The median of a set of histogram data

Lidia Rivoli, Antonio Irpino and Rosanna Verde

Abstract Nowadays, several sources produce very large amount of data for which storage, processing and analysis becomes impractical. An approach for dealing with these data is to summarize the observations by means of synthetic representations. In this paper, we consider data summarized by histograms and we propose a novel distance-based method for defining the median histogram of a set of histograms and, in general, histogram-order statistics.

Key words: histogram data, cumulative distribution function, order statistics.

1 Introduction

Massive datasets are continuously generated by several sources: sensor networks, web traffic logs, financial transactions, security systems. Often, the limits in storage capability, communication bandwidth and computational resources impose to treat these data using suitable summaries. A typical example of such summaries is the histogram: it is parsimonious with respect to storage requirements and it provides an idea of the data underlying distribution. In Symbolic Data Analysis, the *Histogram Variable* is a particular case of symbolic multi-valued modal variable and several techniques (Clustering, Regression, PCA, ...) have been proposed [4] to analyze histogram data.

Lidia Rivoli
Università di Napoli Federico II, e-mail: lidia.rivoli@unina.it

Antonio Irpino
Seconda Università di Napoli Facoltà di Studi Politici e Mediterranei e-mail: antonio.irpino@unina2.it

Rosanna Verde
Seconda Università di Napoli Facoltà di Studi Politici e Mediterranei e-mail: rosanna.verde@unina2.it

In this paper, we deal with data described by means of a set of histograms using a Symbolic Data Analysis (in short SDA) approach with the aim to propose histogram-order statistics for a Histogram Variable. We, specifically, show how to compute the median-histogram and the quartiles-histograms. Basic statistics (like the sample mean or the standard deviation of a histogram variable) have been proposed in [4] and [5].

In general, order statistics depend on the definition of an order relationship and this is not trivial for histograms. Considering a histogram as a particular set-valued description, the issue is very close to the definition of an ordering among multivariate data, while considering histogram as an estimate of density function, we can consider it like a functional data. In multivariate data analysis and in functional data analysis, a proposal for the definition of an ordering among data is based on the concept of *data depth* [3], [2] which measures the degree of centrality of an observation with respect to the whole dataset.

To tackle the ordering definition problem, using the quantile functions associated with each histogram, we propose a method based on the minimization of a ℓ_1 Wasserstein distance based criterion. It extends the properties of the classic median and allows to identify a quantile function whose values correspond to the median quantiles of the set of histograms, in a point-wise way.

2 Definition of the order statistics

Let $\mathbf{Y}$ be a continuous variable defined on a finite support $\mathbf{S} = [\underline{y}; \bar{y}]$, where $\underline{y}$ and $\bar{y}$ are the minimum and maximum values of the domain of $\mathbf{Y}$. The support $\mathbf{S}$ is partitioned into a set of contiguous and non overlapping intervals (bins). Thus given n observations of $\mathbf{Y}$, a histogram H is a representation of $\mathbf{Y}$ consisting of a finite number of pairs $\{(I_k, f_k); k = 1, \ldots, K\}$ where $I_k = [\underline{y}_k, \bar{y}_k)$ (with $\underline{y}_k \leq \bar{y}_k$), are the K bins of the histogram and f_k are the associated relative frequencies (that is, the number of observed values contained in I_k normalized by n).

A *Histogram Variable* $\mathbf{H}$ is a symbolic multi-valued variable whose realizations are histograms [4]. Let's $\{H_j\}_{j=1,\ldots,N}$ be a set of N realizations of $\mathbf{H}$ with $H_j = \{(I_{jk}, f_{jk}); k = 1, \ldots, K_j\}$. Since it is assumed that the values are uniformly distributed within each interval $I_{jk} = [\underline{y}_{jk}, \bar{y}_{jk})$, the cumulative distribution function associated to each H_j is:

$$F_j(x) = \begin{cases} 0 & if\, x < \underline{y}_{j1}, \\ \sum\limits_{l=1}^{k-1} f_{jl} + \frac{x - \underline{y}_{jk}}{\bar{y}_{jk} - \underline{y}_{jk}} f_{jk}, & if\, \underline{y}_{jk} \leq x < \bar{y}_{jk}, \\ 1 & if\ x \geq \bar{y}_{jK_j}. \end{cases} \tag{1}$$

The first order statistics for the set $\{H_j\}_{j=1,\ldots,N}$ which will be defined is the median. Taking into account its proprieties in descriptive statistics and as shown in [1], the *Median histogram* can be defined as the histogram H_M minimizing an appropriate

ℓ_1 distance between histogram data:

$$\min_{H_M} \sum_{j=1}^{N} d_W(H_j, H_M) = \min_{H_M} \sum_{j=1}^{N} \int_0^1 \left| F_j^{-1}(t) - F_M^{-1}(t) \right| dt, \tag{2}$$

where d_W is also known as Wasserstein distance, F_j^{-1} and F_M^{-1} are the inverse functions of the cumulative distribution function F_j and F_M associated to H_j and H_M respectively. For determining H_M or the *Median distribution* F_M, we need to consider the cumulated relative frequencies of each bin I_{jk} that is, $w_{jk} = \sum_{l=1}^{k} f_{jl},\ k = 1, \dots, K_j$. Thus, the set

$$\mathbf{w} = \left\{ w_{11}, ..., w_{1K_1}, \dots, w_{j1}, ..., w_{jK_j} ..., w_{J1}, ..., w_{JK_J} \right\} \tag{3}$$

will consist of cumulated relative frequencies associated to all histograms $H_j,\ j = 1, \dots, N$. It is evident that $F_j(x)$ and $F_{j'}(x)$ can intersect in a point x^* then, the value w^* such that $w^* = F_j(x^*) = F_{j'}(x^*)$ will be put in $\mathbf{w}$. The element of $\mathbf{w}$ are sorted and only different values are kept determining the set:

$$\mathbf{w} = \{w_1, \dots, w_l, \dots, w_L\}, \tag{4}$$

where $w_1 = 0,\ w_L = 1$ and $\max_{1 \le j \le J} K_j \le L \le \sum_{j=1}^{J} K_j - 1$.

For each $l = 1, \dots, L$ and for each $j = 1, \dots, N$ the values of $F_j^{-1}(w_l)$ are known or they can be easily calculated if they are not yet available so that, we can consider the set $S(w_l) = \left\{ F_{[1]}^{-1}(w_l), \dots, F_{[j]}^{-1}(w_l), \dots, F_{[N]}^{-1}(w_l) \right\}$, containing the ordered quantile values for each fixed l. The Median distribution can be defined as the piecewise linear function joining the L medians of $S(w_l)$ that is, the values $F_{\left[\frac{N+1}{2}\right]}^{-1}(w_l),\ \forall l = 1, \dots, L$ if N is odds; otherwise, joins the average of values $F_{\left[\frac{N}{2}\right]}^{-1}(w_l)$ and $F_{\left[\frac{N}{2}+1\right]}^{-1}(w_l)\ \forall l = 1, \dots, L$.

In [1], the authors highlight that the definition of Median histogram with an even number of histogram is non unique. They propose of considering a function joining any quantile enclose between $F_{\left[\frac{N}{2}\right]}^{-1}(w_l)$ and $F_{\left[\frac{N}{2}+1\right]}^{-1}(w_l)\ \forall l = 1, \dots, L$. Instead in our proposal, we extend the definition of median used in descriptive statistics with even number of observations.

Furthermore, it is noteworthy that, according to (2), H_M is the barycenter histogram for the Wasserstein distance so as the Average histogram is the barycenter for the Mallow's distance [5], [6].

Similarly to Median distribution, the *p-th distribution* in the set of the distribution functions $\{F_i(t)\}_{i=1,\dots,N}$ is the piecewise linear function which joins the L values $F_{[p]}^{-1}(w_l)\ \forall l = 1, \dots, L$ that is, the values which are in the p-th position in each set $S(w_l)\ \forall l = 1, \dots, L$. We also give the definition of the other order statistics:

- the *Minimum* is the histogram associated to the piecewise linear function obtained joining the values $\min S(w_l)\ \forall l = 1, \dots, L$;

- the *First Quartile* is the histogram associated to the piecewise linear function joining the L values $F^{-1}_{\left[\frac{N+1}{4}\right]}(w_l)$, $\forall l = 1,\ldots,L$ if J is odds; otherwise, the averages of $F^{-1}_{\left[\frac{N}{4}\right]}(w_l)$ and $F^{-1}_{\left[\frac{N}{4}+1\right]}(w_l)$ $\forall l = 1,\ldots,L$;
- the *Third Quartile* is the histogram associated to the piecewise linear function joining the L values $F^{-1}_{\left[\frac{3(N+1)}{4}\right]}(w_l)$, $\forall l = 1,\ldots,L$ if J is odds; otherwise, the averages of $F^{-1}_{\left[\frac{3N}{4}\right]}(w_l)$ and $F^{-1}_{\left[\frac{3N}{4}+1\right]}(w_l)$ $\forall l = 1,\ldots,L$;
- the *Maximum* is the histogram associated to the piecewise linear function obtained joining the values $\max S(w_l)$ $\forall l = 1,\ldots,L$.

Obviously, the distributions associated to histogram order statistics can belong or not to the set of observed distributions $\{F_j\}_{j=1,\ldots,N}$. Furthermore, extending the five-number summaries into a *five-histogram* summaries it is also possible to extend the classical definition of *box and whiskers* plot [7] for a set of histogram data. For example, we can represent the order statistics of the set of histograms plotting on a graph their cumulative distribution functions.

Future research should concentrate on the investigation of variants and improvements towards a proposal of a histogram box-plot, and especially, on the definition of criteria for the choice of the histogram-whiskers and, consequently, a criterion to identify potential outliers.

References

1. Arroyo, J., Mate, C., Munoz San Roque, A.: Smoothing Methods for Histogram-valued Time Series. An application to Value-at-Risk. Statistical Analysis and Data Mining **4**(2), 216–228 (2011)
2. Sun, Y., Genton, M.G.: Functional boxplots. Journal of Computational and Graphical Statistics **20**, 316–334 (2011)
3. Hyndman, R.J., Shang, H.L.: Rainbow plots, bagplots and boxplots for functional data. Journal of Computational and Graphical Statistics **19**, 29–45 (2010)
4. Billard, L., Diday, E.: From the statistics of data to the statistics of knowledge: Symbolic data analysis. Journal of the American Statistical Association **98**, 470–487 (2003)
5. Irpino A., Verde, R.: Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A., Vichi, M., Advances in computational statistics, pp. 869-876. Physica-Verlag, Heidelberg (2006)
6. Verde, R., Irpino, A.: Dynamic clustering of histogram data: using the right metric. In: Studies in Classification, Data Analysis, and Knowledge Organization Part I, 123–134 (2007)
7. Tukey, John W.: Exploratory Data Analysis. Addison-Wesley (1977)