5-2020

How Expert Knowledge Can Help Measurements: Three Case Studies

Vladik Kreinovich

Comments:
Technical Report: UTEP-CS-20-52

How Expert Knowledge Can Help Measurements: Three Case Studies

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
vladik@utep.edu

Abstract

In addition to measurement results, we often have expert estimates. These estimates provides an additional information about the corresponding quantities. However, it is not clear how to incorporate these estimates into a metrological analysis: metrological analysis is usually based on justified statistical estimates, but expert estimates are usually not similarly justified. One way to solve this problem is to calibrate an expert the same way we calibrate measuring instruments. In the first two case studies, we show that such a calibration indeed leads to useful result. The third case study provides an example of another use of expert knowledge in measurement practice: this knowledge can be used to make semi-empirical measurement models more explainable – and thus, more reliable.

1 Introduction

Using expert knowledge is important, but how? A large amount of information comes from measurements. However, in many areas, it is crucial to also use expert knowledge; for example:

- With all modern medical tests and measurements, doctor's intuition is still crucial.
- In spite of all the successes of self-driving cars, it is still not possible to fully replace a human driver.

It is therefore important to supplement measurement results with expert estimates.

And this is a big problem for metrology, because in metrology, we are accustomed to work with statistically justified estimates, while expert estimates are not similarly justified.

So how can expert knowledge help measurements? In measurement practice:

- we come up with a parametric model of the corresponding class of phenomena,
- we test this model – to make sure that it provides an adequate description of the phenomena, and
- we use measurements to estimate the parameters corresponding to a given situation.

How can experts help?

- experts can provide such a model, and
- experts can provide estimates of the corresponding parameters.

Why is this useful? In terms of a model: the currently used model often comes from a semi-empirical study. Such curve-fitting models are not very convincing, this can be over-fitting. Experts' knowledge and intuition can help separate explainable models from curve-fitting results.

In terms of expert estimations: experts may not be accurate as measurements, but they are often faster and cheaper to use. They also supplement measurement results, this making the resulting estimates more accurate.

But how to incorporate expert knowledge into a metrological framework. From the common sense viewpoint, expert knowledge is useful. But how can include their estimates into a metrological framework, with its precise justifications?

A natural idea is to treat an expert as a measuring instrument: to calibrate the expert. Thus, we can get a statistically justified estimate for the accuracy of expert-generated numbers.

Moreover, we can use this calibration to improve the expert's estimates. This is similar to how, once know the instrument's bias, we can subtract it and get more accurate results.

Three case studies. To illustrate the above general ideas, we provide three case studies.

- In the first case study, we show that application of usual linear calibration to experts can be helpful.
- In the second case study, we provide an example of useful non-linear calibration.
- The third case study explains how expert knowledge can make semi-empirical models more convincing.

Comment. Preliminary results of the three test studies first appeared in [3, 37, 39].

2 First Case Study: Measurement-Type "Calibration" of Expert Estimates Improves Their Accuracy and Their Usability – Pavement Engineering

Experts are often used for estimation. Sometimes, experts are used because no measuring instruments can replace these experts. For example, in dermatology, estimates of a skilled expert are more accurate results than of any algorithm. This is one of the main reasons why, in spite of numerous expert systems, human doctors are still needed and still valued.

In other cases, in principle, we can use automatic systems, but experts are still much cheaper to use. An example of such situation is pavement engineering. In principle, we can use an expensive automatic vision-based system to gauge the condition of the pavement. However, it is much cheaper – and faster – to use human raters.

Expert estimates are often very imprecise. Humans rarely have a skill of accurately evaluating the values of different quantities. For example, it is well known that humans drastically overestimate small probabilities. Correspondingly, humans underestimate the probabilities which are close to 1; see, e.g., [14] and references therein.

Since most people's estimates are very inaccurate, it is difficult to find good expert estimators. It is well known that there is a high competition to get into medical schools. Even in pavement engineering, finding a good rater is difficult.

It is difficult to find good experts: example from pavement engineering. According to a current standard [7], the condition of a pavement is evaluated by using a special index. This Pavement Condition Index (PCI) combines different possible pavement faults. To gauge the accuracy of a rater candidate, many locations across the US use criteria developed by the Metropolitan Transportation Commission (MTC) of California [26].

A crucial part of the rater certification is a field survey exam. In this exam, a rater evaluates 24 test sites that have been previously evaluated by expert raters. Candidate's PCI values are then compared with the PCI values of the expert rater. The expert's values are taken as the ground truth (GT). To certify, the rater must satisfy the following two criteria:

- at least for 50% of the evaluated sites, the difference should not exceed 8 points, and
- at least for 88% of the evaluated sites, the difference should not exceed 18 points.

MTC provided a sample of 18 typical candidates. Out of these candidates, only 5 (28%) satisfy both criteria and thus, pass the exam and can be used as raters.

Problems.

- What can we do to increase the number of available experts?
- And for those who have been selected as experts, can we improve the accuracy of their estimates?

Calibration. We are interested in situations when expert serve, in effect, as measuring instruments.

Measuring instruments are usually much more accurate then human experts. Still, they are sometimes not very accurate. Even when they are originally reasonably accurate, in time, their accuracy decreases.

When the measuring instrument becomes not very accurate, we do not necessarily throw it away. For example, before we step on the scales, they already show 10 pounds. We do not necessarily throw away these scales: instead, we adjust the starting point.

When a household device for measuring blood pressure starts producing weird results, the manufacturers do not advise the customers to throw it away and to buy a new one, they advise the customers to come to a doctor's office and to calibrate the customer's instrument.

In general, calibration is a routine procedure for measuring instruments; see, e.g., [38]. We measure the same quantities:

- by using our measuring instruments – resulting in the values $x_1, \ldots, x_n$, and
- by using a much more accurate ("standard") measuring instrument – resulting in the values $s_1, \ldots, s_n$.

In many cases – like in the above scales example – the main problem is the bias. We compensate for the bias by subtracting the estimated value. The resulting corrected values $x_i + b$ are closer to the ground truth s_i. A reasonable way to estimate the bias is to use the Least Squares method [38, 41]: $\sum\limits_{i=1}^{n}((x_i+b)-s_i)^2 \to \min$.

In some cases, there is also a relative systematic error, when each value is under- or over-estimated by a certain percentage. To compensate for this under- and over-estimation, we need to multiply by an appropriate constant. For example if all the values are overestimated by 10%, then each ground truth value s_i is replaced by the biased value $s_i + 0.1 \cdot s_i = 1.1 \cdot s_i$. To compensate for this relative bias, we thus need to multiply all the measurement results by $1/1.1$.

In general, we need to replace the original measurement results x_i by corrected values $a \cdot x_i$ for some a. In general, to compensate for both absolute and relative biases, we replace x_i with $a \cdot x_i + b$.

The values a and b can be found by the Least Squares method:

$$\sum_{i=1}^{n}((a \cdot x_i + b) - s_i)^2 \to \min.$$

After that, instead of using the original measurement result x produced by the measuring instrument, we calibrate it into a more accurate value

$$x' = a \cdot x + b.$$

In addition to such a linear calibration, it is sometimes beneficial to use non-linear calibration. Sometimes, a quadratic or cubic calibration is used – which leads to more accurate measurement results. In many practical situations, it is also beneficial to use fractional-linear re-scaling $x' = \dfrac{a \cdot x + b}{1 + c \cdot x}$; see, e.g., [16, 17, 18, 24, 25, 28].

Idea: let us calibrate experts. A natural idea is that since experts serve as measuring instruments, we can similarly calibrate the experts. Namely, instead of using the original expert estimates:

- we first re-scale the original expert estimates in accordance with the appropriate calibration function, and then
- we use these re-scaled values instead of the original expert estimates.

As a result – just like for measuring instruments – we will hopefully get more accurate estimates.

In some situations, when for some experts, their original estimates were not very accurate, we may end up with re-scaled estimates of acceptable quality, so we can use them.

Such calibration is indeed helpful. A good example of the efficiency of such calibration is expert's estimations of small probabilities. According to Kahneman and Tversky [**?**], these estimates e_i are way off.

However, the values $e'_i = a \cdot \sin^2(b \cdot e_i)$ are much more accurate; see, e.g., [19, 20, 21, 22]. Namely, for $p_i < 20\%$:

- the worst-case difference between the original estimates e_i and the actual probabilities was 8.6% – more than 40% of the original probability value – while
- the worst-case difference between the re-scaled estimates e'_i and the probabilities p_i is 0.7% – which is 3.5% of the original probability value, and is, thus, an order of magnitude more accurate.

We applied our idea to pavement engineering. We started with the 18 rater candidates from the original MTC sample. In the original test, only five of these candidates passed the exam: rater candidates R6, R8, R9, R14, and R15.

Originally, we compare this rater's ratings r_i with the 24 corresponding ground truth values s_i. Instead, we first found the values a and b that minimize the sum of the squares $\sum\limits_{i=1}^{24}((a \cdot r_i + b) - s_i)^2$. Then, we used the re-scaled values $r'_i = a \cdot r_i + b$ to compare with the ground truth.

As a result, more experts are selected. Based on the re-scaled ratings, four more candidates passed the test: candidates R1, R3, R5, and R11. This means that these four folks can now be used for rating pavement conditions.

Of course, instead of using their original ratings r_i, we first re-scale them to $r'_i = a \cdot r_i + b$ for this rater's a and b. As a result, we can accept 9 raters. Thus, the acceptance rate is now no longer $5/18 \approx 28\%$, it is $9/18 = 50\%$.

For most originally selected experts, re-scaling leads to more accurate estimates. After re-scaling, one of the originally accepted candidates – R9 – no longer fits. For this rater, we use his original ratings.

For the remaining four originally selected raters, re-scaling improves the accuracy of their estimates:

- for R6, the mean square rating error decreases from 11.21 points to 10.01 points – a decrease of 9.9%;
- for R8, the mean square rating error decreases from 10.00 points to 8.66 points – a decrease of 6.4%;
- for R14, the mean square rating error decreases from 8.62 to 6.95 points – a decrease of 19.4%; and
- for R15, the mean square rating error decreases from 6.47 points to 6.21 points – a decrease of 4.0%.

Comment. Similarly good results were consistently achieved for several other groups of rater candidates.

3 Second Case Study: Relationship Between Measurement Results and Expert Estimates of Cumulative Quantities, on the Example of Pavement Roughness

Cumulative quantities. Many physical quantities can be measured directly: e.g., we can directly measure mass, acceleration, force. However, we are often interested in *cumulative* quantities that combine values corresponding to different moments of time and/or different locations. For example, when we are studying public health or pollution or economic characteristics, we are often interested in characteristics describing the whole city, the whole region, the whole country.

Formulation of the problem. Cumulative characteristics are not easy to measure. To measure each such characteristic, we need to perform a large number of measurements, and then to use an appropriate algorithm to combine these results into a single value.

Such measurements are complicated. So, we often have to supplement the measurement results with expert estimates. To process such data, it is desirable to describe both estimates in the same scale:

- to estimate the actual value of the corresponding quantity based on the expert estimate, and
- vice versa, to estimate the expert estimate based on the actual value of the quantity.

Case study: estimating pavement roughness. Estimating road roughness is an important problem. Indeed, road pavements need to be maintained and repaired. Both maintenance and repair are expensive. So, it is desirable to estimate the pavement roughness as accurately as possible.

- If we overestimate the road roughness, we will waste money on "repairing" an already good road.
- If we underestimate the road roughness, the road segment will be left unrepaired and deteriorate further.

As a result, the cost of future repair will skyrocket.

The standard way to measure the pavement roughness is to use the International Roughness Index (IRI); see, e.g., [6, 9, 10, 40]. This measure of roughness is recommended by the US standards [6, 9, 10].

Crudely speaking, IRI describes the effect of the pavement roughness on a standardized model of a vehicle. Measuring IRI is not easy, because the real vehicles differ from this standardized model. As a result, we measure roughness by some instruments and use these measurements to estimate IRI. For example, we can:

- perform measurements by driving an available vehicle along this road segment,
- extract the local roughness characteristics from the effect of the pavement on this vehicle, and then
- estimate the effect of the same pavement on the standardized vehicle.

In view of this difficulty, in many cases, practitioners rely on expert estimates of the pavement roughness. The corresponding measure – estimated on a scale from 0 to 5 – is known as the Present Serviceability Rating (PSR); see, e.g., [5, 11].

Empirical relation between measurement results and expert estimates. The empirical relation between PSR and IRI is described by the formula:

$$\text{PSR} = 5 \cdot \exp(-0.0041 \cdot \text{IRI}).$$

This formula was first proposed by B. Al-Omari and M. Darter in [4], and it still remains actively used in pavement engineering; see, e.g., [8, 11, 34, 35]. It

works much better than many previously proposed alternative formulas, such as

$$\text{PSR} = a + b \cdot \sqrt{\text{IRI}}$$

proposed in [27]. However, it is not clear why namely this formula works so well.

What we do in this section. We propose a possible explanation for the above empirical formula. This explanation will be general: it will apply to all possible cases of cumulative quantities.

We will come up with a general formula $y = f(x)$ that describes how a subjective estimate y of a cumulative quantity depends on the result x of its measurement.

As a case study, we will use gauging road roughness.

Main idea. In general, the numerical value of a *subjective estimate* depends on the scale. In road roughness estimates, we usually use a 0-to-5 scale. In other applications, it may be more customary to use 0-to-10 or 0-to-1 scales.

A usual way to transform between the two scales is to multiply all the values by a corresponding factor. For example, to transform from 0-to-10 to 0-to-1 scale, we multiply all the values by $\lambda = 0.1$. In other transitions, we can use transformations $y \to \lambda \cdot y$ with different re-scaling factors λ.

There is no major advantage in selecting a specific scale. So, subjective estimates are defined modulo such a re-scaling transformation $y \to \lambda \cdot y$.

At first glance, the result of *measuring* a cumulative quantity may look uniquely determined. However, a detailed analysis shows that there is some non-uniqueness here as well. Indeed, the result of a cumulative measurement comes from combining values measured at different moments of time and/or values corresponding to different spatial locations. For each individual measurement, the probability of a sensor's malfunction may be low. However, often, we perform a large number of measurements. So, some of them bound to be caused by such malfunctions and are, thus, outliers.

It is well known that even a single outlier can drastically change the average. So, to avoid such influence, the usual algorithms first filter out possible outliers. This filtering is not an exact science; we can set up slightly different thresholds for detecting an outlier, slightly different threshold for allowed number of remaining outliers, etc.

We may get a computation result that only takes actual signals into account. With a different setting, we may get a different result, affected by a few outliers.

Let's denote the average value of an outlier is L and the average number of such outliers is n. Then, the second scheme, in effect, adds a constant $n \cdot L$ to the cumulative value computed by the first scheme.

Yes, there is also some random deviation. However, when the number n is reasonably large, then, due to the Large Numbers theorem, these deviations average out and we get approximately the mean value (see, e.g., [41]) – just like when we flip a coin many (N) times, the overall number of times when it falls head will be close to $0.5 \cdot N$.

So, the measured value of a cumulative quantity is defined modulo an addition of some value:

$$x \to x + a \text{ for some constant } a.$$

Motivation for invariance. We do not know exactly what is the ideal threshold, so we have no reason to select a specific shift as ideal. It is therefore reasonable to require that the desired formula $y = f(x)$ not depend on the choice of such a shift, i.e., that the corresponding dependence not change if we simply replace x with $x' = x + a$.

Of course, we cannot just require that $f(x) = f(x + a)$ for all x and all a. Indeed, in this case, the function $f(x)$ will simply be a constant, but y increases with x. But this is clearly not how invariance is usually defined. For example, for many physical interactions, there is no fixed unit of time. So, formulas should not change if we simply change a unit for measuring time: $t' = \lambda \cdot t$. The formula $d = v \cdot t$ relating the distance d, the velocity v, and the time t should not change. We want to make this formula true when time is measured in the new units. So, we may need to also appropriately change the units of other related quantities.

In the above example, we need to appropriately change the unit for measuring velocity, so that not only time units are changed, e.g., from hours to second, but velocities are also changed from km/hour to km/sec.

So, if we re-scale x, the formula $y = f(x)$ should remain valid if we appropriately re-scale y. As we have mentioned earlier, possible re-scalings of the subjective estimate y have the form $y \to y' = \lambda \cdot y$. Thus, for each a, there exists $\lambda(a)$ (depending on a) for which $y = f(x)$ implies that $y' = f(x')$, where

$$x' \stackrel{\text{def}}{=} x + a \text{ and } y' \stackrel{\text{def}}{=} \lambda \cdot y.$$

Definition. *A monotonic function $f(x)$ is called* unit-invariant *if for every real number a, there exists a positive real number $\lambda(a)$ for which, for each x and y:*

- *if $y = f(x)$,*
- *then $y' = f(x')$, where $x' \stackrel{\text{def}}{=} x + a$ and $y' \stackrel{def}{=} \lambda(a) \cdot y$.*

Proposition. *A function $f(x)$ is unit-invariant if and only if it has the form*

$$f(x) = C \cdot \exp(-b \cdot x) \text{ for some } C \text{ and } b.$$

Comment. For road roughness, this result explains the empirical formula.

Proof. It is easy to check that every function $y = f(x) = C \cdot \exp(-b \cdot x)$ is indeed unit-invariant.

Indeed, for each a, we have

$$f(x') = f(x+a) = C \cdot \exp(-b \cdot (x+a)) =$$

$$C \cdot \exp(-b \cdot x - b \cdot a) = \lambda(a) \cdot C \cdot \exp(-b \cdot x).$$

Here we denoted $\lambda(a) \stackrel{\text{def}}{=} \exp(-b \cdot a)$. Thus here, indeed, $y = f(x)$ implies that $y' = f(x')$.

Vice versa, let us assume that the function $f(x)$ is unit-invariant. Then, for each a, the condition $y = f(x)$ implies that $y' = f(x')$, i.e., that $\lambda(a) \cdot y = f(x+a)$. Substituting $y = f(x)$ into this equality, we conclude that $f(x+a) = \lambda(a) \cdot f(x)$. It is known (see, e.g., [2]) that every monotonic solution of this functional equation has the form

$$f(x) = C \cdot \exp(-b \cdot x) \text{ for some } C \text{ and } b.$$

The proposition is proven.

Conclusions of this section. In pavement engineering, it is important to accurately gauge the quality of road segments. Such estimates help us decide how to best distribute the available resources between different road segments. So, proper and timely maintenance is performed on road segments whose quality has deteriorated. Thus, to avoid future costly repairs of untreated road segments.

The standard way to gauge the quality of a road segment is International Roughness Index (IRI). It requires a large amount of costly measurements. As a result, it is not practically possible to regularly measure IRI of all road segments. So, IRI measurements are usually restricted to major roads.

For local roads, we need to an indirect way to estimate their quality. To estimate the quality of a road segment, we combine user estimates of different segment properties into a single index known as Present Serviceability Rating (PSR).

There is an empirical formula relating IRI and PSR. However, one of the limitations of this formula is that it purely heuristic. This formula lacks a theoretical explanation and thus, the practitioners may be not fully trusting its results. In this section, we provide such a theoretical explanation. We hope that the resulting increased trust in this formula will help enhance its use. Thus, it will help with roads management.

4 Third Case Study: Normalization-Invariant Fuzzy Logic Operations Explain Empirical Success of Student Distributions in Describing Measurement Uncertainty

Traditional engineering approach to measurement uncertainty. Traditionally, in engineering applications, it is assumed that the measurement error is normally distributed; see, e.g., [38].

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., [32, 33]. 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., [41]), for the large number of components, the resulting distribution is indeed close to Gaussian.

Another explanation: we only have partial information about the distribution. Often, we only know the first and the second moments. The first moment – mean – represents a bias. If we know the bias, we can always subtract it from the measurement result. Thus re-calibrated measuring instrument will have 0 mean. So, we can always safely assume that the mean is 0. Then, the 2nd moment is simply the variance $V = \sigma^2$.

There are many distributions w with 0 mean and given 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 makes sense to select the one which maximally preserves the uncertainty.

Uncertainty can be gauged by average number of binary questions needed to determine x with accuracy ε. It is described by *entropy* $S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx$; see, e.g., [13, 30]. Out of all distributions $\rho(x)$ with mean 0 and given σ, the entropy is the largest for normal $\rho(x)$.

Need for heavy-tailed distributions. For the normal distribution,

$$\rho(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

The "tails" – values corresponding to large $|x|$ – are very light, practically negligible.

Often, $\rho(x)$ decreases much slower, as $\rho(x) \sim c \cdot x^{-\alpha}$; see, e.g., [23, 36]. We cannot have $\rho(x) = c \cdot x^{-\alpha}$, since $\int_0^\infty x^{-\alpha}\, dx = +\infty$, and we want $\int \rho(x)\, dx = 1$.

Often, the measurement error is well-represented by a Student distribution $\rho_S(x) = (a + b \cdot x^2)^{-\nu}$. This is true in geodesy, and in other applications as well. This distribution is even recommended by the International Organization for Standardization (ISO) [12].

What we do. How to explain the empirical success of Student's distribution $\rho_S(x)$? We show that a natural fuzzy-logic-based ([15, 31, 42]) formalization of commonsense requirements leads to $\rho_S(x)$.

Our idea: uncertainty means that the first value is possible, and the second value is possible, etc. Let's select $\rho(x)$ with the largest degree to which all the values are possible.

It is reasonable to use fuzzy logic to describe degrees of possibility. An expert marks his/her degree by selecting a number from the interval $[0, 1]$.

Need for normalization. For "small", we are absolutely sure that 0 is small: $\mu_{\text{small}}(0) = 1$ and $\max\limits_x \mu_{\text{small}}(x) = 1$. For "medium", there is no x

with $\mu_{\rm med}(x) = 1$, so $\max\limits_x \mu_{\rm med}(x) < 1$.

A usual way to deal with such situations is to *normalize* $\mu(x)$ into $\mu'(x) = \dfrac{\mu(x)}{\max\limits_y \mu(y)}$. Normalization is also performed when we get additional information.

Example: we knew that x is small, we learn that $x \ge 5$. Then, $\mu_{\rm new}(x) = \mu_{\rm small}(x)$ for $x \ge 5$ and $\mu_{\rm new}(x) = 0$ for $x < 5$, and $\max\limits_x \mu_{\rm new}(x) < 1$.

Normalization is also needed when experts use probabilities to come up with the degrees. Indeed, the larger $\rho(x)$, the more probable it is to observe a value close to x. Thus, it is reasonable to take the degrees $\mu(x)$ proportional to $\rho(x)$: $\mu(x) = c \cdot \rho(x)$. Normalization leads to $\mu(x) = \dfrac{\rho(x)}{\max\limits_y \rho(y)}$. Vice versa, if we have the result $\mu(x)$ of normalizing a pdf, we can reconstruct $\rho(x)$ as $\rho(x) = \dfrac{\mu(x)}{\int \mu(y)\,dy}$.

How to combine degrees. For each x, we get a degree to which x is possible. We want to compute the degree to which x_1 is possible *and* x_2 is possible, etc. So, we need to apply an "and"-operation (t-norm) to the corresponding degrees.

A natural idea is to use normalization-invariant t-norms. We can compute the normalized degree of confidence in a statement $A \,\&\, B$ in two different ways:

- we can normalize $f_{\&}(a,b)$ to $\lambda \cdot f_{\&}(a,b)$;
- or, we can first normalize a and b and then apply an "and"-operation: $f_{\&}(\lambda \cdot a, \lambda \cdot b)$.

It's reasonable to require that we get the same estimate: $f_{\&}(\lambda \cdot a, \lambda \cdot b) = \lambda \cdot f_{\&}(a,b)$.

It is known that strict Archimedean t-norms $f_{\&}(a,b) = f^{-1}(f(a) + f(b))$ are universal approximators; see, e.g., [29]. So, we can safely assume that $f_{\&}$ is strict Archimedean:

$$c = f_{\&}(a,b) \Leftrightarrow f(c) = f(a) + f(b).$$

Thus, invariance means that $f(c) = f(a) + f(b)$ implies $f(\lambda \cdot c) = f(\lambda \cdot a) + f(\lambda \cdot b)$. So, for every λ, the transformation $T : f(a) \to f(\lambda \cdot a)$ is additive: $T(A + B) = T(A) + T(B)$.

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 $f(a)$, this implies $f(a) = C \cdot a^{-\alpha}$; see, e.g., [29]. So, $f(c) = f(a) + f(b)$ implies $C \cdot c^{-\alpha} = C \cdot a^{-\alpha} + C \cdot b^{-\alpha}$, and $c = f_{\&}(a,b) = (a^{-\alpha} + b^{-\alpha})^{-1/\alpha}$.

Deriving Student distribution. We want to maximize the degree

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

The function $f(a)$ is decreasing. So, maximizing $f_{\&}(\mu(x_1), \ldots)$ is equivalent to minimizing the sum $(\mu(x_1))^{-\alpha} + (\mu(x_2))^{-\alpha} + \ldots$ In the limit, this sum tends to

$I \stackrel{\text{def}}{=} \int (\mu(x))^{-\alpha}\,dx$. So, we minimize I under constrains $\int x \cdot \rho(x)\,dx = 0$ and $\int x^2 \cdot \rho(x)\,dx = \sigma^2$, where $\rho(x) = \dfrac{\mu(x)}{\int \mu(y)\,dy}$. Thus, we minimize $\int (\mu(x))^{-\alpha}\,dx$ under 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 leads to minimizing

$$\int (\mu(x))^{-\alpha}\,dx + \lambda_1 \cdot \int x \cdot \mu(x)\,dx +$$

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

Equating the derivative w.r.t. $\mu(x)$ to 0, we get:

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

Thus, $\mu(x) = (a_0 + a_1 \cdot x + a_2 \cdot x^2)^{-\nu}$.

For $\rho(x) = c \cdot \mu(x)$, we get $\rho(x) = c \cdot (a_0 + a_1 \cdot x + a_2 \cdot x^2)^{-\nu}$. So, $\rho(x) = c \cdot (a_2 \cdot (x - x_0)^2 + c_1)^{-\nu}$. This $\rho(x)$ is symmetric w.r.t. x_0, so, the mean is x_0. We know that the mean is 0, so $x_0 = 0$, and $\rho(x) = \text{const} \cdot (1 + a_2 \cdot x_2)^{-\nu}$: exactly Student's $\rho_S(x)$!

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

The author is greatly thankful to Svetlana Prokopchina for the encouragement.

References

[1] J. Aczél, *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.

[2] J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 2008.

[3] H. Alkhatib, B. Kargoll, I. Neumann, and V. Kreinovich, "Normalization-invariant fuzzy logic operations explain empirical success of Student distributions in describing measurement uncertainty", In: P. Melin, O. Castillo, J. Kacprzyk, M. Reformat, and W. Melek (eds.), *Fuzzy Logic in Intelligent System Design: Theory and Applications*, Springer Verlag, Cham, Switzerland, 2018, pp. 300–306.

[4] B. Al-Omari and M. Darter, "Relationship between International Roughness Index and Present Serviceability Rating", *Transportation Research Record*, 1994, No. 1435, pp. 130–136.

[5] American Association of State Highway and Transportation Officials (AASHTO), *AASHTO Guide for Designing Pavement Structures*, AASHTO, Washington, DC, 1993.

[6] American Society for Testing and Materials (ASTM), *Standard Terminology Relating to Vehicle-Pavement Systems*, ASTM Standard E867-06, West Conshohocken, Pennsylvania, 2012.

[7] ASTM International, *Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys*, International Standard D6433-18, 2018.

[8] S. A. Arhin, E. C. Noel, and A. Ribbiso, "Acceptable International Roughness Index thresholds based on Present Serviceability Rating", *Journal of Civil Engineering Research*, 2015, Vol. 5, No. 4, pp. 93–96.

[9] Federal Highway Administration (FHWA), *Highway Performance Monitoring System (HPMS) Field Manual*, US Department of Transportation, Washington, DC, 2015.

[10] Federal Highway Administration (FHWA), *Highway Performance Monitoring System (HPMS) Field Manual*, US Department of Transportation, Washington, DC, 2016.

[11] Federal Highway Administration (FHWA), *National Performance Measurement Measures: Highway Safety Improvement Program*, Title 23 Code of Federal Regulations (CFR) Part 490, US Department of Transportation, Washington, DC, 2016, with changes made in 2017 and 2018.

[12] 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.

[13] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[14] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus, and Giroux, New York, 2011.

[15] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[16] V. G. Knorring, Y. Ya. Kreinovich, and V. D. Mazin, "Measurement Information: Scales and Conversions", *Measurement Techniques*, 2002, Vol. 45. No. 2, pp. 113–115 (Russian version February 2002, pp. 3-4).

[17] I. N. Krotkov, V. Kreinovich, and V. D. Mazin, "Methodology of designing measuring systems, using fractionally linear transformations," *Measuring Systems. Theory and Applications*, Novosibirsk Electrical Engineering Institute, Novosibirsk, Russia, 1986, pp. 5–14 (in Russian).

[18] I. N. Krotkov, V. Kreinovich, and V. D. Mazin, "General form of measurement transformations which admit the computational methods of metrological analysis of measuring-testing and measuring-computing systems," *Izmeritelnaya Tekhnika*, 1987, No. 10, pp. 8–10 (in Russian); English translation: *Measurement Techniques*, 1987, Vol. 30, No. 10, pp. 936–939.

[19] J. Lorkowski and V. Kreinovich, "Fuzzy Logic Ideas Can Help in Explaining Kahneman and Tversky's Empirical Decision Weights", *Proceedings of the 4th World Conference on Soft Computing*, Berkeley, California, May 25–27, 2014, pp. 285–289.

[20] J. Lorkowski and V. Kreinovich, "Granularity Helps Explain Seemingly Irrational Features of Human Decision Making", In: W. Pedrycz and S.-M. Chen (eds.), *Granular Computing and Decision-Making: Interactive and Iterative Approaches*, Springer Verlag, Cham, Switzerland, 2015, pp. 1–31.

[21] J. Lorkowski and V. Kreinovich, "Fuzzy Logic Ideas Can Help in Explaining Kahneman and Tversky's Empirical Decision Weights", In: L. Zadeh et al. (Eds.), *Recent Developments and New Direction in Soft-Computing Foundations and Applications*, Springer Verlag, 2016, pp. 89–98.

[22] J. Lorkowski and V. Kreinovich, *Bounded Rationality in Decision Making Under Uncertainty: Towards Optimal Granularity*, Springer Verlag, Cham, Switzerland, 2018.

[23] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.

[24] V. D. Mazin and V. Kreinovich, *An important property of fractional-linear transformation functions,* Leningrad Polytechnical Institute and National Research Institute for Scientific and Technical Information (VINITI), 1988, 13 pp. (in Russian).

[25] V. D. Mazin and V. Kreinovich, "A Universal Sensor Model", *Proceedings of the 12th International Conference Sensor'2005*, Nuremberg, Germany, May 10–12, 2005, pp. 317–322.

[26] Metropolitan Transportation Commission (MTC), *MTC Rater Certification Exam*, Streetsaver Academy, San Francisco, California, 2018.

[27] Minnesota Department of Transportation, *An Overview of Mn/DOT's Pavement Condition Rating Procedures and Indices*, Technical Report, 2003.

[28] H. T. Nguyen, V. Kreinovich, C. Baral, and V. D. Mazin, "Group-Theoretic Approach as a General Framework for Sensors, Neural Networks, Fuzzy Control, and Genetic Boolean Networks", *Proceedings of the 10th IMEKO TC7 International Symposium on Advances of Measurement Science*, St. Petersburg, Russia, June 30 – July 2, 2004, Vol. 1, pp. 65–70.

[29] 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.

[30] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.

[31] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[32] P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).

[33] A. I. Orlov, "How often are the observations normal?", *Industrial Laboratory*, 1991, Vol. 57, No. 7, pp. 770–772.

[34] Pavement Tools Consortium, a partnership between several state Department of Transportation (DoTs), the Federal Highway Administration (FHWA), and the University of Washington, *Poughness*, https://www.pavementinteractive.org/reference-desk/pavement-management/pavement-evaluation/roughness/. accessed on September 5, 2019.

[35] J. R. Prasad, S. Kanuganti, P. N. Bhanegaonkar, A. K. Sarkar, and S. Arkatkar, "Development of relationship between roughness (IRI) and visible surface distresses: A Study on PMGSY Roads", In: *Proceedings of the 2nd Conference of Transportation Research Group of India (2nd CTRG)*, *Procedia – Social and Behavioral Sciences*, 2013, Vol. 104, pp. 322–331.

[36] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Varlag, New York, 2007.

[37] E. D. Rodriguez Velasquez, C. M. Chang Albitres, and V. Kreinovich, "Measurement-type 'calibration' of expert estimates improves their accuracy and their usability: pavement engineering case", *Proceedings of the IEEE Symposium on Computational Intelligence for Engineering Solutions CIES'2018*, Bengaluru, India, November 18–21, 2018.

[38] S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, New York, 2005.

[39] E. D. Rodriguez Velasquez, C. M. Chang Albitres, and V. Kreinovich, "Relationship between measurement results and expert estimates of cumulative quantities, on the example of pavement roughness", *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence SSCI'2019*, Xiamen, China, December 6–9, 2019, pp. 881–884.

[40] M. W. Sayers, T. D. Gillespie, and C. A. V. Querioz, *The International Road Roughness Experiment: Establishing Correlation and a Calibration Standard for Measurements*, World Bank Technical Paper No. 45, Washington, DC, 1986.

[41] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[42] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.

A Auxiliary Results for Section 2

First auxiliary result: why 50%? In the MTC procedure, as the first threshold, we consider the accuracy with which we should have at least 50% of the measurements. In other words, we compare the median of the empirical distribution with some threshold. But why 50%? Why not select a value corresponding to, say, 40% or 60%?

The only explanation that MTC provides is that selecting 50% leads to empirically the best results. But why? Here is our explanation.

We want to find a parameter describing how distribution of expert's approximation errors. This may be the standard deviation, this may be some other appropriate parameter. We want the relative accuracy with which we determine this parameters to be as good as possible.

We estimate this parameter based on a frequency f that corresponds to some probability p. It is known (see, e.g., [41]) that, after n observations, $f - p$ is approximately normally distributed, with 0 mean and

$$\sigma[p] = \sqrt{\frac{p \cdot (1-p)}{n}}.$$

We can measure the relative accuracy both:

- with respect to the probability p of the original event and
- with respect to the probability $1 - p$ of the opposite event.

We want both relative accuracies to be as small as possible. The relative accuracy with which we can find the desired probability p is equal to

$$\frac{\sigma[p]}{p} = \sqrt{\frac{1-p}{n\cdot p}} = \sqrt{\frac{1}{n}\cdot\left(\frac{1}{p}-1\right)}.$$

Similarly, the relative accuracy with which we can find the probability $1-p$ is equal to

$$\frac{\sigma[p]}{1-p} = \sqrt{\frac{p}{n\cdot(1-p)}} = \sqrt{\frac{1}{n}\cdot\left(\frac{1}{1-p}-1\right)}.$$

We need to make sure that the largest of these two values is as small as possible. One can check that the largest of these two values is

$$\sqrt{\frac{1}{n}\cdot\left(\max\left(\frac{1}{p},\frac{1}{1-p}\right)-1\right)} =$$

$$\sqrt{\frac{1}{n}\cdot\left(\frac{1}{\min(p,1-p)}-1\right)}.$$

This expression is a decreasing function of $\min(p,1-p)$. Thus, for the relative standard deviation to be as small as possible, $\min(p,1-p)$ must be as large as possible.

This expression grows from 0 to 0.5 when p increases from 0 to 0.5, then decreases to 0. Thus, its maximum is attained when $p = 0.5$ – and this is exactly what MTC recommends. So, we have a theoretical explanation for this empirically successful recommendation.

Why 88%. There are many different independent reasons why an expert estimate may differ from the actual value, so the expert uncertainty can be represented as a sum of a large number of small independent random variables. It is known – see, e.g., [41] – that, under reasonable condition, the distribution of such a sum is close to normal. This result is known as the Central Limit Theorem. Thus, we can safely assume that the distribution of expert uncertainty is normal.

For a normal distribution with 0 mean, if the probability for the value to be within ± 8 is 50%, then the probability for the value to be within ± 18 is indeed close to 88%. This explains the second part of the MTC test.

Comment. In both cases, our explanations seem to be simple and natural. We would not be surprised if it turns out that, when selecting the corresponding numbers, the authors of the MTC test were inspired not only by the empirical evidence, but also by similar simple theoretical ideas.