
Processing Educational Data: From Traditional Statistical Techniques to an Appropriate Combination of Probabilistic, Interval, and Fuzzy Approaches

Olga M. Kosheleva¹ and Martine Ceberio²

¹ Department of Teacher Education, University of Texas at El Paso
olgak@utep.edu

² Department of Computer Science, University of Texas at El Paso
mceberio@cs.utep.edu

Summary. 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 either intervals (e.g., A means anything from 90 to 100) or fuzzy-type perceptions, words from the natural language like “understood well” or ”understood reasonably well”. We show that the use of intervals and fuzzy 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.

Traditional approach to solving this problem. There exist numerous papers that perform this experimental comparison. The vast majority of these papers use traditional statistical techniques (see, e.g., [15, 16]) to process the experimental results.

Namely, usually, the results (grades, degree of satisfaction, etc.) are translated into numbers, and then these numbers are processed by using the standard statistical techniques.

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 student 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 $m = (x_1 + \dots + x_n)/n = 3$, and the population variance is equal to $V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 0$.
- In the second method, the average is equal to $m = (3 + 4)/2 = 3.5$, so for each i , $(x_i - m)^2 = 0.25$, hence the standard deviation is equal to $V = \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 = 0.25$.

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” grades like A, B, and C to the original grades, 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 a grade 80, and half got a grade 88; and

- in the second method, half of the students got a grade 89, and half of the student got a grade 91.

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

- In the first method, the average is equal to $m = (80 + 88)/2 = 84$, so for each i , $(x_i - m)^2 = 16$, hence the standard deviation 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 $m = (89 + 91)/2 = 90$, so for each i , $(x_i - m)^2 = 1$, hence the standard deviation 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: In Brief

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 number grade. 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 $[70, 80]$.

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, 9, 11].

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, 9, 11].

How we can process interval data: case of statistical characteristics. In particular, several algorithms have been developed for the case when 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, 10] 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}_i . Thus, for the average m , the interval takes the form $[\underline{m}, \overline{m}]$, where

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

If all the letter grades are A, B, C, or D, then the width $\overline{x}_i - \underline{x}_i$ of each corresponding interval is 10, so $\overline{m} = \underline{m} + 10$. In this situation, it is sufficient to compute *one* of the bounds \overline{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 \overline{V} is, in general, NP-hard. However, for educational data, the intervals only intersect at a single point. For such data, there exist efficient algorithms for computing \overline{V} .

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

To compute the lower bound \underline{V} , similarly, for every k , we select:

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

We then compute the average m of the selected x_i and check whether this average satisfies the inequality $\underline{x}_k \leq m \leq \overline{x}_k$. 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. For 3 sorted grades C, B, and A, we get $\underline{x}_1 = 70$, $\overline{x}_1 = 80$, $\underline{x}_2 = 80$, $\overline{x}_2 = 90$, $\underline{x}_3 = 90$, $\overline{x}_3 = 100$. For this data, let us first compute \overline{V} . For $k = 1$, we pick $x_1 = \underline{x}_1 = 70$, $x_2 = \overline{x}_2 = 90$, and $x_3 = \overline{x}_3 = 100$. Here, $m = (x_1 + x_2 + x_3)/3 = 86\frac{2}{3}$. Since $\underline{x}_1 = 70 \leq m \leq \overline{x}_2 = 90$, the upper bound \overline{V} is equal to the population variance $(1/n) \cdot \sum (x_i - m)^2$ of the values $x_1 = 70$, $x_2 = 90$, and $x_3 = 100$, hence $\overline{V} = 102\frac{2}{9}$.

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 $[\underline{x}_1, \overline{x}_1] = [70, 80]$, 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, 10].

3 Fuzzy Approach: In Brief

Formulation of the problem. The main idea behind fuzzy uncertainty (see, e.g., [8, 14]) is that, instead of just describing which objects (in our case, grades) are possible, we also describe, for each object x , the degree $\mu(x)$ to which this object is possible. For each degree of possibility α , we can determine the set of objects that are possible with at least this degree of possibility – the α -cut $\{x \mid \mu(x) \geq \alpha\}$ of the original fuzzy set. Vice versa, if we know α -cuts for every α , then, for each object x , we can determine the degree of possibility that x belongs to the original fuzzy set [1, 8, 12, 13, 14].

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 grades, we have a fuzzy set $\mu_i(x)$ of possible grades, 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.

Simplest case: normally distributed grades. Let us first consider the reasonable case when the actual number grades 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 a and σ .

If we knew the values of the number grades x_i , then we could apply the above statistics and estimate a and $\sigma = \sqrt{V}$. In many situations, we do not know the values of the *number* grades, we only know the values of the *letter* grades. How can we then estimate a 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 grade 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:

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

$$\sigma \cdot \psi_0(f + d + c) + m \approx 80; \quad \sigma \cdot \psi_0(f + d + c + b) + m \approx 90,$$

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 $F(x) = F_0((x - m)/\sigma)$, 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 is an acceptable grade. 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

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

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 if 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.

References

1. Bojadziew G and Bojadziew M (1995), *Fuzzy Sets, Fuzzy Logic, Applications*, World Scientific, Singapore.
2. Dhompongsa S, Kreinovich V, and Nguyen HT (2001), How to Interpret Neural Networks In Terms of Fuzzy Logic?, In: Proceedings of the Second Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications VJ-FUZZY'2001, Hanoi, Vietnam, December 7–8, 2001, pp. 184–190.
3. Dubois D, Fargier H, and Fortin J (2005), 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. Ferson S, Ginzburg L, Kreinovich V, Longpré L, and Aviles M (2005), Exact Bounds on Finite Populations of Interval Data, *Reliable Computing*, 11(3):207–233.
5. Jaulin L, Keiffer M, Didrit O, and Walter E (2001), *Applied Interval Analysis*, Springer-Verlag, Berlin.
6. Kearfott RB (1996), *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht.
7. Kearfott RB and Kreinovich V, eds. (1996), *Applications of Interval Computations*, Kluwer, Dordrecht.
8. Klir G, Yuan, B (1995), *Fuzzy sets and fuzzy logic*, Prentice Hall, New Jersey.
9. Kreinovich V, Berleant D, Koshelev M, website on interval computations <http://www.cs.utep.edu/interval-comp>
10. Kreinovich V, Xiang G, Starks SA, Longpré L, Ceberio M, Araiza R, Beck J, Kandathi R, Nayak A, Torres R, and Hajagos J (to appear) Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity, *Reliable Computing*.
11. Moore RE (1979), *Methods and Applications of Interval Analysis*, SIAM, Philadelphia.
12. Moore RE and Lodwick WA (2003), Interval Analysis and Fuzzy Set Theory, *Fuzzy Sets and Systems*, 135(1):5–9.
13. Nguyen HT and Kreinovich V (1996), Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications, In [6], pp. 245–290
14. Nguyen HT and Walker EA (1999), *First course in fuzzy logic*, CRC Press, Boca Raton, Florida.
15. Sheskin DJ (2004), *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida.
16. Wadsworth HM Jr ed. (1990) *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., New York.

Appendix

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 is a crisp value if $e_0 \& e_0 = e_0$.

This implies that $e_0 = 1$.

Selecting implication and negation. From the commonsense 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 $1/a$.