Soft Computing manuscript No.
(will be inserted by the editor)

Interval or Moments: Which Carry More Information?

Michael Beer · Vladik Kreinovich

Received: April 1, 2012/ Accepted: date

Abstract In many practical situations, we do not have enough observations to uniquely determine the corresponding probability distribution, we only have enough observations to estimate two parameters of this distribution. In such cases, the traditional statistical approach is to estimate the mean and the standard deviation. Alternatively, we can estimate the two bounds that form the range of the corresponding variable and thus, generate an interval. Which of these two approaches should we select? A natural idea is to select the most informative approach, i.e., an approach in which we need the smallest amount of additional information (in Shannon's sense) to obtain the full information about the situation. In this paper, 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, the moments approach is more informative.

Keywords intervals · moments · information

M. Beer
Institute for Risk and Uncertainty
School of Engineering
University of Liverpool
Liverpool L69 3BX, UK
E-mail: mbeer@liverpool.ac.uk

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

1 Formulation of the Problem

It is important to take measurement uncertainty into account. Most information about the physical world comes either directly from measurements, or by processing measurement results.

Because of the importance of measurement results, it is extremely important to take into account that the result $\widetilde{x}$ of measuring a quantity x is, in general, different from the actual (unknown) value of this quantity. In other words, we need to take into account that, in general, we have a non-zero *measurement error* $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$; see, e.g., Rabinovich (2005).

How to get information about measurement uncertainty. A usual way to get information about the measurement uncertainty of a given measuring instrument (MI) is to *calibrate* this MI, i.e., to perform several measurements of the same quantity (or quantities) with this MI and with a much more accurate MI that serves as a "standard" (Rabinovich (2005)). Since the standard MI has much higher accuracy than our MI, the value $\widetilde{x}_s$ obtained by this standard MI is much closer to the actual value x that the original measurement result $\widetilde{x}$. Since $|\widetilde{x}_s - x| \ll |\widetilde{x} - x|$, we have $\Delta x = \widetilde{x} - x \approx \widetilde{x} - \widetilde{x}_s$. Thus, the difference $\widetilde{x} - \widetilde{x}_s$ between the measurement results can serve as a good approximation for the (unknown) measurement error Δx.

How to describe measurement uncertainty: ideal case. The more information we have about the measurement error Δx, the better. Ideally, it is desirable to know, for each measuring instrument, which value Δx are possible, and what is the frequency with which different possible values of Δx occur. In precise terms, ideally,

we would like to know the probability distribution on the set of all possible values Δx.

Practical limitations. Real-life measuring instruments have different probability distributions, often non-Gaussian; see, e.g., Novitskii et al. (1991); Orlov (1991). To exactly describe the corresponding distribution, we need to know the values of several parameters describing this distribution.

The more parameters we need to determine, the more observations we need to determine the values of all these parameters (Sheskin (2007)). In practice, it is rarely possible to have many calibrations, so usually, we can determine two parameters; see, e.g., Rabinovich (2005).

Which two parameters should we select: interval or moments? 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 et al. (2001); Moore (2009).

Which of the approaches should we choose?

Practical example. A geotechnical engineering example – in which such a choice is needed – is described, in detail, in Beer et al. (2011). Let us give a brief description of this example. In this example, we want to build a structure on a layered ground: a sand layer on top of the clay layer on top of the solid (rock) layer. The pressure caused by the structure leads to the compression of the clay layer; this compression, in turn, causes the structure's *settlement* – i.e., its vertical downward shift.

The settlement may vary from one part of the structure to the other; so, to avoid structural problems, it is usually required that this settlement is limited to a small amount (e.g., $\leq$ 6-7 cm). To estimate the settlement amount, we need to know such randomly changing parameters as the thickness of the clay layer, the compression index for the clay layer, etc.

For all these parameters, we only have a reasonably small number of measurement results, not enough to determine the actual probability distribution of these parameters. So, we have a choice: we can either estimate the moments (and use statistical methods to process them), or we can estimate the intervals (and use interval techniques to process them).

What we do in this paper. In Beer et al. (2011), the problem of selecting one of the two alternative representations of uncertainty was solved for the above specific geotechnical problem. In this paper, we answer this question in the *general* context, by analyzing which of these two approaches contains more information.

Remark 1 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.

2 How to Gauge Amount of Information: A Brief Reminder

General idea. Let us first recall how the amount of information is usually gauged.

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 $q \stackrel{\text{def}}{=} \lceil \log_2(n) \rceil$.

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.

Remark 2 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.

Definition 1

- 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

$$\sigma_i \stackrel{\text{def}}{=} \sqrt{p_i \cdot (1-p_i) \cdot N}.$$

- 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.*

Proposition 1 *When the number of events N tends to infinity, the average number of questions tends to*

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

Remark 3

- 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 we promised, 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. For reader's convenience, a new proof (first presented in Kreinovich et al. (2010)) is reproduced in the Proofs section.

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).

Definition 2

- 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.*

Definition 3 By the *entropy* $S(P)$ of a probabilistic knowledge P, we mean the largest possible entropy among all distributions $p \in P$; $S(P) \stackrel{\text{def}}{=} \max\limits_{p \in P} S(p)$.

Proposition 2 *When the number of events N tends to infinity, the average number of questions tends to the entropy $S(P)$.*

Remark 4

- This proposition was also first proved in Kreinovich et al. (2010); its proof is reproduced in the Proofs section.
- 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. Due to Proposition 1, 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

A problem: reminder. 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 a manned spaceflight, when a minor technical problem can lead to a disaster, 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.

Now, we are ready to perform the necessary computations.

Remark 5 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 $\frac{1}{x}$. Since $\log_2(x) = \frac{\ln(x)}{\ln(2)}$, 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.

Remark 6

- 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 et al. (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, and to the anonymous referees for useful suggestions.

References

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

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

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

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

Moore RE, Kearfott RB, Cloud MJ (2009) Introduction to Interval Analysis, SIAM Press, Philadelphia, Pennsylvania.

Novitskii PV, Zograph IA (1991) Estimating the Measurement Errors. Energoatomizdat, Leningrad (in Russian).

Orlov AI (1991) How often are the observations normal? Industrial Laboratory 57(7):770–772.

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

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

Walley P (1991) Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, New York.

5 Proofs

Proof of Shannon's result. Let's first fix some values N_i, that are consistent with the given probabilistic distribution. Due to the inequalities that express the consistency demand, the ratio $f_i = \dfrac{N_i}{N}$ tends to p_i as $N \to \infty$. Let's count the total number C of results, for which for every i the number of events with outcome i is equal to this N_i. Once we know C, we will be able to compute $N_{\rm cons}$ by adding these C's.

Actually we are interested not in $N_{\rm cons}$ itself, but in $Q(N) = \lceil \log_2(N_{\rm cons}) \rceil$, and moreover, in $\lim \left(\dfrac{Q(N)}{N} \right)$. So we'll try to estimate not only C, but also $\log_2(C)$ and $\lim \dfrac{\log_2(C)}{N}$.

To estimate C means to count the total number of sequences of length N, in which there are N_1 elements, equal to 1, N_2 elements, equal to 2, etc. It is known that this number is equal to

$$C = \frac{N!}{N_1! \cdot N_2! \cdot \ldots \cdot N_n!}$$

To simplify computations, we can use the well-known Stirling formula

$$k! \sim \left(\frac{k}{e} \right)^k \cdot \sqrt{2\pi \cdot k}.$$

Then, we get

$$C \approx \frac{\left(\dfrac{N}{e} \right)^N \sqrt{2\pi \cdot N}}{\left(\dfrac{N_1}{e} \right)^{N_1} \cdot \sqrt{2\pi \cdot N_1} \cdot \ldots \cdot \left(\dfrac{N_n}{e} \right)^{N_n} \cdot \sqrt{2\pi \cdot N_n}}$$

Since $\sum N_i = N$, terms e^N and e^{N_i} cancel each other.

To get further simplification, we substitute $N_i = N \cdot f_i$, and correspondingly $N_i^{N_i}$ as $(N \cdot f_i)^{N \cdot f_i} = N^{N \cdot f_i} \cdot f_i^{N \cdot f_i}$. Terms N^N is the numerator and

$$N^{N \cdot f_1} \cdot N^{N \cdot f_2} \cdot \ldots \cdot N^{N \cdot f_n} = N^{N \cdot f_1 + N \cdot f_2 + \ldots + N \cdot f_n} = N^N$$

in the denominator cancel each other. Terms with $\sqrt{N}$ lead to a term that depends on N as $c \cdot N^{-(n-1)/2}$. So, we conclude that

$$\log_2(C) \approx -N \cdot f_1 \cdot \log_2(f_1) - \ldots - N \cdot f_n \log_2(f_n) - \frac{n-1}{2} \cdot \log_2(N) - \text{const}.$$

When $N \to \infty$, we have $\dfrac{1}{N} \to 0$, $\dfrac{\log_2(N)}{N} \to 0$, and $f_i \to p_i$, therefore

$$\frac{\log_2(C)}{N} \to -p_1 \cdot \log_2(p_1) - \ldots - p_n \cdot \log_2(p_n),$$

i.e., $\dfrac{\log_2(C)}{N}$ tends to the entropy of the probabilistic distribution.

Now, that we have found an asymptotics for C, let's compute $N_{\rm cons}$ and $\dfrac{Q(N)}{N}$. For a given probabilistic distribution $\{p_i\}$ and every i, possible values of N_i form an interval of length $L_i \stackrel{\rm def}{=} 2k_0 \cdot \sqrt{p_i \cdot (1-p_i)} \cdot \sqrt{N}$. So there are no more than L_i possible values of N_i. The maximum value for $p_i \cdot (1-p_i)$ is attained when $p_i = \dfrac{1}{2}$, therefore $p_i \cdot (1-p_i) \le \dfrac{1}{4}$, and hence $L_i \le 2k \cdot \sqrt{\dfrac{N}{4}} = \dfrac{k_0}{2} \cdot \sqrt{N}$. For every i from 1 to n there are at most $\dfrac{k_0}{2} \cdot \sqrt{N}$ possible values of N_i, so the total number of possible combinations of $N_1, \ldots, N_n$ is smaller than $\left(\dfrac{k_0}{2} \cdot \sqrt{N} \right)^n$. Let us denote this number of combinations by $N(p)$.

The total number $N_{\rm cons}$ of consistent results is the sum of $N(p)$ different values of C (values that correspond to $N(p)$ different combinations of N_1, N_2, …, N_n). Let's denote the biggest of these values C by $C_{\max}$. Since $N_{\rm cons}$ is the sum of $N(p)$ terms, and each of these terms is not larger than the largest of them $C_{\max}$, we conclude that $N_{\rm cons} \le N(p) \cdot C_{\max}$. On the other hand, the sum $N_{\rm cons}$ of non-negative integers is not smaller than the largest of them, i.e., $C_{\max} \le N_{\rm cons}$. Combining these two inequalities, we conclude that

$$C_{\max} \le N_{\rm cons} \le N(p) \cdot C_{\max}.$$

Since $N(p) \le \left(\dfrac{k_0}{2} \cdot \sqrt{N} \right)^n$, we conclude that

$$C_{\max} \le N_{\rm cons} \le \left(\frac{k_0}{2} \cdot \sqrt{N} \right)^n \cdot C_{\max}.$$

Turning to logarithms, we find that

$$\log_2(C_{\max}) \le \log_2(N_{\text{cons}}) \le$$

$$\log_2(C_{\max}) + \frac{n}{2} \cdot \log_2(N) + \text{const}.$$

Dividing by N, tending to the limit $N \to \infty$ and using the fact that $\frac{\log_2(N)}{N} \to 0$ and the (already proved) fact that $\frac{\log_2(C_{\max})}{N}$ tends to the entropy S, we conclude that $\lim \frac{Q(N)}{N} = S$. The proposition is proven. □

Proof of Proposition 2. By definition, a result is consistent with the probabilistic knowledge P if and only if it is consistent with one of the distributions $p \in P$. Thus, the set of all the results which are consistent with P can be represented as a union of the sets of all the results consistent with different probability distributions $p \in P$. In the proof of Shannon's theorem, we have shown that for each $p \in P$, the corresponding number is asymptotically equal to $\exp(N \cdot S(p))$.

To be more precise, for every N, the number C of results with given frequencies $\{f_j\}$ ($f_j \approx p_j$) has already been computed in the proof of Shannon's theorem: $\lim \frac{\log_2(C)}{N} = -\sum f_j \cdot \log_2(f_j)$.

The total number of the results N_{cons} which are consistent with a given probabilistic knowledge P is equal to the sum of N_{co} different values of C that correspond to different f_j. For a given N, there are at most $N+1$ different values of $N_1 = N \cdot f_1$ (namely, values $N_1 = 0, 1, \ldots, N$), at most $N+1$ different values of N_2, etc., totally at most $(N+1)^n$ different sets of $\{f_j\}$. So, we get an inequality $C_{\max} \le N_{\text{cons}} \le (N+1)^n \cdot C_{\max}$, from which we conclude that $\lim \frac{Q(N)}{N} = \lim \frac{\log_2(C_{\max})}{N}$.