ATLANTIS PRESS
Atlantis Studies in Uncertainty Modelling, volume 3
Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)

In Search of a Precise Estimator Based on Imprecise Data

*Przemyslaw Grzegorzewski[a,b] and Joanna Goławska[b]
[a]Systems Research Institute PAS, Newelska 6, 01-447 Warszawa, Poland, pgrzeg@ibspan.waw.pl
[b]Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland, j.k.golawska@gmail.com

Abstract

Statistics with interval-valued data are getting less interest from practitioners than it really deserves. This is partly because the solutions it offers are often too conservative and hence do not fully meet the expectations of potential users. Thus it is necessary to develop methods which, despite imprecise input data, will lead to more precise final statistical decisions. In the paper we discuss several refinement-oriented methods that may be useful in estimation based on interval-valued data.

Keywords: Imprecise observations, Interval-valued data, EM algorithm, Middle censoring, Symbolic data.

1 Introduction

The famous statistician, Churchill Eisenhart, defined the "practical power" of a mathematical procedure as "the product of the mathematical power by the probability that the procedure will be used". This sentence is sometimes quoted to emphasize the gap between theoretical achievements and what practitioners expect to meet their needs. It happens sometimes that academia offers sophisticated mathematical constructions which surely deserve respect because of their complexity, generality or just beauty. However, these merits do not automatically guarantee their practical usefulness. Indeed, practitioners often look for such tools which are simple and in some sense compact, i.e. can admit somewhat reduced properties "for what is practically important" [20].

It seems that the situation in question is somehow typical for statistics with imprecise data, where two kinds of uncertainty – randomness and imprecision – coexist. For ages uncertainty was perceived as a flaw in a perfect construction of science. It was only a hundred years ago that people realized that uncertainty was an indelible feature and they had to get used to living with it. However, while randomness alone described by probability theory and classical statistics was tamed relatively early and boasts a huge number of implementations, the situation is different in the case of uncertainty when randomness and imprecision occur together. Actually, even in the case of the simplest model of interval-valued data we can talk about quite a large number of theoretical models but definitely little interest from practitioners. Why? Obviously, there are many reasons such as getting used to traditional crisp methods or insufficient knowledge of new models. But it seems that quite often the most common reason for the lack of interest in statistics with interval-valued data are too conservative models that are proposed. This excessive conservatism comes from the fact that we are not yet able to cope well with overlapping uncertainty of both types (i.e. randomness and imprecision) and at the same time we want to provide models with a low risk of error. Such caution gives solutions that are safe, but at the same time ineffective in practice, which discourages potentially interested users.

Obviously, it is not possible to provide a unique and universal solution to the aforementioned problem. In this paper we restrict our attention to some estimation problems based on interval data and consider several scenarios of how to deal with the problem of excessive uncertainty in this context. Thus we show how to make at least a step forward in precisiation of imprecise outputs (see [23]).

The paper is organized as follows: a brief introduction to interval-valued data is given in Sec. 2, while in Sec. 3 we summarize basic ideas and problems connected with statistical inference based on epistemic interval data. Next, in Sec. 4, we discuss several ways one can follow to obtain precise estimators from interval data. The considered approaches are compared and summarized in Sec. 5, while in Sec. 6 we conclude the paper.

2 Interval data

Interval-valued data appear naturally in many areas of science and engineering. However, one should be conscious that an interval may serve for modeling two kinds of information: either the imprecise description of a point-valued quantity or the precise description of a set-valued entity.

Indeed, quite often the experimental results are imprecisely observed or so uncertain that they are recorded as intervals containing the precise outcomes. It may also happen the exact values of a variable are hidden deliberately for some confidentiality reasons. In all such cases intervals are considered as disjunctive sets representing incomplete information so they correspond to the **epistemic view**, according to [2]. Thus an **epistemic interval** A contains an ill-known actual value of a point-valued quantity x, so we can write $x \in A$. Please, notice that interval A represents here the epistemic state of an agent, so it does not exist per se.

There are also situations when data appear as essentially interval-valued which describe a precise information, such as a range of fluctuations of some physical measurements. Such intervals correspond to the **ontic view** [2], where an interval is the precise representation of an objective entity, so A is just a value of a set-valued variable X and hence we can write $X = A$.

Awareness of what type of data we are dealing with is strongly important. Although the notation and basic operations on intervals do not depend on whether they are perceived as epistemic or ontic, there are significant differences in statistical inference with data perceived from those two perspectives. In particular, in the epistemic approach we deal with a usual random variable which attributes a real value to each random event, but its perception is not known precisely but is exact to the interval value only. On the other hand, in the ontic approach, we deal with random intervals.

Further on in this paper we restrict our attention to epistemic data only.

Let $\mathscr{K}_c(\mathbb{R}) = \{[a,b] : a,b \in \mathbb{R},\, a \leqslant b\}$ denote the family of all non-empty closed and bounded intervals in the real line $\mathbb{R}$. Each compact interval $\xi \in \mathscr{K}_c(\mathbb{R})$ can be expressed by its endpoints, i.e. $\xi = [a,b]$, or alternatively, by its mid-point (center) and spread (radius).

Interval arithmetic which is a special case of set arithmetic was independently invented by Warmus [22], Sunaga [19] and Moore [14]. Natural calculations on $\mathscr{K}_c(\mathbb{R})$ are defined by the Minkowski addition and the scalar multiplication, i.e.

$$\xi_1 + \xi_2 = \{x + y : x \in \xi_1, y \in \xi_2\},$$

$$C \cdot \xi = \{C \cdot x : x \in \xi\},$$

for any $\xi, \xi_1, \xi_2 \in \mathscr{K}_c(\mathbb{R})$ and $C \in \mathbb{R}$.

Using the endpoints of the intervals the aforementioned operations, i.e. addition and subtraction of two intervals $\xi_1 = [a_1, b_1]$ and $\xi_2 = [a_2, b_2]$ and the scalar multiplication of $\xi = [a,b]$ and $C \in \mathbb{R}$ are given by

$$\begin{aligned} \xi_1 + \xi_2 &= [a_1 + a_2, b_1 + b_2] \\ \xi_1 - \xi_2 &= [a_1 - b_2, b_1 - a_2], \\ C \cdot \xi &= [\min\{C \cdot a, C \cdot b\}, \max\{C \cdot a, C \cdot b\}]. \end{aligned}$$

It should be noted that the space $(\mathscr{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semi-linear, due to the lack of the opposite element with respect to the Minkowski addition. Indeed, in general, $\xi + (-1) \cdot \xi \neq \{0\}$, unless $\xi = \{a\}$ is a singleton. Consequently, $\xi_3 = \xi_1 - \xi_2$ does not guaranty, in general, that $\xi_2 + \xi_3 = \xi_1$. To overcome this drawback the so-called Hukuhara difference was defined as follows: $\xi_3 = \xi_1 -_h \xi_2$ if and only if $\xi_2 + \xi_3 = \xi_1$. Unfortunately, the Hukuhara difference does not exist for any two intervals $\xi_1, \xi_2 \in \mathscr{K}_c(\mathbb{R})$ but only for such $\xi_1 = [a_1, b_1]$ and $\xi_2 = [a_2, b_2]$ that $b_1 - a_1 \geqslant b_2 - a_2$.

Generally, any function defined on real values can be extended to intervals in a straightforward way. In particular, the extension of a function $f : \mathbb{R} \to \mathbb{R}$ to intervals results in the set of all possible values that could be obtained from f by supplying real-valued arguments from the respective interval, i.e. for $\xi = [a,b]$ we obtain

$$f(\xi) = \{f(x) : x \in [a,b]\}. \tag{1}$$

Such operations like logarithm, square root, powers, etc. are generalized to interval arguments in this way. In a case where (1) is not an interval because it has "holes" (i.e., values between the smallest and largest values of (1) which are themselves not possible results of the function), it can be made into an interval by taking a convex hull and replacing the set of results by an interval defined by the smallest and largest result values.

It is worth remembering that following (1) in a "naive" way may lead to mathematically rigorous results, i.e. surely enclosing the true range, but which fails to be best-possible because it is wider than it needs to be. It often happens when values belonging to intervals enter into the calculation more than once. In particular, the square of $\xi = [a,b]$ might be perceived as $\xi^2 = \xi \cdot \xi = \{x \cdot : x \in [a,b], y \in [a,b]\}$ or as $\xi^2 = \{x^2 : x \in [a,b]\}$ and both results generally differ. Thus, e.g., in calculating the sample variance one should calculate squares as follows (see, e.g., [12, 15])

$$\xi^2 = \begin{cases} [0, \max\{a^2, b^2\}], & \text{if } 0 \in [a,b], \\ [\min\{a^2, b^2\}, \max\{a^2, b^2\}], & \text{if } 0 \notin [a,b]. \end{cases}$$

Another difficulty we encounter in processing interval-valued data is the problem with ranking intervals, which prevents the use of rank tests or procedures based on order statistics. Indeed, the family $\mathscr{K}_c(\mathbb{R})$ is not linearly ordered and one can define order relations between intervals in many ways. Finally, let us mention that the interval type (whether they are narrow or broad, nesting, overlapping, binned, less or more scattered, of the same or not the same precision, etc.) may also cause bigger or smaller computational problems. In descriptive statistics with interval-valued data we sometimes distinguish just two basic types of intervals: "skinny" (i.e. the intervals that are so narrow as to not overlap each other) or "puffy" (intervals which are generally wider and exhibit a lot of overlap). For more details we refer to [5].

3 Statistics with epistemic interval-valued data

Suppose, we observe a sample of real-valued random variables $\mathbb{X} = (X_1,\dots,X_n)$ from the distribution F_θ, where $\theta \in \Theta$ stands for the unknown parameter. However, as an output of the experiment we actually receive a sequence of n interval-valued observations $\xi_1 = [a_1,b_1],\dots,\xi_n = [a_n,b_n]$. We assume that the true value of $X_i \in \xi_i$, i.e. each $x_i \in [a_i,b_i]$ is a possible value of X_i.

Let $\widehat{\theta} = \widehat{\theta}(\mathbb{X})$ denote a classical (i.e. based on the real-valued sample) estimator of θ. However now, having interval data only, we may consider different possible values of $\widehat{\theta}$, i.e.

$$\widehat{\theta}(\xi_1,\dots,\xi_n) = \{\widehat{\theta}(x_1,\dots,x_n) : x_i \in \xi_i,\, i = 1,\dots,n\}. \tag{2}$$

It is not always possible (or easy) to find the actual range of $\widehat{\theta}(\xi_1,\dots,\xi_n)$. Thus we try to compute its enclosure, i.e. an interval $\widetilde{\theta}$ such that $\widetilde{\theta} \supseteq \widehat{\theta}(\xi_1,\dots,\xi_n)$. If $\widetilde{\theta} = \widehat{\theta}(\xi_1,\dots,\xi_n)$ we say that the enclosure is exact.

When the estimator $\widehat{\theta}$ is in some sense regular (e.g., continuous or monotonic), to determine $\widehat{\theta}(\xi_1,\dots,\xi_n)$ it is usually enough to identify the smallest and largest value of $\widehat{\theta}$ denoted by $\widehat{\theta}_{\min}$ and $\widehat{\theta}_{\max}$, respectively. Finding the exact (or satisfactory) enclosures is not easy in general. Moreover, in some cases, it is even impossible in a reasonable time (e.g., determining the sample variance for an arbitrary sample of the interval data perceived from the epistemic perspective is the NP-hard problem, see [16]).

However, even if obtaining the desired estimator (2) involve no calculation problems, statistics based on interval data may not be satisfying for practitioners, especially if the range of $\widehat{\theta}(\xi_1,\dots,\xi_n)$, i.e. the distance between $\widehat{\theta}_{\max}$ and $\widehat{\theta}_{\min}$ is too large. To be more specific let us consider the following example.

Example 1. Consider the following interval-valued sample: [2.35, 7.33], [6.02, 8.59], [9.74, 11.64], [7.59, 13.03], [23.11, 27.42], [12.95, 17.93], [27.08, 30.37] [5.93, 9.22], [32.63, 34.53], [13.85, 19.28] illustrating imprecise recordings of the time to failure (in hours) of some devise. The upper and lower bounds of the empirical distribution function for this sample, i.e.

$$\widehat{F}_n^L(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(a_i \leqslant t), \quad \widehat{F}_n^U(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(b_i \leqslant t),$$

are given in Fig. 1 (the dashed rectangles between $\widehat{F}_n^U$ and $\widehat{F}_n^L$ have no meaning here but they are drawn just to visualize better the distance between these two bounds, while indices L and U applied in the notation correspond to the stochastic order between two borderline random variables, see [6, 7]).

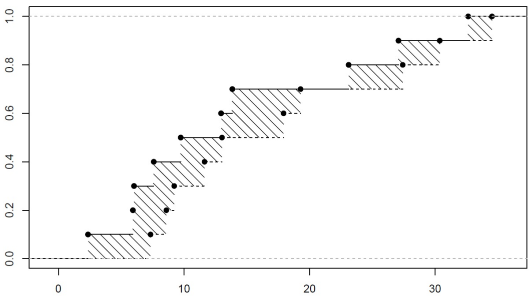

Figure 1: The empirical distribution function for interval-valued data.

Suppose our data come from the exponential distribution and our goal is to estimate its MTTF, i.e. the mean time to failure. If we have a real-valued sample $\mathbb{X} = (X_1,\dots,X_n)$ our estimator would be equal to the average of the observations. In our case of interval data we estimate MTTF by the mean of intervals which results in $[14.125, 17.934]$. Although this conclusion is perfectly rigorous from the point of view of interval arithmetic, it is of a slight interest to engineers because of its too high imprecision. This solution becomes much more unaccepted if they need a confidence interval. Indeed, the classical formula for the $(1-\alpha)100\%$ confidence interval is given by

$$\frac{2n\bar{x}}{\chi^2_{n,1-\alpha/2}} < \text{MTTF} < \frac{2n\bar{x}}{\chi^2_{2n,\alpha/2}},$$

where $\chi^2_{k,\nu}$ is the $100\cdot\nu$ percentile of the chi-squared distribution with k degrees of freedom. This formula extended through (2) for the interval data provides the

95% confidence interval of the form: $8.99 < \text{MTTF} < 33.06$, which makes no practical sense for engineers.

Obviously, problems with interval-valued data do not restrict to estimation. Suppose we verify a null hypothesis H_0 against the alternative H_1 and let T denote a desired test statistic. Given interval-valued data $\xi_1,\dots,\xi_n$ we can determine the set $T(\xi_1,\dots,\xi_n)$ of all possible test statistics values. Hence, to make a decision whether reject or accept H_0 we have to compute a p-value which is a subset of the interval $[p^L, p^U]$, where

$$p^L = \mathbb{P}_{H_0}\big(T \geqslant \max T(\xi_1,\dots,\xi_n)\big),$$
$$p^U = \mathbb{P}_{H_0}\big(T \geqslant \min T(\xi_1,\dots,\xi_n)\big)$$

(see [8, 9, 10, 17]). Assuming the significance level α the decision algorithm is as follows

- if $p^U < \alpha$, then we reject H_0,
- if $\alpha < p^L$, then we do not reject (accept) H_0,
- otherwise (i.e. if $\underline{p} \leqslant \alpha \leqslant \overline{p}$), then we abstain from the decision.

Unfortunately, if $T(\xi_1,\dots,\xi_n)$ is too broad, then the most expected result is abstention. If it happens too often it becomes very discouraging for practitioners.

4 Precise estimation based on interval data

In this section we discuss methods of estimating parameters of a distribution under study when we have a sample in the form of interval data.

4.1 EM algorithm

The EM (*Expectation-Maximization*) was first proposed by Dempster et al. [3]. It is one of the most widely used algorithms for iteratively computing the Maximum Likelihood Estimator (MLE) for incomplete data. It can be used at various levels of generality, making it useful in many of the problems of missing data, censored data, estimation for mixtures of distributions, etc. The great advantage of the EM algorithm is numerical stability, reliable global convergence to the local extreme and simplicity of implementation, although it is also not without limitations.

Suppose we have a random sample $X_1,\dots,X_n$ with a density $g(x\,|\,\theta) = g_\theta(x)$. We are interested in estimating with the maximum likelihood method on the basis of $x_1,\dots,x_n$ observations the value of the θ parameter. However, some of the observations (or perhaps all) of x_i are incomplete, imprecise, etc. Therefore, we will introduce additional variables $Z_1,\dots,Z_n$, the observations of which $z_1,\dots,z_n$ are already complete (precise) data. Write the total density of $X_1,\dots,X_n$ and $Z_1,\dots,Z_n$ as $g(\mathbb{x},\mathbb{z}\,|\,\theta) = g_\theta(\mathbb{x},\mathbb{z})$. Obviously, $g_\theta(\mathbb{x},\mathbb{z}) = g_\theta(\mathbb{x}\,|\,\mathbb{z})\cdot g_\theta(\mathbb{z})$. Thus the complete-data likelihood is given by

$$L^c(\theta) := L^c(\theta\,|\,\mathbb{x},\mathbb{z}) = g_\theta(\mathbb{x},\mathbb{z}). \tag{3}$$

Now let us denote by $\theta^{(q)}$ the estimator of θ obtained in q-th iteration. Next, we define the expected value of the complete log-likelihood function, assuming observed $\mathbb{Z} = \mathbb{z}$

$$\begin{aligned} Q_{EM}(\theta\,|\,\theta^{(q)}) :&= \mathbb{E}_{\theta^{(q)}}\big[\log L^c(\theta)\,\big|\,\mathbb{z}\big], \\ &= \int_{\mathscr{X}} \big[\log g_\theta(\mathbb{x},\mathbb{z})\big]\, g_{\theta^{(q)}}(\mathbb{x}\,\big|\,\mathbb{z})\,d\mathbb{x}. \end{aligned} \tag{4}$$

The parameters set at the input of the algorithm are the start value $\theta^{(0)}$ of the searched parameter and the threshold value ε for stopping the algorithm. The EM algorithm consists of two steps repeated at each iteration:

Step E (*Expectation*): In q-th iteration we compute the value of the expected log-likelihood

$$Q_{EM}(\theta\,|\,\theta^{(q)}) = \mathbb{E}_{\theta^{(q)}}\big(\log L^c(\theta)\,|\,\mathbb{Z})\big). \tag{5}$$

Step M (*Maximization*): While maximizing the expected log-likelihood, we are looking for the next parameter value θ, i.e.

$$\theta^{(q+1)} = \arg\max_\theta Q_{EM}(\theta\,|\,\theta^{(q)}). \tag{6}$$

We repeat steps E and M until the stopping condition is met. It can be defined in various ways, e.g. $|\,\theta^{(q+1)} - \theta^{(q)}\,| < \varepsilon$ or $|\,L(\theta^{(q+1)}) - L(\theta^{(q)})\,| < \varepsilon$.

Now let $X_1,\dots,X_n$ denote i.i.d. random variables from the exponential distribution $\text{Exp}(\lambda)$ with the density function $g(x) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$. Our goal is to estimate the unknown parameter $\lambda > 0$. However, the aforementioned real-valued random variables are not observed precisely but instead of their precise values $x_1,\dots,x_n$ we observe intervals $\xi_i = [a_i,b_i]$, such that $x_i \in [a_i,b_i]$ for each $i = 1,\dots,n$. Here the complete-data likelihood is given by

$$L^c(\lambda\mid z) = \lambda^n \exp\big(-\lambda\sum_{i=1}^n z_i\big). \tag{7}$$

Thus, following (3)-(5) we obtain

$$
\begin{aligned}
Q_{EM}(\lambda\,|\,\lambda^{(q)}) &= \mathbb{E}_{\lambda^{(q)}}\left(\log L^c(\lambda)\,|\,z\right) \\
&= n\log\lambda - \frac{n\lambda}{\lambda_{(q)}} \\
&\quad - \lambda\sum_{i=1}^{n}\frac{a_i\exp\left(-\lambda^{(q)}a_i\right) - b_i\exp\left(-\lambda^{(q)}b_i\right)}{\exp\left(-\lambda^{(q)}a_i\right) - \exp\left(-\lambda^{(q)}b_i\right)},
\end{aligned}
$$

while the value of the estimated parameter in the $(q+1)$-th iteration is given by

$$
\lambda_{EM}^{(q+1)} = \frac{n}{\frac{n}{\lambda^{(q)}} + \sum\limits_{i=1}^{n}\frac{a_i\exp\left(-\lambda^{(q)}a_i\right) - b_i\exp\left(-\lambda^{(q)}b_i\right)}{\exp\left(-\lambda^{(q)}a_i\right) - \exp\left(-\lambda^{(q)}b_i\right)}}. \tag{8}
$$

4.2 IEM algorithm

The IEM (*Interval Expectation-Maximization*) is a generalization of the EM algorithm especially for interval data. It was originally proposed by Su et al. [18] but actually it might be perceived as an adaptation of the fuzzy expectation-maximization (FEM) algorithm [4] for interval data. Suppose $X_1,\ldots,X_n$ is a random sample from the distribution with a density $g(x\,|\,\theta) = g_\theta(x)$. Now, assuming that each true (but unknown) observation x_i belongs to (known) interval $\xi_i = [a_i, b_i]$ let us define the likelihood function as follows

$$
L(\theta;\xi) = \int\limits_{\mathscr{X}} \mathbb{1}(\mathbb{x}) g_\theta(\mathbb{x})\,d\mathbb{x} = \mathbb{E}_\theta\left[\mathbb{1}(\mathbb{x})\right],
$$

where $\mathbb{1}(\mathbb{x}) = \mathbb{1}_{\xi_1}(x_1)\cdot\ldots\cdot\mathbb{1}_{\xi_n}(x_n)$ denotes the product indicator functions corresponding to interval observations. Then the expected log-likelihood defined as

$$
\begin{aligned}
Q_{IEM}(\theta,\theta^{(q)}) &= \mathbb{E}_{\theta^{(q)}}\left(\log\left[L(\theta;\xi)\right]\,|\,\xi\right) \\
&= \frac{\int\limits_{\mathscr{X}}\log\left[L(\theta;\xi)\right]\mathbb{1}(\mathbb{x}) g_{\theta^{(q)}}(\mathbb{x})\,d\mathbb{x}}{L(\theta^{(q)};\xi)},
\end{aligned} \tag{9}
$$

is maximized and updated until

$$
\frac{\left|L(\theta^{(q+1)};\xi) - L(\theta^{(q)};\xi)\right|}{L(\theta^{(q)};\xi)} \leqslant \varepsilon.
$$

Now, applying this approach to interval data from the exponential distribution with likelihood function (7) we obtain the expected log-likelihood (9) of the form

$$
\begin{aligned}
Q_{IEM}(\lambda,\lambda^{(q)}) &= n\log\lambda - \lambda\sum_{i=1}^{n}\mathbb{E}_{\lambda^{(q)}}\left(X_i \mid \mathbb{1}_{\xi_i}(x_i)\right) \\
&= n\log\lambda - \lambda\sum_{i=1}^{n}\alpha_i(q)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_i(q) &= \mathbb{E}_{\lambda^{(q)}}\left(X_i \mid \mathbb{1}_{\xi_i}(x_i)\right) \\
&= \mathbb{E}_{\lambda^{(q)}}\left(X_i \mid g(\mathbb{x} \mid \mathbb{1}_{\xi_i}(x_i);\lambda^{(q)})\right) \\
&= \frac{\int\limits_{\mathscr{X}} x_i\, g(x_i;\lambda^{(q)})\mathbb{1}_{\xi_i}(x_i)\,dx_i}{\int\limits_{\mathscr{X}} g(x_i;\lambda^{(q)})\mathbb{1}_{\xi_i}(x_i)\,dx_i} \\
&= \frac{\int_{a_i}^{b_i} x_i\,\lambda^{(q)}\exp\left(\lambda^{(q)}x_i\right)dx_i}{G_{\lambda^{(q)}}(b_i) - G_{\lambda^{(q)}}(a_i)},
\end{aligned} \tag{10}
$$

where $G_{\lambda^{(q)}}$ stands for the cumulative distribution function of the exponential distribution with parameter $\lambda^{(q)}$. Finally,

$$
\begin{aligned}
Q_{IEM}(\lambda,\lambda^{(q)}) &= n\log\lambda \\
&\quad - \lambda\sum_{i=1}^{n}\frac{\exp\left(-\lambda a_i\right)\left(a_i+\frac{1}{\lambda}\right) - \exp\left(-\lambda b_i\right)\left(b_i+\frac{1}{\lambda}\right)}{G_{\lambda^{(q)}}(b_i) - G_{\lambda^{(q)}}(a_i)},
\end{aligned}
$$

so by solving the equation $\frac{dQ(\lambda,\lambda^{(q)})}{d\lambda} = \frac{n}{\lambda} - \sum\limits_{i=1}^{n}\alpha_i(q)$ we obtain the following value of the estimated parameter in the $(q+1)$-th iteration

$$
\lambda_{IEM}^{(q+1)} = \frac{n}{\sum\limits_{i=1}^{n}\alpha_i(q)}. \tag{11}
$$

4.3 Middle Censoring algorithm

Another approach for estimation based on interval data, called *Middle Censoring Algorithm* (MCA) was proposed by Iyer et al. [11]. It combines both left-censoring and right-censoring methods well-known in classical statistics, especially in survival analysis and reliability. Since here we have to assume that the borders of the observed intervals are the outputs of some real-valued random variables, instead of describing the method in a general way we restrict our attention directly to the exponential distribution.

As before we assume that $X_1,\ldots,X_n$ are nonobserved i.i.d. random variables from the exponential distribution $\mathrm{Exp}(\lambda)$. However, now the intervals $[a_1,b_1],\ldots,[a_n,b_n]$ we observe are supposed to have some probabilistic structure. In particular, we assume that the lower bounds $a_1,\ldots,a_n$ are realizations of the random samples $A_1,\ldots,A_n$ from the exponential distribution $\mathrm{Exp}(\alpha)$, while the upper bounds $b_1,\ldots,b_n$ are given by random variables $B_1,\ldots,B_n$ such that each difference $B_i - A_i$ is also exponentially distributed from $\mathrm{Exp}(\beta)$. Moreover, we assume that the unknown hiperparameters α and β are independent of λ and that random variables A_i and $(B_i - A_i)$ are independent of X_i for each $i = 1,\ldots,n$.

Under such assumptions the log-likelihood function $l = \log L$ has a very simple form

$$l(\lambda) = \log c + \sum_{i=1}^{n} \log\left(\exp(-\lambda a_i) - \exp(-\lambda b_i)\right),$$

but the equation of interest

$$\frac{\partial}{\partial \lambda} l(\lambda) = \sum_{i=1}^{n} \frac{b_i - a_i}{\exp\left(-\lambda(a_i - b_i)\right) - 1} - \sum_{i=1}^{n} a_i = 0$$

has no analytic solution. Instead, we may consider the iterative procedure, where

$$\lambda_{MC}^{(q+1)} = \frac{\lambda^{(q)}}{\sum_{i=1}^{n} a_i} \cdot \sum_{i=1}^{n} \frac{(b_i - a_i)\exp\left(-\lambda^{(q)}(b_i - a_i)\right)}{1 - \exp\left(-\lambda^{(q)}(b_i - a_i)\right)}. \tag{12}$$

Starting from the inverse of the mean of the centers of intervals, i.e.

$$\lambda^{(1)} = \frac{2n}{\sum_{i=1}^{n}(a_i + b_i)}.$$

We continue the iterative algorithm until

$$\left|\lambda^{(q+1)} - \lambda^{(q)}\right| \leqslant \varepsilon.$$

It can be shown that the estimator obtained by MCA is consistent and asymptotically normal.

4.4 Induced likelihood method

While discussing all the previous methods, we assumed that we had a priori knowledge of the distribution of a random variable. In fact, we often don't have such information. In the case of precise data, a number of tests are known, thanks to which we can check whether a given theoretical distribution does not differ significantly from the empirical distribution from the sample. Unfortunately, in the case of imprecise data, we do not have such possibilities - the information we have is somewhat distorted by the fact that it is in the form of a range. Here, we describe the maximum likelihood estimation method that does not use unverifiable information about the distribution of a variable from which the data comes. It was proposed by Le-Redemacher and Billard in [13], but here we present it in a slightly modified form. We will discuss two variants of these methods, assuming independence and dependence of the estimated parameters.

Following the idea of Bertrand and Goupil [1] developed for the symbolic data we may define the empirical distribution based on n interval observations $[a_1, b_1], \ldots, [a_n, b_n]$ as follows

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{k=1}^{n} \frac{t - a_k}{b_k - a_k} + \frac{\#\{k : t \geqslant b_k\}}{n},$$

However, for interval observations $[a_1, b_1], \ldots, [a_n, b_n]$ the so-called induced likelihood function is constructed as a composition of two distributions, say g_1 and g_2, characterizing the internal mean Θ_{i1} and the internal variance Θ_{i2} of the observed interval X_i, respectively, with realizations $\theta_{1i} := \frac{a_i + b_i}{2}$ and $\theta_{2i} := \frac{(b_i - a_i)^2}{12}$ for $i = 1, \ldots, n$. Let $\Theta_i = (\Theta_{i1}, \Theta_{i2})$ denote the join random variable while $\tau = (\lambda, \beta)$ let stand for the join parameter of the distributions g_1 and g_2 (where λ corresponds to the first distribution and β for the second one). If the distributions of Θ_{i1} and Θ_{i2} are independent then the induced likelihood is defined as

$$L(\tau, \theta) = \prod_{i=1}^{n} \left[g_1\left(\frac{a_i + b_i}{2}; \lambda\right) \cdot g_2\left(\frac{(b_i - a_i)^2}{12}; \beta\right) \right]$$

If both distributions are exponential with parameters λ and β, respectively, we obtain

$$L(\tau, \theta) = \prod_{i=1}^{n} \lambda \exp\left(-\lambda \frac{a_i + b_i}{2}\right) \times \beta \exp\left(-\beta \cdot \frac{(b_i - a_i)^2}{12}\right).$$

Next we maximize the log-likelihood and finally through simple calculations lead to the following results

$$\hat{\lambda}_{IL} = \frac{n}{\sum_{i=1}^{n} \frac{a_i + b_i}{2}}, \tag{13}$$

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n} \frac{(b_i - a_i)^2}{12}}.$$

5 Methods comparison

The goal of this section is to see which of the estimation methods presented in Sec. 4 is the best: primarily in terms of the estimator quality (like its unbiasendess and spread) and, secondly, in speed of algorithms convergence (in the case of iterative methods).

Firstly, some straightforward calculations show that coefficients (10) for the exponential distribution can be expressed by

$$\alpha_i(q) = \frac{1}{\lambda} + \frac{a_i \exp(-\lambda a_i) - b_i \exp(-\lambda b_i)}{\exp(-\lambda a_i) - \exp(-\lambda b_i)},$$

and substituted into (11) we receive (8), which means that both EM and IEM algorithms in the case of the exponential distribution produce identical estimators.

Hence, a natural question arises whether obtained equality of estimators happens only for some distributions, like the exponential, or is it a general rule. The following theorem can be proved.

Theorem 5.1 *Let $X_1,\dots,X_n$ denote a sample of i.i.d. random variables form the distribution with a density $g(x\,|\,\theta) = g_\theta(x)$, however, instead of the real-valued realizations x_i we observe intervals $\xi_i = [a_i, b_i]$, where $i = 1,\dots,n$. Then, assuming the same starting points both the EM algorithm and the IEM algorithm produce the same estimators in q-th iteration, i.e.*

$$\theta_{EM}^{(q+1)} = \theta_{IEM}^{(q+1)}.$$

Proof: Since $X_1,\dots,X_n$ are i.i.d. we have $g_\theta(\mathbb{x}) = \prod_{i=1}^n g_\theta(x_i)$. Moreover, assuming x_i are not directly observed and the only available data are intervals $\xi_i = [a_i, b_i]$, where $i = 1,\dots,n$, substituting these information into (4) we obtain

$$\begin{aligned} Q_{EM}(\theta\,|\,\theta^{(q)}) &= \mathbb{E}_{\theta^{(q)}}\left[\log L^c(\theta)\,|\,\mathbb{z}\right], \\ &= \int_{\mathscr{X}} \log\left[g_\theta(\mathbb{x})\right] \cdot \prod_{i=1}^{n} \frac{g_{\theta^{(q)}}(z_i)\mathbb{1}_{\xi_i}(z_i)}{\int_{\mathscr{X}} g_{\theta^{(q)}}(z_i)\mathbb{1}_{\xi_i}(z_i)\,dz_i}\,d\mathbb{x} \\ &= \frac{\int_{\mathscr{X}} \log\left[g_\theta(\mathbb{x})\right] \cdot g_{\theta^{(q)}}(\mathbb{x})\mathbb{1}(\mathbb{x})\,d\mathbb{x}}{\int_{\mathscr{X}} g_{\theta^{(q)}}(z_i)\mathbb{1}_{\xi_i}(z_i)\,dz_i}, \\ &= \frac{\int_{\mathscr{X}} \log\left[L(\theta;\xi)\right]\mathbb{1}(\mathbb{x})g_{\theta^{(q)}}(\mathbb{x})\,d\mathbb{x}}{L(\theta^{(q)};\xi)} = Q_{IEM}(\theta,\theta^{(q)}). \end{aligned}$$

Hence,

$$\begin{aligned} \theta_{EM}^{(q+1)} &= \arg\max_\theta Q_{EM}(\theta\,|\,\theta^{(q)}) \\ &= \arg\max_\theta Q_{IEM}(\theta\,|\,\theta^{(q)}) = \theta_{IEM}^{(q+1)}. \end{aligned}$$

■

To sum up, both the EM and IEM algorithms, despite their different backgrounds, come down to the same estimator. However, the advantage of the IEM estimator is its more natural and straightforward interpretation which identifies an interval sample with indicator functions representing real-valued observations originated from the uniform probability distribution.

When comparing the IEM and the Middle Censoring algorithms we have to underline a fundamental difference between the way of constructing their likelihood functions. In the case of MCA we have some additional limitation like that boundaries of the intervals should be independent and should come from exponential distributions (otherwise, one should be able to express the density and cdf of the borders in an analytical form). Extended simulations showed that under aforementioned assumptions the MCA algorithm turned out to work similarly as the IEM algorithm. However, usually more iterations were needed for the MCA to obtain the convergence than for the IEC algorithm.

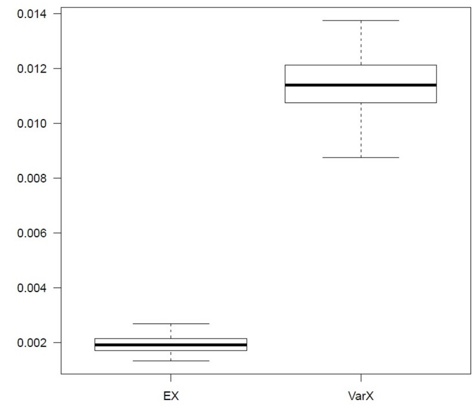

Figure 2: Box-plots for the difference of estimators of the mean and variance obtained with the induced likelihood method and the IEM method.

An obvious advantage of the induced likelihood method is that it is not an iterative procedure, therefore it works much “faster”. However, the simulation study showed that a weak point of the induced likelihood method is the variance estimation, which was significantly less stable than for the IEM method. Box-plots for the difference between estimators of the mean and variance obtained using both methods shown in Fig. 2. As it is seen, in the case of the variance the induced likelihood method generates systematically greater results than the IEM. estimator.

6 Conclusions

In the paper we discussed several methods of refinement for estimation with interval-valued data. It turned out that the most reasonable choice of all the methods presented is the IEM algorithm. Although our considerations were restricted to point estimation, one can apply the results for constructing more precise confidence intervals or tests indicating not triple but binary decisions.

However, the main goal of the paper was rather to emphasize the need for statistical tools for imprecise data which produce results precise enough to be satisfying for practitioners. This will not only increase the interest in statistics with interval data, but will also be in line with the general direction indicated by John W. Tukey: “All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" [21].

References

[1] P. Bertrand, F. Goupil, Descriptive Statistics for symbolic Data, in: H.H. Bock, E. Diday, Analysis of Symbolic Data, Springer, 2000, pp. 106–119.

[2] I. Couso, D. Dubois, Statistical reasoning with set-valued information: Ontic vs. epistemic views, International Journal of Approximate Reasoning 55 (2014) 1502–1518.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977), 1–38.

[4] T. Denœux, Maximum likelihood estimation from fuzzy data using the EM algorithm, Fuzzy Sets and Systems 183 (2011), 72–91.

[5] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, L. Ginzburg, Experimental uncertainty estimation and statistics for data having interval uncertainty, Sandia Report, SAND2007-0939, 2007.

[6] P. Grzegorzewski, The Kolmogorov goodness-of-fit test for interval-valued data, 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017), 2017, pp. 1–6.

[7] P. Grzegorzewski, The Kolmogorov-Smirnov goodness-of-fit test for interval-valued data, in: E. Gil et al. (Eds.), The Mathematics of the Uncertain, Springer, 2018, pp. 615–627.

[8] P. Grzegorzewski, M. Śpiewak, The sign test for interval-valued data, in: Ferraro M.B. et al. (Eds.), Soft Methods for Data Science, Springer, 2017, pp. 269–276.

[9] P. Grzegorzewski, M. Śpiewak, The Mann-Whitney test for interval-valued data, in: J. Kacprzyk et al. (Eds.), Advances in Fuzzy Logic and Technology, Springer, 2017, pp. 188–199.

[10] P. Grzegorzewski, M. Śpiewak, The sign test and the signed-rank test for interval-valued data, International Journal of Intelligent Systems 34 (2019) 2122–2150.

[11] S.K. Iyer, S. Rao Jammalamadaka, D. Kundu, Analysis of middle-censored data with exponential lifetime distributions, Journal of Statistical Planning and Inference 138 (2008) 3550–3560.

[12] A. Kołacz, P. Grzegorzewski, Asymptotic algorithm for computing the sample variance of interval data, Iranian Journal of Fuzzy Systems 16 (2019) 83–96.

[13] J. Le-Redemacher, L. Billard, Likelihood functions and some maximum likelihood estimators for symbolic data, Journal of Statistical Planning and Infernce 141 (2011) 1593–1602.

[14] R.E. Moore, Automatic error analysis in digital computation, Technical Report Space Div. Report LMSD 84821, Lockheed Missiles and Space Co., 1959.

[15] R.E. Moore, R.B. Kearfott, M.J. Cloud, Introduction to Interval Analysis, SIAM, 2009.

[16] H.T. Nguyen, V. Kreinovich, B. Wu, G. Xiang, Computing Statistics under Interval and Fuzzy Uncertainty, Springer, 2012.

[17] J. Perolat, I. Couso, K. Loquin, O. Strauss, Generalizing the Wilcoxon rank-sum test for interval data, International Journal of Approximate Reasoning 56 (2015) 108–121.

[18] Z.G. Su, P.H. Wang, Y.G. Li, Z-K. Zhou, Parameter estimation from interval–valued data using the expectation–maximization algorithm, Journal of Statistical Computation and Simulation 85 (2015) 320–338.

[19] T. Sunaga, Theory of interval algebra and its application to numerical analysis, RAAG Memoirs, Ggujutsu Bunken Fukuy-kai, Tokyo, 1958, pp. 29–46, 547–564.

[20] J.W. Tukey, Quick, compact, two-sample test to Duckworth's specifications, Technometrics 1 (1959) 31–48.

[21] J.W. Tukey, The future of data analysis, Annals of Mathematical Statistics 33 (1962) 1–67.

[22] M. Warmus, Calculus of approximations, Bulletin de l'Academie Polonaise de Sciences 4 (1956) 253–257.

[23] L.A. Zadeh, Computing with Words, Springer, 2013.