Normalization-Invariant Fuzzy Logic Operations Explain Empirical Success of Student Distributions in Describing Measurement Uncertainty

Hamza Alkhatib, Boris Kargoll, Ingo Neumann, and Vladik Kreinovich

Abstract In engineering practice, usually measurement errors are described by normal distributions. However, in some cases, the distribution is heavy-tailed and thus, not normal. In such situations, empirical evidence shows that the Student distributions are most adequate. The corresponding recommendation – based on empirical evidence – is included in the International Organization for Standardization guide. In this paper, we explain this empirical fact by showing that a natural fuzzy-logic-based formalization of commonsense requirements leads exactly to the Student's distributions.

1 Formulation of the Problem

Traditional engineering approach to measurement uncertainty. Traditionally, in engineering applications, it is assumed that the measurement error is normally distributed; see, e.g., [12].

This assumption makes perfect sense from the practical viewpoint, it has been shown that for the majority of measuring instruments, the measurement error is indeed normally distributed; see, e.g., [10, 11]. It also makes sense from the theoretical viewpoint, since in many cases, the measurement error comes from a joint effect of many independent small components, and, according to the Central Limit Theorem (see, e.g., [14]), for the large number of components, the resulting distribution is indeed close to Gaussian.

Hamza Alkhatib, Boris Kargoll, and Ingo Neumann
Geodätisches Institut, Leibniz Universität Hannover, Nienburger Strasse 1,
30167 Hannover, Germany, e-mail: alkhatib@gih.uni-hannover.de, kargoll@gih.uni-hannover.de, neumann@gih.uni-hannover.de

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University,
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Yet another explanation for the normal distribution comes from the fact that usually, we only have partial information about the distribution. For example, for the measurement error, we only know the first and the second moments of the corresponding distributions. The first moment – mean – represents a bias. If we know the bias, we can always subtract it from the measurement result, and thus re-calibrated measuring instrument will have 0 mean. Thus, we can always safely assume that the mean is 0. In this case, the second moment is simply the variance $V = \sigma^2$.

There are many different distributions with 0 mean and given standard deviation σ. For example, we can have a distribution in which we have σ and $-\sigma$ with probability 1/2 each. However, such a distribution creates a false certainty – that no other values of x are possible. Out of all such distributions, it therefore makes sense to select the one which maximally preserves the original uncertainty. Uncertainty can be naturally measured by the average number of binary questions needed to determine the value with a given accuracy. It is described by *entropy* $S = -\int \rho(x) \cdot \log_2(\rho(x))\,dx$, where $\rho(x)$ is the probability density function (pdf); see, e.g., [4, 8]. One can easily check that out of all distributions $\rho(x)$ with mean 0 and given standard deviation σ, the entropy is the largest exactly for the normal distribution.

Sometimes, we encounter heavy-tailed distributions. For the Gaussian (normal) distribution, the probability density function $\rho(x) = \dfrac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\dfrac{x^2}{2\sigma^2}\right)$ gets to practically 0 very fast when $|x|$ increases. In other words, the "tails" of this distribution – i.e., values corresponding to large $|x|$ – are very light, practically negligible.

In practice, however, we sometimes encounter distributions with heavy tails, for which $\rho(x)$ decreases much slower, often as a power of x: $\rho(x) \sim c \cdot x^{-\alpha}$; see, e.g., [6, 13].

Power law is not a probability distribution. At first glance, we may want to have $\rho(x) = c \cdot x^{-\alpha}$ for all x. However, the integral of such a function is always infinite – for small α, it is infinite at infinity; for larger α, it is infinite at 0. So, we need expressions for the probability density function $\rho(x)$ which are asymptotically equal to $c \cdot x^{-\alpha}$ but for which $\int \rho(x)\,dx = 1$.

In such cases, Student distributions work well. Our experience of geodetic applications shows that in many such cases, the distribution of the measurement error is well-represented by a Student distribution $\rho(x) = (a + b \cdot x^2)^{-\nu}$ for some a, b, and ν. This empirical observation clearly applies to other application areas as well, since the use of the Student distributions is recommended by the International Organization for Standardization (ISO) [3].

What we do in this paper. In this paper, we explain this empirical fact by showing that a natural fuzzy-logic-based ([5, 9, 15]) formalization of commonsense requirements leads exactly to the Student's distributions.

2 Let Us Use Normalization-Invariant Fuzzy Logic Operations

Our main idea. Informally, uncertainty means that the first value is possible, and the second value is possible, etc. So, when we select a distribution, it makes sense to select a one for which the degree to which all the values are possible is the largest. Let us describe this idea in precise terms.

Fuzzy logic and normalization: a brief reminder. Fuzzy logic was motivated by the fact that many expert statements are formulated by using imprecise (fuzzy) words from natural language, such as "small" (or, in our case, "possible"). To describe such terms, for every possible value x of the corresponding quantity, we ask the expert to estimate the degree $\mu(x)$ to which this value satisfies this quantity (e.g., "is small"). An expert can mark his/her degree of confidence by selecting a number from the interval $[0,1]$, so that 1 means full confidence, 0 means no confidence, and intermediate values indicate partial confidence. The resulting function $\mu(x)$ is called a *membership function*.

For properties like "small", there are values (e.g., $x=0$) for which are absolutely sure that this value is small. For such values, we have $\mu(x)=1$, so the maximum of the corresponding membership function is equal to 1.

For other properties – e.g., "medium" – we may not have such values. In this case, the maximum of $\mu(x)$ may be smaller than 1. A usual way to deal with such property is to *normalize* the corresponding membership function, i.e., to consider a new function $\mu'(x)=\dfrac{\mu(x)}{\max\limits_y \mu(y)}$ for which $\max\limits_x \mu'(x)=1$.

Normalization is also performed when we get an additional information about the property. For example, we knew that x is small, now we learn that $x \geq 5$. In this case, if simply keep the previous values of $\mu(x)$ for $x \geq 5$ and set all the values $\mu(x)$ for $x<5$ to 0, we get a new membership function whose maximum is smaller than 1. So, we normalize it.

Finally, normalization is a must when experts use the available information about probabilities to come up with the corresponding degrees. Indeed, if we know the probability density function $\rho(x)$, this means that the larger $\rho(x)$, the more probable it is to observe a value close to x. Thus, in this case, it is reasonable to take, as degrees $\mu(x)$, either the values $\rho(x)$ themselves, or some values proportional to $\rho(x)$: $\mu(x)=c\cdot\rho(x)$ for some constant c. In this case, normalization leads to the membership function $\mu(x)=\dfrac{\rho(x)}{\max\limits_y \rho(y)}$. Vice versa, if we have the result $\mu(x)$ of normalizing a pdf, we can reconstruct the original pdf $\rho(x)$ if we multiply $\mu(x)$ by an appropriate constant - the constant to be determined from the requirement that $\int \rho(x)\,dx=1$; thus: $\rho(x)=\dfrac{\mu(x)}{\int \mu(y)\,dy}$.

Fuzzy logic operations: a reminder. The need for logical operations comes from the fact that answers to questions of interest often depend on several expert's statements. This is exactly our case: we are interested in knowing to what extent the first value is possible *and* the second value is possible, etc.

Thus, in addition to knowing the experts' degrees of confidence in different statements A, B, etc., we also need need to know the expert's degree of confidence in different logical combinations of these statements, such as $A\,\&\,B$ and $A \vee B$.

In our case, we do not just want to know to what extend each value is possible, we also want to know to what extend the value is possible *and* the second value is possible, etc.

Ideally, we should elicit these degrees from the experts, but there are exponentially many such combinations, so such an elicitation is not feasible. Thus, we need to estimate the expert's degree of confidence $d(A\,\&\,B)$ in a composite statement like $A\,\&\,B$ based only on his/her degrees of confidence a and b in statements A and B. The corresponding estimate for $d(A\,\&\,B)$ is called an *"and"-operation* (or a *t-norm*) and is denoted by $f_\&(a,b)$.

Since $A\,\&\,B$ and $B\,\&\,A$ mean the same, it makes sense to require that our estimates for these two statements are the same, i.e., that the operation $f_\&(a,b)$ is commutative: $f_\&(a,b) = f_\&(b,a)$.

Similarly, the fact that $A\,\&\,(B\,\&\,C)$ and $(A\,\&\,B)\,\&\,C$ mean the same encourages us to require that $f_\&(a, f_\&(b,c)) = f_\&(f_\&(a,b),c)$, i.e., that the operation $f_\&(a,b)$ is associative.

For associative operations, we can define $f_\&(a,b,\ldots,c)$ by induction, as the result of applying the "and"-operation in any order: e.g., as $f_\&(\ldots(f_\&(a,b),\ldots,c)$.

It also makes sense to require that if A is false, then $A\,\&\,B$ is false, i.e., that $f_\&(0,b) = 0$ for all b, and that if we increase our degree of confidence in A and/or in B, our confidence in $A\,\&\,B$ will not decrease, i.e., that the function $f_\&(a,b)$ is (non-strictly) increasing in each of its variables.

Since 1 is usually interpreted as full confidence, if A is absolutely true, then $A\,\&\,B$ is equivalent to B for all B, i.e., $f_\&(1,b) = b$ for all b.

From traditional fuzzy operations to normalization-invariant ones. In some cases, the degree 1 means absolute confidence, but in other cases the degree 1 comes from normalization and thus, corresponds to less-than-absolute confidence. In such cases, it does not make sense to require that $f_\&(1,b) = b$, since our degree of confidence in $A\,\&\,B$ may be smaller than our degree of confidence in the original statement B.

It therefore makes sense to consider a more general class of "and"-operations: we still keep commutativity, associativity, monotonicity, and the property that $f_\&(0,b) = 0$, but we no longer require that $f_\&(1,b) = b$ for all b.

What should we require? A natural requirement is that the "and"-operation should be preserved under normalization. To be more precise, we can compute the normalized degree of confidence in a statement $A\,\&\,B$ in two different ways:

- we can take the original degree $f_\&(a,b)$ and normalize it, by multiplying it by an appropriate constant λ;

- alternatively, we can first normalize the degrees of confidence in A and B, getting $\lambda \cdot a$ and $\lambda \cdot b$, and then apply an "and"-operation to the new degrees, resulting in the value $f_{\&}(\lambda \cdot a, \lambda \cdot b)$.

It is reasonable to require that these two ways lead to the same estimate. Thus, we arrive at the following definition.

Definition 1. *By a* normalization-invariant *"and"-operation, we means a function $f_{\&}(a,b)$ which is commutative, associative, (non-strictly) increasing in each of the variables, and satisfies the properties $f_{\&}(0,b) = 0$ and*

$$f_{\&}(\lambda \cdot a, \lambda \cdot b) = \lambda \cdot f_{\&}(a,b)$$

for all $\lambda \geq 0$, $a \geq 0$, and $b \geq 0$.

Let us describe all possible normalization-invariant "and"-operations. Similar to the case of the usual "and"-operations [7], one can prove that for every normalization-invariant "and"-operation and for every $\varepsilon > 0$, there exists a normalization-invariant "and"-operation of the type $f_{\&}(a,b) = f^{-1}(f(a) + f(b))$ for some strictly decreasing function $f(x)$. Thus, for all practical purposes, we can safely assume that our operation has this form.

For such functions, $c = f_{\&}(a,b)$ is equivalent to $f(c) = f(a) + f(b)$. Thus, scale-invariance means that $f(c) = f(a) + f(b)$ implies $f(\lambda \cdot c) = f(\lambda \cdot a) + f(\lambda \cdot b)$. Thus, for every λ, the transformation T from $f(a)$ to $f(\lambda \cdot a)$ is additive: if $C = A + B$, then $T(C) = T(A) + T(B)$, i.e., in other words, $T(A+B) = T(A) + T(B)$. It is known (see, e.g., [1, 2]) that every monotonic additive function is linear. Thus, $f(\lambda \cdot a) = c(\lambda) \cdot f(a)$ for all a and λ. For monotonic functions $f(a)$, the only solution for this functional equation is $f(a) = C \cdot a^{-\alpha}$ for some C and α [1, 2].

For this function, the equality $f(c) = f(a) + f(b)$, i.e., $C \cdot c^{-\alpha} = C \cdot a^{-\alpha} + C \cdot b^{-\alpha}$, is equivalent to $c^{-\alpha} = a^{-\alpha} + b^{-\alpha}$, i.e., to $c = f_{\&}(a,b) = (a^{-\alpha} + b^{-\alpha})^{-1/\alpha}$.

3 Resulting Derivation of the Student Distributions

We want to select a membership function $\mu(x)$ which is the best fit with our requirement that all possible values x are indeed possible. In other words, we want to maximize the degree to which x_1 is possible, and x_2 is possible, etc. For each value x_i, the degree to which this value is possible is equal to $\mu(x_i)$.

Now that we have a general formula for the normalization-invariant "and"-operation, we can describe the degree to which x_1 is possible and x_2 is possible as

$$f_{\&}(\mu(x_1), \mu(x_2), \ldots) = ((\mu(x_1))^{-\alpha} + (\mu(x_2))^{-\alpha} + \ldots)^{-1/\alpha}.$$

Maximizing this degree is equivalent to minimizing the sum $(\mu(x_1))^{-\alpha} + (\mu(x_2))^{-\alpha} + \ldots$ In the limit, when we take a denser and denser grid of values x_i and make them cover a longer and longer interval, this sum turns into an integral $\int (\mu(x))^{-\alpha}\, dx$.

We need to find the smallest possible value of this integral under the constraints that the mean is 0 and that the variance is equal to a given value σ^2. These constrains have the form $\int x\cdot\rho(x)\,dx=0$ and $\int x^2\cdot\rho(x)\,dx=\sigma^2$, where $\rho(x)=\dfrac{\mu(x)}{\int\mu(y)\,dy}$, i.e., the form $\int x\cdot\dfrac{\mu(x)}{\int\mu(y)\,dy}\,dx=0$ and $\int x^2\cdot\dfrac{\mu(x)}{\int\mu(y)\,dy}\,dx=\sigma^2$. These equalities can be simplified into $\int x\cdot\mu(x)\,dx=0$ and $\int x^2\cdot\mu(x)\,dx-\sigma^2\cdot\int\mu(x)\,dx=0$. Thus, we arrive at the following constraint optimization problem:
Minimize $\int(\mu(x))^{-\alpha}\,dx$ under the constraints

$$\int x\cdot\mu(x)\,dx=0 \text{ and } \int x^2\cdot\mu(x)\,dx-\sigma^2\cdot\int\mu(x)\,dx=0.$$

Lagrange multiplier method reduces this constraint optimization problem to the unconstrained optimization one

$$\int(\mu(x))^{-\alpha}\,dx+\lambda_1\cdot\int x\cdot\mu(x)\,dx+\lambda_2\cdot\left(\int x^2\cdot\mu(x)\,dx-\sigma^2\cdot\int\mu(x)\,dx\right)\to\min.$$

Differentiating the left-hand side with respect to $\mu(x)$ and equating the derivative to 0, we conclude that

$$-\alpha\cdot(\mu(x))^{-\alpha-1}+\lambda_1\cdot x+\lambda_2\cdot x^2-\lambda_2\cdot\sigma^2=0,$$

i.e., that $\mu(x)=(a_0+a_1\cdot x+a_2\cdot x^2)^{-\nu}$ for some a_i and ν, i.e., equivalently, the form $\mu(x)=c\cdot(1+a_1\cdot x+a_2\cdot x^2)^{-\nu}$.

The pdf $\rho(x)=\dfrac{\mu(x)}{\int\mu(y)\,dy}$ differs from the membership function by a multiplicative constant, so we also have $\rho(x)=\text{const}\cdot(1+a_1\cdot x+a_2\cdot x^2)^{-\nu}$. The quadratic expression inside can be described as $a_2\cdot(x-x_0)^2+\text{const}$ for some x_0. This formula is symmetric with respect to x_0 thus its mean is x_0. Since we know that the mean should be 0, we get $x_0=0$, hence $\rho(x)=\text{const}\cdot(1+a_2\cdot x^2)^{-\nu}$.

Taking into account that we should have $\int\rho(x)\,dx=1$, we get exactly Student distributions – so we indeed get the desired justification!

Acknowledgments

This work was performed when Vladik was a visiting researcher with the Geodetic Institute of the Leibniz University of Hannover, a visit supported by the German Science Foundation. This work was also supported in part by NSF grant HRD-1242122.

References

1. J. Aczél, *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.
2. J. Aczél and H. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 1989.
3. International Organization for Standardization (ISO), *ISO/IEC Guide 98-3:2008, Uncertainty of Measurement – Part 3: Guide to the Expression of Uncertainty in Measurement (GUM:1995)*, 2008.
4. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
5. G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.
6. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.
7. H. T. Nguyen, V. Kreinovich, and P. Wojciechowski, "Strict Archimedean t-norms and t-conorms are universal approximators, *International Journal of Approximate Reasoning*, 1998, Vol. 18, Nos. 3–4, pp. 239-249.
8. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
9. H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.
10. P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).
11. A. I. Orlov, "How often are the observations normal?", *Industrial Laboratory*, 1991, Vol. 57, No. 7, pp. 770–772.
12. S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, Berlin, 2005.
13. S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Varlag, New York, 2007.
14. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
15. L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.