we will represent letter grades by numbers, with 1 representing the lowest passing grade and G represents the highest one. Hence, for the purpose of our examples $G = 10$ and 10 represents A. A grading method is an assignment to each finite multi-set[1] S of scores in $[0, 1]$ a non-decreasing function $g_S : S \to \{1, ..., G\}$.

In this context it is interesting to consider mathematical conditions necessary for a grading method to be "fair". We see two such conditions:

(1) if S' is obtained from S by increasing the value of a certain element from s to s' then $g_S(s) \leq g_{S'}(s')$.
(2) removing an element s from S does not increase the value of $g_S(s')$ of any $s' > s$ and does not decrease the value of $g_S(s'')$ for any $s'' < s$.

Remark 2. *In our terminology, a grading method is criterion-referenced iff $g_S(s) = g_{\{s\}}(s)$ (i.e. $g_S(s)$ does not depend on S containing s). In that situation, condition (1) above implies, that there are grade thresholds $t_2, ..., t_G$ such that a score s translates into a grade g which is the largest number such that $s \geq t_g$.*

Example 3. *In countries using A-D scale, 90% is often a threshold for A, 80% threshold for B, and so on, with some possible modifications allowing for A-, B+, and so on.*

5. **Bell curve grading method.** Usually norm-referenced grading systems rely on "grading on a curve" which consists of assigning grades on a basis of a pre-determined distribution of grades. When the distribution of grades resembles the (truncated) normal distribution, we will (somewhat loosely) call that grading method a bell curve method. We will see below that these methods are hard to implement in practice and never optimal. (A different, much better grading "on a curve" method will be discussed in Sec. 6.)

Let us start with the most straightforward and seemingly "scientific" version of the above, to distribute grade thresholds evenly at intervals of d standard deviations of length, measured from the untruncated score mean.

We say that the bell curve is "centered" at the value $g_c \in \{1, ..., G\}$ if g_c is the unique grade of highest frequency. A bell curve can be also centered at a half-integer, $g_c \in \{\frac{1}{2}, \frac{3}{2}, ..., G - \frac{1}{2}\}$, which indicates that there are two adjacent grades $g_c \pm 1/2$ of equal highest frequency.

A simple math shows that if the bell curve is "centered" at the value g_c as above then the score threshold for grade g is

$$\mu + d\sigma(g - g_c - 0.5),$$

where μ is the class' scores mean, σ is the class' scores standard deviation and d is a chosen constant, cf. Fig 5. We will call grading with these thresholds strict bell curve grading method. This perhaps unintuitive formula should

[1]A multi-set is a sequence in which order of elements does not matter.

make more sense in the following examples: For $G = 10$ and $g_c = 5$ the thresholds are:

Threshold	$\mu + 3.5d\sigma$	$\mu + 2.5d\sigma$	$\mu + 1.5d\sigma$	$\mu + 0.5d\sigma$	$\mu - 0.5d\sigma$	...
Grade	A	A-	B+	B	B-	...

For $G = 10$, $g_c = 4.5$ the thresholds are:

Threshold	$\mu + 4d\sigma$	$\mu + 3d\sigma$	$\mu + 2d\sigma$	$\mu + d\sigma$	μ	$\mu - d\sigma$	...
Grade	A	A-	B+	B	B-	C+	...

Note now that the thresholds for the lowest passing grade (D in our examples) is $\mu + d\sigma(0.5 - g_c)$ but that by our assumption means

$$\mu + d\sigma(0.5 - g_c) = m,$$

implying that

$$d = \frac{\mu - m}{\sigma(g_c - 0.5)}.$$

In our example (quite typical for freshmen-junior math college courses), $\mu = 0.75$ and $\sigma = 0.14$ imply that $d = 0.29$. For $g_c = 5.5$ (middle of the scale between 1 and 10) the distribution of grades is as in Fig. 5(b).

No such example!

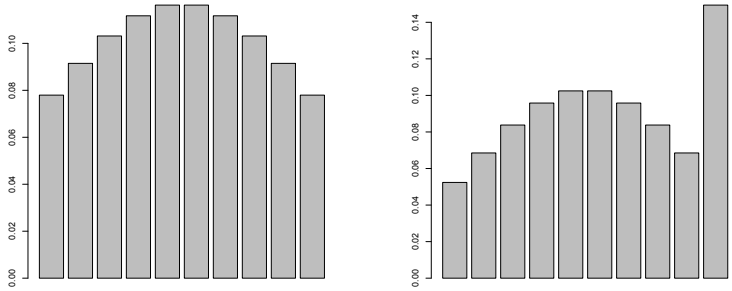

FIGURE 3. (a) An example of a bell curve, (b) An attempt to construct a bell curve by considering score intervals of equal length for $G = 10$, and $g_c = 5.5$, $m = 0.55$, $\mu = 0.75$, $\sigma = 0.14$.

As you can see that the distribution of grades does not resemble the bell curve because of the very high frequency of As. The reason for that is that in the corresponding normal distribution the probability of A (score above $\mu + 4d\sigma = 0.89$) is way higher than the probability of $A-$ (score in between $\mu + 3d\sigma = 0.86$ and $\mu + 4d\sigma = 0.89$.

Now note that if we increase g_c then we only increase the frequencies of As. So to make the distribution of grades more "bell curve" like, we need

to significantly decrease the value of g_c. In our example above, one needs to have $g_c \leq 3.5$ but that would force an unusually low grade of C- for an average student in a class.

That problem persists for other values of score averages and standard deviations unless either (a) the standard deviation is much lower than those we have seen in the scores available to us or (b) the cut-off passing score m well below 0.5.

To recap, the strict bell curve grading method does not distribute grades according to the bell curve shape, because of the unusually high number of As. But perhaps it is still a good grading method (and we should just disregard the traditional expectation of the bell curve shape of grade frequencies)? We will conclude the answer in negative to this question in Sec. 9.

Remark 4. *One can achieve a "real" bull curve $\mathcal{N}(\mu, \sigma^2)$ like shaped grade frequencies for arbitrary mu and σ by assigning each grade with a predetermined frequency, computed for example by the following R command yielding the grade frequencies of Fig. 5(a):*

$$7.8\%, 9.1\%, 10.3\%, 11.2\%, 11.6\%, 11.6\%, 11.2\%, 10.3\%, 9.1\%, 7.8\%$$

That means that the the top 7.8% *receives an A, the next 9.1% receives A-, and so on...*

```
#Frequencies for a bell curve
mu=5; s=5;
sapply(c(0:9), function(x) dtruncnorm(x+1/2, a=0,b=10,mean=mu,sd=s))
```

6. **Even distribution grading method.** If the number of students in a class, N, is divisible by G then the simplest efficient grading method assigns each grade to precisely N/G students, i.e. the top N/G students gets an A, the following N/G students gets $A-$ and so on. (Two students with almost identical scores may end up with different grades by this method. We will discuss this fairness issue in Section 11.)

There are various ways of extending this method to the case when N is not divisible by G, the simplest of which is perhaps this:

Let us denote the integral quotient of N by G by $\lfloor N/G \rfloor$ and the remainder of that division by r.

The even distribution grading method assigns grades $1, ..., G-r$ to $\lfloor N/G \rfloor$ students each and grades $G-r+1, ..., G$ to $\lfloor N/G \rfloor + 1$ students each. Intuitively, that means that the grades are distributed evenly, except for a top segment of grade scale being assigned one more the than rest of the grades. For example, for the four grade scale, A,B,C,D, a class of 5 students will be assigned grades A,A,B,C,D, since $\lfloor N/G \rfloor = r = 1$ in this case and a class of 11 will be assigned A,A,A,B,B,B,C,C,C,D,D since $\lfloor N/G \rfloor = 2$, $r = 3$.

Note however that the even distribution grading method is grading on a curve.

A generalization of the even distribution grading method called "exponential grading method" will be defined in Sec. 13.

7. **Quantile grading method.** The following grading method will play important role in this paper. In fact, it will be shown to be the best among grading methods discussed in this paper:

Definition 5. *The quantile grading method assigns grades* $1, ..., G$ *with thresholds* $q_1 = m, q_2, ..., q_G$ *given by the* G*-quantiles of a truncated normal distribution,* $N_{[m,1]}(\mu, \sigma^2)$*, best fitting the distribution of passing scores. Specifically, for each* q_g *is defined by the condition* $P(X < q_g) = \frac{g-1}{G}$ *for a random variable* X *with such distribution. The parameters* μ *and* σ *are computed by the maximum likelihood method applied to the passing scores.*

We will see in Sections XXX and XXX that quantile grading method is overall an excellent grading method, exceeding all other ones known to us in terms of its informative value (entropy) and exhibiting a high consistency (fairness).

Although there are no closed formulas for the G-quantiles (for $G > 2$) of truncated normal distributions, $N_{[m,1]}(\mu, \sigma^2)$, in terms of m, μ and σ, these quantiles can be computed with a good precision by the "R" command

```
library(truncnorm)
qtruncnorm((g-1)/G, a=0.55, b=1, mean=mu, sd=sigma)
```

For example for $G = 10$ and a typical grade distribution $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ (cf. Sec. 3) we obtain the grade thresholds

(1) 61.4%, 65.7%, 69.3%, 72.6%, 75.7%, 78.9%, 82.2%, 86.1%, 91.1%.

The grade of a student with score s can be also computed directly by

```
ceiling(ptruncnorm(s, a=0.55, b=1, mean=mu, sd=sigma)*G)
```

where mu and sigma are given by the maximum likelihood method, with an "R" code provided in Sec. 3.

By its definition, the quantile grading method leads to approximately even distribution of grades. However, unlike for the even distribution method, the distributions of grades by this method will vary (even for a given fixed N). We will comment more on it in Sec. XXX

8. **Grading inspired by clustering methods.** The process of grading can be seen as a division of the students in a given class into G groups – each composed of students of comparable skills or abilities. That point of view places the theory of grading in the context of cluster analysis. There are many clustering methods which lead to promising grading methods. In particular, an adaptation of the k-means clustering method to grading yields the following method for $N > G$:

Definition 6. *The k-means clustering grading method (for $N > G$) sets up grade thresholds which minimize the sum of variances,*

$$M = \sum_{g=1}^{G} \sigma^2_{S_g},$$

where S_g is the multi-set of scores assigned grade g.

We assume $\sigma_S = 0$ for empty and one element sets S.

In theory, it it possible that such clusters $S_1, ..., S_G$ are not unique. In practice, one may use an "R" command `kmeans` and consider its output as the definition of this grading method.

Finally, one can consider more exotic grading systems. Here is just one simple example out of infinity of possibilities:

Example 7. *Teamwork grading method assigns the same grade to all students, determined by which of the pre-assigned thresholds the mean class score fits in.*

9. **Shannon Entropy and Grading method Entropy.** For the purpose of designing a grading method which carries the largest possible amount of information about the students' mastery of contents, reflected by their scores, it is useful to turn to the tools of information theory for a quantitative measure of the volume of information. (For information theory see [MJ]).

An amount of information provided by random quantity X taking values in an n element set with probabilities $p_1, ..., p_n$, is given by Shannon entropy,

$$E(X) = -\sum_{i=1}^{n} p_i \cdot log_2\, p_i. \tag{2}$$

To motivate that concept consider a non-negative integer X written in the binary base system with N digits determined by N coin tosses – heads representing 1. Such X takes any integral value between zero and $2^N - 1$ with an equal probability $p_0 = ... = p_{2^N-1} = 2^{-N}$. Consequently,

$$E(X) = -\sum_{i=0}^{2^N-1} 2^{-N} log_2 2^{-N} = -log_2 2^{-N} = N.$$

This is consistent with the fact that X contains precisely N bits of information.

Assume that passing student scores have a certain fixed distribution. (As discussed earlier, it is $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ for the purpose of this paper.) Then the formula (2) involving grade frequencies $p_1, ..., p_G$ of a grading method is called its entropy. (That entropy will usually depend on the class size.)

Imagine now a person receiving information about a student's grade. That person, say a potential employer or a university admission officer, will usually be aware of grades of other students, occurring with certain frequencies

$p_1, ..., p_G$. Therefore, the amount of information carried by a single grade equals the entropy of a grading method.

Example 8. *A grading method consisting of assigning the same grade to every student has entropy*

$$-1 \cdot log_2 1 = 0.$$

Indeed, this grading method contains no information about students scores.

We will analyze entropies of other grading methods using the following fundamental result belonging to folk knowledge. (For completeness, we enclose a proof of this result in the Appendix at the end of the paper.)

Theorem 9. *A random variable with distribution of values given by $p_1, ..., p_G$ achieves the largest possible Shannon entropy for given G iff*

$$p_1 = ... = p_G = 1/G.$$

The above result provides an easy upper bound on the entropy of grading method with G grades:

$$E_G^{max} = -\sum_{i=1}^{G} p_i \cdot log_2 p_i = -G \cdot \frac{1}{G} \cdot log_2 \frac{1}{G} = log_2 G.$$

For example, for $G = 10$, $E_{10}^{max} = 3.32$.

Definition 10. *We call a grading method efficient (with respect to a given G and N and score distribution ρ) iff for a random N scores (distributed according to ρ – usually we assume truncated normal distribution as above) the expected values of the frequencies of all grades, $p_1, ..., p_G$ are $1/G$. That is equivalent to saying that the entropy of the grading method is E_G^{max}.*

Note that the mean grade of an efficient grading method is $\frac{G+1}{2}$. (That is between C and $C+$ for the traditional grade scale with $G = 10$.)

Example 11. *Since the even distribution grading method assigns grades $1, ..., G$ to a class of $N = G \cdot k + r$ students with probabilities $p_1 = ... = p_{G-r} = k/N$ and $p_{G-r} = ... = p_G = (k+1)/N$, its entropy is*

$$E_{N,G}^{s.e.d} = -(G-r)k/N \cdot ln\,(k/N) - r(k+1)/N \cdot ln\,((k+1)/N) =$$
$$(G-r)k/N \cdot ln\,(N/k) + r(k+1)/N \cdot ln\,(N/(k+1)).$$

Note that for N not divisible by G this method is not efficient, but it becomes very close to one for large N.

On the other hand grade frequencies are not equal in the bell curve method and, therefore, that method is never efficient. For example, the entropy of the bell curve grade distribution of Fig. 5(a) is 2.97.

Although the entropies of other grading methods defined above cannot be given by an explicit formula, we have computed them for $G = 10$ and various N using the Monte Carlo method. The results of our computations are shown in the following table. (As before, the below computation assumes

the distribution of passing scores given by the normal distribution with $\mu = 0.75$ and $\sigma = 0.14$ truncated to $[0.55, 1]$.)Recall that $E_{10}^{max} = 3.32$.

TABLE 1. Entropies of the even distribution grading method, quantile grading method, and K-means clustering method. $G = 10$

Method	Number of students									
	1	2	3	4	5	6	7	8	9	10
EvenDist	0	1	1.58	2.00	2.32	2.58	2.81	3.00	3.17	3.32
Quantile	2.80	2.485	3.22	3.29	3.31	3.31	3.31	3.32	3.32	3.32
K-means										3.32

Method	Number of students									
	11	12	15	20	30	50	100	200	500	1000
EvenDist	3.28	3.25	3.24	3.32	3.32	3.32	3.32	3.32	3.32	3.32
Quantile	3.32	3.32	3.32	3.32	3.32	3.32	3.32	3.32	3.32	3.32
K-means	3.32	3.32	3.32	3.32	3.32	3.32	3.315	3.31	3.30	2.93

Note that the quantile grading method appears to have the highest entropy overall among the methods considered here. It converges to E_{10}^{max} for $N \to \infty$. Indeed, that convergence can be verified theoretically: as N increases, the passing scores distribution will be better and better approximated by the best fitting truncated normal distribution. Since the grade thresholds in this method are given by the G-quantiles of that truncated normal distribution, the frequencies of individual grades will all converge to $1/G$ as $N \to \infty$.

Note that the K-means method appears to be efficient for $N \leq 50$ and that the entropy of this method slightly decreases for large N. It is quite likely though that this is an afterfact of R's inefficiency in performing an optimal k-means clustering on very large class scores, which is a daunting task. If that is the case, then it is that the K-means method is efficient for every N. It would desirable to find a theoretical proof of that fact.

Let us consider now criterion-referenced grading, that is grading on the basis of predetermined grade thresholds. As stated in Example 3, traditionally they are 90% for A, 80% for B, 70% for C and 60% for D, with some additional thresholds for "+" and "-" versions of these grades – for example two extra percentage points for "+" and two less for "-": 88% for A- and 82% for B+ and so on. We will call it the predetermined threshold grading method. For any class mean μ, by the central limit theorem, there will be many more students with the grade bracket of μ than any other grades, that is many more Cs then Bs. For example, under our typical score distribution $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ the grade frequencies assuming the above thresholds are

A 11.8%, A- 3.9%, B+ 14.9%, B 5.9%, B- 6.17%, C+ 19.1%, C 6.17%

and so on. The significant unevenness of these grade frequencies implies a low entropy of the grading method.

On the other hand, a teacher being able to predict a distribution of the final scores in advance may assign thresholds which will result in equal grade frequency. However, it is not easy to make such prediction for small and medium size classes.

Example 12. *Suppose that teachers* 20 *person class had score mean* 0.75 *and the (sample) standard deviation* 0.14 *last year and that the teacher predicts that this years student cohort should be comparable and on the basis of that prediction he establishes the quantile thresholds for* $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ *as in (1). However, the sample mean is not the true mean and the standard error of is* $0.14/\sqrt{20}$ *which means that this years mean should be outside the interval* 0.75 ± 0.03 *with chances around 22%. If the mean is above* $0.75 + 0.03$ *then the frequency of As increases from 10% to 13% and if below* $0.75 - 0.03$ *then the frequency of As decreases to 7% – clearly resulting in an uneven grade distribution.*

10. **The maximal entropy for different** G**.** Any grading system which does not fully utilize the grading scale is inefficient. For example the maximum entropy of a grading method which utilizes grades A,A-,B+, and B only is $lg_2 4 = 2$, which is 40% less than the maximal entropy, 3.32, of the full grade scale A-D.

More generally, the maximal entropy, E_G^{max}, of grading methods increases with G (corresponding to "finer" scales). Therefore, it is useful to allow for scales with large numbers of grades. Indeed, many standardized tests, like SAT and GRE, assigns grades on the scale 0%,1%,...,100% which has maximal entropy $E_{100}^{max} = 6.64$ (for $N \geq 100$)– much higher than that of the A-D scale.

In certain circumstances a small grade scale is sufficient – for example for driver license tests, which are designed to have a pass-fail output. Currently professional bar exams (eg. medical, law) fall into this category as well, however, one might argue that a knowledge of a grade in a larger (finer) grade scale might benefit the future clients and employers of a professional who takes these exams.

11. **Grading consistency and fairness.** Let us start the discussion of grading fairness and consistency with the following simple example:

Example 13. *Imagine a course with four students enrolled with final scores*

94%, 93%, 85%, 75%.

is to be graded by the even distribution grading method with A,B,C,D scale ($G = 4$)*. Obviously each of the students receives a different grade then. These grades seem unfair to the student with 93% score as her score is very close to the top student's score. The same phenomenon will occur for small*

classes graded with other grading methods. However, the scale of that phenomenon does depend on the method used. For example, the quantile grading (based on the normal distribution truncated to $[0.55, 1]$*) assigns grades A,A,C,D to the students in this class.*

In order to analyze the above phenomena we consider a quantitative index of grading method inconsistency, MGI, defined below, which in our view is the second most important characteristic of a grading method, aside from entropy.

To motivate the concept of "(in)consistency" consider a random student, whom we will name Alice, in a class (i.e. the focal cohort) of N students. Suppose Alice's score is x. Her grade will depend on the scores of other $N-1$ passing students in her class (although it may be a null dependence). As usual we will assume that these scores are distributed by a truncated normal distribution. Then Allice's grade, g_{Alice}, is a random variable depending on the grading method and on the values of her $N-1$ peers' scores. The grade inconsistency for score x, denoted by $\mathrm{GI}(x)$ is the the standard deviation of that random variable g_{Alice} divided by $G-1$. That division normalizes our measure of inconsistency making it comparable for different grade scales. Indeed, grades 1 and 3 mean "worst" and "best" on a three grade scale, as 1 to 9 do on the nine grade scale. Hence the grade difference of two on a three grade scale is comparable with the grade difference of eight on the nine grade scale. (We will motivate that value $G-1$ better soon.)

For every grading method discussed in this paper, $\mathrm{GI}(x)$ achieves its maximum for x being the mean class score value. For the score distribution, Really? $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$ it is 0.76, cf. for example Fig. 4 and R code below. There is a simple explanation for that – the closer Alice's grade x is to 1 the higher the chances that Alice receives an A, and hence the uncertainty of her grade decreases. Similarly, if x is close to m then her grade is 0 (F) with high probability and, hence, low inconsistency.

But what if we want to measure the inconsistency of grading of the entire class? One natural approach is to take the maximum of GI, denoted by us by MGI. By the discussion above, that is $\mathrm{GI}(x_M)$ for the class mean x_M.

Another possibility is to consider the average of $\mathrm{GI}(x)$ over all students in class. Under the standard assumption of the truncated normal distribution x, that average is given by

$$\mathrm{AGI} = \int_m^1 \mathrm{GI}(x) \cdot \rho \, dx,$$

where ρ is the probability density of the truncated normal distribution. We call it the average grade method inconsistency.

These two metrics, MGI and AGI are very similar, and since the first one is much easier to compute we are going to use it below.

Note that the "criterion-referenced grading", that is the one based on predetermined grade thresholds is the most consistent one. Indeed, $\mathrm{GI}(x) = 0$ for any x in for that grading!

On the other hand the most inconsistent is the "madman" grading method assigning each student either 1 or G at random, each with probability 1/2. We invoke it here for the purpose of example even though it does not strictly satisfy our definition of a grading method. It is easy to compute that it has inconsistency 1/2. (It is reasonable, to define grade inconsistency alternatively as twice that of ours so that the madman method is 100% inconsistent.)

Let $pg(x, g)$ be the probability that Alice whose score is x receives grade g. Then

$$\mathrm{GI}(x) = \frac{1}{G-1}\sqrt{\sum_{g=1}^{G} pg(x,g)(g-\bar{g})^2},$$

where

$$\bar{g} = \sum_{g=1}^{G} pg(x,g)g.$$

Let us consider the even distribution grading method now. Let d be the result of the integral division of N by G, $d = \lfloor N/G \rfloor$, and r be the remainder of that division. Suppose that there are k students with scores lower than x (Alice's score). She will receive grade g if

$$(g-1)\cdot d + max(g-G+r-1, 0) \leq k < g\cdot d + max(g-G+r, 0). \quad (3)$$

Assuming (as before) that the scores are distributed by $\mathcal{N}_{[m,1]}(\mu, \sigma^2)$, the probability of a random passing student receiving a score less than x is $\Psi_{m,\mu,\sigma}(x)$, where Ψ is the cumulative distribution function for $\mathcal{N}_{[m,1]}(\mu, \sigma^2)$. Consequently,

$$pg(x,g) = \sum \binom{N-1}{k} \Psi_{m,\mu,\sigma}(x)^k (1-\Psi_{m,\mu,\sigma}(x))^{N-1-k},$$

where the range of k is given by (3).

The above approach to computing GI and MGI of the even distribution method is implemented in an R code presented in Appendix 2.

Our computations (and Fig. 4) show that highest grade inconsistency happens for $x = 0.76$ which is the mean of $\mathcal{N}_{[0.55,1]}(0.75, 0.14^2)$, cf. Sec. 3. For $N = 10$ it is 0.16 which means that the standard deviation of a Alice's grade with the score 0.76 on $G = 10$ scale is $0.16 \cdot 9 = 1.44$. Approximating the probability distribution of this grade by normal distribution, that means that the probability of Alice's grade being between 3 and 6 (roughly the one standard deviation interval $(4.5 - 1.44, 4.5 + 1.44)$) is roughly only 70% The 30% probability that it may be outside that range is not very reassuring to Alice!

On the other hand $\mathrm{GI}(1) = 0$ since score 1 will be assigned grade G with probability 1, i.e. with zero inconsistency. $\mathrm{GI}(0.55) = 0$ for an analogous reason.

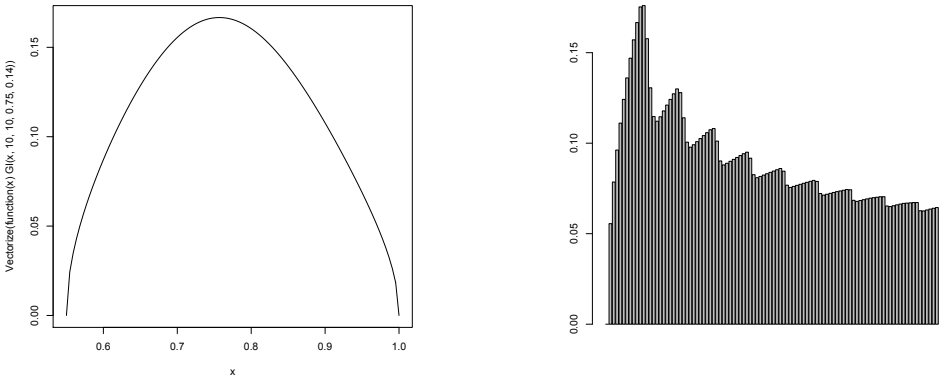

FIGURE 4. Left: Plot of GI(x) function for even distribution grading method, for $N = G = 10$, $\mu = 0.75$, $\sigma = 0.14$.

FIGURE 5. Plot of MGI values for $G = 10$ and $N = 1...100$ for the even distribution method.

The inconsistency peaks at $N = 10$. (That should not be surprising since the even distribution grading does not utilize the full grading scale for $N < 10$.) The plot in Fig. 5 illustrates that the grading inconsistency decreases with an increasing N.

Repetition?

Fig. 5 shows the dependence of MGI on N. The values for $N < 10$ are artificially low, since AGI is normalized under assumption that the full G grade scale is used, which is not the case for the even distribution grading method for $N < G$. For example, a student in a class of one will be assigned an A for certain and, thus, with zero inconsistency. We see however that AGI is decreasing for $N > G$ as expected.

For a class of 100 students, the standard deviation of Alice's grade (assuming again the score of 0.76) reduces to $\mathrm{GI}_{100}(0.76) \cdot (G-1) = 0.064 \cdot 9 = 0.58$, which is almost a three-fold decrease.

Benefits of cohorts?

However, the benefits of large student cohorts become more incremental as N increases – $\mathrm{GI}_{100}(0.76) = 0.064$ while $\mathrm{GI}_{1000}(0.76) = 0.049$ a 24% reduction only!

Let us analyze now the inconsistency of other grading methods of this paper. Unlike the even distribution grading method above, there are no formulas for GI(x) for them. Therefore, we resolve to computing them through

Cite

a Monte-Carlo method, cf. R code in Appendix 3.

The results of the computations are summarized in Tables 11 and 11. Overall, the quantile grading method appears the most consistent of the methods, with even distribution coming close second.

TABLE 2. MGI percentages

Method	Number of students											
	1	2	3	4	5	6	7	8	9	10	11	12
EvenDist	0	6%	8%	10%	11%	12%	14%	15%	16%	17%	18%	18%
Quantile	34%	34%	30%	26%	23%	20%	19%	18%	17%	16%	15%	14%
K-means											16%	16%

TABLE 3. MGI percentages

Method	Number of students									
	13	14	15	20	30	50	100	200	500	1000
EvenDist	16%	13%	11%	12%	10%	8%	6%	5%	4%	3%
Quantile	14%	13%	12%	11%	9%	7%	5%	5%	5%	5%
K-means	16%	16%	15%	15%	14%	13%	12%	10%	8%	7%

It is somewhat surprising to see that the K-means method is a little less consistent than the other two, since it may appear the most consistent from the teacher's point of view. Indeed, many teachers look for natural gaps in the scores to divide them into different grade groups. That leads to fewer borderline cases and, hence, "consistency" from the teacher's point of view. But our computations show that this method is little less fair to students then the other ones!

12. **Improving grading consistency for small classes.** Table 11 and the discussion of the previous section show that the grading consistency is particularly challenging for classes of small size. This issue can be addressed by grading on the basis of comparison of students with a larger cohort of students which we will call a comparison cohort. Consider the following example: Suppose a teacher is to grade students in Math141 (Calculus I) at University at Buffalo. We will refer to them as the focal cohort. The student score mean varies from section to section of Math141; there are several of them each semester. Assuming that all current and recent sections have comparable scoring system, it is reasonable to grade the focal cohort students by comparing them not just to their section peers but to all students in concurrent and recent calculus sections. These students form the comparison cohort.

Since the score mean in the focal section is more prone to random fluctuation than the mean in the much larger comparison cohort, it is more fair to assign grades on the comparison basis with the entire comparison cohort.

If the comparison cohort mean, μ_c, and standard deviation, σ_c are known, then the quantile grading method thresholds can be easily computed in R to reflect these numbers:

```
qtruncnorm((g-1)/G, a=0.55, b=1, mean=mu_c, sd=sigma_c)
```

We call it the quantile grading method with a calibration by comparison cohort.

Obviously, μ_c and σ_c may shift the mean of the focal cohort away from $(G+1)/2$.

Assume $\sigma_c = \sigma_f$. Explain why.

Theorem 14. *If the focal cohort has its scores mean μ_f and standard deviation σ_f then the resulted grades will have the mean*

$$\overline{g} \simeq (G+1)/2 - \alpha(\mu_f, \sigma_f, G) \cdot \frac{\mu_c - \mu_f}{\sigma_f}, \tag{4}$$

where $\alpha(\mu_f, \sigma_f, G) = \frac{1}{\sqrt{2\pi}} \sum_{g=1}^{G-1} e^{-q_g^2/2}$, CORRECT THAT and $q_1, ..., q_{G-1}$ are the G-quantiles for $\mathcal{N}_{[m,1]}(\mu_f, \sigma_f)$.

Proof. PROOF NEEDS EDITING. Denote by p_g the percentage of students in the focal cohort who received the grade $\leq g$, for some $g = 1, ..., G$ by the quantile grading method with the comparison cohort $\mathcal{N}_{[m,1]}(\mu_c, \sigma_c)$. Since the probability of a student receiving grade g is $p_{g+1} - p_g$ and $p_0 = 0$, $p_G = 1$,

$$\overline{g} = \sum_{g=1}^{G} (p_g - p_{g-1}) \cdot g = G - \sum_{g=1}^{G-1} p_g.$$

p_g represents the fraction of the students with scores below the g-th quantile of $\mathcal{N}_{[m,1]}(\mu_c, \sigma_c)$, which is $q_g \sigma_c + \mu_c$. For the purpose of the computation of the average mean, one can take the average over multiple focal cohorts, which boils down to taking a single large focal cohort. For such cohort, one can assume a normal distribution of scores and, hence,

$$p_g = \Psi\left(\frac{q_g \sigma_c + \mu_c - \mu_f}{\sigma_f}\right) = \Psi\left(\frac{\sigma_c}{\sigma_f} q_g + d\right),$$

where d and Ψ are as in the statement.

$d = \frac{\mu_c - \mu_f}{\sigma_f}$ $G - 1 - \sum_{g=1}^{G-1} \Psi\left(c \cdot q_g + d\right)$, $c = \frac{\sigma_c}{\sigma_f}$, Ψ is the cumulative density function of $\mathcal{N}(0,1)$.

For c near 1 and d near 0, the above formula can be approximated by

$$\overline{M}(c, d, G) \simeq \overline{M}(1, 0, G) + \frac{\partial \overline{M}}{\partial c}(1, 0, G) \cdot (c-1) + \frac{\partial \overline{M}}{\partial d}(1, 0, G) \cdot d =$$

$$\frac{G-1}{2} - \sum_{g=1}^{G-1} \Psi'(q_g)(q_g \cdot (c-1) + d).$$

Since $\Psi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is an even function and $q_g = -q_{G-g}$,

$$\sum_{g=1}^{G-1} \Psi'(q_g) q_g = 0$$

and the desired equality follows. □

$\alpha(\mu_f, \sigma_f, G)$ is given by the R function:

```
 CORRECT THAT
Mbar <- function(c,d,G)
  {G-1-sum(pnorm(sapply(c(1:(G-1)),
                   function(i) c*qnorm(i/G)+d)))}
betaG <- function(G)
 {s<-0; for(g in 1:(G-1)) s<-s+exp(-qnorm(g/G)^2/2); s/sqrt(2*pi)}
```

For example, $\alpha(0.75, 0.14, 10) = 2.77$???

When a grade calibration is desired and the mean and the standard deviations of the comparison cohort are unknown, a teacher may use the above formula to compute...

However, it may be more practical to choose appropriate μ_c, σ_c to fit a desired grade average $\overline{M}$.

When a particular grade average is desired, formula (4) proves particularly useful as it implies that $\sigma_c = \sigma_f$ and

$$\mu_c = \mu_f - (\overline{M} - (G-1)/2) \cdot \sigma_f/\beta_G \tag{5}$$

will result in the focal cohort grade average near $\overline{M}$.

The thresholds for D+,C-,...,A-,A in the above grading method will be given by the following 10-quantiles

```
quant<-qnorm(c(1:(G-1))/G, mean=mufc, sd=sigmafc)
```

where mufc and sigmafc stand for μ_c and σ_c. We call it the
even distribution grading method with pre-determined grade average.

Example 15. *Suppose that the focal cohort has $\mu_f = 0.75$ and $\sigma_f = 0.13$. Suppose that $G = 10$ and we intend to grade students in that cohort with the grade average $M = 6$ (corresponding to B). By (5), we should choose*

$$\sigma_c = 0.13 \text{ and } \mu_c = 0.75 - 1.5 \cdot 0.13/2.77 = 0.68.$$

Since $d = (0.68 - 0.75)/0.13 = -0.54$, formula (??) CORRECT THAT computed by

```
Mbar(1,-0.54,10)
```

yields $\overline{M} = 5.96$, which indeed is very close to the desired value 6. Consequently, the thresholds for D+,C-,...,A-,A in the above grading method will be given by the following 10-quantiles

```
quant<-qnorm(c(1:9)/10, mean=0.68, sd=0.13)
```

Obviously, this choice of μ_c does not guarantee the grade average 5.96 in every actual cohort, since that average will depend on the specific distribution of scores which will never be perfectly normal. However, $\overline{M} = 5.96$ will be the average grade average of grades for cohorts with the above μ_f and σ_f.

To see the a concrete application of this discussion, we have simulated numeric scores a cohort of 1000 students with normal distribution with $\mu_f = 0.75$, $\sigma_f = 0.13$ by

```
v <- sort(rnorm(1000,mean=0.75,sd=0.13))
```

We can compute the frequencies of different grades in this cohort by the following:

```
frequencies<-function(s) {
  j<-1; c<-1; p<-rep(0,10)
  for(i in 1:9) {
    while(j< length(s) & s[j]<quant[i]) j<- j+1
    p[i]<- j-c;  c<- j
  if(j>=length(s)) {p[i]<-p[i]+1; break}
  }
p[10]<-length(s)-c+1; p/length(s)
}
```

Now,

```
grades<-freq(v); sum(grades*c(0:9))
```

yield the grade distribution in our simulated cohort with the grade average of 6.05.

We repeated these calculations for a simulated cohort of 10,000 students. This time we obtained grade average 5.96*, exactly as predicted by formula (**??**). The actual grade distribution was*

3.5 5.1 6.0 7.0 8.1 9.1 10.1 12.5 15.4 23

(meaning, 23% of As).

13. **Maximal entropy grading with a pre-determined grade average.** There may be times that the teacher doesn't know μ_c, but wants to assign grades with a pre-determined grade average. Also, the teacher may wish to assign grades with an average other than G+1, following what is customary in the other similar courses, as discussed above.

Sometimes there may be reasons for a teacher wanting to assign grades with a pre-determined grade average. For example, grade averages are usually B (or higher) in graduate school courses. As we have observed earlier, having that high of an average is not efficient from the point of view of informative value of grades. However, addressing this problem requires an across the board systemic change in grading customs. An individual teacher enacting an efficient grading system on her own with a lower average than customary will disappoint students and potentially mislead school administrators.

This leads to the question of finding the most informative grading method with a given pre-determined grade average. We turn again to the notion of Shannon entropy for answers.

Theorem 16. *Among random variables X with values in $\{1, ..., G\}$, with given, fixed, mean $\bar{X} = M$, $1 < M < G$, the one with the highest Shannon entropy has distribution $P(X = g) = \frac{1-r}{1-r^G} \cdot r^{g-1}$, for $g \in \{1, ..., G\}$, where r is a unique positive number such that*

Chg $G-1$ to G

$$\frac{1-r}{1-r^G}\sum_{g=1}^{G} g\cdot r^{g-1} = M. \tag{6}$$

Proof. The statement follows from Theorem 5.11 in [Co]. Indeed, Conrad's n is our G and his $E_1, ..., E_n$ are $1, ..., G$ respectively. Denote Conrad's $e^{-\beta}$ by r. Then $\sum_{i=1}^{G} e^{-\beta E_i} = r\frac{1-r^G}{1-r}$ and Conrad's theorem implies the existence of a unique value r for which the probability distribution $p_g = \frac{1-r}{1-r^G}\cdot r^{g-1}$, for $g = 1, ..., G$, has mean M.

For completeness, we outline a self-contained proof of Theorem 16 in the appendix. □

We call a grading method with the above frequencies of grades the exponential grading method.

Note that for $r = 1$ the exponential grading method yields an even distribution.

The grade frequencies $p_1, ..., p_G$ for the exponential grading method can be computed by the following R function:

```
expfrequencies <- function(M,G) {
 s<-function(r) sum(sapply(c(0:(G-1)), function(i) r^i))
 f<- function(r) {y<-0; for(g in 1:G) y<- y+g*r^(g-1); y-s(r)*M}
 r<-uniroot(f, interval=c(1,1000))[[1]]
 sapply(c(1:G), function(g) r^(g-1)/s(r))}
```

Example 17. *For $G = 10$ and $M = 6$ (corresponding to grade B),*

```
expfrequencies(6,10)
```

returns following grade distribution: 20.5% A, 16.9% A-, 14.0% B+, 11.6% B, 9.5% B-, 7.8% C+, 6.5%C, 5.3% C-, 4.4% D+, and 3.6% D for the exponential grading method. ($r = 1.21$ in this example.)

QUESTION: Does quantile method with a shifted score mean have a maximal entropy with a given mean, i.e. exponential grade distribution?

Appendix 1

14. **Proof of Thm. 9:** For a given $c > 0$ consider the function $f_c(x) = x log_2 x + (c-x) log_2 (c-x)$, for $0 \le x \le c$. Since

$$f_c'(x) = log_2 x - log_2 (c-x) = log_2 \left(\frac{c}{c-x} - 1\right),$$

we conclude that $f_c(x)$ is decreasing for $x < c/2$ and increasing for $x > c/2$.

Suppose that FINISH. To prove the statement of the theorem it is enough to show that any distribution with $p_i > p_j$ for some i, j, has a lower entropy than an identical distribution with n_i reduced by 1 and n_j increased by 1. Note that the difference of entropies of these two distributions is

$$-\left(\frac{n_i-1}{N} log_2 \frac{n_i-1}{N} + \frac{m_j+1}{N} log_2 \frac{m_j+1}{N}\right) + \left(\frac{n_i}{N} log_2 \frac{n_i}{N} + \frac{m_j}{N} log_2 \frac{m_j}{N}\right) =$$

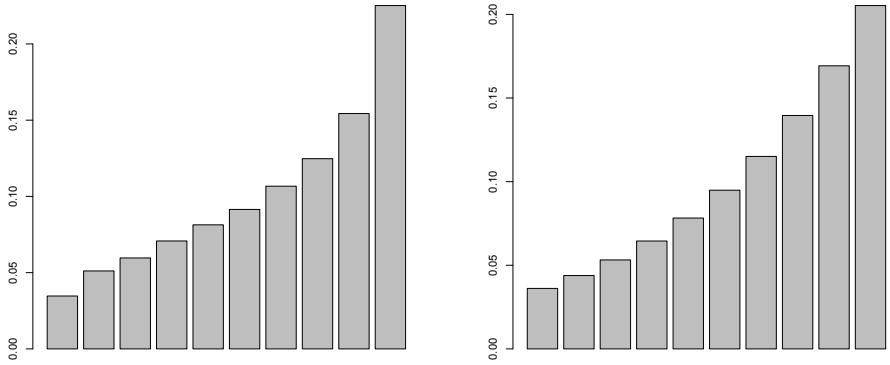

FIGURE 6. Left: Grade frequencies in our computer simulation for the even distribution grading method with predetermined mean $\overline{M} = 6$. Right: Grade frequencies for the exponential grading method $\overline{M} = 6$ WHAT IS THAT?

$$-f_c\left(\frac{n_i - 1}{N}\right) + f_c\left(\frac{n_i}{N}\right)$$

for $c = \frac{n_i + m_j}{N}$. Since $\frac{n_i - 1}{N}, \frac{n_i}{N} > c/2$ this expression is positive by the discussion above. □

15. **Proof of Thm. 16:** For completeness, we enclose the main idea of the proof of Theorem 16, without relaying on the theorem of Conrad. Specifically, we will prove the above result under the assumption that $p_0, ..., p_{G-1} \in (0, 1)$ for the distribution of X maximizing Shannon entropy with $\bar{X} = M$. (The possibility that some of the p_i are zero or one, are disproved in [Co].)

We need to maximize entropy $E = -\sum_{i=0}^{G-1} p_i \cdot log_2\, p_i$, where $p_0, ..., p_{G-1} \in [0, 1]$ are subject to conditions $\sum_{i=0}^{G-1} p_i = 1$ and $\sum_{i=0}^{G-1} p_i \cdot i = M$. By the method of Lagrange multipliers, the G-tuple $(p_0, ..., p_{G-1})$ maximizing E as above is a critical point of

$$L = \sum_{i=0}^{G-1} p_i \cdot log_2\, p_i - \alpha\left(\sum_{i=0}^{G-1} p_i - 1\right) - \beta\left(\sum_{i=0}^{G-1} p_i \cdot i - M\right),$$

where α, β are formal variables. The critical points of L satisfy the equations:

$$\partial L/\partial p_i = 0, \text{ for } i = 0, ..., G-1, \text{ and } \partial L/\partial\alpha = \partial L/\partial\beta = 0$$

which lead to

$$log_2\, p_i + \frac{1}{ln\, 2} - \alpha - \beta \cdot i = 0, \text{ for } i = 0, ..., G-1,$$

and

$$\sum_{i=0}^{G-1} p_i = 1 \text{ and } \sum_{i=0}^{G-1} p_i \cdot i = M.$$

Let $c = 2^{\alpha - \frac{1}{ln\,2}}$ and $r = 2^{\beta}$. Then

$$p_i = 2^{\alpha+\beta i - \frac{1}{ln\,2}} = c \cdot r^i,$$

and

$$\sum_{i=0}^{G-1} c \cdot r^i = 1, \quad \sum_{i=0}^{G-1} c \cdot r^{i-1} i = M. \tag{7}$$

implying $c = \frac{1-r^G}{1-r}$ and equation (6). (The uniqueness of r satisfying the second equation of (7) is proved in [Co].) □

Appendix 2: R code for the even distribution method

```
 CHANGE NOTATION. NO ‘‘AGI’’
m <- 0.55; mu <- 0.75; s <- 0.14
pg <- function(x,g,N,G,mu,s) {
 d<-floor(N/G); r<-N-G*d;
 if (g<=G-N) 0 else
  sum(dbinom(((g-1)*d+max(g-G+r-1,0)):(g*d+max(g-G+r,0)-1),
    N-1,ptruncnorm(x, a=m, b=1, mean=mu, sd=s)))
}

gbar <- function(x,N,G,mu,s) {sum(sapply(c(1:G), function(g)
pg(x,g,N,G,mu,s))*c(1:G))}

GI<-function(x,N,G,mu,s) {
mi <-gbar(x,N,G,mu,s);
sqrt(sum(sapply(c(1: G), function(g) pg(x,g,N,G,mu,s)*(g-mi)^2)))/(G-1)
}

AGI<-function(N,G,mu,s) {integrate(Vectorize(function(x)
 GMIF(x,N,G,mu,s)*dtruncnorm(x,a=m,b=1,mean=mu,sd=s),m,1)[[1]]}
```

Appendix 3: R code for the XXX

```
m = 0.55; mu = 0.75; s = 0.14
#Quantile grading method grade between 0 and G -- CHANGE TO 1-G.
#for score x assuming that peers have scores sc.
grade <- function(sc,x,G){
mv <- tryCatch({ml<-mle.tmvnorm(cbind(append(sc,x)),
lower=m,upper=1, lower.bounds=c(m,0),
upper.bounds=c(1,Inf), method="L-BFGS-B");
c(coef(ml)[[1]],coef(ml)[[2]])}, error=function(cond)
```

```
{c(mean(sc),sd(sc)^2)});
min(floor(ptruncnorm(x, a=m, b=1, mean=mv[1],
sd=sqrt(mv[2]))*G),G-1)
}

#K-means method grade between 0 and G
#for score x assuming that peers have scores sc.
grade<-function(sc,x,G){
km<-kmeans(append(sc,x),G);
rank(km[[2]])[km[[1]][length(sc)+1]]-1
}

#Frequencies of grades in a vector gvec.
freq <- function(gvec,G) {
num <- rep(0,G); for(j in 1:length(gvec))
num[gvec[j]+1]<-num[gvec[j]+1]+1;
num/length(gvec)}

#Computes the frequencies of grades assigned
#r=number of Monte-Carlo tries.
pgdistx <- function(x,N,G,r){
 freq(sapply(c(1:r), function(i) grade(rtruncnorm(
          N-1, a=m, b=1, mean=mu, sd=s),x,G)),G)
}

#Computes GMI(x)
GMIF <- function(x,N,G,r) {
pgd<-pgdistx(x,N,G,r); me<-sum(pgd*c(0:(G-1)))
t<-sum(pgd*(c(0:(G-1))-me)^2)
sqrt(t)/(G-1)}

plot(Vectorize(function(x) GMIF(x,10,10,1000)),m,1)

GMI<-function(N,G,m){
   sum(GMIF(c(1:9)/10,N,G,r)*rho(x))/10
}
```

References

[Ba] F.B. Baker, The Basics Of Item Response Theory, ERIC Clearinghouse on Assessment and Evaluation, 2001.

[Co] K. Conrad, Probability Distributions And Maximum Entropy, http://www.math.uconn.edu/~kconrad/blurbs/analysis/entropypost.pdf

[Fe] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed. New York: Wiley, 1968.

[FHPR] Michael T. French, Jenny F. Homer, Ioana Popovici, and Philip K. Robins, What You Do in High School Matters: High School GPA, Educational Attainment, and

Labor Market Earnings as a Young Adult, *Eastern Economic Journal* **41**(2015), 370–386.

[Gl] R. Glaser, Instructional technology and the measurement of learning outcomes, *American Psychologist* **18** (1963), 510–522.

[HF] William C. Hiss, Valerie W. Franks, Defining Promise: Optional Standardized Testing Policies In American College And University Admissions.

[KR] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, An Introduction to Cluster Analysis, 1990, Wiley, New York.

[MJ] N.F.G. Martin, W. James, Mathematical Theory of Entropy. Cambridge University Press, 2011.

[Oe] P. Oehrlein, Determining Future Success of College Students, Undergraduate Economic Review, Vol. 5(1), 2009.

[R] The R Project for Statistical Computing, `http://www.r-project.org`

[Si] A.S. Sikora, Mathematical theory of incentives, in preparation.

244 Math Bldg, University at Buffalo, SUNY, Buffalo, NY 14260
E-mail address: `asikora@buffalo.edu`