

Towards Interval Techniques for Processing Educational Data

Olga Kosheleva^{1,2}, Vladik Kreinovich^{1,3}, Luc Longpré^{1,3},
Mourat Tchoshanov^{1,2,4}, and Gang Xiang^{1,3}

¹NASA Pan-American Center for Earth
and Environmental Studies

²Department of Teacher Education

³Department of Computer Science

⁴Department of Mathematical Sciences

University of Texas at El Paso

El Paso TX 79968, USA

contact olgak@utep.edu, vladik@utep.edu

Abstract

There are many papers that experimentally compare effectiveness of different teaching techniques. Most of these papers use traditional statistical approach to process the experimental results. The traditional statistical approach is well suited to numerical data but often, what we are processing is intervals (e.g., A means anything from 90 to 100). We show that the use of interval techniques leads to more adequate processing of educational data.

1. Formulation of the Problem

Practical problem: comparing teaching techniques.

Teaching is very important, and teaching is not always very effective. There exist many different pedagogical techniques that help teach better, and new teaching techniques are being invented all the time.

To select the techniques which are the most efficient for a given educational environment, we must experimentally compare effectiveness of different teaching techniques in this and similar environments.

Setting up such an experiment in a meaningful way is a very difficult task. One needs to make sure that the students assigned to two different methods represent the same population, that the topics selected for the course are the same for both methods – otherwise, we will not get a fair and convincing comparison. In statistics, and especially in applications of statistics to social phenomena like teaching, there is a vast literature on experiment design.

In this paper, we will assume that the experiment has already been designed in a proper way, so that the comparison is reasonably fair, and we will concentrate on the next problem: how can we process the results of this experiment? based on these results, what conclusions can we make about the compared techniques?

A natural way to compare teaching techniques: compare the grades.

A natural way to measure the effectiveness of a technique on an individual student is by recording the grade that this student received when being taught by this particular technique. To describe which method is better in general, it is therefore reasonable to describe the distribution of the grades for students taught under different techniques.

To get a complete picture, it is desirable to know the full grade distributions; however, in practice, researchers usually compute the mean and standard deviation.

In some cases, we have meaningful system of numerical grades.

In some situations, there is a well-developed standardized test which provides a reasonably objective numerical measure of the student knowledge. For example, in the USA, there is the SAT test which is used to gauge the student's degree of preparation for undergraduate studies, the GRE test which is used to gauge the student's preparedness for graduate studies, the TOEFL test which gauges the degree to which non-native speakers know English, etc. Such tests provide a numerical grade from, say 0 to 800, and the difference between 640 and 650 is indeed meaningful.

The development of such a standardized test is a very difficult task. Moreover, once the test is given, and its answers are widely known, it is not possible to re-use it; so,

next time this test needs to be given, a new version of this test has to be designed. As a result, situations in which such standardized tests exist are extremely rare.

In most pedagogical situations, there is only a small number of possible grades. Typically, instructors assign grades from a small discrete set of possible grades. For example, in the US system, typical grades are A (excellent), B (good), C (satisfactory), D (sometimes used as passing grade), and F (fail); they are called *letter grades*.

In another system with which several of us are very familiar, namely, the Russian system, typical grades are 5 (excellent), 4 (good), 3 (satisfactory), 2 (bad), and 1 (sometimes used for awful). These grades are not described by letters; however, they serve the exact same purpose as the US letter grades. So, to enhance the useful distinction between meaningful numerical grades (like SAT scores) and grades from a small set, we will follow the US tradition and call grades from a small set *letter grades*.

Both in the US and in the Russian grading systems, there are sometimes refined versions of these systems with the possibility of adding + or -: e.g., 4+ is somewhat better than 4, A- is somewhat worse than A.

Letter grades often comes from points. One of the main methods of assigning these letter grades is based on so-called *points*. Specifically, an instructor assigns points for different assignments and tests. These points add up during the course. At the end, the total amount of points determines the letter grade. There are many ways to translate from points to resulting grades. In the US, a typical translation is as follows:

- 90 points or above means A;
- at least 80 but less than 90 means B;
- at least 70 but less than 80 means C;
- at least 60 but less than 70 means D;
- less than 60 means F.

Why not keep the original points? The amount of points gained provides a much more refined description of the student knowledge than the resulting letter grade. So, a natural question is: why not simply use the original number of points instead of the letter grades?

The answer to this question is actually contained in the above explanation of why we cannot use too refined a scale: the main reason is the need to make the grades more objective. When several instructors teach different sections of a class, they spend a lot of effort trying to make sure that they have the same criteria for A, B, and C. When a department is accredited, accreditors look at samples of A, B, and C

papers to make sure that these sample would indeed get the corresponding letter grade at other schools as well.

This unification of letter grades is very difficult, so difficult that beyond letter grades, there is definitely no uniformity. A student who get a solid A (say, 95) with one of the instructors will most probably still get an A with others, but that A may vary from 91 to 99, depending on the instructor.

Because of this well-understood subjectivity, the points are usually not archived and not used to compare students taught by different instructors. Only letter grades are recorded.

Alternative approach: grades based on perceptions. In some disciplines, it is easy to reasonably objectively assign points to individual assignments: e.g., a well-formulated mathematical problem is either solved or not, its solution is either correct or not, and if there is a partial solution, it is possible to come up with an agreed-upon objective scale of points.

In other disciplines, however, e.g., in writing essays about a difficult philosophical concept, it is difficult to assign points. In such disciplines, it makes sense to make judgements like “understands well”, “understands the main concept reasonably well, but still has misconceptions”, etc. These judgments are then translated into points or directly into letter grades.

Again, instructors try their best to make these judgments objective. This unification is very difficult, so difficult that there is no way to extend this unification beyond the small scale of letter grades, into something more refined.

Statistical processing of letter grades: a problem. Let us get back to our problem. We want to compare two (or more) teaching techniques. So, we apply two different techniques to two groups of students, and we compare the results.

Because of the above explanations, at the end of the experiment, we only have letter grades: we have letter grades of students who were taught according to the first technique, and we have letter grades of students taught according to the second technique. To compare the techniques, we must perform a statistical analysis of these letter grades.

Statistical processing of letter grades: traditional approach. The traditional statistical approach to statistical processing of the grades is motivated by the existing ways of processing these data.

For example, students who study well receive different honors. These honors are usually based on simply averaging the grades: the higher the average, the higher the honors. In the Russian system, “letter grades” are actually numbers so they can be directly averaged. In the US system, there is a standard translation of letter grades into numbers: A means 4, B means 3, C means 2, D means 1, and F means 0. After this translation, letter grades become numbers, so we can easily compute their average.

Similarly, to gauge the degree to which a class learns a material, we can compute the average grade of this class.

In accordance with this idea, the vast majority of these papers that experimentally compare different teaching techniques by treating them as numbers:

- first, we translate the grades into numbers; and
- then, we process these numbers by using traditional statistical techniques; see, e.g., [18, 21].

Problems with the traditional approach: general description. The traditional statistical approach is well suited for processing numerical data. However, in processing educational data, often what we are processing is:

- either *intervals*: e.g.
 - the A grade usually means anything from 90 to 100,
 - the B grade means anything between 80 and 90, and
 - the C grade mean anything between 70 and 80;
- or *fuzzy*-type perceptions, words from the natural language like “understood well” or “understood reasonably well”.

Problems with the traditional approach: example. In selecting a teaching method, it is important not only to make sure that the *average* results m are good – e.g., that the average grade on a standard test is good – but also to ensure that the results are *consistently* good – i.e., in statistical terms, that the standard deviation σ of the grade is low.

If the standard deviation σ is high, that would mean while some students learn really well under this technique, there are many others who are left behind, and we cannot afford that.

So, to compare several teaching techniques based on the grades the student got, we must compare not only their averages, but also the standard deviations.

The following simple example will show that when we replace an interval with a single value, we lose important information that could influence the computation of the standard deviation, and we could get erroneous results.

Suppose that in one method, all the students got Bs, while in the other method, half of the students got Bs and half of the students got As. Which of the two methods shows more stable results, with a smaller standard deviation?

In the traditional statistical approach, we interpret A as 4 and B as 3.

- In the first method, the resulting grades are $x_1 = \dots = x_n = 3$, so the average grade is equal to

$$m = \frac{x_1 + \dots + x_n}{n} = 3,$$

the population variance is equal to

$$V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 0,$$

and the standard deviation is equal to $\sigma = \sqrt{V} = 0$;

- In the second method, the average is equal to

$$m = \frac{3 + 4}{2} = 3.5,$$

so for each i , $(x_i - m)^2 = 0.25$, hence the population variance is equal to

$$V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 0.25,$$

and the standard deviation is equal to $\sigma = \sqrt{V} = 0.5$.

So, if we use the traditional statistical approach, we conclude that while the second method has a higher average, it is less stable than the first one.

In reality, if we go back from the “interval” letter grades like A, B, and C to the original numbers of points, it may turn out the second method is not only better on average, but also much more stable. Indeed, suppose that:

- in the first method, half of the students got 80 points, and half got 88 points; and
- in the second method, half of the students got 89 points, and half of the student got 91 points.

In terms of As and Bs, this is exactly the situation as described above. However, when we compute the standard deviation for these numbers of points, we get a different result than when we process the letter grades:

- In the first method, the average is equal to

$$m = \frac{80 + 88}{2} = 84,$$

so for each i , $(x_i - m)^2 = 16$, hence the variance is equal to

$$V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 16.$$

- In the second method, the average is equal to

$$m = \frac{89 + 91}{2} = 90,$$

so for each i , $(x_i - m)^2 = 1$, hence the variance is equal to

$$V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 1 \ll 16.$$

What needs to be done. It is desirable to develop techniques for processing educational data that would take into account that the grades are not exactly equal to the corresponding numerical values but may differ from these values.

In other words, we need techniques that would provide guaranteed answers to questions like: Is the first method better than the second one?

It is OK to have an answer “we do not know”, but if the answer is “yes”, we want to be sure that no matter what additional information we learn about these experiments the answer will remain the same.

Such techniques are outlined in this paper.

2. Interval Approach

Processing interval data: analysis of the situation. The main reason why we had the above problem is that letter grade ℓ represents not a *single* value of the number grade x , but rather an *interval* $\mathbf{x} = [\underline{x}, \bar{x}]$ of possible values of the numbers of points. For example:

- the letter grade A represents the interval [90, 100];
- the letter grade B represents the interval [80, 90];
- the letter grade C represents the interval [70, 80].

So, for the educational data, instead of the exact value x of each number of points, we often only know the intervals $[\underline{x}, \bar{x}]$ corresponding to the letter grade ℓ . This is true for the American system, this is true for the Russian system, this is true for any grading system in which the number of points is used to describe the resulting “letter grade”.

Processing interval data: formulation of the problem. Our objective is, given a set of letter grades ℓ_1, \dots, ℓ_n , to compute a certain statistical characteristic C such as average, standard deviation, correlation with other characteristics (such as the family income or the amount of time that a student spends on homeworks), etc.

The desired statistical characteristic is defined in terms of numerical values, as $C = C(x_1, \dots, x_n)$. For example,

the average is defined as $m = \frac{x_1 + \dots + x_n}{n}$, the variance is defined as $V = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2$, etc.

For the educational data, instead of the *exact* values x_i , we often only know the *intervals* \mathbf{x}_i corresponding to the letter grade ℓ_i . For different possible values $x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n$, we get different values of the corresponding characteristic C .

Our objective is to find the range of possible values of the desired characteristic when $x_i \in \mathbf{x}_i$, i.e., the interval

$$\mathbf{C} = \{C(x_1, \dots, x_n) \mid x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n\}.$$

This problem is a particular case of the general problem of interval computations. The need to perform computations under interval uncertainty occurs in many areas of science and engineering. In many such areas, we therefore face the following problem:

- we know:
 - n intervals $\mathbf{x}_1, \dots, \mathbf{x}_n$ and
 - an algorithm $y = f(x_1, \dots, x_n)$ that transforms n real numbers (inputs) into a single number y (result of data processing);
- we must estimate the range of possible values of y , i.e., the interval

$$\mathbf{y} = \{f(x_1, \dots, x_n) \mid x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n\}.$$

This problem is called the main problem of *interval computations*; see, e.g., [5, 6, 7, 10, 13].

We can therefore conclude that the problem of processing educational data under interval uncertainty is a particular case of the more general problem of interval computations.

How we can process interval data: general description. Many efficient techniques have been developed to solve generic interval computations problems; see, e.g., [5, 6, 7, 10, 13].

How we can process interval data: case of statistical characteristics. In particular, several algorithms have been developed for the case when the the function $f(x_1, \dots, x_n)$ is one of the standard statistical characteristics such as average m or standard deviation V ; see, e.g. [4, 11, 12] and references therein.

Computing average under interval uncertainty. In particular, since the average is a monotonic function of each of its variables, its value is the largest when each x_i attains the

largest possible value $x_i = \bar{x}_i$, and its value is the smallest when the variance attains its smallest possible value \underline{x}_j . Thus, for the average m , the interval takes the form $[\underline{m}, \bar{m}]$, where

$$\underline{m} = \frac{\underline{x}_1 + \dots + \underline{x}_n}{n}; \quad \bar{m} = \frac{\bar{x}_1 + \dots + \bar{x}_n}{n}.$$

If all the letter grades are A, B, C, or D, then the width $\bar{x}_i - \underline{x}_i$ of each corresponding interval is 10, so $\bar{m} = \underline{m} + 10$. In this situation, it is sufficient to compute *one* of the bounds \bar{m} or \underline{m} , the other bound can be easily reconstructed from this one.

If one of the grades is a F grade, for which the interval of possible values is $[0, 60]$ with a width $60 > 10$, then we must compute *both* bounds.

Computing variance under interval uncertainty. For the variance V , there exist efficient algorithms for computing the lower bound \underline{V} , but the problem of computing the upper bound \bar{V} is, in general, NP-hard. However, for educational data, the intervals either coincide or intersect at a single point; as a result, none of the intervals is a subset of the interior of any other: $[\underline{x}_i, \bar{x}_i] \not\subseteq (\underline{x}_j, \bar{x}_j)$. For the case when the intervals satisfy this *subset property*, there exists an efficient algorithm for computing \bar{V} ; see, e.g., [12].

Specifically, to compute \bar{V} for intervals which satisfy the subset property, we first sort the intervals in lexicographic order

$$[\underline{x}_i, \bar{x}_i] \preceq [\underline{x}_j, \bar{x}_j] \leftrightarrow \underline{x}_i < \underline{x}_j \vee (\underline{x}_i = \underline{x}_j \ \& \ \bar{x}_i \leq \bar{x}_j).$$

For the points intervals, this simply means that we sort the letter grades into an increasing sequence. As a result, we get $\underline{x}_1 \leq \underline{x}_2 \leq \dots \leq \underline{x}_n$ and $\bar{x}_1 \leq \bar{x}_2 \leq \dots \leq \bar{x}_n$. For every k from 1 to n , we pick $x_i = \underline{x}_i$ for $i \leq k$ and $x_i = \bar{x}_i$ for $i > k$; then, we compute the average $m = \frac{\underline{x}_1 + \dots + \underline{x}_k + \bar{x}_{k+1} + \dots + \bar{x}_n}{n}$ of the selected x_i , and check whether this average satisfies the inequality $\underline{x}_k \leq m \leq \bar{x}_{k+1}$. If it does, then the population variance of the corresponding sequence x_1, \dots, x_n is exactly the desired upper bound \bar{V} .

According to [4, 12], to compute the lower bound \underline{V} , similarly, for every k , we select:

- $x_i = \bar{x}_i$ when $\bar{x}_i \leq \underline{x}_k$, and
- $x_i = \underline{x}_i$ when $\underline{x}_i \geq \bar{x}_k$.

We then compute the average m of the selected x_i and check whether this average satisfies the inequality $\underline{x}_k \leq m \leq \bar{x}_k$. If it does, then we assign $x_i = m$ for all the un-assigned value i , and the population variance of the corresponding sequence x_1, \dots, x_n is exactly the desired lower bound \underline{V} .

Numerical example: computing \bar{V} . For 3 sorted grades C, B, and A, we get $\underline{x}_1 = 70$, $\bar{x}_1 = 80$, $\underline{x}_2 = 80$, $\bar{x}_2 = 90$,

$\underline{x}_3 = 90$, $\bar{x}_3 = 100$. For this data, let us first compute \bar{V} . For $k = 1$, we pick $x_1 = \underline{x}_1 = 70$, $x_2 = \bar{x}_2 = 90$, and $x_3 = \bar{x}_3 = 100$. Here, $m = (x_1 + x_2 + x_3)/3 = 86\frac{2}{3}$. Since $\underline{x}_1 = 60 \leq m \leq \bar{x}_2 = 90$, the upper bound \bar{V} is equal to the population variance $\frac{1}{n} \cdot \sum (x_i - m)^2$ of the values $x_1 = 70$, $x_2 = 90$, and $x_3 = 100$, hence $\bar{V} = 155\frac{5}{9}$.

Numerical example: computing \underline{V} . For \underline{V} , we also start with $k = 1$. For this k , in accordance with the above algorithm, we assign the values $x_2 = \underline{x}_2 = 80$ and $x_3 = \underline{x}_3 = 90$. Their average $m = 85$ is outside the interval $[\underline{x}_1, \bar{x}_1] = [70, 80]$, so we have to consider the next k .

For $k = 2$, we assign $x_1 = \bar{x}_1 = 80$ and $x_3 = \underline{x}_3 = 90$. The average $m = 85$ of these two values satisfies the inequality $\underline{x}_2 = 80 \leq m \leq \bar{x}_2 = 90$; hence we assign $x_2 = 85$, and compute \underline{V} as the population variance of the values $x_1 = 80$, $x_2 = 85$, and $x_3 = 90$, hence $\underline{V} = 16\frac{2}{3}$.

Computing other statistical characteristics under interval uncertainty. Similar algorithms are known for other statistical characteristic such as median, higher moments, covariance, etc. [4, 11, 12].

3. Fuzzy Approach: In Brief

Formulation of the problem. In the interval approach, each letter grade is characterized by a set (interval) of possible values of points. For example, a letter grade A corresponds to the interval $[90, 100]$, etc.

On average, 90 is a reasonable and convenient threshold, but in reality, this threshold is somewhat arbitrary. Instructors often do not cut off at 90 sharp and assign A to all those who got 90.1 and B to those who got 89.9. In many cases, if there are good students whose grades are almost 90 (say, 89.1), they will get an A grade. In such cases, the threshold is made flexible: the instructor looks for a gap between clearly A and clearly B students (such a gap usually exists), and assigns A to all the students whose grades are higher than this gap and B to those students whose grades are below this gap.

As a result, even if a student has an A, we cannot say with 100% confidence that this student's number of points was above 90. The resulting situation can be described by the technique of fuzzy uncertainty see, e.g., [8, 16]. In this technique, for each number of points x and for each letter grade (e.g., A), we have a degree $\mu_A(x) \in [0, 1]$ with which x corresponds to A.

- When $x \geq 90$, we are absolutely sure that the letter grade is A, so $\mu_A(x) = 1$.

- When $x \leq 87$, we are absolutely sure that the letter grade is not A, so $\mu_A(x) = 0$.
- When $87 < x < 90$, there is a possibility that A was assigned as a letter grade, so we get $\mu_A(x) \in (0, 1)$.

This value $\mu_A(x)$ is called a *membership degree* – the degree to which the value of x points is a member of the (fuzzy) set of all the values which correspond to the A grade.

To find these membership degrees, we can, e.g., use linear interpolation, and define $\mu_A(x)$ on the interval $[87, 90]$ as a linear function which takes the value 0 on $x = 87$ and the value 1 for $x = 90$.

For each membership α , we can determine the set of values x that are possible with at least this degree of possibility – the α -cut $\{x \mid \mu_A(x) \geq \alpha\}$ of the original fuzzy set. Vice versa, if we know α -cuts for every α , then, for each object x , we can determine the membership degree $\mu_A(x)$ with which x belongs to the original fuzzy set as the largest values α for which x belongs to the corresponding α -cut [1, 8, 14, 15, 16].

A fuzzy set can be thus viewed as a nested family of its α -cuts.

How we can process fuzzy data: general idea. If instead of a (crisp) interval \mathbf{x}_i of possible numbers of points, we have a fuzzy set $\mu_i(x)$ of possible grades, where i is A, B, C, D, or F, then we can view this information as a family of nested intervals $\mathbf{x}_i(\alpha)$ – α -cuts of the given fuzzy sets.

Our objective is then to compute the fuzzy number corresponding to this the desired characteristic $C(x_1, \dots, x_n)$.

In this case, for each level α , to compute the α -cut of this fuzzy number, we can apply the interval algorithm to the α -cuts $\mathbf{x}_i(\alpha)$ of the corresponding fuzzy sets. The resulting nested intervals form the desired fuzzy set for C .

How we can process fuzzy data: case of statistical characteristics. For statistical characteristics such as variance, more efficient algorithms are described in [3].

4. Towards Combining Probabilistic, Interval, and Fuzzy Uncertainty

Need for such a combination. In the case of interval uncertainty, we consider all possible values of the grades, and do not make any assumptions about the probability of different values within the corresponding intervals. However, in many cases, we can make commonsense conclusions about the frequency of different grades.

For example, if a student has almost all As but only one B, this means that this is a strong student, and most probably this B is at the high end of the B interval. On the other hand, if a student has almost all Cs but only one B, this means that this is a weak student, and most probably this B is at

the lower end of the B interval. It is desirable to take such arguments into account when processing educational data.

Let us describe how we can do this.

Comment. To avoid misunderstanding, it is worth mentioning that commonsense conclusions are not always possible: in some cases, we can make such commonsense conclusions; in some cases, we cannot. For example, if the first student has one A, two Bs and one C, and the second student has two As and two Cs, it is not clear which of two students is better, so it is difficult to make any conclusions about the quality of these grades.

Simplest case: normally distributed grades. Let us first consider the reasonable case when the actual points are normally distributed, with an (unknown) mean m and an unknown standard deviation σ . In other words, we assume that the cumulative probability distribution (cdf) $F(x) \stackrel{\text{def}}{=} \text{Prob}(\xi < x)$ has the form $F_0\left(\frac{x-m}{\sigma}\right)$, where $F_0(x)$ is the cdf of the standard Gaussian distribution with 0 mean and unit standard deviation. Our objective is to determine the values m and σ .

If we knew the values of the points x_i , then we could apply the above statistics and estimate m and $\sigma = \sqrt{V}$. In many situations, we do not know the values of the *points*, we only know the values of the *letter grades*. How can we then estimate m and σ based on these letter grades?

Case of normally distributed grades: towards an algorithm. Based on the letter grades, we can find, for the threshold values 60, 70, etc., the frequency with which we have the number of points smaller than this threshold. If we denote by f the proportion of F grades, by d the proportion of D grades, etc., then the frequency of $x < 60$ is f , the frequency of $x < 70$ is $f + d$, the frequency of $x < 80$ is $f + d + c$.

It is well known that the probability can be defined as a limit of the corresponding frequency when the sample size n increases. Thus, when the sample size is large enough, we can safely assume that the corresponding frequencies are close to the corresponding probabilities, i.e., to the values $F(x)$. In other words, we conclude that:

$$F_0\left(\frac{60-m}{\sigma}\right) \approx f; \quad F_0\left(\frac{70-m}{\sigma}\right) \approx f+d;$$

$$F_0\left(\frac{80-m}{\sigma}\right) \approx f+d+c; \quad F_0\left(\frac{90-m}{\sigma}\right) \approx f+d+c+b.$$

If we denote by $\psi_0(t)$ the function that is inverse to $F_0(t)$, then, e.g., the first equality takes the form $(60-m)/\sigma \approx \psi_0(f)$, i.e., $\sigma \cdot \psi_0(f) + m \approx 60$. Thus, to find the unknowns m and σ , we get a system of linear equations:

$$\sigma \cdot \psi_0(f) + m \approx 60; \quad \sigma \cdot \psi_0(f+d) + m \approx 70;$$

$\sigma \cdot \psi_0(f+d+c) + m \approx 80$; $\sigma \cdot \psi_0(f+d+c+b) + m \approx 90$, which can be solved, e.g., by using the Least Squares Method.

Comment. In some cases, the distribution is non-Gaussian, and we know its shape, i.e., we know that

$$F(x) = F_0\left(\frac{x-m}{\sigma}\right),$$

where $F_0(t)$ is a known function, and m and σ are unknown parameters. In this case, we can use the same formulas as above.

Simplified case when all the grades are C or above. In many cases, only C and above are acceptable letter grades. In such situations, $f = d = 0$ and $c + b + a = 1$, so we get a simplified system of two linear equations with two unknowns:

$$\sigma \cdot \psi_0(c) + m = 80; \quad \sigma \cdot \psi_0(c+b) + m = 90.$$

Subtracting the first equation from the second one, we conclude that

$$\sigma = \frac{10}{\psi_0(b+c) - \psi_0(c)}.$$

This formula can be further simplified if the distribution $F_0(x)$ is symmetric (e.g., Gaussian distribution is symmetric), i.e., for every x , the probability $F_0(-x)$ that $\xi \leq -x$ is equal to the probability $1 - F_0(x)$ that $\xi \geq x$. Thus, we can conclude that $\psi_0(1-x) = -\psi_0(x)$ for every x . In particular, since $c + b + a = 1$, we conclude that

$$-\psi_0(c+b) = \psi_0(1-(c+b)) = \psi_0(a).$$

Thus, the formula for σ takes the form:

$$\sigma = -\frac{10}{\psi_0(a) + \psi_0(c)}. \quad (1)$$

Similarly, if we divide the equation $(90-m)/\sigma = \psi_0(b+c)$ by $(80-m)/\sigma = \psi_0(c)$, we conclude that

$$\frac{90-m}{80-m} = \frac{\psi_0(b+c)}{\psi_0(c)} = -\frac{\psi_0(a)}{\psi_0(c)},$$

hence

$$m = 80 + \frac{10}{1 + \frac{\psi_0(a)}{\psi_0(c)}}. \quad (2)$$

Relation to fuzzy logic. As we can see from the formulas (1) and (2), the standard deviation is an increasing function of the sum $\psi_0(a) + \psi_0(c)$, while the mean m is monotonically increasing with the ratio $|\psi_0(a)/\psi_0(c)|$. This makes sense of we take into account that $\psi_0(a)$ monotonically depends on the proportion a of grades in the A range: the more

grades are in the A range and the fewer grades are in the C range, the larger the average grade m , so m should be kind of monotonically depending on the degree to which it is true that we have A grades and not C grades.

It is worth mentioning that the operations of sum as “or” and ratio as “ a and not c ” appear when we try to interpret neural networks in terms of fuzzy logic [2]; see also Appendix.

5. Conclusions and Future Work

Processing incomplete pedagogical data: formulation of the problem and what we did (a brief summary). In the above text, we considered the following pedagogical situation:

- we have two (or more) techniques for teaching the same material;
- we have a way to gauge the degree to which students learned this material;
- we need to select a techniques for which, on average, the students learn better.

In this paper, we consider the situation in which the degree to which a student learned the material is determined by this student’s letter grade for this class. This letter grade is, of course, an incomplete description of the student’s knowledge. In this paper, we described new algorithms which take into account this incompleteness.

Natural next question: how can we make the pedagogical data more complete? From the mathematical and computational viewpoint, the above setting already leads to non-trivial computational problems. From this viewpoint, the natural next idea may be to improve these algorithms, make them more efficient and more general – in other words, how to best process the existing incomplete data.

From the pedagogical viewpoint, however, a natural next question is how we can *supplement* the letter grades to get a more complete picture of the students’ knowledge.

How to make pedagogical data more complete: main idea. For most classes, a large part of the material studied in the class is important not (or not only) only by itself, but (also) because it provides a basis for the important things which will be learned in the following classes.

This fact was emphasized by Vygotsky (see, e.g., [20]), according to whom the class’ success is determined not only by the students’ mastery of the class material, but also by the increased student abilities to learn new related material. To enhance transition to next classes, it is desirable to prepare students for future courses by appropriate examples.

Example. Success in high school mathematics can be judged not only by the students' grades on the corresponding subjects, but also by how prepared the students are in the long run for studying advanced topics such as calculus. From this viewpoint, it is desirable to include simple optimization and area-computational exercises in algebra and geometry – as examples for which later-learned calculus techniques will work much faster [19].

Related future work. To test the success of this technique, to compare different techniques of teaching lower-level classes (e.g., algebra or geometry), we must take into account not only the student grades in these classes, but also the students' grades in the following more advanced classes (e.g., calculus).

In view of this need, we must extend our statistical techniques in such a way that they not only take into account interval, fuzzy, and probabilistic uncertainty in the letter grades for the current class, but also the corresponding uncertainty in letter grades for the future classes.

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, Texas Department of Transportation grant No. 0-5453, and the University Research Institute grant from the University of Texas at El Paso.

The authors are thankful to participants of SCAN'06 for valuable discussions, and to the anonymous referees for important suggestions.

References

- [1] G. Bojadziev and M. Bojadziev. *Fuzzy Sets, Fuzzy Logic, Applications*, World Scientific, Singapore, 1995.
- [2] S. Dhompongsa, V. Kreinovich, and H. T. Nguyen. How to interpret neural networks in terms of fuzzy logic?, In: *Proceedings of the Second Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications VJFUZZY'2001*, Hanoi, Vietnam, December 7–8, 2001, pp. 184–190.
- [3] D. Dubois, H. Fargier, and J. Fortin J. The empirical variance of a set of fuzzy intervals, In: *Proceedings of the 2005 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2005*, Reno, Nevada, May 22–25, 2005, pp. 885–890.
- [4] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles. Exact bounds on finite populations of interval data. *Reliable Computing*, 11(3):207–233, 2005.
- [5] L. Jaulin, M. Keiffer, O. Didrit, and E. Walter. *Applied Interval Analysis*, Springer-Verlag, London, 2001.
- [6] R. B. Kearfott. *Rigorous Global Search: Continuous Problems*. Kluwer, Dordrecht, 1996.
- [7] R. B. Kearfott and V. Kreinovich (eds.). *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.
- [8] G. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, New Jersey, 1995.
- [9] O. M. Kosheleva and M. Ceberio. Processing educational data: from traditional statistical techniques to an appropriate combination of probabilistic, interval, and fuzzy approaches. In: *Proceedings of the International Conference on Fuzzy Systems, Neural Networks, and Genetic Algorithms FNG'05*, Tijuana, Mexico, October 13–14, 2005, pp. 39–48.
- [10] V. Kreinovich, D. Berleant, and M. Koshelev, website on interval computations <http://www.cs.utep.edu/interval-comp>
- [11] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos. Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases. *Journal of Computational and Applied Mathematics*, 199(2):418–423, 2007.
- [12] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing*, 12(6):471–501, 2006.
- [13] R. E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.
- [14] R. E. Moore and W. A. Lodwick. Interval analysis and fuzzy set theory. *Fuzzy Sets and Systems*, 135(1):5–9, 2003.
- [15] H. T. Nguyen and V. Kreinovich. Nested intervals and sets: concepts, relations to fuzzy sets, and applications. In: R. B. Kearfott and V. Kreinovich (eds.). *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.
- [16] H. T. Nguyen and E. A. Walker. *A First Course in Fuzzy Logic*. CRC Press, Boca Raton, Florida, 2005.
- [17] M. Q. Patton. *Qualitative Research and Evaluation Methods*, Sage Publ., Thousand Oaks, California, 2002.
- [18] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.
- [19] M. Tchoshanov, S. Blake, and A. Duval. Preparing teachers for a new challenge: Teaching Calculus concepts in middle grades. In: *Proceedings of the Second International Conference on the Teaching of Mathematics (at the undergraduate level)*, Hersonissos, Crete, Greece, 2002.
- [20] L. Vygotsky. *Thought and Language*. M.I.T. Press, Cambridge, Massachusetts, 1962.
- [21] H. M. Wadsworth Jr (ed.) *Handbook of Statistical Methods for Engineers and Scientists*. McGraw-Hill Publishing Co., New York, 1990.

A. Statistical Formulas Have a Fuzzy Meaning: Detailed Explanation

Reminder. In the main text, we have derived statistical formulas for the mean and standard deviation, and we mentioned that these formulas can be interpreted in terms of fuzzy logic – if we use sum as “or” and ratio as “ a and not c ”.

Comment. In principle, we could use arbitrary fuzzy t-norms, t-conorms, and fuzzy negations, but then we would

then not get the statistical formulas. Our point is not that we can use different fuzzy operations, but that *statistical* formulas can be interpreted in terms of reasonable *fuzzy* operations.

Selecting an “or” operation. The degree of belief a in a statement A can be estimated as proportional to the number of arguments in favor of A . In principle, there exist infinitely many potential arguments, so in general, it is hardly probable that when we pick a arguments out of infinitely many and then b out of infinitely many, the corresponding sets will have a common element. Thus, it is reasonable to assume that every argument in favor of A is different from every argument in favor of B . Under this assumption, the total number of arguments in favor of A and arguments in favor of B is equal to $a + b$. Hence, the natural degree of belief in $A \vee B$ is proportional to $a + b$.

Selecting an “and” operation. Different experts are reliable to different degrees. Our degree of belief in a statement A made by an expert is equal to $w \& a$, where w is our degree of belief in this expert, and a is the expert’s degree of belief in the statement A . What are the natural properties of the “and”-operation?

First, since $A \& B$ means the same as $B \& A$, it is reasonable to require that the corresponding degrees $a \& b$ and $b \& a$ should coincide, i.e., that the “and”-operation be commutative.

Second, when an expert makes two statements B and C , then our resulting degree of belief in $B \vee C$ can be computed in two different ways:

- We can first compute *his* degree of belief $b \vee c$ in $B \vee C$, and then use the “and”-operation to generate our degree of belief $w \& (b \vee c)$.

- We can also first generate our degrees $w \& b$ and $w \& c$, and then use an “or”-operation to combine these degrees, arriving at $(w \& b) \vee (w \& c)$.

It is natural to require that both ways lead to the same degree of belief, i.e., that the “and”-operation be distributive with respect to \vee .

It is also reasonable to assume that the value $w \& a$ is a monotonically (non-strictly) increasing function of each its variables.

It can be shown [2] that every commutative, distributive, and monotonic operation $\& : R \times R \rightarrow R$ has the form $a \& b = C \cdot a \cdot b$ for some $C > 0$. This expression can be further simplified if we introduce a new scale of degrees of belief $a' \stackrel{\text{def}}{=} C \cdot a$; in the new scale, $a \& b = a \cdot b$.

Selecting a crisp truth value. We know that “true” and “true” is “true”, and that “false” and “false” is “false”. Thus, it is reasonable to call a positive degree of belief e_0 a crisp value if $e_0 \& e_0 = e_0$.

This implies that $e_0 = 1$.

Selecting implication and negation. From the common-sense viewpoint, an implication $A \rightarrow B$ is a statement C such that if we add C to B , we get A . Thus, it is natural to define an *implication operation* as a function $\rightarrow : R \times R \rightarrow R$ for which, for all a and b , we have $(a \rightarrow b) \& a = b$. One can easily check that $a \rightarrow b = b/a$.

Negation $\neg A$ can be viewed as a particular case of implication, $A \rightarrow F$, for a crisp (specifically, false) value F . Thus, we can define negation operation as $a \rightarrow e_0$, i.e., as $1/a$.