AStA Adv Stat Anal (2016) 100:475–494
DOI 10.1007/s10182-016-0274-z

ORIGINAL PAPER

Data generation processes and statistical management of interval data

Angela Blanco-Fernández[1] · **Peter Winker**[2]

Received: 2 September 2014 / Accepted: 24 June 2016 / Published online: 30 June 2016

Abstract Statistical methods for dealing with interval data have been developed for some time. Real intervals are the natural extension of real point values. They are commonly considered to generalize the nature of the experimental outcomes from the classical scenario to a more imprecise situation. Interval data have been mainly treated in the context of fuzzy models, as a particular case of increasing the level of imprecision of the data. However, specific methods to deal explicitly with interval data have also been developed. It is described which experimental settings might result in interval-valued data. Some of the major statistical procedures used to deal with interval data are presented. Given the quite different data generation processes resulting in interval data, it is discussed which method appears most appropriate for specific types of interval data. Some practical applications demonstrate the link between data generation processes, specific type of interval data, and statistical methods used for the analysis of these data.

Keywords Experimental data generation process · Interval data · Epistemic view · Ontic view · Possibilistic approach · Probabilistic approach

✉ Angela Blanco-Fernández
blancoangela@uniovi.es

Peter Winker
Peter.Winker@wirtschaft.uni-giessen.de

1 Department of Statistics and Operational Research, Oviedo University, C/ Calvo Sotelo s/n, 33007 Oviedo, Spain

2 Department of Statistics and Econometrics, Justus-Liebig University, Licher Strasse 64, 35394 Giessen, Germany

1 Introduction

A careful analysis of the data generation process associated with the realization of any random experiment is the key step to start the statistical modeling of the problem. Depending on the nature of the data generation process underlying an experimental setup, different types of data might be obtained. Most classical statistical methods deal with real-valued random variables, which model real point-valued outcomes. Such data are also labeled *crisp* data. Often, the set of possible outcomes underlies further restrictions, e.g., to only positive or integer values. Such restrictions also have to be taken into account when selecting an appropriate statistical model.

However, in the last decades, new types of variables have been considered to describe more general experimental scenarios. In particular, the outcomes of experiments might be better described by subsets of the real line instead of a point value. Then, real-valued intervals are a natural extension to be considered. Historically, such interval random variables have been treated in the context of *fuzzy* and set-valued models (see Matheron 1975). When considering fuzzy random variables (see Zadeh 1965; Puri and Ralescu 1986), in addition to the randomness of the data generation process, a potentially imprecise nature of the data is also taken into account. Consequently, fuzzy random variables are useful for modeling experimental settings for which the outcomes can be measured only with some imprecision, which is reflected by a real interval rather than by a point value. However, interval data might also be the direct outcome of specific experimental setups. Then, fuzzy modeling might not be the most appropriate statistical tool. Recently, a number of methods have been developed for this class of models (see Diamond 1990; Gil et al. 2001; D'Urso and Giordani 2004; Blanco-Fernández et al. 2011 among others). A detailed review on the relation between interval- and fuzzy approaches can be found in Corral et al. (2011).

In this paper, we describe some typical experimental settings and the corresponding data generating processes (DGPs), which result in interval data (Sect. 2). In Sect. 3, we introduce some methods used for the statistical analysis of interval-valued data and discuss which method might be most appropriate depending on the specific data generation processes. We conclude with an outlook to promising fields of research both with regard to the statistical methods and their application to data generating processes being interval-valued in Sect. 4.

2 Interval data as result of different DGPs

Interval-valued data—or interval data for short—might be the outcome of a random experiment in quite different situations. Depending on the characteristics of the study, including aspects such as the number of measurements obtained with regard to a single variable value, the applications of interval data models might be more appropriate for the further statistical analysis than the aggregation to point data and the usage of classical modeling tools. To understand better the properties of interval data and the indication for an explicit interval data approach, we provide a classification of DGPs which might result in interval data.

A first major distinction can be made with regard to the interpretation of interval data. In some cases, interval data are used to represent uncertainty or imprecision of measurements, which sometimes is reflected by the outcome of repeated measurements of the same quantity. We will describe DGPs of this type in Sect. 2.1. A different class comprises DGPs, where there is a genuine interest in the interval data themselves, e.g., reflecting the minimum and maximum over some (sub)sample. DGPs generating this type of data will be considered in Sect. 2.2.

This major classification coincides with the *epistemic* vs. *ontic* distinction—or disjunctive/conjunctive perception—of set-valued (or more generally fuzzy-valued) information in statistical reasoning, considered in the literature. Analogously to the preceding distinction, an epistemic—or disjunctive—set represents incomplete knowledge about an underlying precise object, whereas an ontic—or conjuntive—set represents itself the object of interest. For a further discussion on both the epistemic and the ontic points of view, from a wider and more philosophical perspective, readers are encouraged to look at Dubois and Couso (2014) and Blanco-Fernández et al. (2014), respectively.

2.1 Uncertainty and imprecision reflected in interval data

There exist situations in which the main interest focuses on a real-valued (crisp) magnitude, but it is impossible to observe the experimental outputs, i.e., the random realizations of the data generation process, since they are imprecisely identified. This may be due to different reasons.

Sometimes the exact value of a variable is not available to the researcher due to confidentiality issues; for example, personal answers to medical details, classified information about companies or people, etc. In such situations, it is common to group possible values and to allow the respondent to provide an interval which contains a representative value for his/her answer instead of the precise datum. The same procedure might also be applied to reduce the effort for filling a questionnaire. Thus, although the real value is available in principle, it will not be provided to the researcher, but given as an interval including this value. Just to provide a few examples of this situation, for instance, Jahanshahloo et al. (2008) consider a sample of inputs and outputs provided in interval form by several commercial bank branches in Iran. Alternatively, in medical studies as provided by Kristiansen et al. (2008) or Hodge et al. (2011), the answers to sensitive questions, such as the frequency of alcohol intake, are not provided in the form of precise values but only in the form of intervals. In recent social or educational surveys, as ALL-BUS (http://www.gesis.org/en/allbus), PASS (http://www.gl-assessment.co.uk/products/pass), PISA (http://www.oecd.org/pisa/), among others, this process is usually followed to collect some information from respondents who refuse to answer a direct question. Besides, some questions in these surveys, as the income, are answered by either a precise value or an interval on the respondent's own choice. It seems that lower incomes are likely to be reported as point values, and the higher the income, the more imprecision to report an exact value, so that an interval is provided. Some statistical methods for coarse or grouped data, as, for instance, *coarse at random* (CAR) assumptions, would fail in this situation, as commented next.

The exact values of a variable might be also not available because of the use of a non-sufficiently precise measurement device. When the possibility of errors in the observation of the experimental data is a fact, which is not easily described by a specific random mechanism allowing to provide distributional information, it is suitable to propagate the uncertainty around the observed value, leading to interval data. This technique is very common in topographic measurements; for instance, the localization of an object through the GPS coordinates is expressed in interval form in Abdallah et al. (2007). It is important to remark that in these situations, each interval represents a precise value imprecisely located in the interval. Thus, intervals provide an effective way to include in the statistical processing the uncertainty of measurement of the real (but unobservable) magnitude.

Rounding and censoring processes might also lead to this type of intervals. On one hand, rounded data arises when a (typically continuous) variable X is measured in a discrete manner, rounded to a certain level of accuracy, due to either simplicity or the imprecise nature of the measuring instrument. Rounded data, usually denoted by $x^* = x + \delta$, where x is the (unobservable) value of the variable X and δ is the associated rounding error, can be equivalently treated as interval representations of the value $x \in [x^* \pm |\delta|]$. Schneeweiss et al. (2010) present a clarifying review on rounding processes. On the other hand, interval censoring arises when a failure time T cannot be observed, but can only be determined to lie in an interval obtained from a sequence of examination times, i.e., $T_i \in (T_{i,L}, T_{i,U})$. Censoring processes usually appear in survival analysis and epidemiologic studies (see Huang and Wellner 1997 for a review on the topic). A more general class of incomplete data is considered with the so-called *coarse at random* (CAR) data, in which the rounding/grouping/censoring process fulfils that the conditional probability of the observed data is the same for all possible values of coarsened data, i.e., for all point values being consistent with the observed coarse data. Although this experimental situation seems to be quite rarely satisfied in practice, some motivating examples in medical and epidemiologic studies are provided by Heitjan and Rubin (1991). The crucial point is whether or not the event of being coarsened depends on the true value itself. If that happens, then the CAR assumption is violated, as already commented in the example of the income question in ALL-BUS or PASS surveys. Heitjan and Rubin (1991) show another illustrative example in this context, regarding the number of cigarettes smoked daily, whose exact/coarsened outcomes seem to depend on the lower/higher the true exact values are.

2.2 Explicit modeling of interval data

When the objects of analysis are the intervals themselves, we can differentiate between two situations, as they appear in experimental settings. In the first case, the interval data are obtained by aggregating several realizations of the DGP, while the second case covers processes which directly result in interval outcomes. We will discuss both settings in some more detail and provide examples in this subsection.

2.2.1 Interval data from aggregation

The aggregation of point data is done for several reasons. Let us introduce some examples in this context.

Symbolic data
Sometimes the aggregation is used to summarize information stored in large data sets, resulting in a smaller and more manageable data set which desirably preserves the essential information. This process leads to the so-called symbolic data. Depending on the groups of information being considered, the summarized data can be represented, e.g., by lists, distributions, or sets. Consequently, different kinds of symbolic data are employed: categorical multi-valued, histogram-valued, set-valued, and, in particular, interval-valued data. We refer Bock and Diday (2000) and Billard and Diday (2003) for detailed results on symbolic data analysis. The interval coding of symbolic data often represents descriptions of biological species or technical specifications. For instance, in Domingues et al. (2010), a sample of several physical characteristics of some mushroom species is summarized by means of intervals which have been obtained by aggregating individual mushrooms according to the kind of specie. In a similar way, some characteristics of different species of Italian peppers are described by intervals in Lauro and Palumbo (2005). In Duarte and Brito (2006), a list of technical specifications for some car models is summarized by several symbolic variables, some of them coded by intervals, as the height, the width, or the price of the cars corresponding to each model.

It is important to remark that in the symbolic data analysis, the intervals are considered to allow multiple values for each variable, by implicitly assuming an uniform distribution in each interval. They are viewed as a *description* of the (assumed uniform) behaviour of a real-valued variable on a group of individuals. This implicit assumption of an underlying uniform distribution seems rather ad hoc, and it entails troublesome feelings when applying statistics with it in mind. In Sect. 3.2, some drawbacks on the statistical analysis in the context of symbolic interval data are described.

Range of variation of a magnitude: longitudinal analysis
There are situations in which the statistical study does indeed focus on the range of variation of the values of a magnitude over a certain period of time or within a crossection. Let us consider, for example, the daily—equivalently monthly, yearly,...—fluctuation of a numeric random variable. It might occur that the DGP leading to point values of the variable executes only once per day—equivalently month, year, etc. However, it often happens that the variable is measured several times for each randomly selected individual in the considered period of time. In such cases, the classical procedure is to describe the behaviour of the variable in each individual during this time with a real number, the averaged value of all his/her registered values, in general. This clearly entails a great loss of information about the behaviour of the variable during the selected period. It is also possible to consider the range of variation of the variable for each observational unit, as the set of values between the minimum and the maximum registered during that period of time. This process leads to interval data, which can be considered as experimental outcomes of the (interval-valued) random variable *fluctuation of the magnitude over the fixed period of time*.

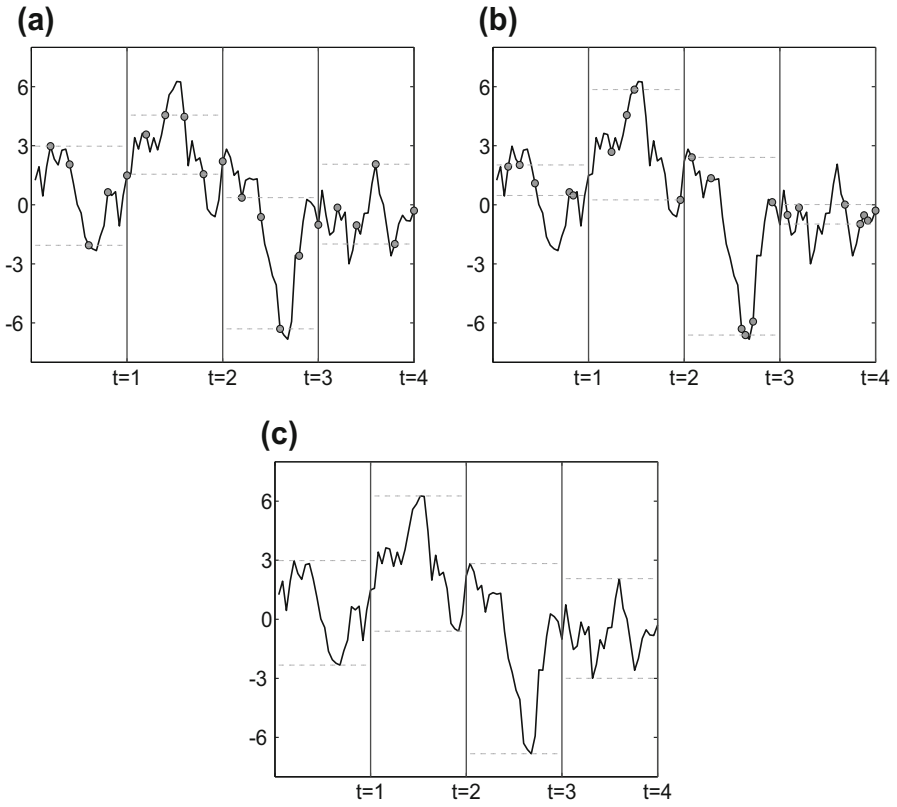

Fig. 1 Interval outcomes for the fluctuation of a magnitude: fixed (**a**), random (**b**), or continuous (**c**) number of measurements

Different approaches may result in this kind of interval variable depending of the number of measurements. Three typical situations are presented in Fig. 1 and will be discussed in the following. The solid lines in Fig. 1 show the (known or unknown) continuous behaviour of the real magnitude over time. The points in panels (a) and (b) represent the registers of the magnitude in certain moments of the time window—five fixed moments in (a) and a not-fixed but a random number of registers in (b). Finally, the dashed lines indicate the resulting measures of lower and upper bounds of the interval value of the fluctuation in each window time. Obviously, the three approaches lead to different interval outcomes for the same underlying DGP. The diverse descriptions of the variable might lead to different statistical conclusions when executing a statistical method to the interval-valued sample data set. The choice of the approach for the modeling of the DGP depends on the available measurement device.

First, each interval outcome is computed from a fixed number of registers of the magnitude during the period of time, i.e., from the minimum and the maximum of a set of a fixed number of random variables [see panel (a) of Fig. 1]. The following example illustrates a simple practical situation when this is the case.

Example 1 Consider the RAM consumption of the computers in an office. Suppose that it is measured for any computer at each of the eight working hours each day. By defining X_i = "RAM consumption of the computer at hour i", for $i = 1, \ldots, 8$, then $\widetilde{X}$ = "daily fluctuation of the RAM consumption of the computer", such that $\widetilde{X}(\omega) = [\min_{i=1,\ldots,8} X_i(\omega), \max_{i=1,\ldots,8} X_i(\omega)]$, where ω is a generic computer of the office, and is an interval-valued random variable.

A real application following this approach can be found in Ramos-Guajardo and González-Rodríguez (2013). It deals with the statistical analysis of the tidal daily fluctuation in a specific area of Spain, defining the interval outcomes by the minimum and maximum values over the set of heights of the tides registered every 10 min in the day.

Second, the number of times the magnitude is registered during a preset period may be random itself [see panel (b) of Fig. 1]. Thus, each outcome of the interval-valued variable is computed from the minimum and the maximum of a set of a random number of random variables. This situation is common, e.g., in medical monitoring. A typical example is the following.

Example 2 During the hospitalization period of a patient, his/her basic physical parameters, such as temperature, pulse, blood fluctuation, sugar level, etc. are measured at different times during a day, either by a nurse or an electronic device monitoring the patient's behaviour. In general, a different number of such indicators are gathered for each patient in a day. Moreover, some devices are programmed to record not the whole set of values of a parameter over a day, but only the lowest and highest values registered during the day. Thus, the outcomes of the interval variable $\widetilde{X}$= "daily fluctuation of a physical parameter of a patient" are defined in this case as $\widetilde{X}(\omega) = [\min_{i=1,\ldots,r_\omega} X_i, \max_{i=1,\ldots,r_\omega} X_i]$, where ω denotes a patient and r_ω is a random number of measurements of the parameter for patient ω during the day.

In Gil et al. (2001), González-Rodríguez et al. (2007), Blanco-Fernández et al. (2011), among other works on the development of statistical methods for interval-valued data, a medical case study on this scenario is addressed. It deals with the systolic and diastolic blood pressure daily fluctuations of a sample of patients in a hospital. The daily fluctuation of the pulse rate of the patients is additionally considered in González-Rodríguez et al. (2007).

Third, a further situation can appear when the real magnitude is measured continuously on the pre-fixed period of time, i.e., the consideration of interval data resulting as minimum and maximum of a time series within a certain time span [see panel (c) of Fig. 1]. In contrast to the previous examples, a full set of measurements might be available, but cannot be used due to data protection issues or cost of provision. The following example illustrates this case.

Example 3 In the situation of Example 2, if the electronic device controlling the physical behaviour of the patient monitors his/her status continuously over the day, then it is possible to define the interval variable $\widetilde{X}$ as $\widetilde{X}(\omega) = [\min_{t\in T}, X(t), \max_{t\in T}, X(t)]$, with T being the window time (the considered day in this case).

This approach is common in studies of financial markets, in which the statistical analysis of prices is based on data sampled at high frequency, ideally at the level of

individual transactions. The full set of measurements, however, is costly to obtain, not readily available for past periods, and prone to all sorts of micro-structure noise that may bias estimates, in general. When dealing with inter-daily data—equivalently monthly, yearly, this problem is usually overcome by replacing the intra-daily data to successive closing prices. By also considering the range, i.e., the interval spanned by the lowest and the highest recorded daily price, more intra-daily information might be effectively used. Consequently, the use of interval-valued data provides an alternative way to model and forecast the unobservable volatility of stock market returns with respect to the usual techniques. Some early applications in this setting are provided, e.g., by Hu and He (2007), Xu et al. (2008), and Fischer et al. (2015).

Range of variation of a magnitude: cross-sectional analysis
A range of variation for a variable of interest may also arise from a cross-sectional setting. In particular, with regard to expectations, from a cross section of individuals, the data can be aggregated by providing the range of their expectations with regard to some variable of interest. This aggregation is made either to maintain confidentiality of individual opinions or to deal with the situation of frequent changes in the composition of the cross section, e.g., due to non response. For instance, in Fischer et al. (2013), a data set of the range of members' forecasts for key macroeconomic variables published by the Federal Open Market Committee (FOMC) is employed. Two different interval forecasts are provided by FOMC; the full ranges of all individual forecasts and the truncated central tendency intervals (disregarding the three highest and the three lowest individual forecasts). Analogously to what happens in the previously described longitudinal setting, the general approach considers the midpoint of one of these intervals as the FOMC's "consensus" forecast and applies classical techniques for *crisp* data. However, in Fischer et al. (2013), it is shown that considering the midpoint of the interval of forecasts only, one effectively discards the information about how much dispersion surrounds this "consensus" forecast as given by the variation in the FOMC members' views. It is shown that the consideration of this variation of individual projections improves forecasts of future inflation.

2.2.2 Interval outcomes from the DGP

There are situations in which the experimental outcomes obtained from the execution of the DGP are real intervals themselves. As for the preceding case, different experimental scenarios may be considered.

Behaviour of a characteristic on groups of individuals
Sometimes the interval data corresponds to the direct result of some physical estimates. For instance, when the interest focuses on the study of the behaviour of a characteristic on groups of individuals in a population instead of single individuals, in this case, the DPG is usually executed in two steps; first, samples of each group in the population are chosen, and the characteristic of interest is measured on them. Second, the mean value and the standard deviation of the variable on each sample are computed. For normally distributed data, the interval [mean $\pm$ sd] summarizes effectively the characteristics in each group. Thus, these intervals are the experimental outcomes of the interval-valued

variable in focus, which might be described as the *behaviour of the characteristic on a group of individuals of the population*.

Many applications in this setting are found in the literature. For example, in Kallithraka et al. (2001), some essence parameters of wines of several geographical areas in Greece are considered (minerals, phenols, sensory values, etc.). A sample data set for 33 wine codes—from different vintage areas—is recorded through the intervals defined by the mean value and the standard deviation computed from the observation of a sample of wines of each variety. Medical studies usually handle this kind of data too. In Beunza et al. (2010), a statistical analysis on a sample of 10.376 participants of a study population in Spain is addressed. Some medical and physical characteristics are observed, and the behaviour of these characteristics on the individuals of the sample belonging to certain groups—pre-defined in terms of a score punctuation—is represented by intervals with the structure [mean $\pm$ sd].

Personal perceptions

Experimental data regarding personal perceptions or valuations can be also recorded as intervals. The most frequently employed technique to collect data about valuations in surveys is the well-known Likert scale. Recent studies have proposed more flexible and informative scales by employing fuzzy or interval data (see Gil and González-Rodríguez 2012; De la Rosa de Sáa et al. 2015). The use of a free-interval-valued response format instead of a pre-fixed and discrete list of answers enhances the variability and informational content of the experimental data. Intervals allow to describe the personal perception/valuation/judgement of the respondents in a rich and expressive way in many real-life situations as the following example demonstrates.

Example 4 The fire risk for vegetation is generally measured in terms of an index varying from a *very low* risk, to *low*, *moderate*, *high* or an *extreme* fire risk, codified with the integers ranging from 1 to 5, respectively (see Spano et al. 2003). Sometimes a continuous scale of fire risk ranging from 0 to 100 is alternatively used (see Gonzalez-Calvo et al. 2005). The fire risk index is evaluated for a certain place and time by an expert from several perspectives and factors; type of vegetation, meteorological conditions, water stress, etc. Some of these parameters can be precisely measured, but other factors entail a certain degree of subjectivity or imprecision on the expert perception. Thus, the expert must derive a precise value for the risk from some imprecise or subjective perceptions on the area. Recently, an experimental application using an interval-valued fire risk has been developed in the *Institute of Natural Resources and Territorial Planning* (INDUROT), in Asturias, Spain. Instead of representing their perception about the fire risk with a number, experts model their perception with an interval within the range 0–5 (0 meaning no risk and 5 extreme risk). In this case, the whole interval acts as the expert valuation, modeling jointly both the location and the range of the personal perception of fire risk. This is the main difference with the examples regarding personal answers presented in Sect. 2.1, in which the answers are exact, but unobservable and then *imprecisely measured* as intervals. The statistical analysis differs in each context, as it will be described in Sect. 3. In Table 1, some of the collected data are shown. In general, experts feel more comfortable by giving an interval response to their risk perception than by fixing a precise 1-,2-,3-,4-, or 5

Table 1 Interval-valued fire risk index in several areas in Asturias on 21/03/2012

Area	Fire risk index
Vegadeo	1.7–2
Grandas	3–4
Cangas Narcea	3.5–4.5
Luarca	2–2.4
Tineo	2.1–3.1
Salas	1–2
Pravia	2.8–3.2
Grado	2–2.5
Lena	1.5–2
Siero	2–2.5
Gijón	2–2.5
Laviana	2–2.2
Villaviciosa	1.5–2
Cangas Onís	2.5–3
Llanes	2–2.5

indices. Moreover, the interval-valued responses allow more versatility on the expert valuations; for example, two different experts giving a risk of 2 to the same area could provide different interval risk valuations [1.7, 2.1] and [2, 2.8]. Clearly, the perception of risk from the second expert is slightly greater and more imprecise than the first one.

3 Statistical methods for interval data

When the data generation process leads to interval data, the statistical analysis must be obviously adapted to work with this kind of data. Otherwise, e.g., when only the midpoint of the intervals is treated as if it was a crisp value, relevant information is lost. Since intervals are obtained from different experimental situations, as discussed in Sect. 2, the appropriate statistical methods for modeling these interval data may be also different.

The main distinction on the existing methods is the consideration of a *possibilistic* or a *probabilistic* approach.

3.1 Possibilistic approach

When intervals are considered as imprecise measurements of precise unobservable values, the loss of information about the point data is transferred to the statistical process. Thus, any statistical analysis, as parameter estimation, regression methods, inferential studies, and so on, also provides imprecise results. For instance, the computation of the sample mean in this framework is illustrated in the following example.

Example 5 Let $\{x_i\}_{i=1}^n$ be a sample of point values for which only a sample of intervals $\{X_i\}_{i=1}^n$, such that $x_i \in X_i$ is available. Obviously, the sample mean of $\{x_i\}_{i=1}^n$ cannot

be computed precisely, and so it might be defined as the set of all possible values for the sample mean of n values ranging over the intervals $X_1, \ldots, X_n$, respectively; that is

$$\widehat{\mu} = \left\{ \sum_{i=1}^{n} \tilde{x}_i / n : \tilde{x}_i \in X_i \right\} .$$

In general, parameter estimates are no longer point values but sets of *all possible* values of the corresponding precise estimate obtained from points belonging to the intervals. It is considered a *possibilistic* version of the estimator under interval uncertainty.

It is clear that this approach makes sense for interval data arising from the execution of a DGP in one of the situations described in Sect. 2.1, and under the epistemic or disjunctive point of view as presented in Dubois and Couso (2014). The statistical analysis focuses on the real magnitude being imprecisely measured; since it is affected by the loss of information on the data, the statistical conclusions regarding the underlying variable are also imprecise. Different statistical problems in this line have been developed, such as parameter estimation (Horowitz et al. 2003), distributions (Joslyn 1997), clustering (Pimentel and Souza 2014), or regression problems (Černý et al. 2013), to name but a few. A detailed review on some statistical methods in the framework of the epistemic view is presented in Dubois and Couso (2014).

Similar methods are applied in the context of partially identified models, with main applications in econometric studies (see Tamer 2010 for a detailed survey on the field). The partial (interval-valued) identification of the experimental data is transferred to the statistical model. Thus, the parameters are no longer identified by single points but by identification regions, i.e., sets of parameters compatible with the data and the model assumptions. One option is to consider the *possibilistic* collection of values, i.e., collecting all predictions based on all possible values compatible with the interval information, as illustrated in Example 5. Alternative methods are also proposed in recent literature to obtain identification regions for moments and distributions of variables (see Manski 2003; Stoye 2010). In particular for partially identified regression models based on interval data, different types of identification regions for the regression parameters are suggested in Manski and Tamer (2002), Cerquera et al. (2012), Schollmeyer and Augustin (2015), among others.

Engineering problems dealing with uncertainties on both initial states and system parameters usually employ algorithms, whose design is based on interval computation to obtain reliable enclosures to the corresponding states and parameters. Some applications in various fields of engineering are shown in Hofer and Rauh (2006).

When data comes from coarsing processes—rounded, grouped, and censored, CAR data as particular cases—specific statistical methods are available. A huge amount of literature is at hand. In general, these methods work by relating the statistical characteristics of the rounded/censored/coarsed variable with the corresponding characteristics of the original variable. For instance, if $X^* = X + \delta$ is a rounded variable from X (as presented in Sect. 2.1), analytic expressions for some univariate moments, as the expected value, the variance, and in general, the characteristic function of X^*, as well as linear regression estimates, are related to those of X (see Schneeweiss et al. 2010).

When working with CAR data, Heitjan and Rubin (1991) show that many statistical methods—based on likelihood and Bayesian inferences—are equivalent to those for grouped but non-random coarsed data, i.e., the stochastic nature of the coarsening process can be ignored when applying the technique.

3.2 Descriptive interval-extended approach

When the analysis is focused on a magnitude which is not real-valued any more, but exhibits an explicit interval-valued behaviour, the intervals obtained through the execution of the data generation process are considered as the experimental data to be statistically analyzed. This is the case for the DGPs described in Sect. 2.2. Depending on the theoretical scenario describing the experimental setting, different approaches might be considered.

The descriptive analysis of the experimental interval-valued data can be done without taking into account probabilistic assumptions on the random experiment. This is the case of symbolic data analysis. In symbolic data analysis, the interval variables are considered as multi-valued variables uniformly distributed over the interval. The statistical methods are generally based on a representation of each interval with certain real values (its lower and upper bounds, $A = [A_L, A_U]$, or its midpoint and range, $A = [A^c \pm A^s]$), assuming an uniform distribution for all the point values on the intervals, and then applying classical techniques to those real values. Linear regression problems (Domingues et al. 2010; Lima Neto and de Carvalho 2010), principal component analysis (Lauro and Palumbo 2005; Zuccolotto 2012), discriminant analysis (Duarte and Brito 2006), or clustering methods (Chavent et al. 2006) have been developed under this point of view. These methods do not take into account in general the interval arithmetic, and they usually fail with regard to the coherency of the results with the interval nature of the data. For instance, in Domingues et al. (2010), the estimation of a linear regression function for a sample of intervals $\{(X_i, Y_i)\}_{i=1}^n$ is solved through the classical fitting of the real-valued linear equations between the midpoints and the ranges of the intervals, i.e., $Y_i^c = \beta^c X_i^c + \varepsilon_i^c$ and $Y_i^s = \beta^s X_i^s + \varepsilon_i^s$, respectively. The prediction rule is then defined for any interval $\widehat{Y} = [Y_L, Y_U]$ as $\widehat{Y}^c = \widehat{\beta}^c X_i^c$ and $\widehat{Y}^s = \widehat{\beta}^s X_i^s$, and so $\widehat{Y}_L = \widehat{Y}^c - \widehat{Y}^s$ and $\widehat{Y}_U = \widehat{Y}^c + \widehat{Y}^s$. However, it cannot be excluded that $\widehat{Y}_L > \widehat{Y}_U$ for the prediction, i.e., that $\widehat{Y}$ is not a well-defined interval. Important efforts to overcome these drawbacks are still needed. Some attempts can be found in Lima Neto and de Carvalho (2010) and references therein. Another distinctive feature of the statistical methods developed in symbolic data analysis is that they do not fix a probabilistic scenario, i.e., they do not take into account the random execution of the DGP leading to the experimental outcomes of a random variable taking values on a certain probability space, but rather develop descriptive fitting methods for the available interval data set. With no probability assumptions on the statistical models, the definition of classical theoretical concepts, such as moments of the variables, stochastic independence, induced distribution, and so on, is not properly stated (but *ad hoc*). Finally, given that lack of statistical model, there is no basis for further inferential analysis after fitting the model to the data.

3.3 Probabilistic interval-extended approach

An alternative line of research tries to extend the classical techniques developed for real-valued random variables to the statistical treatment of random variables with interval experimental outcomes. These methods generally try to stick as close as possible to the classical statistical process while guaranteeing the interval nature of the experimental data.

3.3.1 Basic concepts

Given a probability space $(\Omega, \mathcal{A}, P)$, a random interval $\mathcal{X} : \Omega \to I$ (with I being the space of intervals in the real line) is defined as a Borel-measurable mapping with respect to the σ-field generated by the topology induced by a certain metric on the space I (see Matheron 1975). It is immediate to see the analogy of this concept with the one for a real random variable $X : \Omega \to \mathbb{R}$ defined on the same probability space. The unique distinction is the space of outcomes, interval or real values, respectively. The Borel measurability of the mapping $\mathcal{X}$ guarantees that one can properly refer to concepts, such as expectation or variance of a random interval, stochastic independence of random intervals, and so on (see Körner and Näther 2002; Molchanov 2005). Analogously to the descriptive approach in Sect. 3.2, random intervals can also be characterized by means of real-valued representatives, but they are soundly defined within the probabilistic setting in this case. Let $\mathcal{X} : \Omega \to I$ be a random interval. Following the previous notations, for each $\omega \in \Omega$, $\mathcal{X}(\omega) = [\mathcal{X}(\omega)_L, \mathcal{X}(\omega)_U]$ or equivalently, $\mathcal{X}(\omega) = [\mathcal{X}(\omega)^c \pm \mathcal{X}(\omega)^s]$. We define $\inf \mathcal{X} : \Omega \to \mathbb{R}$, such that $(\inf \mathcal{X})(\omega) = \mathcal{X}(\omega)_L$, for each $\omega \in \Omega$. Analogously $(\sup \mathcal{X})(\omega) = \mathcal{X}(\omega)_U$, $(\mathrm{mid}\mathcal{X})(\omega) = \mathcal{X}(\omega)^c$ and $(\mathrm{spr}\mathcal{X})(\omega) = \mathcal{X}(\omega)^s$, for all $\omega \in \Omega$.

$\mathcal{X} : \Omega \to I$ is a random interval if, and only if, one of the following conditions are fulfilled:

- $\inf \mathcal{X}, \sup \mathcal{X} : \Omega \to \mathbb{R}$ are real random variables verifying $\inf \mathcal{X} \leq \sup \mathcal{X}$ a.s.—[P]—i.e. $P(\{\omega \in \Omega : \inf \mathcal{X}(\omega) \leq \sup \mathcal{X}(\omega)\}) = 1$.
- $\mathrm{mid}\mathcal{X}, \mathrm{spr}\mathcal{X} : \Omega \to \mathbb{R}$ are real random variables such that $\mathrm{spr}\mathcal{X} \geq 0$ a.s.-[P].

The classical notion of probability established on Ω allows us to compute precise probabilities $P(\mathcal{X} \in A) = P(\{\omega \in \Omega : \mathcal{X}(\omega) \in A\})$, for any $A \in \sigma_I - \sigma_I$ denoting the σ field on I considered in the definition of $\mathcal{X}$. Thus, the classical condition of independence arises naturally: two random intervals $\mathcal{X}$ and $\mathcal{Y}$ associated with $(\Omega, \mathcal{A}, P)$ are independent if, and only if, $P(\mathcal{X} \in A, \mathcal{Y} \in B) = P(\mathcal{X} \in A)P(\mathcal{Y} \in B)$, for all $A, B \in \sigma_I$. The two most relevant summary measures associated with the distribution of a random interval are the mean and the variance. Their definition is usually based on Fréchet's generalization for abstract metric spaces (see Körner and Näther 2002), leading to the so-called *Aumann* expectation and the real-valued associated variance. Namely

$$E(\mathcal{X}) = \arg\min_{A \in I} E(d^2(\mathcal{X}, A)) \,, \tag{1}$$

where d is an L^2-type metric on the space I, and

$$\mathrm{Var}(X) = \min_{A \in I} E(d^2(\mathcal{X}, A)) = E(d^2(\mathcal{X}, E(\mathcal{X}))) \,, \tag{2}$$

whenever this arg min exists. It has been shown that $E(\mathcal{X})$ exists if, and only if, $\inf \mathcal{X}, \sup \mathcal{X} \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$, the class of integrable functions with respect to $(\Omega, \mathcal{A}, P)$, and it can be computed in that case as $E(\mathcal{X}) = [E(\inf \mathcal{X}), E(\sup \mathcal{X})] \in I$ (see Blanco-Fernández et al. 2014). It is easy to see that this concept of expectation for random intervals generalizes intuitively the classical expectation for real random variables, which is also defined, besides its usual expression, as the number minimizing the averaged squared distance to the observations.

The methodology for the statistical analysis of these measures as well as further results for random intervals is then based on the use of appropriate metrics for intervals. The well-known Hausdorff metric, defined as $d_{\mathrm{H}}(A, B) = \max\{\max_{a \in A} \min_{b \in B} |a - b|, \max_{b \in B} \min_{a \in A} |a - b|\}$, is L^1 type, and it only takes into account distances between the end points of the considered intervals. Therefore, Bertoluzza et al. (1995) introduced an L^2-type metric, which implies very good properties in connection with least squares and other optimization methods, and it additionally takes into account all the convex combinations between the end points of the intervals. Bertoluzza's metric has been initially defined as $d_{\mathrm{W}}(A, B) = \sqrt{\int_{[0,1]} (t(\inf A - \in B) + (1-t)(\sup A - \sup B))^2 dW(t)}$, where W is a (Borel) probability measure on $[0, 1]$. Recently, more intuitive and operative—but equivalent—expressions for this metric have been proposed. For instance, in Blanco-Fernández et al. (2011), it is employed $d_\theta(A, B) = \sqrt{(\mathrm{mid}A - \mathrm{mid}B)^2 + \theta(\mathrm{spr}A - \mathrm{spr}B)^2}$, with $\theta > 0$. A crucial point to highlight is that, despite their differences, both d_{H} and d_{W} induce the same topology on I, so the concept of random interval and further statistical properties can be equivalently addressed. In González-Rodríguez et al. (2009), a detailed discussion on the metrics and their equivalences is shown. Furthermore, in Blanco-Fernández et al. (2014), statistical methods based on a general expression of the original Bertoluzza's metric are presented. Some statistical procedures do not depend on the parameters defining the metric; for instance, the regression problem addressed in Blanco-Fernández et al. (2011) is solved irrespectively of the constant θ considered for the metric $d_\theta(A, B)$. Other statistical results indeed depend on the parameters defining the metric (see Blanco-Fernández et al. 2014 for some detailed examples).

Once the probability setting is formalized, the DGP is executed to obtain a simple random sample $\{\mathcal{X}_i\}_{i=1}^n$ from $\mathcal{X}$, and so descriptive and inferential studies for the variable $\mathcal{X}$ can be conducted.

3.3.2 General statistical treatment and interval arithmetic

The statistical treatment of the random sample of intervals $\{\mathcal{X}_i\}_{i=1}^n$ is done by means of the natural interval arithmetic, which is composed of the Minkowski addition and the product by scalars. Namely

$$A + B = \{a + b : a \in A, b \in B\}\ ,\ \text{and}\ \ \lambda A = \{\lambda a : a \in A\}\ ,$$

for any real intervals A, B and $\lambda \in \mathbb{R}$. They are inner operations in the space of real compact intervals; it is straightforward to see that $A + B = [\inf A + \inf B, \sup A + \sup B]$ and $\lambda A = [\lambda \inf A, \lambda \sup A]$ if $\lambda \geq 0$, $\lambda A = [\lambda \sup A, \lambda \inf A]$ for $\lambda < 0$. Thus, these operators conform to a natural arithmetic between intervals. However, the space is not linear but semilinear, due to the lack of a symmetric element with respect to the addition, in general. It is straightforward to see that $A + (-A) = \{0\}$ if, and only if, A is a singleton; otherwise, $A + (-A) = [-a, a]$, $a > 0$.

Besides, a universally acceptable complete ordering of intervals cannot be established. The statistical processing of experimental interval data must take into account these features, and the methods must guarantee the coherency of estimation and inferential results with the interval space. This is the main reason why the direct application of multivariate statistical methods to the real components of the intervals, as the linear regression problem presented in Sect. 3.2, usually fails.

For the sake of comparison with the possibilistic approach, in Example 6, the computation of the sample mean of a random interval is illustrated.

Example 6 Let $(\Omega, \mathcal{A}, P)$ be a probability space, $\mathcal{X} : \Omega \to I$ be a random interval, and $\{\mathcal{X}_i\}_{i=1}^n$ be a simple random sample of data obtained from $\mathcal{X}$. The sample mean of $\mathcal{X}$ is defined as

$$\overline{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i\ . \tag{3}$$

This expression mimics the classical definition for the mean of a sample of real values, being defined in terms of the interval arithmetic. It is straightforward to see that it is indeed an interval, which describes the average behaviour of the sample intervals $\mathcal{X}_1, \ldots, \mathcal{X}_n$. It can be equivalently computed in terms of classical sample means as $\overline{\mathcal{X}} = [\overline{\inf \mathcal{X}}, \overline{\sup \mathcal{X}}] = [\overline{\text{mid}\mathcal{X}} \pm \overline{\text{spr}\mathcal{X}}]$. However, different from the descriptive approach in Sect. 3.2, the coherency with the interval arithmetic and the probabilistic framework are kept, so the statistical analysis of $\mathcal{X}$ from $\{\mathcal{X}_i\}_{i=1}^n$ remains sound. In fact, the sample mean in (3) fulfils the Strong Law of Large Numbers with respect to the Aumman expectation in (1) (see Colubi et al. 1999). This property is crucial for statistical inference on expected values of random intervals—some methods are described next.

Some statistical problems in this line have been addressed already: e.g., estimation of moments (Körner and Näther 2002; Sinova et al. 2010), regression analysis (Gil et al. 2001; González-Rodríguez et al. 2007; Blanco-Fernández et al. 2011), principal component analysis (D'Urso and Giordani 2004; Giordani and Kiers 2004), ANOVA methods (Nakama et al. 2010), confidence sets (Blanco-Fernández et al. 2012), and hypothesis testing (Gil et al. 2007; Ramos-Guajardo and González-Rodríguez 2013). As commented before, the general procedure for the developments in this line tries to mimic classical techniques, by taking into account the particularities of the space of intervals.

3.3.3 Statistical inference

Since no realistic parametric model for describing the distribution of a random interval has been widely accepted until now, the exact distributions for the parameter estimators are not available. Thus, inferential studies in this setting are often based on asymptotic or bootstrap approaches. The limit distributions of the estimators must be obtained and then used for statistical inference (see Blanco-Fernández et al. 2012). As expected, asymptotic techniques provide accurate results for very large sample sizes, but they often lose accuracy for small or moderate samples. In these cases, bootstrap methods are known to improve the results. Furthermore, bootstrap techniques do not need the distribution of the estimator to be known, but they mimic that distribution empirically through the generation of a large number of samples of the variable based on a random and with replacement selection of individuals from the original sample (see Shao and Tu 1995 for detailed explanations).

For the sake of illustration of these techniques, and following the introduction of the expectation measure in (1) and the corresponding estimation in (3), some methods for statistical inference on the expected value of a random interval $\mathcal{X}$ are briefly described. More details can be found in Ramos-Guajardo and González-Rodríguez (2013), Colubi (2009), and Nakama et al. (2010). Let $\{\mathcal{X}_i\}_{i=1}^n$ be a simple random sample from $\mathcal{X}$. The construction of a confidence region for $E(\mathcal{X})$ at a fixed confidence level $\alpha \in (0,1)$ is done by looking for a value $\delta > 0$, such that $P(d(E(\mathcal{X}), \overline{\mathcal{X}}) < \delta) = \alpha$. The confidence region is then defined as the (interval) ball given by

$$\{A \in I : d(A, \overline{\mathcal{X}}) < \delta\}. \tag{4}$$

The radius δ is computed by bootstrapping (see Ramos-Guajardo and González-Rodríguez 2013). The algorithm is summarized as follows: Fix a large enough number of bootstrap replications B.

1. Obtain B bootstrap samples $\{\mathcal{X}_i^*\}_{i=1}^n$ by re-sampling uniformly and with replacement from $\{\mathcal{X}_i\}_{i=1}^n$, and compute the bootstrap sample mean $\overline{\mathcal{X}_b^*}$ for each one.
2. Compute the distance between the sample mean and each bootstrap sample mean, $d(\overline{\mathcal{X}}, \overline{\mathcal{X}^*}_b)$, for each $b = 1, \ldots, B$.
3. Choose δ as one of the α-quantiles of the sample $\{d(\overline{\mathcal{X}}, \overline{\mathcal{X}_b^*})\}_{b=1}^B$, i.e., the value so that at least $100\alpha\%$ of the computed distances are smaller than or equal to δ and at least $100(1-\alpha)\%$ of the computed distances are greater than or equal to δ.

Approaches to hypothesis testing for $E(\mathcal{X})$ are also available in the literature. Given $A \in I$, one can test $H_0 : E(\mathcal{X}) = A$ vs. $H_1 : E(\mathcal{X}) \neq A$ from the information provided by the sample $\{\mathcal{X}_i\}_{i=1}^n$ (see Colubi 2009). In addition, a recent work by Ramos-Guajardo et al. (2014) generalizes the *equality* test to an *inclusion* test, testing if $E(\mathcal{X})$ is (completely or to a certain degree) included in the interval A. The basic bootstrap equality-test procedure works as follows:

1. Compute the value of the test statistic $T = d^2(A, \overline{\mathcal{X}})$.
2. Obtain B bootstrap samples $\{\mathcal{X}_i^*\}_{i=1}^n$ from $\{\mathcal{X}_i\}_{i=1}^n$, and compute the bootstrap counterpart of the statistic $T_b^* = d^2(\overline{\mathcal{X}}, \overline{\mathcal{X}_b^*})$ for each $b = 1, \ldots, B$.

3. Compute the bootstrap p-value to test $H_0 : E(\mathcal{X}) = A$ as the proportion of values in $\{T_b^*\}_{b=1}^B$ being greater than T.

The equality test has been extended to the multiple case, when testing the equality of expectations of k random intervals in ANOVA problems (see Nakama et al. 2010).

3.4 Alternatives to model uncertainty based on a probabilistic approach

There exist in the literature some approaches based on standard probabilistic modeling that also take into account heterogeneity or variation in the data. For instance, mixed-effect models in longitudinal studies, containing both fixed and random effects, show some advantages in dealing with missing values, so that they are often preferred over more traditional approaches (see Liu et al. 2015). Measurement error models also deal with possible unobservable response and/or regressors, and they are usually estimated by means of probabilistic estimation methods based on distributional assumptions for the variables and the error involved in the model. For instance, Schneeweiss and Augustin (2006) propose several alternatives to estimate the conditional mean function of a regression model, $E(y \mid \varepsilon) = m(\varepsilon, \beta)$, when the regressor ε is unobservable (with density $h(\varepsilon, \gamma)$, unknown parameter γ), but it is related to an observable variable x through a measurement model $\varepsilon = x + \delta$, with conditional distribution $g(x \mid \varepsilon, \alpha)$ and the random error δ being normally distributed with zero mean and independent of ε and y (alternatively, Berkson models consider δ independent of x instead of ε). Although the idea of unobservable data in these models comes from the imprecision or uncertainty in the experimental observations, the statistical analysis is different from the possibilistic approaches in Sect. 3.1, since they make use of sound and sophisticated probabilistic distributional approaches.

4 Conclusions

Interval data appear quite often in different fields of sciences ranging from medical monitoring to surveys. It is of high importance to understand well the DGP of such interval data before selecting appropriate statistical modeling techniques. In fact, the stochastic processes leading to interval data differ substantially. This paper provides a classification of such processes providing also a substantial number of examples.

Given the specific nature of interval data, alternative methods are available for the further analysis, including both traditional methods, e.g., modeling only the midpoint of the intervals, and recently developed methods dealing explicitly with the interval nature of the data. Different approaches are presented and discussed. It is argued that choosing the appropriate statistical model as a function of the DGP is crucial also in the context of interval data. It might be even more important in this setting as compared to traditional statistical analysis given the substantial heterogeneity of DGPs resulting eventually in the observation of interval data. Some references to recent applications of such methods highlight the potential gains of explicit modeling instead of risking information loss when considering only a standard approach, e.g., by modeling the midpoint of the intervals, or when imposing inappropriate distributional assumptions.

Future research will concentrate on three issues: a refinement of the classification proposed for interval data, new methods for the collection of genuine interval data, e.g., in surveys, and further developments of estimation and inference methods, e.g., in the time series domain.

Acknowledgments Financial support from Spain's Ministerio de Economía y Competitividad through Grant MTM2013-44212-P and Ministerio de Ciencia e Innovación through Acciones Integradas Hispano-Alemanas 2012—cofinanced by the German Academic Exchange Service (DAAD), PPP Spanien 2012, Research Grant 54367957, is kindly acknowledged. Authors are grateful to the anonymous reviewer whose valuable suggestions have helped to improve the paper.

References

Abdallah, F., Gning, A., Bonnifait, P.: Adapting particle filter on interval data for dynamic state estimation. IEEE Int. Conf. Acoust. Speech Signal Proc. **ICASSP**(2), 1153–1156 (2007)

Bertoluzza, C., Corral, N., Salas, A.: On a new class of distances between fuzzy numbers. Mathware Soft Comput. **2**(2), 71–84 (1995)

Beunza, J., Toledo, E., Hu, F., Bes, M., Serrano, M., Sanchez, A., Martinez, J.A., Martinez, M.A.: Adherence to the Mediterranean diet, long-term weight change, and incident overweight or obesity: the Seguimiento Universidad de Navarra (SUN) cohort. Am. J. Clin. Nutr. **92**, 1484–1493 (2010)

Billard, L., Diday, E.: From the statistics of data to the statistics of knowledge: symbolic data analysis. J. Am. Stat. Assoc. **98**(462), 470–487 (2003)

Blanco-Fernández, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput. Stat. Data Anal. **55**(9), 2568–2578 (2011)

Blanco-Fernández, A., Colubi, A., González-Rodríguez, G.: Confidence sets in a linear regression model for interval data. J. Stat. Plan Inference **142**(6), 1320–1329 (2012)

Blanco-Fernández, A., Casals, R., Colubi, A., Corral, N., García-Bárzana, M., Gil, M.A., González-Rodríguez, G., López, T., Lubiano, A., Montenegro, M., Ramos-Guajardo, A., de la Rosa de Sáa, S., Sinova, B.: A distance-based statistical analysis of fuzzy number-valued data. Int. J. Approx. Reason. **55**, 1487–1501 (2014)

Bock, H.H., Diday, E.: Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Heidelberg (2000)

Černý, M., Antochb, J., Hladík, M.: On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data. Inf. Sci. **244**, 26–47 (2013)

Cerquera, D., Laisney, F., Ullrich, H.: Considerations on Partially Identified Regression Models. Working Papers of BETA No. 2012-07, ZEW. Centre for European Economic Research Discussion Paper No. 12-024 (2012). ftp://ftp.zew.de/pub/zew-docs/dp/dp12024

Chavent, M., Carvalho, F.A.T., Lechevallier, Y., Verde, R.: New clustering methods for interval data. Comput. Stat. **21**, 211–229 (2006)

Colubi, A., López-Díaz, M., Domínguez-Menchero, J.S., Gil, M.A.: A generalized strong law of large numbers. Probab. Theory Relat. **114**, 401–417 (1999)

Colubi, A.: Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data. Fuzzy Set Syst. **160**, 344–356 (2009)

Corral, N., Gil, M.A., Gil, P.: Interval and Fuzzy-valued approaches to the Statistical Management of Imprecise Data. In: Pardo, L., et al. (eds.) Modern Mathematical Tools and Techniques in Capturing Complexity. Understanding Complex Systems, pp. 453–468. Springer, Heidelberg (2011)

De la Rosa de Sáa, S., Gil, M.A., González-Rodríguez, G., López, M.T., Lubiano, M.A.: Fuzzy rating scale-basedquestionnaires and their statistical analysis. IEEE T Fuzzy Syst. **23**, 111–126 (2015)

Diamond, P.: Least squares fitting of compact set-valued data. J. Math. Anal. Appl. **147**, 531–544 (1990)

Domingues, M.A.O., de Souza, R., Cysneiros, F.J.A.: A robust method for linear regression of symbolic interval data. Pattern Recogn. Lett. **31**, 1991–1996 (2010)

Duarte, A.P., Brito, P.: Linear discriminant analysis for interval data. Comput. Stat. **21**(2), 289–308 (2006)

Dubois, D., Couso, I.: Statistical reasoning with set-valued information: Ontic vs. epistemic views. Int. J. Approx. Reason. **55**, 1502–1518 (2014)

D'Urso, P., Giordani, P.: A least squares approach to principal component analysis for interval valued data. Chemom. Intell. Lab. Syst. **70**(2), 179–192 (2004)
Fischer, H., García-Bárzana, M., Tillmann, P., Winker, P.: Evaluating FOMC forecast ranges: an interval data approach. Empir. Econ. **47**(1), 365–388 (2013)
Fischer, H., Blanco-Fernández, A., Winker, P.: Predicting stock return volatility: can we benefit from regression models for return intervals? J. Forecast. (2015) (forthcoming)
Gil, M.A., López-García, M.T., Lubiano, M.A., Montenegro, M.: Regression and correlation analyses of a linear relation between random intervals. Test **10**, 183–201 (2001)
Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval-valued data. Comput. Stat. Data Anal. **51**(6), 3002–3015 (2007)
Gil, M.A., González-Rodríguez, G.: Fuzzy vs Likert Scales in Statistics. In: Trillas, E., et al. (eds.) Combining Experimentation and Theory. A Hommage to Abe Mamdani. Studies in Fuzziness and Soft Computing 271, pp. 407–420. Springer, Heidelberg (2012)
Giordani, P., Kiers, H.A.L.: Three-way component analysis of interval valued data. J. Chemometr. **18**(5), 253–264 (2004)
Gonzalez-Calvo, A., Hernandez-Leal, P.A., Arbelo, M.: Forest Fire Risk Dynamic Index. In: De la Riva, J. et al. (eds.) Proceedimgs of 5th International Workshop on Remote Sensing and GIS Applications to Forest, pp. 125–129 (2005)
González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A.: Least squares estimation of linear regression models for convex compact random sets. Adv. Data Anal. Classif. **1**, 67–81 (2007)
González-Rodríguez, G., Trutschnig, W., Colubi, A.: Confidence regions for the mean of a fuzzy random variable. In: Abstracts of IFSA World Congress/EUSFLAT Conference (IFSA-EUSFLAT 2009, Lisbon, Portugal)
Hofer, E.P., Rauh, A.: Applications of Interval Algorithms in Engineering. In: Luther, W., Otten, W. (eds.) International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (12th GAMM—IMACS 2006, Germany). IEEE Computer Society Conference Publishing Services (2006)
Horowitz, J.L., Manski, C.F., Ponomareva, C.F., Stoye, J.: Computation of bounds on population parameters when the data are incomplete. Reliab. Comput. **9**(6), 419–440 (2003)
Heitjan, D.F., Rubin, D.B.: Ignorability and coarse data. Ann. Stat. **19**(4), 2244–2253 (1991)
Hodge, A.M., English, D.R., Itsiopoulos, C., ODea, K., Giles, G.G.: Does a mediterranean diet reduce the mortality risk associated with diabetes: evidence from the Melbourne Collaborative Cohort Study. Nutr. Metab. Cardiovasc. Dis. **21**, 733–739 (2011)
Hu, C., He, L.: An application of interval methods to stock market forecasting. Reliab. Comput. **13**, 423–434 (2007)
Huang, J., Wellner, J.: In: Lin, D.Y., Fleming, T.R. (eds.) Proceedings of First Seattle Symposium in Biostatistics, Lecture Notes in Statistics. Interval Censored Survival Data: A Review of Recent Progress, pp. 123–169. Springer, New York (1997)
Jahanshahloo, G.R., Lotfi, F.H., Malkhalifeh, M.R., Namin, M.A.: A generalized model for data envelopment analysis with interval data. Appl. Math. Model. **33**, 3237–3244 (2008)
Joslyn, C.: Measurement of possibilistic histograms from interval data. Int. J. Gen. Syst. **26**, 9–33 (1997)
Kallithraka, S., Arvanitoyannis, I.S., Kefalas, P., El-Zajouli, A., Soufleros, E., Psarra, E.: Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin. Food Chem. **73**, 501–514 (2001)
Körner, R., Näther, W.: On the Variance of Random FuzzyVariables. Statistical Modelling, Analysis and Management of FuzzyData, pp. 22–39. Springer, Berlin (2002)
Kristiansen, L., Gronbaek, M., Becker, U., Tolstrup, J.-S.: Risk of pancreatitis according to alcohol drinking habits: a population-based cohort study. Am. J. Epidemiol. **168**(8), 932–937 (2008)
Lauro, N.C., Palumbo, F.: Principal component analysis for non-precise data. In: Vichi et al. (eds.) New developments in classification and data analysis, pp. 173–184. Springer (2005)
Lima Neto, E.A., de Carvalho, F.A.T.: Constrained linear regressionmodels for symbolic interval-valued variables. Comput. Stat. Data Anal. **54**, 333–347 (2010)
Liu, J., Liu, W., Wu, L., Yan, G.: A flexible approach for multivariate mixed-effects models with non-ignorable missing values. J. Stat. Comput. Simul. **85**, 3727–3743 (2015)
Manski, C.F., Tamer, E.: Inference on regressions with interval data on a regressor or outcome. Econometrica **70**(2), 519–546 (2002)
Manski, C.F.: Partial Identification of Probability Distributions. Springer, New York (2003)

Matheron, G.: Random Sets and Integral Geometry. Wiley, New York (1975)
Molchanov, I.: Theory of Random Sets. Probability and its Applications. Springer, London (2005)
Nakama, T., Colubi, A., Lubiano, M.A.: Two-way analysis of variance for interval-valued data. In: Borgelt, C., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing 77, pp. 475–482. Springer, Heidelberg (2010)
Pimentel, B.A., de Souza, M.C.R.: Possibilistic clustering methods for interval-valued data. Int. J. Uncertain. Fuzzy **22**, 263–291 (2014)
Puri, M., Ralescu, D.: Fuzzy random variables. J. Math. Anal. Appl. **114**, 409–422 (1986)
Ramos-Guajardo, A.B., González-Rodríguez, G.: Testing the variability of interval data: an application to tidal fluctuation. In: Borgelt, C., et al. (eds.) Towards Advanced Data Analysis by Combining Soft Computing and Statistics. Studies in Fuzziness and Soft Computing 285, pp. 65–74. Springer, Heildelberg (2013)
Ramos-Guajardo, A.B., Colubi, A., González-Rodríguez, G.: Inclusion degree tests for the Aumann expectation of a random interval. Inf. Sci. **288**(20), 412–422 (2014)
Schneeweiss, H., Augustin, T.: Some recent advances in measurement error models and methods. Allgemeines Statistisches Archiv AStA **90**, 183–197 (2006)
Schneeweiss, H., Komlos, J., Ahmad, A.S.: Symmetric and asymmetric rounding: a review and some new results. Adv. Stat. Anal. **94**, 247–271 (2010)
Schollmeyer, G., Augustin, T.: Statistical modeling under partial identification: distinguishing three types of identification regions in regression analysis with interval data. Int. J. Approx. Reason. **56**, 224–248 (2015)
Shao, J., Tu, D.: The Jackknife and Bootstrap. Springer, New York (1995)
Sinova, B., Casals, M.R., Colubi, A., Gil, M.A.: The Median of a Random Interval. In: Borgelt, C., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing, 77, pp. 575–583. Springer, Heildeberg (2010)
Spano, D., Georgiadis, T., Duce, P., Rossi, F., Delitala, A., Dessy, C., Bianco, G.: A fire index for mediterranean vegetation based on micrometeorological and ecophysiological measurements. Am. Meteorol. Soc. 3.1 (2003). https://ams.confex.com/ams/pdfpapers/65497.pdf
Stoye, J.: Partial identification of spread parameters. Quant. Econ. **1**, 323–357 (2010)
Tamer, E.: Partial identification in Econometrics. Annu. Rev. Econ. **2**, 167–195 (2010)
Xu, S., Chen, X., Han, A.: Interval/Probabilistic Uncertainty and Non-classical Logics. In: Huynh, V.N., et al. (eds.) Interval Forecasting of Crude Oil Price, pp. 353–363. Springer, Heidelberg (2008)
Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
Zuccolotto, P.: Principal component analysis with interval imputed missing values. Adv Stat Anal **96**, 123 (2012)