Adding Constraints to Situations When, In Addition to Intervals, We Also Have Partial Information about Probabilities

Martine Ceberio
Vladik Kreinovich
Gang Xiang
Dept. of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
contact vladik@utep.edu

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, NY 11733, USA
scott@ramas.com

Cliff Joslyn
Distributed Knowl. Syst. Team
Computer Research Group
Los Alamos National Lab
Mail Stop B265
Los Alamos, NM 87545, USA
joslyn@lanl.gov

Abstract

In many practical situations, we need to combine probabilistic and interval uncertainty. For example, we need to compute statistics like population mean $E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ *or population variance* $V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2$ *in the situations when we only know intervals* $\mathbf{x}_i$ *of possible values of* x_i*. In this case, it is desirable to compute the range of the corresponding characteristic.*

Some range computation problems are NP-hard; for these problems, in general, only an enclosure is possible. For other problems, there are efficient algorithms. In many practical situations, we have additional information that can be used as constraints on possible cumulative distribution functions (cdfs). For example, we may know that the actual (unknown) cdf is Gaussian. In this paper, we show that such constraints enable us to drastically narrow down the resulting ranges – and sometimes, transform the originally intractable (NP-hard) computational problem of computing the exact range into an efficiently solvable one.

This possibility is illustrated on the simplest example of an NP-problem from interval statistics: the problem of computing the range $\mathbf{V}$ *of the variance* V*.*

We also describe how we can estimate the amount of information under such combined intervals-and-constraints uncertainty.

1. Formulation of the Problem

Statistical analysis is important. Statistical analysis of measurement and observation results is an important part of data processing and data analysis. When faced with new data, engineers and scientists usually start with estimating standard statistical characteristics such as the mean E, the variance V, the cumulative distribution function (cdf) $F(x)$ of each variable, and the covariance and correlation between different variables. In the traditional statistical analysis, we estimate the value of each characteristic by computing the corresponding statistic $C(x_1, \ldots, x_n)$, such as:

- population mean $E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$;
- population variance $V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2$;
- histogram cdf $F_n(x) = \frac{\#i : x_i \le x}{n}$ (where $\#i : P(i)$ denotes "the number of i for which $P(i)$ is true");
- population covariance

$$C_{x,y} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y).$$

Limitations of traditional statistical techniques and the need to consider interval uncertainty. Traditional methods of statistical analysis assume that the measured values $\widetilde{x}_1, \ldots \widetilde{x}_n$ are the actual values $x_1, \ldots, x_n$ of the measured quantities. These methods work well if the variability of each variable is much higher than the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. For example, the accuracy with which we measure a person's height (≈ 1 cm) is much smaller than the variability in height between different people.

In many practical situations, however, the measurement errors are of the same order of magnitude as variability and therefore, cannot be ignored. Often, the only information that we have about the measurement errors is that they are upper bounded by Δ_i – and we have no information about the probabilities of different values $\Delta x_i \in [-\Delta_i, \Delta_i]$. In such situations, the only information we have after the measurements about the (unknown) actual value x_i is that x_i belongs to the intervals $\mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

In the case of interval uncertainty, instead of the actual (exact) values x_i, we only know the intervals $\mathbf{x}_i$ of possible values of x_i. In this case, we must find the range

$$\mathbf{C} = C(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=}$$

$$\{C(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}$$

of possible values of the given statistic.

Adding interval uncertainty to statistical techniques: what is known. There is a lot of research for computing such ranges. Some range computation problems are NP-hard – even the problem of computing the range for the population variance is, in general, NP-hard; see, e.g., [6, 9]. For such problems, in general, we can only compute an enclosure for the desired range.

For other problems, there are efficient algorithms; see, e.g., [2, 6, 9, 11] and references therein. For example, efficient algorithms are possible:

- for computing the range $\mathbf{E}$ of the population mean E,
- for computing the lower endpoints $\underline{V}$ of the range of variance,
- for computing the upper endpoint $\overline{V}$ of the range of variance when the intervals $\mathbf{x}_i$ are not contained in each other, i.e., $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$ for all i and j,
- etc.

Limitations of the existing approach. To explain the main limitation of the existing approach, let us briefly summarize what this approach is doing:

- we start with a statistic $C(x_1, \ldots, x_n)$ for estimating a given characteristic S;
- we evaluate this statistic under interval uncertainty, resulting in $\mathbf{C} = C(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

The main limitation of this idea is that a statistic is only an approximation to the desired statistical characteristic, i.e., $C(x_1, \ldots, x_n) \approx S$. For example, the population mean is only approximately equal to the mean value of the random quantity; similarly, the population variance is only an approximation to the actual variance, etc.

The approximation error $C(x_1, \ldots, x_n) - S \neq 0$ is not always taken into account when we take the interval range $\mathbf{C}$ as the range of the actual values of S.

For example, in this approach, if the values x_i are known exactly, then as a range of the population average, we will get a single number $E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ – while in reality, the actual mean can differ from this population average.

Seemingly natural solution can lead to excess width. A natural solution is that, instead of the original statistic C, we consider the bounds C^- and C^+ of the corresponding confidence interval $[C^-, C^+]$.

By definition of the confidence interval, this interval contains the actual value of the characteristic S with an appropriate certainty. For example, under reasonable assumptions (e.g., if the distribution is Gaussian), the interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$, where $\sigma \stackrel{\text{def}}{=} \sqrt{V}$ and k_0 (usually, 2, 3, or 6) is a given constant.

Thus, if we compute the interval range $[\underline{C}^-, \overline{C}^-]$ and $[\underline{C}^+, \overline{C}^+]$ for the statistics C^- and C^+, then the corresponding interval $[\underline{C}^-, \overline{C}^+]$ is an enclosure for S (with appropriate certainty). The ranges for C^- and C^+ can indeed be often efficiently computed [1, 5, 6, 9].

The problem with this idea is that a confidence interval is often defined so as to contain the actual value – but not necessarily as the narrowest interval that contains this value. As a result, the interval $[\underline{C}^-, \overline{C}^+]$ may contain excess width.

New idea. Let us instead find the actual range

$$\mathbf{S} = \{S(F) : F \text{ is possible}\}$$

of the characteristic S. Estimating this range is the main problem that we will be solving in this paper.

To solve this main problem, we must be able to solve the following closely related problem: how to describe class $\mathcal{F}$ of all the probability distributions F which are consistent with the given observations $[\underline{x}_i, \overline{x}_i]$?

2. How to Describe Possible Probability Distributions: p-Boxes

Case of exactly known probability distribution. The class of all probability distributions is infinite-dimensional; thus, to exactly describe a probability distribution, we need infinitely many parameters. In a computer, we can only store and process finitely many numbers; thus, if we want to represent probability distributions in a computer, we must select finitely many characteristics that will actually be representing this distribution.

To make this representation useful in practical applications, we must select characteristics which are practically useful. In many practical example, there is a critical threshold x_0 after which some undesirable event happens: a chip delays too much, a panel cracks, etc. In such situations, we want to make sure that the probability of exceeding x_0 is small. The resulting characteristic $\mathrm{Prob}(x_i \le x_0)$ is the value of the cumulative distribution function (cdf) $F(x)$ for $x = x_0$.

Thus, from the practical viewpoint, it is beneficial to describe a probability distribution by its cdf $F(x)$.

Case of partially known probability distribution. When, for every x, we know the exact value of $F(x)$, we thus know the actual probability distribution exactly. So, when we only have partial information about the probability distribution, this means that we do not know the exact values of $F(x)$. Instead, we may know, for every x, an interval $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$ that contains the actual (unknown) value $F(x)$.

Thus, a natural way to describe partial information about a probability distribution is to describe, for every x, a function $x \to \mathbf{F}(x)$. This function is called a *p-box* [2].

3. Estimates for Statistical Characteristics Based on the Use of p-Boxes

New idea (reminder). We have several observations $x_1, \ldots, x_n$ of a given random variable. These observations may be exact – in which case, we know the exact values of x_i – or, more generally, they may consist of known intervals $\mathbf{x}_i$ which contain the actual (unknown) values x_i of the observed quantity.

Our objective is to estimate the value of a statistical characteristic S based on these observations. Our new idea is that we estimate S in two steps:

- first, we describe the class of all probability distributions which are consistent with the given observations; since we agreed to represent such classes as p-boxes, we must transform observations $x_1, \ldots, x_n$ into a p-box;
- second, we estimate the range of the desired characteristic S based on this p-box.

Kolmogorov-Smirnov (KS) p-box. In statistics, there is a known way to produce bounds on cdfs (i.e., a p-box) from observations: use Kolmorogov-Smirnov (KS) inequalities; see, e.g., [7, 10].

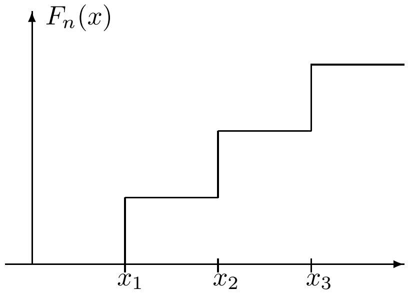
The main idea behind KS inequalities is rather straightforward. Namely, for each x_0, we have

- the actual (unknown) probability $p = F(x_0)$ that $x \le x_0$, and
- the observed frequency $F_n(x_0) = \dfrac{\#i : x_i \le x_0}{n}$.

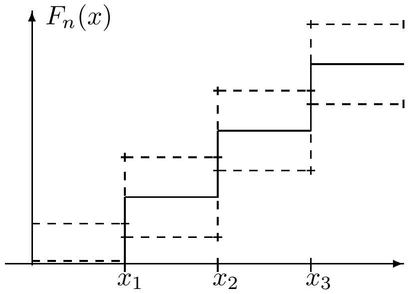
It is known that when n tends to infinity, then the distribution for the frequency tends to normal. Thus, for large n, this distribution is approximately normal. Hence, with given certainty α, we have $p - k \cdot \sigma \le F_n(x_0) \le p + k \cdot \sigma$, where $\sigma = \sqrt{\dfrac{p \cdot (1-p)}{n}}$ is the standard deviation of this simple random variable and $k = k(\alpha)$ is a factor that determines the confidence level. So, with certainty α, we get bounds on $p = F(x_0)$ in terms of $F_n(x_0)$.

We can now use these bounds for $x_0 = x_1, \ldots, x_0 = x_n$, and use monotonicity of the cdf $F(x)$ to get bounds $[F_n(x) - \varepsilon, F_n(x) + \varepsilon]$ for all $x \in [x_i, x_{i+1}]$.

Graphically, for a histogram



the Kolmogorov-Smirnov p-box takes the form:



For interval-valued data $[\underline{x}_i, \overline{x}_i]$, instead of single histogram, we have a p-box $[\underline{F}_n(x), \overline{F}_n(x)]$ formed by:

- the histogram $\underline{F}_n(x)$ generated by the values $\overline{x}_1, \ldots, \overline{x}_n$, and
- the histogram $\overline{F}_n(x)$ generated by the values $\underline{x}_1, \ldots, \underline{x}_n$.

To get a guaranteed bound (with appropriate certainty), we perform the same ε-enlargement to this p-box, producing a new p-box

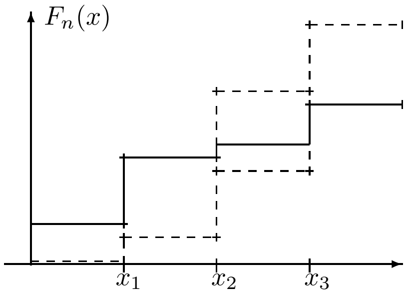
$$\mathbf{F}(x) = [\max(\underline{F}_n(x) - \varepsilon), 0), \min(\overline{F}_n(x) + \varepsilon, 1)].$$

Computing bounds for variance based on the KS p-box. Most known algorithms for computing the lower and upper bounds for the population variance under the interval uncertainty (see, e.g., [6, 9]) use the results of the calculus-type analysis of optimal values. Specifically, we use the following facts:

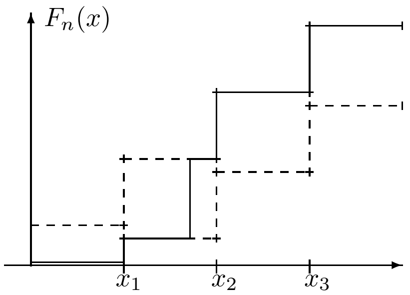
- if the function V attains a maximum or minimum for some value x_i which is strictly inside the interval $[\underline{x}_i, \overline{x}_i]$, then $\dfrac{\partial f}{\partial x_i} = 0$;
- if V attains a maximum for $x_i = \underline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \le 0$;
- if V attains a minimum for $x_i = \underline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \ge 0$;
- if V attains a maximum for $x_i = \overline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \ge 0$;
- if V attains a minimum for $x_i = \overline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \le 0$.

For the actual variance $V = \int x^2\,dF(x) - \left(\int x\,dF(x)\right)^2$, a similar reasonably simple calculus-type analysis leads to the following conclusions:

- the minimum $\underline{V}$ of the variance V is attained when the cdf $F(x) \in \mathbf{F}(x)$ first follows the upper cdf $\overline{F}(x)$, then stays horizonal, and then follows the lower cdf $\underline{F}(x)$:



- the maximum $\overline{V}$ of the variance V is attained when the cdf $F(x) \in \mathbf{F}(x)$ first follows the lower cdf $\underline{F}(x)$, and then jumps (vertically) to the upper cdf $\overline{F}(x)$:



Comment. The only difference with the case of population variance – in which we have finitely many variables $x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n$ – is that now we have an unknown function $F(x) \in \mathbf{F}(x)$ – i.e., in effect, infinitely many variables $F(x) \in \mathbf{F}(x)$ corresponding to different values x.

Computational complexity of computing $\underline{V}$ and $\overline{V}$. For the bounds on the *population* variance:

- we can compute $\underline{V}$ in linear time $O(n)$ [12];
- computing $\overline{V}$ is, in general, NP-hard;
- when $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$, we can compute $\overline{V}$ in linear time [12].

For the actual variance, if we use KS p-box, then the only remaining question is when to make a jump. For n data points, there are n possible interval containing this jump. For each interval, finding the best location is an easy-to-solve (quadratic) optimization problem with one variable, so its complexity does not depend on n. Thus, by applying the above observation:

- we can compute $\underline{V}$ in linear time $O(n)$, and
- we can compute $\overline{V}$ in linear time $O(n)$;

Conclusion. When we go from computing the range of the *population* variance to computing the range of the *actual* variance, we not only make our estimates more adequate – we also, in general, make computations much faster.

4 How to Handle Additional Constraints

Possibility of additional information. In the previous text, we assumed that the only information we have about the cdf $F(x)$ is that it is contained in the given p-box: $F(x) \in \mathbf{F}(x)$. However, often, we have additional information about $F(x)$.

This information that can be used as constraints on possible cdfs $F(x) \in \mathbf{F}(x)$. It is desirable to use these constraints when estimating statistical characteristics – similarly to the way constraints can be combined with traditional interval computations; see, e.g., [3].

Types of additional information. Often, we sometimes know the *shape* of $F(x)$, i.e., we know that $F(x) = F_0(x, a_1, \ldots, a_n)$ for a known function F_0 and for some parameters a_i. Usually, we do not know the *exact* values of each of these parameters; we may know the intervals $\mathbf{a}_i = [\underline{a}_i, \overline{a}_i]$ that contain the actual (unknown) values of these parameters.

A typical situation is when this dependence is linear in a_i, i.e., when

$$F(x) = F_0\left(\sum_{i=1}^{n} a_i \cdot e_i(x)\right).$$

To be more precise, the known dependence may not be linear in terms of the *given* parameters, but it may be described in this form if we use *appropriate* parameters.

For example, if we know that the actual distribution is Gaussian, this means that $F(x) = F_0\left(\dfrac{x-a}{\sigma}\right)$ for some parameters a and σ. With respect to the given parameters a and σ, this dependence is not linear, but if we select new parameters $a_1 \stackrel{\text{def}}{=} \dfrac{1}{\sigma}$ and $a_2 \stackrel{\text{def}}{=} -\dfrac{a}{\sigma}$, then we get the desired linear form: $F(x) = F_0(a_1 \cdot x + a_2)$.

How to take this additional information into account: first seemingly natural solution. We have mentioned that a natural way to represent a class of probability distributions is to find an appropriate p-box. Thus, it seems natural to find a p-box containing this class, i.e., for every x, to find the interval of possible values of $F(x)$ corresponding to the given class.

Once the p-box is found, we can then estimate the range of the desired characteristic – e.g., of the variance V – based on this p-box.

Limitations of the above seemingly natural approach. This approach indeed provided a guaranteed bound (enclosure) for the desired range – but it may also have excess width.

For example, if we start with the family of all normal distributions with 0 average and standard deviation σ from a given interval $[\underline{\sigma}, \overline{\sigma}]$, then the actual mean is always 0. However, as one can easily check, the corresponding p-box has non-zero width; as a result, it contains distribution with non-zero mean – and thus, the enclosure for the mean computed by using this p-box will contain non-zero values.

Towards exact estimates. Once we have a KS p-box $[\underline{F}(x), \overline{F}(x)]$ based on observations, we know that the actual (unknown) cdf $F(x)$ must be within this interval for all x:

$$\underline{F}(x) \le F(x) \le \overline{F}(x).$$

By definition, the KS p-box is obtained from the values at $x = x_i$ via monotonicity: when $x_i < x < x_{i+1}$, we take $\underline{F}(x) = \underline{F}(x_i)$ and $\overline{F}(x) = \overline{F}(x_{i+1})$. Thus, to guarantee that $F(x) \in \mathbf{F}(x)$ for all x, it is sufficient to check that this enclosure occurs for $x = x_1, \ldots, x_n$, i.e., that

$$\underline{F}(x_i) \le F(x_i) \le \overline{F}(x_i).$$

We know that $F(x) = F_0\left(\sum\limits_{i=1}^{n} a_i \cdot e_i(x)\right)$, so we can conclude that

$$\underline{F}(x_i) \le F_0\left(\sum_{i=1}^{n} a_i \cdot e_i(x)\right) \le \overline{F}(x_i).$$

Since the cdf $F_0(x)$ is monotonic, we can apply the inverse function F_0^{-1} to all the sides and get an equivalent inequality:

$$F_0^{-1}(\underline{F}(x_i)) \le \sum_{i=1}^{n} a_i \cdot e_i(x) \le F_0^{-1}(\overline{F}(x_i)). \qquad (1)$$

Thus, if we know the dependence $S(a_1, \ldots, a_n)$ of the desired characteristic S on the parameters a_i, then we can find the range of this characteristic by finding the minimum and the maximum of the corresponding function $S(a_1, \ldots, a_n)$ under the constraints (1) and $\underline{a}_i \le a_i \le \overline{a}_i$.

In particular, if the dependence $S(a_1, \ldots, a_n)$ is linear in a_i, then the problems of finding the minimum $\underline{S}$ and the maximum $\overline{S}$ are linear programming problems – i.e., problems which can be efficiently solved by known feasible algorithms.

Example. In practice, there are examples when the actual dependence $S(a_1, \ldots, a_n)$ is not linear, but this dependence can be reduced to linear by an appropriate transformation.

For example, for the case of the Gaussian distribution, we may be interested in the variance $V = \sigma^2$. In this case, as we have mentioned, $a_1 = \dfrac{1}{\sigma}$, hence $\sigma = \dfrac{1}{a_1}$, and $V = \dfrac{1}{a_1^2}$. This dependence is non-linear; however, this dependence is strictly increasing. Thus:

- finding the minimum of V is equivalent to finding the maximum of a_1 and
- finding the maximum of V is equivalent to finding the minimum of a_1.

The problem of finding the minimum and maximum of a_1 under linear constraints is already a linear programming problem.

Conclusion. The use of additional information about the probability distribution not only eliminates the excess width; it may also transform the originally NP-hard problem of estimating the range of the variance into a feasible one.

5. Gauging Amount of Uncertainty

Formulation of the problem and a seemingly natural solution. Every time we have uncertainty, an important question is how to gauge the amount of uncertainty; see, e.g., [4]. In the traditional statistical approach, the uncertainty in a probability distribution is usually described by Shannon's entropy

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

where $\rho(x) = F'(x)$ is the probability density function of this distribution.

We have already mentioned that in the situations when we have partial information about the probability distribution $F(x)$ – e.g., when we only know that $F(x)$ belongs to a non-degenerate p-box $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$, a reasonable estimate for an arbitrary statistical characteristic S is the range of possible values of S over all possible distributions $F(x) \in \mathbf{F}(x)$.

It therefore seems natural to apply this approach to entropy as well – and return the range of entropy as a gauge of uncertainty of a p-box; see, e.g., [4, 13].

Limitations of the above (seemingly natural) solution. The problem with the above approach is that every non-degenerate p-box includes discrete distributions, i.e., distributions which take discrete values $x_1, \ldots, x_n$ with finite probabilities. For such distributions, Shannon's entropy is $-\infty$.

Thus, for every non-degenerate p-box, the resulting interval $[\underline{S}, \overline{S}]$ has the form $[-\infty, \overline{S}]$. Thus, once the distribution with the largest entropy $\overline{S}$ is fixed, we cannot distinguish between a very narrow p-box or a very thick p-box – in both case, we end up with the same interval $[-\infty, \overline{S}]$.

It is therefore desirable to develop a new approach that would enable us to distinguish between these two cases.

Our idea: go back to the foundations. To design this new characteristic, let us go back to the foundations, check how Shannon came up with his measure of uncertainty, and see how Shannon's derivations can be modified to the case of p-boxes.

Traditional approach to gauging amount of information: reminder. 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.

Discrete case, when we have 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 $q \stackrel{\text{def}}{=} \lceil \log_2(n) \rceil$.

Comment. The value $\lceil x \rceil$ is the smallest integer which is larger than or equal to x. It is called the *ceiling* of the number x.

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 $2^q \ge n$, i.e., $q \ge \log_2(n)$. To complete the derivation, it is let us 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 $2^q \ge n$, we can this 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 $\log(n)$ 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 $N_i = n_1 + n_2 + \ldots + n_N$, where n_k equals to 1 if in k-th event the output is i and 0 otherwise. The average $E(n_k)$ of n_k equals to $p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i$. The mean square deviation $\sigma[n_k]$ is determined by the formula $\sigma^2[n_k] = p_i \cdot (1 - E(n_k))^2 + (1 - p_i) \cdot (0 - E(n_k))^2$. If we substitute here $E(n_k) = p_i$, we get $\sigma^2[n_k] = p_i \cdot (1 - p_i)$. The outcomes of all these events are considered independent, therefore n_k are independent random variables. Hence the average value of N_i equals to the sum of the averages of n_k: $E[N_i] = E[n_1] + E[n_2] + \ldots + E[n_N] = Np_i$. The mean square deviation $\sigma[N_i]$

satisfies a likewise equation $\sigma^2[N_i] = \sigma^2[n_1] + \sigma^2[n_2] + \ldots = N \cdot p_i \cdot (1 - p_i)$, 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 $a - k \cdot \sigma$ to $a + k \cdot \sigma$, where k is 2, 3, or 4. So in our case we can conclude that N_i lies between $N \cdot p_i - k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$ and $N \cdot p_i + k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$. Now we are ready for the formulation of Shannon's result.

Comment. In this quality control example the choice of k matters, but, as we'll see, in our case the results do not depend on k 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 $N \cdot p_i - k \cdot \sigma_i \leq N_i \leq N + k \cdot \sigma_i$, 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 $N_{cons}(N)$. The number $\lceil \log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$. The fraction $Q(N)/N$ will be called the *average number of questions.*

Theorem (Shannon; see, e.g., [8]). *When the number of events N tends to infinity, the average number of questions tends to*

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

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 $|x - r| \leq \varepsilon$.

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 $(\underline{x} + \overline{x})/2$ 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 ($x_{i+1} = x_i + 2\varepsilon$), 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 $\rho(x)$.

Case of p-boxes: description of the situation. Our main motivation is that the traditional approach of interval-valued entropy does not allow us to distinguish between narrow and wide p-boxes. For a wide p-box, it is OK to make a wide interval like $[-\infty, \overline{S}]$, but for narrow p-boxes, we would like to have narrower estimates. Let us therefore consider narrow p-boxes.

Since entropy is defined for smooth (differentiable) cdfs $F(x)$, it is reasonable to start with the case when the central function of a p-box is also smooth. In other words, we consider p-boxes of the type

$$\mathbf{F}(x) = [F_0(x) - \Delta F(x), F_0(x) + \Delta F(x)],$$

where $F_0(x)$ is differentiable, with derivative $\rho_0(x) \stackrel{\text{def}}{=} F_0'(x)$, and $\Delta F(x)$ is small.

Formulation of the problem. For each $\varepsilon > 0$ and for each distribution $F(x) \in \mathbf{F}(x)$, we can use the above formulas to estimate the average number $S_\varepsilon(F)$ of "yes"-"no" question that we need to ask to determine the actual value with accuracy ε. Our objective is to compute the range

$$[\underline{S}, \overline{S}] = \{S_\varepsilon(F) : F \in \mathbf{F}\}.$$

Known result. It is known (see, e.g., [8]) that asymptotically,

$$\overline{S} \sim -\int \rho_0(x) \cdot \log_2(\rho_0(x))\, dx - \log_2(2\varepsilon).$$

New result. Our new result is that

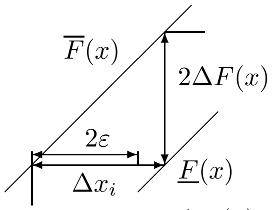
$$\underline{S} \sim -\int \rho_0(x) \cdot \max(2\Delta F(x), 2\varepsilon \cdot \rho_0(x))\, dx.$$

Comment. This result holds when ε and the width of ΔF both tends to 0. If instead we fix the width ΔF and let $\varepsilon \to 0$, then $\overline{S} \to \infty$ but $\underline{S}$ remains finite.

Idea of the proof. When we discretize the distribution, we get $p_i \approx \rho_0(x_i) \cdot \Delta x_i$, hence

$$-\sum p_i \cdot \log_2(p_i) \approx -\int \rho_0(x) \cdot \log(\rho_0(x) \cdot \Delta x)\, dx.$$

To minimize the entropy, we can take the discrete distribution with values $x_1, \ldots, x_n$ as far away from each other as possible. A distribution which is located at x_i and x_{i+1} and has 0 probability to be in between is described by a cdf $F(x)$ which is horizontal on $[x_i, x_{i+1}]$. Thus, we must select a cdf $F(x) \in \mathbf{F}(x)$ for which these horizontal segments are as long as possible. The length of a horizontal segment is bounded by the geometry of the p-box:



Thus, this length cannot exceed $\dfrac{2\Delta F(x)}{\rho_0(x)}$. If this length is $> 2\varepsilon$, then we can take this interval between the sequential values x_i. If this length is $< 2\varepsilon$, then we can still take $\Delta x_i = 2\varepsilon$. Thus, in general, we take $\Delta x_i = \max\left(\dfrac{2\Delta F(x)}{\rho_0(x)}, 2\varepsilon\right)$. Substituting this expression into the above asymptotic formula, we get the desired asymptotic for $\underline{S}$.

6 Conclusions

This paper mainly deals with the following problem: When we have several observations of given random variables, these observations being known as intervals, how is it possible to compute enclosures for some characteristics such as mean values or standard deviation of the true population.

In our paper, we describe feasible algorithms for solving several practically important cases of this problem.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

The authors are thankful to participants of SCAN'06 for valuable discussions, and to the anonymous referees for important suggestions.

References

[1] E. Dantsin, A. Wolpert, M. Ceberio, G. Xiang, and V. Kreinovich. Detecting outliers under interval uncertainty: a new algorithm based on constraint satisfaction. In: *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006, pp. 802–809.

[2] S. Ferson. *RAMAS Risk Calc 4.0*. CRC Press, Boca Raton, Florida, 2002.

[3] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.

[4] G. J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*. J. Wiley, Hoboken, New Jersey, 2005.

[5] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg. Outlier detection under interval uncertainty: algorithmic solvability and computational complexity. *Reliable Computing*, 11(1):59–76, 2005.

[6] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos. Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases. *Journal of Computational and Applied Mathematics*, 199(2):418–423, 2007.

[7] V. Kreinovich, E. J. Pauwels, S. Ferson, and L. Ginzburg. A feasible algorithm for locating concave and convex zones of interval data and its use in statistics-based clustering. *Numerical Algorithms* 37:225–232, 2004.

[8] V. Kreinovich, G. Xiang, and S. Ferson. How the concept of information as average number of "yes-no" questions (bits) can be extended to intervals, p-boxes, and more general uncertainty. In: *Proceedings of the 24th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pp. 80–85.

[9] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing*, 12(6):471–501, 2006.

[10] H. M. Wadsworth, Jr. (ed.). *Handbook of statistical methods for engineers and scientists*. McGraw-Hill Publishing Co., New York, 1990.

[11] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, New York, 1991.

[12] G. Xiang, M. Ceberio, and V. Kreinovich. *Computing Population Variance and Entropy under Interval Uncertainty: Linear-Time Algorithms*. University of Texas at El Paso, Department of Computter Science, Technical Report UTEP-CS-06-28b, November 2006.

[13] G. Xiang, O. Kosheleva, and G. J. Klir. Estimating information amount under interval uncertainty: algorithmic solvability and computational complexity. In: *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006, pp. 840–847.