Exploring Distribution of Fuzzy Data

Eugenia Stoimenova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bontchev str. block 8
1113 Sofia, Bulgaria
Email: jeni@math.bas.bg

***Abstract*—This paper is concerned with the nonparametric estimation of a density function when the data are incomplete due to fuzzyness. The aim is to contribute to better description and presentation of density distribution of non-precise data. The histogram density estimator is modified to allow description of such data. An interactive R application is developed to explore different estimates.**

***Index Terms*—fuzzy data, histogram density plot, nonparametric density estimation**

I. Modeling of distribution

Let X be a continuous random variable and f its probability density function (pdf). Specifying the pdf we have a natural description of the distribution of X on the universe $\mathcal{U}$. From the pdf we can also calculate the mean and variance of X (if they exist) and the probability that X will take on values in a certain interval. The pdf is, thus, very useful to characterize the distribution of the random variable X.

In practice, the pdf of some observable random variable X is in general unknown. When it cannot be specified, an estimate of this density may be performed by using a sample of n observations independent and identically distributed $(X_1, \ldots, X_n)$ of X. We shall assume that the n observations are independent and that they all indeed come from the same distribution, namely $f(x)$. That is, we will be concerned with estimating $f(x)$ based on such a sample of data.

Nonparametric density estimation allows the form of the density to be determined entirely by the data. The resulting estimator is called empirical density function (edf). Common methods assume the observations to be *exact* numbers or vectors. That is, assuming continuous distribution, we "believe" that observed data are real numbers. Nevertheless it is not possible in practice, data rounding to rational points is usually small enough and most of the mathematical models are applicable without loss of precision. The literature on nonparametric density estimation spans the modern era of statistics. Several excellent books are devoted to the theory and practice of nonparametric density estimation with usual continuous assumption, i.e. [4], [6] and [5]. Many of the important applications of density estimation are to description of bivariate data, but since all the multivariate methods are generalizations of univariate methods, it is worth justifying the univariate case first.

The exact observing is often not realistic assumption because measurement results of continuous quantities are always not precise numbers but more or less non-precise. The imprecision could be quite large regarding variable values. This kind of uncertainty is different from errors and variability where the errors are relatively small. Whereas errors and variability can be modeled by stochastic variables and probability distributions, imprecision is another kind of uncertainty, called fuzziness.

Unlike fuzzy logic where the interest is logic with fuzzy concepts, in this paper we are facing statistics with fuzzy data. Based on fuzzy data description, density estimation has to be adapted. For a quantitative description of fuzzy data the most up-to-date method is to use fuzzy numbers and fuzzy vectors which are special fuzzy models (see [1]).

By fuzzy data we mean imprecise data which are recorded as interval-valued observations in which data consist of intervals rather than points of some domain of interest. Interval-valued observations can arise in different contexts. For example, rounding of a number with fixed decimal digits is an interval (which is symmetrical around the rounding point and contains the true value). Here, we are interested in the observed intervals themselves.

A fuzzy set is characterized by its characteristic function that indicate the membership grade of an element of $\mathcal{U}$ in the set in question. Larger values denote higher degrees of set membership. Such a function is called membership function, and the set defined by it a fuzzy set.

Our aim is to involve non-precise data in a classical histogram density estimator in order to extent this notion for fuzzy data. This is possible and explained in the paper. Clearly if fuzzy sets can be defined they should generalize crisp sets. The concept of membership in crisp sets will be extended to accommodate for partial membership, i.e. with degrees of membership varying gradually from 0 to 1.

In section II we describe the statistical meaning of fuzzy data. In section III the classical histogram density estimator is described. Then in section IV the modified estimator is proposed to allow histogramming of fuzzy data.

In the last section a real data set will be used to illustrate the method. It comprises the dates of origin of 807 mediaeval manuscripts.

The R implementation of proposed histogram is in interactive graphic environment. The program uses the tcltk package [3] to provide simple functions for building of a control panel

for the graphics. The user can change number of bins by a slider and explore different shapes of the distribution.

II. Coarse data and fuzzy sets

Coarse data refer to data with low quality. We run into this type of data in situations such as missing data, censor data (say, in survival analysis), and grouped data. It is the difficulty in observing accurate data that leads to coarse data. This might happen in performing random experiments as well as observing random phenomena in nature.

Specifically, suppose the random vector X of interest cannot be observed with accuracy. The statistician then tries to extract as much as possible useful information about the values of X, say, by localizing them in "observable" subsets of $\mathcal{U}$. In other words, when we cannot subsets of X with accuracy, we coarsen its space of values, i.e., replacing $\mathcal{U}$ with some collection of subsets of it, called a coarsening scheme.

Formally, suppose our random variable X of interest takes values in $\mathcal{U}$. Suppose X cannot be observed directly but its outcomes can be located in a finite partition $\tau_1, \tau_2, \ldots$ of $\mathcal{U}$ which is independent of X. A such partition is referred to as a coarsening of X. Define

$$R = \inf\{\tau_j : \tau_j \geq X\}, \qquad L = \sup\{\tau_j : \tau_j \leq X\}.$$

The conditional density of X given the interval $Y = [L, R]$ is then

$$f(x|Y) = \frac{f(x)}{\int_Y f(u)\,du}, \quad x \in Y. \tag{1}$$

For each X_i the corresponding interval Y_i may be defined on a different partition of $\mathcal{X}$. The conditional density is itself unknown.

Now, let the random sample $X_1, X_2, \ldots, X_n$ from X cannot be observed, but each X_i can be located in one of the elements A_j of a measurable partition $\{A_j,\ j = 1, 2, \ldots, m\}$ of $\mathbb{R}$. Thus the coarse data associated with $X_1, X_2, \ldots, X_n$ is $A_1, A_2, \ldots, A_n$ (where $X_i \in A_i$, almost sure). Put it differently, let $A_1, A_2, \ldots, A_n$ be a partition of $\mathbb{R}$. Suppose each value of X can be located in exactly one of the A_i's. Then the problem of coarse data is modeled by a (finite) random set S taking values in $\{A_1, A_2, \ldots, A_n\}$ with

$$P(S = A_i) = P(X \in A_i), \quad , i = 1, 2, \ldots, n.$$

It is clear that the relation between the variable X and its coarsening scheme S is that X is an almost sure (as) selector of S, i.e., $P(X \in S) = 1$.

Fuzzy sets are generalization of classic (crisp) sets. In the classical set theory an element x is either a member or nonmember of B, subset of universe $\mathcal{U}$. A set is characterized by its indicator function (also called characteristic function), that declares which elements of $\mathcal{U}$ are members of the set and which are not. Set A is defined by its characteristic function $\mu_B : \mathcal{U} \to \{0, 1\}$, as follows:

$$\mu_B(x) = \begin{cases} 1, & \text{if } x \in B \\ 0, & \text{if } x \notin B \end{cases}$$

A characteristic function of a crisp set assign a value of either 1 or 0 to each individual in universal set, thereby discriminating between members and nonmembers of the crisp set under consideration.

The histogram density estimators is based on accumulation of indicator events, linked to the ability to decide whether the observation X belongs to a subset B of $\mathcal{U}$, the universe, or not. This decision is equivalent to the question whether it is true that $X \in B$ or not (this is a binary question). However, in many practical cases, this question cannot be precisely answered. A reasonable solution consists in using a scale whose elements would express various degrees of truth of $X \in B$, and A thus becomes a fuzzy set. The most commonly used range of membership values of membership functions is the unit interval $[0,\ 1]$.

In view of the bijection between sets and their membership functions, we should try to formalize the concepts and use these generalize membership functions to define fuzzy sets. The concept of membership in crisp sets will be extended to accommodate for partial membership, i.e. with degree of membership varying gradually between 0 and 1. In section IV an appropriate membership function is defined to allow extension of the standard histogram estimator.

The relation between values of a variable X, which is not directly observable, and fuzzy sets is that we can define a density estimator that compromises fuzzy histogram bins and membership functions of the X-values with respect the probability distribution of X.

III. Standard histogram

In the study of the empirical density function (e.d.f.) the first difficulty is to find its best definition. Revesz [2] gives a number of possible definitions of the e.d.f. based on a sample with *exact* data from the underlying distribution.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed random variables with a density f, i.e. $X_1, X_2, \ldots, X_n$ is a sample of size n. Further, let $[a,\ b]$ be an interval of the real line and denote the number of the elements of the sample $X_1, X_2, \ldots, X_n$ lying in the interval $[a,\ b]$ by $K_n(a, b)$. The probability that an observation of X will fall into the bin $[b_j - \frac{h}{2}, b_j + \frac{h}{2})$ is given by

$$P\left(X \in \left[b_j - \frac{h}{2}, b_j + \frac{h}{2}\right)\right) = \int_{b_j - \frac{h}{2}}^{b_j + \frac{h}{2}} f(u)\,du \tag{2}$$

which is just the shaded area under the density between $b_j - \frac{h}{2}$ and $b_j + \frac{h}{2}$. This area can be approximated by a bar with height $f(b_j)$ and width h.

Then the probability $\int_a^b f(t)\,dt$ can be estimated by the relative frequency $\frac{1}{n}K_n(a, b)$ (if n is big enough) while the value $f(x)$, $(a < x \leq b)$ can be estimated by $\frac{1}{b-a}\int_a^b f(t)\,dt$. Therefore an empirical density function (e.d.f.) in the interval $[a, b]$ can be defined by

$$\hat{f}(x) = \frac{K_n(a, b)}{n(b - a)}, \qquad a < x \leq b.$$

The most commonly used nonparametric density estimator is the *histogram*. Its historical advantage is the ease of calculation – it can be reduced to putting registration notes into bins. We assume a distribution F on the real line with density $f(x)$ and look at plots based on independent sample points $X_1, \ldots, X_n$ from this distribution f.

The construction of a histogram is fairly simple. It consists in partitioning a given reference interval into specified number of bins B_j and in counting the number of observations belonging to each cell B_j. These are the steps:

- Select an origin $x_0 = 0$ and divide the real line into bins of binwidth h:
$$B_j = [x_0 + (j-1)h, x_0 + jh), \quad j \in \mathbb{Z}.$$
- Count how many observations fall into each bin. Denote the number of observations that fall into bin j by n_j.
- For each bin divide the frequency count by the sample size n (to convert them into relative frequencies, the sample analog of probabilities), and by the binwidth h (to make sure that the area under the histogram is equal to one):
$$f_j = \frac{n_j}{nh}.$$
- Plot the histogram by erecting a bar over each bin with height f_j and width h.

More formally, the histogram is given by

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} \mathbb{I}_{B_j}(X_i)\mathbb{I}_{B_j}(x), \tag{3}$$

where

$$\mathbb{I}_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Note that formula (3) (as well as its corresponding graph, the histogram) gives an estimate of f for all x. Denote by b_j the center of the bin B_j. It is easy to see from formula (3) that the histogram assigns each x in $B_j = [b_j - \frac{h}{2}, b_j + \frac{h}{2})$ the same estimate for f, namely $\widehat{f}_h(b_j)$.

The indicator function $\mathbb{I}_B$ of a subset B of U is a membership function in the sense that its values indicate whether a point u in U is a member of A or not. Here, there are only two degrees of membership: 1 for full membership and 0 for non-membership. The correspondence between B (as a specified collection of elements) and its membership function A is a bijection. Since B is a mathematical entity well defined, the use of $\mathbb{I}_B$ is often just a matter of convenience, i.e. without $\mathbb{I}_B$ we still can manipulate B. The situation is different for fuzzy concepts. Now that in probability theory, sets are used to describe events.

This is the natural way to define the empirical density function.

The indicator function $\mathbb{I}_{B_j}(\cdot)$ of a bin B_j is a membership function in the sense that its values indicate whether a point x in $\mathcal{U}$ is a member of B_j or not. Here, there are only two degrees of membership: 1 for full membership and 0 for non-membership. The correspondence between A and its membership function $\mathbb{I}_{B_j}$ is a bijection. The situation is different for for fuzzy concepts. Now that in probability theory, sets are used to describe events.

Fuzzy data are defined to be the result of observing continuous variables only up to the intervals containing the true values. Such data arise in a number of natural ways. In observing natural phenomena, we might not be able to record correctly the values of locate them with some degree of accuracy, or more generally, locate them in some regions (subsets) of the sample space. Thus the observations are present but not known! For any particular recorded value we know that the true value belongs to some known interval around this point. For details, see [1].

IV. Empirical density function for fuzzy data

We will follow Revesz' [2] construction to define e.d.f. based on a fuzzy data. The usual histogram density estimator is modified to allow description of such fuzzy data. An interactive R application is developed to explore different estimates.

Let $Y_1, \ldots, Y_n$ be a sample of observed intervals, where $Y_i = [L_i, R_i]$. We assume that each X_i has an equal impact in the sample so the corresponding interval $Y_i = [L_i, R_i]$ should have. A common approach is to assume the conditional distribution over the interval Y_i to be uniform, i.e. we suppose that X_i occurs at any point of the observed interval $[L_i, R_i]$ equally likely and the total mass in the observed interval is $1/n$.

Further, let B be a possible histogram bin, i.e. an interval of the real line, and denote by $M_n(B)$ the restricted mass of all elements of the sample $Y_1, \ldots, Y_n$ into this interval. More precisely, the restriction function is defined as follow:

Notation: For a non-zero interval $Y = (c, d)$ the restriction function over an interval $B = (a,\ b)$ is denoted by $\mathbb{J}_B(Y)$ and equals

$$\mathbb{J}_B(Y) = \begin{cases} \dfrac{\min(d,b) - \max(c,a)}{b-a} & \text{if } B \cap Y \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

For two intercept intervals, the restriction function gives the ratio of the length of the common part of the intervals to the length of the second one (here (a, b)). Then $M_n(B)$ is the sum of all parts of the observed intervals:

$$M_n(B) = \sum_{i=1}^{n} \mathbb{J}_B(Y_i).$$

Definition of histogram: Let

$$\ldots < b_{-1} < b_0 < b_1 < \ldots$$

be a partition of the real line such that

$$b_{i+1} - b_i = h \qquad (i = 0, \pm 1, \pm 2, \ldots).$$

The partition defines a bin sets $B_j = (b_j, b_{j+1})$. Then the histogram can be defined by

$$\hat{f}_c(x) = \frac{M_n(B_j)}{h} \quad \text{for} \quad x \in B_j. \tag{4}$$

It can be easily verified that the area of a histogram is indeed equal to one, a property that we certainly require from any reasonable estimator of a pdf.

Therefore an empirical density function (e.d.f.) in the interval $[a, b]$ can be defined by

$$\hat{f}(x) = \frac{M_n(a,b)}{n(b-a)}, \qquad a < x \leq b.$$

Fig. 1 gives an illustrative interpretation of equation 4 for two observations, Y_1 and Y_2. On the left the observed mass is plotted for each observation. The levels at vertical axes are $\frac{1}{R_i - L_i}$, $i = 1, 2$, and assure unit mass for each of them.

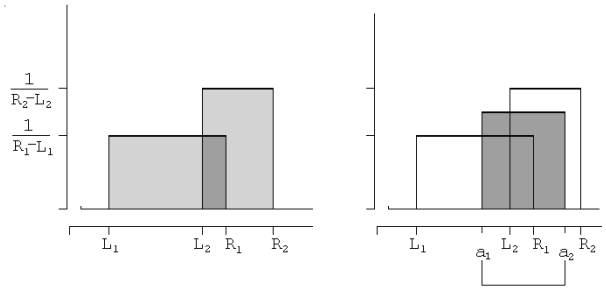

Fig. 1. Restricting observed mass in a bin.

Further, the method calculates an estimate of the density function in any bin $[a_1;\ a_2]$ as follows. let a histogram bin is such that $L_1 < a_1 < L_2$ and $R_1 < a_2 < R_2$ (see right plot on Fig. 1). Both observed intervals , Y_1 and Y_2, partially cover the bin $[a;\ b]$. The part from Y_1 is $\frac{R_1 - a_1}{R_1 - L_1}$ and the part from Y_2 is $\frac{a_2 - L_2}{R_2 - L_2}$. Therefore

$$\hat{f}(x) = \frac{R_1 - a_1}{R_1 - L_1} + \frac{a_2 - L_2}{R_2 - L_2}, \quad a \leq x \leq b$$

is the total mass of the sample into interval $[a_1;\ a_2]$.

The two observed intervals Y_1 and Y_2 cover partially $[a_1;\ a_2]$. The remaining mass out this bin is used to estimate f in the neighbor bins. Two adjacent bins $[a_0;\ a_1]$ and $[a_2;\ a_3]$ are partially covered by Y_1 and Y_2. Fig. 2) gives an illustration of the density estimate in the three bins which contain part of observed mass of Y_1 and Y_2. The mass in $[a_0;\ a_1]$ is $\frac{a_2 - L_1}{R_1 - L_1}$ while the mass in $[a_2;\ a_3]$ is $\frac{R_2 - a_3}{R_2 - L_2}$. The total observed mass over these 3 bins is obviously 2.

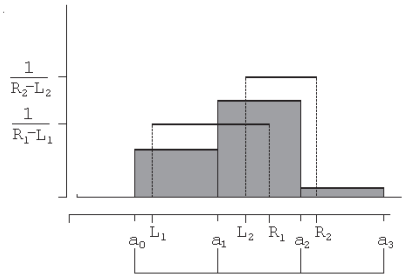

Fig. 2. Restricting observed mass in three bins.

The special case when an observed interval entirely belongs to an histogram bin, the restriction gives usual counting of 1 for this bin.

V. Interactive choice of binwidth

In this section the method considered in this paper is applied to a sequence of 807 intervals of "Notbefore" and "Notafter" dates of origin of mediaeval manuscripts[1]. Figure 3 represents the data as line segments with length equal to the observed interval length $R - L$ and height approximately proportional to $log(1/R - L)$. The two values "Notbefore" and "Notafter" are common in catalogue descriptions and specifies the interval $[L, R]$ for possible origin of the item. The exact dated documents are less that 25% of the data and are depicted at the top level on the Figure 3. The histogram density estimation is applied to a sequence of 807 intervals of this type.

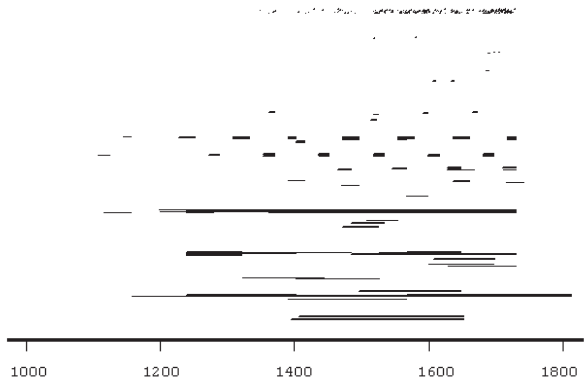

Fig. 3. Raw interval data for medieval manuscripts.

A very natural desire is in the informal investigation of the distribution of manuscripts over time. Here density estimates can give valuable indication of such features as skewness and multimodality in the data. Multimodality of an estimate is of interest in describing chronology data, since they specify possible historical upward and downward trend periods of society development.

The performance of $\hat{f}(x)$ depends critically on the choice of binwidth h that controls the smoothness of the estimates. There are various methods proposed for automatic selecting an optimal choice for complete data and often it minimizes some error functional. Note, that any automatic choice of the smoothing parameter should be view as a benchmark, and need to be adjusted based on subjective impression.

The R implementation of proposed histogram is in interactive graphic environment. The program uses the tcltk package [3] to provide simple functions for building of a control panel for the graphics. The user can change number of bins by a slider and explore different shapes of the distribution.

Figure 4 illustrates the real-time evolution of the fuzzy histogram density estimator. A snapshot of the estimators obtained for 12 is shown.

[1]The data were collected during the work on the KT-DigiCult-Bg project (MTKD-CT-509754, Marie Curie Programme, FP6). I am most grateful to Milena Dobreva for making this data set available to me.

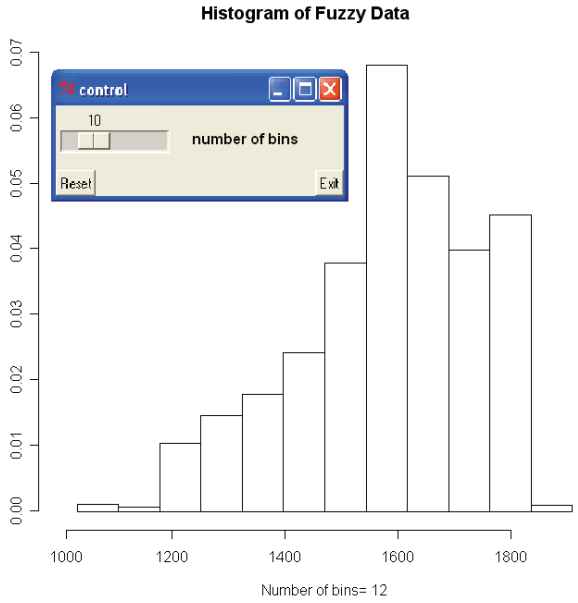

Fig. 4. Interactive exploring of different fuzzy histograms.

The general aim of exploratory data analysis is to find interesting features in data, and to bring them to human perception. It depends on the context and on our intentions to say what is an interesting feature. Using appropriate density plots, i.e. appropriate density estimates, we might get information about the relative location of the mean, or about the skewness, or many other details. We should try to incorporate all available information from data, i.e. fuzziness of the observation to derive good enough estimate of the density or other distribution properties. Full information is not presented in any single graph, but we can examine many graphs by some easy interactive plot tool.

Finally, the drawback of nonparametric methods for density estimation is, that they do not reduce the space, but instead all data points have to be kept in computations. This is undesirable, as especially in image retrieval the matching has to be performed very often. Exploring different density plots, i.e. different density estimates, we aim to suggest appropriate parametric model of the distribution. In the parametric density estimation the model is fully specified by a parametric family and only a finite set of parameters is unknown. Hence, estimating the density is a problem equivalent to estimating the parameters an consequently the computational issues in parametric model are much less than in nonparametric.

REFERENCES

[1] H. T. Nguyen and B. Wu, *Fundamentals of Statistics with Fuzzy Data*, Studies in Fuzziness and Soft Computing, Vol. 198, Springer, 2006.
[2] P. Revesz. (1972) On empirical density function, *Period. Math. Hung.*, Vol. 2, 85-110.
[3] L. Tierney, (2006) tkrplot: simple mechenism for placing R graphs in tk widget", Available: http://cran.r-project.org/.
[4] B. W. Silverman, *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. London - New York: Chapman and Hall, 1986.
[5] J. S. Simonoff, *Smoothing Methods in Statistics*. Springer, New York, 1996.
[6] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization.*, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York, NY: Wiley, 1992.