Comparing intervals and moments for the quantification of coarse information

M. Beer
Institute for Risk & Uncertainty
University of Liverpool, Liverpool, UK

V. Kreinovich
Department of Computer Science
University of Texas at El Paso, El Paso, TX, USA

ABSTRACT: In this paper the problem of the most appropriate modeling of scarce information for an engineering analysis is investigated. This investigation is focused on a comparison between a rough probabilistic modeling based on the first two moments and interval modeling. In many practical cases, the available information is limited to such an extent that a more thorough modeling cannot be pursued. The engineer has to make a decision regarding the modeling of this limited and coarse information so that the results of the analysis provide the most suitable basis for conclusions. We approach this problem from the angle of information theory and propose to select the most informative model (in Shannon's sense). The investigation reveals that the answer to question of model choice depends on the confidence, which is needed for the engineering results in order to make informed decisions.

1 INTRODUCTION

Practical engineering problems often involve the quantification of uncertainty on the basis of quite limited information. This is associated with two problems. First, it is desirable to represent the available information as completely as possible in the theoretical uncertainty model. Second, the modeling should not rely on assumptions, which cannot be justified and so introduce artificial information. These issues need to be addressed properly in order to arrive at realistic results in a structural, reliability or other analysis. In practice, a complete satisfaction of both requirements is virtually impossible, but a good balance can be obtained in many cases. If the available information is very limited, the specification of probability distributions is critical, but coarse models may still be suitable. In the probabilistic framework, one may then estimate the first two moments to work towards a rough probabilistic approximation solution. Alternatively, one could estimate two bounds to form the range of the corresponding variable and thus, generate an interval.

These two options were compared in (Beer et al. 2013) from a practical point of view in a geotechnical context. They were examined in view of (i) an appropriate modeling of the information actually available in practical cases, (ii) the transfer of the uncertainty to the computational results, and (iii) the interpretation of the results. Specific emphasis was put on the interpretability of the results as the basis to derive informed engineering decisions if only scarce information is available. It was found that interval modeling and fuzzy modeling provide certain advantages if the available information is very limited. These models cover the worst case scenario for the input parameters. In a reliability analysis this provides information regarding the magnitude of possible exceedance of the limit state as an indication for the severity of the worst case. Emphasizing extreme events, interval analysis can be quite helpful when low-probability-but-high-consequence events are concerned in a risk assessment, whilst probabilistic methods may fail to identify these events. The feature of covering the worst case (conditional on the requirements set to specify the interval bounds) makes interval results quite conservative. The degree of conservatism is comparable with conclusions based on Chebyshev's inequality, and which approach is more conservative depends on the problem. The difference between results from interval analysis and probabilistic analysis is controlled by the assumption of the probabilistic model for the latter analysis. This difference may be only small if the assumed probabilistic model leads to larger failure probabilities than other probabilistic models plausible according to the available information. As the critical regions of the limit state surface are, in many practical cases, particularly affected by the assumption of the probabilistic models for the input variables, it was concluded that it may be advisable to consider an interval solution if probabilistic input information is very vague or only available in a bounded form. The application of very rough probabilistic approximations such as Chebyshev's inequality may lead to even more conservative conclusions compared to interval analysis, whilst the interval analysis results still satisfy the requirement of revealing the worst case on a pre-defined confidence level.

On the basis of the identified features of interval analysis in comparison with a probabilistic analysis in the case of limited information a rough recommendation was provided to decide whether to model the problem using intervals or to employ probabilistic models. The borderline case is determined by the specification of the first two moments of a variable with sufficient confidence. If this is not possible, intervals provide the better alternative. From a statistical perspective, this conclusion hinges on a sufficiently large sample size for variance estimation. Since this depends on the problem and on the experts' judgment, a general cut-off sample size cannot be defined. Also, if expert knowledge is available to pursue a Bayesian estimate, it depends on the quality of the expert knowledge expressed as prior distributions. It was concluded that a sample size of less than seven may be understood as a strong indication towards using intervals, and that a sample size beyond 30 should be sufficient, for most practical cases, to specify a probabilistic model. For cases in between, advice was given to use both approaches and to go with the more conservative decision from these options.

This provides some support to the engineer to make the most appropriate model choice in the specific case, but it still leaves some remaining indeterminacy in "intermediate" cases. In order to address the remaining indeterminacy, we investigate the model alternatives in view of their information content. A natural idea is to select the most informative approach, i.e., that approach in which we need the smallest amount of additional information (in Shannon's sense) to obtain the full information about the situation. We follow this idea and come up with the following conclusion: in practical situations in which a 95% confidence level is sufficient, interval bounds are more informative; however, in situations in which we need higher confidence, a probabilistic model based on the first two moments is more informative.

2 AMOUNT OF INFORMATION

In this paper, we describe a general approach, in which we compare the information contained in two alternative representations – crudely speaking, by simply counting the bits. In this approach, all parts of missing information are (implicitly) assumed to be of equal importance. In specific applications, we may care more about about some parts of the missing information and less about other parts. In such applications, when deciding which representation is the best, we may need into the account the relative value of different parts of information.

In order to specify a most realistic uncertainty model, we need to find the values of several parameters describing this model. From a statistical perspecive; the more parameters we need to determine, the more observations we need to find the values of all these parameters (Sheskin 2007). In practice, it is rarely possible to have many calibrations, so that we are often limited to determine only two parameters; see, e.g., (Rabinovich 2005). Usually, the parameters that we select are:

- either the first two moments of the distribution (or, equivalently, the mean E and the standard deviation σ),
- or the smallest and the largest values, i.e., the range $[\underline{x}, \overline{x}]$ of possible values.

Crudely speaking, moments correspond to the statistical approach to uncertainty, while the range corresponds to the interval approach to uncertainty; see, e.g., (Jaulin, Kieffer, Didrit, & Walter 2001, Moore, Kearfott, & Cloud 2009).

In order to investigate the amount of information contained in these two alternative representations we employ the concept of Shannon's entropy. The traditional Shannon's notion of the amount of information is based on defining information as the (average) number of "yes"-"no" (binary) questions that we need to ask so that, starting with the initial uncertainty, we will be able to completely determine the object.

After each binary question, we can have 2 possible answers. So, if we ask q binary questions, then, in principle, we can have 2^q possible results. Thus, if we know that our object is one of n objects, and we want to uniquely pinpoint the object after all these questions, then we must have $2^q \geq n$. In this case, the smallest number of questions is the smallest integer q that is $\geq \log_2(n)$. This smallest number is called a *ceiling* and denoted by $\lceil \log_2(n) \rceil$.

For discrete probability distributions, we get the standard formula for the average number of questions $-\sum p_i \cdot \log_2(p_i)$. For the continuous case, we can estimate the average number of questions that are needed to find an object with a given accuracy ε – i.e., divide the whole original domain into sub-domains of radius ε and diameter 2ε.

For example, if we start with an interval $[a, b]$ of width $b - a$, then we need to subdivide it into

$$n \sim \frac{b-a}{2\varepsilon}$$

sub-domains, so we must ask

$$\log_2(n) \sim \log_2(b-a) - \log_2(\varepsilon) - 1$$

questions. In the limit, the term that does not depend on ε leads to $\log_2(b-a)$. For continuous probability distributions, we get the standard Shannon's expression $\log_2(n) \sim S - \log_2(2\varepsilon)$, where

$$S = -\int \rho(x) \cdot \log_2 \rho(x)\, dx.$$

Let us describe this idea in more detail.

Discrete case: no information about probabilities. Let us start with the simplest situation when we know that we have n possible alternatives $A_1, \ldots, A_n$, and we have no information about the probability (frequency) of different alternatives. Let us show that in this case, the smallest number of binary questions that we need to determine the alternative is indeed $q \overset{\text{def}}{=} \lceil \log_2(n) \rceil$.

We have already shown that the number of questions cannot be smaller than $\lceil \log_2(n) \rceil$; so, to complete the derivation, we need to show that it is sufficient to ask q questions. Indeed, let's enumerate all n possible alternatives (in arbitrary order) by numbers from 0 to $n-1$, and write these numbers in the binary form. Using q binary digits, one can describe numbers from 0 to $2^q - 1$. Since $2^q \geq n$, we can describe each of the n numbers by using only q binary digits. So, to uniquely determine the alternative A_i out of n given ones, we can ask the following q questions: "is the first binary digit 0?", "is the second binary digit 0?", etc, up to "is the q-th digit 0?".

Case of a discrete probability distribution. Let us now assume that we also know the probabilities $p_1, \ldots, p_n$ of different alternatives $A_1, \ldots, A_n$. If we are interested in an individual selection, then the above arguments show that we cannot determine the actual alternative by using fewer than $\log_2(n)$ questions. However, if we have many (N) similar situations in which we need to find an alternative, then we can determine all N alternatives by asking $\ll N \cdot \log_2(n)$ binary questions.

To show this, let us fix i from 1 to n, and estimate the number of events N_i in which the output is i. This number N_i is obtained by counting all the events in which the output was i, so

$$N_i = n_{i1} + n_{i2} + \ldots + n_{iN},$$

where n_k equals to 1 if in k-th event the output is i and 0 otherwise. The average $E(n_{ik})$ of n_{ik} equals to $p_i \cdot 1 + (1-p_i) \cdot 0 = p_i$. The mean square deviation $\sigma[n_{ik}]$ is determined by the formula

$$\sigma^2[n_{ik}] = p_i \cdot (1 - E(n_{ik}))^2 + (1-p_i) \cdot (0 - E(n_{ik}))^2.$$

If we substitute here $E(n_{ik}) = p_i$, we get $\sigma^2[n_{ik}] = p_i \cdot (1-p_i)$. The outcomes of all these events are considered independent, therefore n_{ik} are independent random variables. Hence the average value of N_i equals to the sum of the averages of n_{ik}:

$$E[N_i] = E[n_{i1}] + E[n_{i2}] + \ldots + E[n_{iN}] = N \cdot p_i.$$

The mean square deviation $\sigma[N_i]$ satisfies a likewise equation

$$\sigma^2[N_i] = \sigma^2[n_{i1}] + \sigma^2[n_{i2}] + \ldots = N \cdot p_i \cdot (1-p_i),$$

so $\sigma[N_i] = \sqrt{p_i \cdot (1-p_i) \cdot N}$.

For big N the sum of equally distributed independent random variables tends to a Gaussian distribution (the well-known *Central Limit Theorem*), therefore for big N, we can assume that N_i is a random variable with a Gaussian distribution. Theoretically a random Gaussian variable with the average a and a standard deviation σ can take any value. However, in practice, if, e.g., one buys a voltmeter with guaranteed 0.1V standard deviation, and it gives an error 1V, it means that something is wrong with this instrument. Therefore it is assumed that only some values are practically possible. Usually a "k-sigma" rule is accepted that the real value can only take values from $a - k_0 \cdot \sigma$ to $a + k_0 \cdot \sigma$, where k_0 is 2, 3, or 4. So in our case we can conclude that N_i lies between

$$N \cdot p_i - k_0 \cdot \sqrt{p_i \cdot (1-p_i) \cdot N}$$

and

$$N \cdot p_i + k_0 \cdot \sqrt{p_i \cdot (1-p_i) \cdot N}.$$

Now we are ready for the formulation of Shannon's result. In this quality control example, the choice of the parameter k_0 matters, but, as we'll see, in our case the results do not depend on k_0 at all.

- Let a real number $k > 0$ and a positive integer n be given. The number n is called *the number of outcomes*.

- By a *probability distribution*, we mean a sequence $\{p_i\}$ of n real numbers, $p_i \geq 0$, $\sum p_i = 1$. The value p_i is called a *probability* of i-th event.

- Let an integer N is given; it is called *the number of events*.

- By a *result of N events* we mean a sequence r_k, $1 \leq k \leq N$ of integers from 1 to n. The value r_k is called the *result of k-th event*.

- The total number of events that resulted in the i-th outcome will be denoted by N_i.

- We say that the result of N events is *consistent* with the probability distribution $\{p_i\}$ if for every i, we have $N \cdot p_i - k_0 \cdot \sigma_i \leq N_i \leq N \cdot p_i + k_0 \cdot \sigma_i$, where
$$\sigma_i \overset{\text{def}}{=} \sqrt{p_i \cdot (1-p_i) \cdot N}.$$

- Let's denote the number of all consistent results by $N_{\text{cons}}(N)$.

- The number $\lceil \log_2(N_{\text{cons}}(N))\rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$.

- The fraction $\dfrac{Q(N)}{N}$ will be called the *average number of questions.*

- The limit of the average number of questions when $N \to \infty$ will be called the *information.*

When the number of events N tends to infinity, the average number of questions tends to

$$S(p) \stackrel{\text{def}}{=} -\sum p_i \cdot \log_2(p_i).$$

This Shannon's result says that if we know the probabilities of all the outputs, then the average number of questions that we have to ask in order to get a complete knowledge equals to the entropy of this probabilistic distribution. As expected, this average number of questions does not depend on the threshold k_0. Since we somewhat modified Shannon's definitions, we cannot use the original Shannon's proof and refer to (Kreinovich and Xiang 2010) for the new proof relevant to the present investigation.

Case of a continuous probability distribution. After a finite number of "yes"-"no" questions, we can only distinguish between finitely many alternatives. If the actual situation is described by a real number, then, since there are infinitely many different possible real numbers, after finitely many questions, we can only get an approximate value of this number.

Once we fix the accuracy $\varepsilon > 0$, we can talk about the number of questions that are necessary to determine a number x with this accuracy ε, i.e., to determine an approximate value r for which $|x - r| \le \varepsilon$.

Once an *approximate* value r is determined, possible *actual* values of x form an interval $[r - \varepsilon, r + \varepsilon]$ of width 2ε. Vice versa, if we have located x on an interval $[\underline{x}, \overline{x}]$ of width 2ε, this means that we have found x with the desired accuracy ε: indeed, as an ε-approximation to x, we can then take the midpoint $\dfrac{\underline{x} + \overline{x}}{2}$ of the interval $[\underline{x}, \overline{x}]$.

Thus, the problem of determining x with the accuracy ε can be reformulated as follows: we divide the real line into intervals $[x_i, x_{i+1}]$ of width 2ε (so that $x_{i+1} = x_i + 2\varepsilon$), and by asking binary questions, find the interval that contains x. As we have shown, for this problem, the average number of binary question needed to locate x with accuracy ε is equal to $S = -\sum p_i \cdot \log_2(p_i)$, where p_i is the probability that x belongs to i-th interval $[x_i, x_{i+1}]$.

In general, this probability p_i is equal to $\int_{x_i}^{x_{i+1}} \rho(x)\,dx$, where $\rho(x)$ is the probability distribution of the unknown values x. For small ε, we have $p_i \approx 2\varepsilon \cdot \rho(x_i)$, hence $\log_2(p_i) = \log_2(\rho(x_i)) + \log_2(2\varepsilon)$. Therefore, for small ε, we have

$$S = -\sum \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot 2\varepsilon -$$

$$\sum \rho(x_i) \cdot 2\varepsilon \cdot \log_2(2\varepsilon).$$

The first sum in this expression is the integral sum for the integral

$$S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(x)\,dx$$

(this integral is called the *entropy* of the probability distribution $\rho(x)$); so, for small ε, this sum is approximately equal to this integral (and tends to this integral when $\varepsilon \to 0$). The second sum is a constant $\log_2(2\varepsilon)$ multiplied by an integral sum for the interval $\int \rho(x)\,dx = 1$. Thus, for small ε, we have

$$S \approx -\int \rho(x) \cdot \log_2(x)\,dx - \log_2(2\varepsilon).$$

So, the average number of binary questions that are needed to determine x with a given accuracy ε, can be determined if we know the entropy of the probability distribution $\rho(x)$.

Partial information about probability distribution: discrete case. In many real-life situations, as we have mentioned, instead of having *complete* information about the probabilities $p = (p_1, \ldots, p_n)$ of different alternatives, we only have *partial* information about these probabilities – i.e., we only know a *set* P of possible values of p.

If it is possible to have $p \in P$ and $p' \in P$, then it is also possible that we have p with some probability α and p' with the probability $1 - \alpha$. In this case, the resulting probability distribution $\alpha \cdot p + (1 - \alpha) \cdot p'$ is a convex combination of p and p'. Thus, it it reasonable to require that the set P contains, with every two probability distributions, their convex combinations – in other words, that P is a convex set; see, e.g., (Walley 1991).

- By a *probabilistic knowledge*, we mean a convex set P of probability distributions.

- We say that the result of N events is *consistent* with the probabilistic knowledge P if this result is consistent with one of the probability distributions $p \in P$.

- Let's denote the number of all consistent results by $N_{\text{cons}}(N)$.

- The number $\lceil \log_2(N_{\text{cons}}(N))\rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$.

- The fraction $\dfrac{Q(N)}{N}$ will be called the *average number of questions.*

- The limit of the average number of questions when $N \to \infty$ will be called the *information*.

By the *entropy* $S(P)$ of a probabilistic knowledge P, we mean the largest possible entropy among all distributions $p \in P$; $S(P) \stackrel{\text{def}}{=} \max\limits_{p \in P} S(p)$. When the number of events N tends to infinity, the average number of questions tends to the entropy $S(P)$; this proposition was also first proved in (Kreinovich and Xiang 2010). It is worth mentioning that when N goes to infinity, the probability set reduces to a single element – namely, to the distribution related to the long-run relative frequencies.

Partial information about probability distribution: continuous case. In the continuous case, we also often encounter situations in which we only have partial information about the probability distribution. In such situations, instead of a knowing the *exact* probability distribution $\rho(x)$, we only know a (convex) class P that contains the (unknown) distribution. In such situations, we can similarly ask about the average number of questions that are needed to determine x with a given accuracy ε.

Once we fix an accuracy ε and a subdivision of the real line into intervals $[x_i, x_{i+1}]$ of width 2ε, we have a discrete problem of determining the interval containing x. For this discrete problem, the average number of "yes"-"no" questions is equal to the largest entropy $S(p)$ among all the corresponding discrete distributions $p_i = \int_{x_i}^{x_{i+1}} \rho(x)\,dx$. As we have mentioned, for small ε, we have $S(p) \sim S(\rho) - \log_2(2\varepsilon)$, where $S(\rho) = -\int \rho(x) \cdot \log_2(\rho(x))\,dx$ is the entropy of the corresponding continuous distribution. Thus, the largest discrete entropy $S(p)$ comes from the distribution $\rho(x) \in P$ for which the corresponding (continuous) entropy $S(\rho)$ attains the largest possible value.

3 ANALYSIS OF THE PROBLEM

We want to find out which of the two representations is more informative: a representation by the first two moments (or, equivalently, by the mean E and standard deviation σ) and a representation by an interval $[\underline{x}, \overline{x}]$. For both representations, in order to uniquely determine the actual value x, we need to gather additional information. So, in which of these two representations do we need to gather more information?

Toward a reformulation of the problem in precise terms. As we have mentioned in the previous section, the amount of information can be naturally gauged by the average number of questions that we need to ask to determine the actual situation. According to the above results, once we know the class P of possible probability distributions, this average number of questions $S(P)$ can be determined as the largest entropy $S(\rho)$ of all probability distributions ρ from the given class P. So, to answer our question, it is sufficient to compare the values $S(P)$ corresponding to the two representations.

To make a comparison, we need to relate the bounds $\underline{x}$ and $\overline{x}$ with the values E and σ. In the case of normal distribution, with confidence 95%, the actual value of the random variable x is contained in the confidence interval $[E - 2\sigma, E + 2\sigma]$. With confidence 99.9%, the actual value is contained in the interval $[E - 3\sigma, E + 3\sigma]$. With confidence $1 - 10^{-8}$, the actual value is in the six-sigma interval $[E - 6\sigma, E + 6\sigma]$; see, e.g., (Rabinovich 2005, Sheskin 2007). Thus, it makes sense to consider an interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$, for some appropriate value k_0.

In many practical problems, the two-sigma level of confidence is reasonable. The corresponding 5% level is a threshold that is used in many practical applications – to decide when a new medicine is better than the previous one, to decide whether the new medicine or, more generally, a new strategy has an effect, to decide whether a new theory is confirmed by observations, etc.; see, e.g., (Sheskin 2007).

However, there are problems in which a higher level of confidence is needed. For example, in the design of civil engineering structures, when a technical problem can lead to collapse with fatalities, we need at least 3σ level corresponding to $< 0.1\%$ probability of errors. In chip design, the confidence in individual chip elements should be even higher: the reliability of computer means that all the cells are reliable, and to make sure that all millions of cells work correctly, we need to make sure that the probability of failure of an individual cell is $\ll 10^{-6}$. In such situations, the six-sigma level of confidence is used.

For Gaussian distributions, it makes sense to take $k_0 = 2$, $k_0 = 3$, or $k_0 = 6$, depending on the confidence level with which we want to bound the possible values. As we have mentioned, in practice, the distribution is often non-Gaussian. In this case, we may have *heavy tails*, i.e., distributions for which the probability of high deviations is much larger than for the Gaussian distribution. In this case, to cover all possible values of x with a given confidence, we need to consider larger values k_0.

On this basis we can perform the necessary computations. When we look for the distribution with the largest entropy in a given class, a natural way to find the largest value is to differentiate the expression for the entropy and to equate the corresponding derivative to 0. From this viewpoint, instead of the binary logarithms $\log_2(x)$, it is more convenient to use natural logarithms $\ln(x)$, because the natural logarithm is easier to differentiate: its derivative is $\dfrac{1}{x}$. Since $\log_2(x) = \dfrac{\ln(x)}{\ln(2)}$, these two logarithms – and thus, the corresponding values of entropy – differ by a constant factor. When we compare two entropies, multiplying both by a positive constant does not change which one is better. With this in mind, in the following text, we will use a version of Shannon's entropy that uses natural logarithms.

Estimating $S(P)$: interval case. Let us start with the interval case, when all we know is that the actual value x belongs to the interval $[\underline{x}, \overline{x}] = [E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$. In this case, the class P consists of all possible probability distributions $\rho(x)$ which are located on this interval, i.e., for which $\rho(x) = 0$ for all values x outside this interval.

It is known that in this class, the distribution ρ with the largest entropy $S(\rho)$ is the uniform distribution; see, e.g., (Jaynes 2003). Indeed, we need to maximize the entropy $S(\rho) = -\int \rho(x) \cdot \ln(\rho(x))\,dx$ under the constraints $\rho(x) \geq 0$ and $\int \rho(x)\,dx = 1$. The unknown here are the values $\rho(x)$ corresponding to different points x. We can use Lagrange multiplier method to reduce the constraint optimization problem to the unconstrained optimization problem of maximizing the combination

$$-\int \rho(x) \cdot \ln(\rho(x))\,dx + \lambda \cdot \left(\int \rho(x)\,dx - 1\right)$$

for an appropriate value λ. Differentiating this objective function with respect to $\rho(x)$ and equating the derivative to 0, we get $-\ln(\rho(x)) - 1 + \lambda = 0$, hence $\ln(\rho(x)) = 1 - \lambda$ and $\rho(x) = \exp(1 - \lambda)$. This value is the same for all x, so this is indeed a uniform distribution.

From the condition $\int_{\underline{x}}^{\overline{x}} \rho(x)\,dx = 1$ (that the total probability is 1) we conclude that $(\overline{x} - \underline{x}) \cdot \rho(x) = 1$, hence $\rho(x) = \dfrac{1}{\overline{x} - \underline{x}}$. For this probability distribution, the entropy has the form

$$-\int_{\underline{x}}^{\overline{x}} \rho(x) \cdot \ln(\rho(x)) = \int_{\underline{x}}^{\overline{x}} \frac{1}{\overline{x} - \underline{x}} \cdot \ln(\overline{x} - \underline{x})\,dx =$$

$$(\overline{x} - \underline{x}) \cdot \frac{1}{\overline{x} - \underline{x}} \cdot \ln(\overline{x} - \underline{x}) = \ln(\overline{x} - \underline{x}).$$

Describing the range in terms of E and σ, we conclude that in the interval case,

$$S_{\text{int}}(P) = \ln(2 \cdot k_0 \cdot \sigma) = \ln(\sigma) + \ln(2 \cdot k_0).$$

Estimating $S(P)$: case of moments. In the moments case, the class P consists of all probability distributions with given first and second moments $E = \int x \cdot \rho(x)\,dx$ and $M = E^2 + \sigma^2 = \int x^2 \cdot \rho(x)\,dx$.

It is known that in this class, the distribution ρ with the largest entropy $S(\rho)$ is the normal distribution; see, e.g., (Jaynes 2003). Indeed, we need to maximize the entropy $S(\rho) = -\int \rho(x) \cdot \ln(\rho(x))\,dx$ under the constraints $\rho(x) \geq 0$, $\int \rho(x)\,dx = 1$, $\int x \cdot \rho(x)\,dx = E$, and $\int x^2 \cdot \rho(x)\,dx = M$. We can use Lagrange multiplier method to reduce the constraint optimization problem to the unconstrained optimization problem of maximizing the combination

$$-\int \rho(x) \cdot \ln(\rho(x))\,dx + \lambda_0 \cdot \left(\int \rho(x)\,dx - 1\right) +$$

$$\lambda_1 \cdot \left(\int x \cdot \rho(x)\,dx - E\right) +$$

$$\lambda_2 \cdot \left(\int x^2 \cdot \rho(x)\,dx - M\right)$$

for appropriate values λ_i. Differentiating this objective function with respect to $\rho(x)$ and equating the derivative to 0, we get

$$-\ln(\rho(x)) - 1 + \lambda_0 + \lambda_1 \cdot x + \lambda_2 \cdot x^2 = 0,$$

hence

$$\ln(\rho(x)) = 1 - \lambda_0 - \lambda_1 \cdot x - \lambda_2 \cdot x^2,$$

and $\rho(x)$ is, thus, a Gaussian distribution. Since we know the mean E and the standard deviation, this distribution takes the form

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

Shannon's entropy $S(\rho)$ is an expected value of

$$\psi(x) \stackrel{\text{def}}{=} \ln(\rho(x)).$$

For the above Gaussian distribution,

$$\psi(x) = \ln(\sqrt{2 \cdot \pi} \cdot \sigma) + \frac{1}{2} \cdot \frac{(x - E)^2}{\sigma^2}.$$

Here, E is the mean, so the expected value of $(x - E)^2$ is, by definition, the variance σ^2. Thus, the expected value $S(P)$ of the function $\psi(x)$ takes the form

$$S(P) = \ln(\sqrt{2 \cdot \pi} \cdot \sigma) + \frac{1}{2} \cdot \frac{\sigma^2}{\sigma^2} = \ln(\sqrt{2 \cdot \pi} \cdot \sigma) + \frac{1}{2}.$$

Thus, we arrive at the following expression for $S(P)$ for the moments case:

$$S_{\text{mom}}(P) = \ln(\sigma) + \ln(\sqrt{2 \cdot \pi}) + \frac{1}{2}.$$

Resulting comparison. We want to choose a representation for which the remaining number of binary questions is the smallest possible. Thus, we should select the moments if and only if $S_{\text{mom}}(P) < S_{\text{int}}(P)$. Substituting the above expressions for $S_{\text{mom}}(P)$ and

$S_{\text{int}}(P)$, we conclude that the moments method is better if and only if

$$\ln(\sigma) + \ln(\sqrt{2 \cdot \pi}) + \frac{1}{2} < \ln(\sigma) + \ln(2 \cdot k_0),$$

i.e., if and only if

$$\ln(\sqrt{2 \cdot \pi}) + \frac{1}{2} < \ln(2 \cdot k_0).$$

By applying $\exp(x)$ to both sides of this inequality, we can obtain the following equivalent simpler inequality:

$$\sqrt{2 \cdot \pi} \cdot \sqrt{e} < 2 \cdot k_0,$$

i.e.,

$$k_0 > \sqrt{\frac{\pi \cdot e}{2}} \approx 2.066.$$

So, when $k_0 = 2$, the interval representation is better; when $k_0 \geq 3$, the moments representation is more informative.

The above conclusion is based on the assumption that we select a *symmetric* confidence interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$. In principle, we can consider *asymmetric* confidence intervals $[\underline{x}, \overline{x}]$ corresponding to the same confidence level. For such intervals, the width $\overline{x} - \underline{x}$ is larger than for the symmetric ones; thus, the corresponding value of $S(P) = \ln(\overline{x} - \underline{x})$ is also larger. So, in comparison to such intervals, the moments representation may be better.

For normal distributions, $k_0 = 2$ corresponds to 95% confidence intervals, meaning that the probability of being further than 2σ from the mean does not exceed 5%. In some practical situations, the distributions are not normal; see, e.g., (Novitskii & Zograph 1991, Orlov 1991). For example, often, we have *heavy-tailed* distributions, in which the probability of large deviations is much larger than for the normal distribution. For such distributions, 95% confidence intervals correspond to $k_0 \gg 2$. Therefore, for such distributions, even when 95% confidence is satisfactory, the moments representation is more informative.

4 CONCLUSIONS

We are interested in selecting the most informative representation. It turns out that from this viewpoint, which of the two representation to use – the moments representation or the interval representation – depends on what is the desired level of confidence. In practical problems in which the probability distribution is close to normal, and 95% confidence is satisfactory, an interval representation is more informative. To be more precise, interval representation is only slightly more informative, but still more informative, and in many situations, when measurements are difficult and we want to extract as much information from them as possible, any possibility to gain additional information is welcome. On the other hand, in problems in which we need higher levels of confidence – or in which we have a heavy-tailed distribution – the moments representation is more informative.

Acknowledgements. This work was partly supported by the National Science Foundation grants HRD-0734825 and DUE-0926721, and by Grant 1 T36 GM078000-01 from the National Institutes of Health. The authors are thankful to all the participants of the Dagstuhl 2011 seminar *Uncertainty Modeling and Analysis with Intervals: Foundations, Tools, Applications* for valuable discussions.

REFERENCES

Beer, M., Y. Zhang, S. T. Quek, & K. K. Phoon (2013). Reliability analysis with scarce information: Comparing alternative approaches in a geotechnical engineering context. *Structural Safety 41*, 1–10.

Jaulin, L., M. Kieffer, O. Didrit, & E. Walter (2001). *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*. London: Springer.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge, Massachusetts: Cambridge University Press.

Kreinovich, V. & G. Xiang (2010). Estimating information amount under uncertainty: algorithmic solvability and computational complexity. *International Journal of General Systems 39*(4), 349–378.

Moore, R., R. Kearfott, & M. Cloud (2009). *, 2009. Introduc-tion to Interval Analysis*. Philadelphia, Pennsylvania: SIAM Press.

Novitskii, P. & I. Zograph (1991). *Estimating the Measurement Errors*. Leningrad: Energoatomizdat. (in Russian).

Orlov, A. (1991). How often are the observations normal? *Industrial Laboratory 57*(7), 770–772.

Rabinovich, S. (2005). *Measurement Errors and Uncertainties: Theory and Practice*. New York: Springer Verlag.

Sheskin, D. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Florida: Chapman & Hall/CRC.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London, New York: Chapman & Hall.