Article can be accessed online at http://www.publishingindia.com

Analytically Yours

Interval-Valued Data Analysis

Arnab Kumar Laha*

Interval-valued data arise in real-life in many different ways. Weather data published daily in newspapers contains only the maximum and minimum temperature readings during a day for a city, stock market data contains the highest and lowest traded price of a stock or a stock-index on a day etc. In the Big Data context often data are aggregated for certain features of interest giving rise to interval-valued data. For example, a credit card company need not store the entire history of credit card transactions of its customers but instead may store only the minimum and maximum amounts transacted by them. Thus for each customer the bank stores an interval which is the range of the amounts transacted by the customer.

You must now be wondering about the kind of decision problems that can be solved by using such interval data. Here are two examples:

1) Imagine yourself to be an investor in the stock market. You may like to know the usual range of variation of returns on a certain stock during the course of a day to decide whether to invest in that stock or not.

2) Imagine yourself to be a functionary of a credit card company in-charge of preventing credit card frauds. You would be interested in the usual range of amounts transacted by customers of different categories. If the transaction amount for a transaction falls outside the range specified for customers of that category, then you may take necessary action to verify the credentials of the person making the transaction before allowing the same.

Interval-valued data is a special type of Symbolic data. There can be many other forms of Symbolic data as discussed in Billard, 2011. In this short note we discuss analysis of interval-valued data with some applications.

Let $[a_i, b_i]$, i=1,...,n be a random sample of n intervals. We are interested in providing a summary of the data by providing a representative "mean interval". A naive approach is to use $[\tilde{a},\tilde{b}]$ as the "mean interval" where $\tilde{a}=\frac{a_1+...+a_n}{n}$ and $\tilde{b}=\frac{b_1+...+b_n}{n}$. While this method is expected to perform well for datasets in which there are no outliers, it may perform quite badly in case the dataset has outliers. A robust alternative can be to use the interval $[a',b']$ where $a'=\text{median}\{a_1,...,a_n\}$ and $b'=\text{median}\{b_1,...,b_n\}$. A third approach described by Le-Rademacher and Billard (2011) is described below.

Assuming that the variable of interest (say, temperature) of the i-th day is uniformly distributed in the interval $[a_i, b_i]$ we can compute the mean $\Theta_{1i} = (a_i+b_i)/2$ and the variance $\Theta_{2i} = (b_i - a_i)^2/12$. The random variables $\boldsymbol{\Theta}_i = (\Theta_{1i}, \Theta_{2i})$ are referred to as the internal parameters of the i-th interval $[a_i, b_i]$. Le-Rademacher and Billard (2011) assumes that the random variables Θ_{1i} and Θ_{2i} are independent with Θ_{1i} distributed as $N(\mu, \sigma^2)$ and Θ_{2i} distributed as $\text{Exp}(\beta)$ (where $\beta = E(\Theta_{2i})$) for all i =1,..., n. They obtain the MLEs of μ, σ and β based on the observed interval-valued data. Then the "mean interval" $[\hat{a},\hat{b}]$ is computed by solving $\hat{\mu}=\frac{\hat{a}+\hat{b}}{2}$ and $\hat{\beta}=\frac{(\hat{b}-\hat{a})^2}{12}$. This yields $\hat{a}=\hat{\mu}-\sqrt{3\hat{\beta}}$ and $\hat{b}=\hat{\mu}+\sqrt{3\hat{\beta}}$. The standard errors of $\hat{a}$ and $\hat{b}$ can be easily computed and they are $\text{se}(\hat{a})=\text{se}(\hat{b})=\sqrt{\{\text{se}(\hat{\mu})\}^2+3\left\{\text{se}\left(\sqrt{\hat{\beta}}\right)\right\}^2}$. Note that $\text{Cov}(\hat{\mu},\sqrt{\hat{\beta}})=0$ since the Θ_{1i} and Θ_{2i} are assumed to be

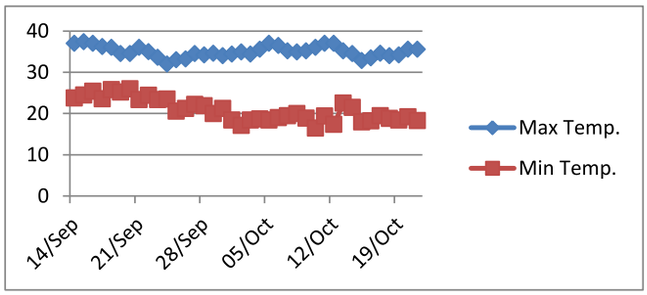
* Professor, Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iimahd.ernet.in

independent for all i=1,...,n. It is not difficult to see that the results can be easily extended to those cases where the internal parameters may have distributions other than those postulated in Le-Rademacher and Billard (2011).

A natural question to ask would be that which of these three "mean intervals" gives us the best representation for a given interval-valued dataset. For that purpose we define for any proposed interval [x,y] a measure of inaccuracy as $L([x,y]) = \frac{1}{n}\sum_{i=1}^{n}(|x-a_i|+|y-b_i|)$. Note that L([x,y]) = 0 is the best case which is attained only when $a_i = x$ and $b_i = y$ for all i = 1,..., n. Further observe that, an interval which yields a lower value of L gives a better representation of the data than another which yields a higher value of L. We compute the value of this inaccuracy measure L for the above three intervals and the one which gives the least value of L is chosen as the mean interval for representing the given interval dataset.

We illustrate the above methodology with two real life datasets. The first dataset correspond to the maximum and minimum temperatures at Ahmedabad during the period 14th September 2015 to 21st October 2015. The data is described graphically in Figure 1 below.

Fig 1: Maximum and Minimum Temperatures at Ahmedabad for the Period 14th September, 2015 to 21st October, 2015



Simple computations give $[\tilde{a},\tilde{b}]$=[20.8, 35.0] and $[a',b']$ = [20, 34.9]. To compute $[\hat{a},\hat{b}]$ first the internal parameters (Θ_{1i}, Θ_{2i}) are computed for i=1,...n. The distribution of the internal parameters did not satisfy the assumptions made in the paper by Le-Rademacher and Billard (2011). However that does not cause any hindrance to the computation of the mean interval. We use the sample mean as an estimate of the population mean and compute $\hat{a} = \frac{1}{n}\sum_{i=1}^{n}\Theta_{1i} - \sqrt{3\left(\frac{1}{n}\sum_{i=1}^{n}\Theta_{2i}\right)}$ and $\hat{b} = \frac{1}{n}\sum_{i=1}^{n}\Theta_{1i} + \sqrt{3\left(\frac{1}{n}\sum_{i=1}^{n}\Theta_{2i}\right)}$. Simple computations now yield $[\hat{a},\hat{b}]$= [20.7, 35.1]. Thus we find that the three methods give very similar estimates of the mean interval. Now to choose one amongst the three we use the measure of inaccuracy L. Again, simple computations yield L([20.8, 35.0])= 3.426, L([20, 34.9])= 3.368 and L([20.7, 35.1])= 3.426. Hence, the interval [20, 34.9] gives the best representation of the variation of temperature during the given period.

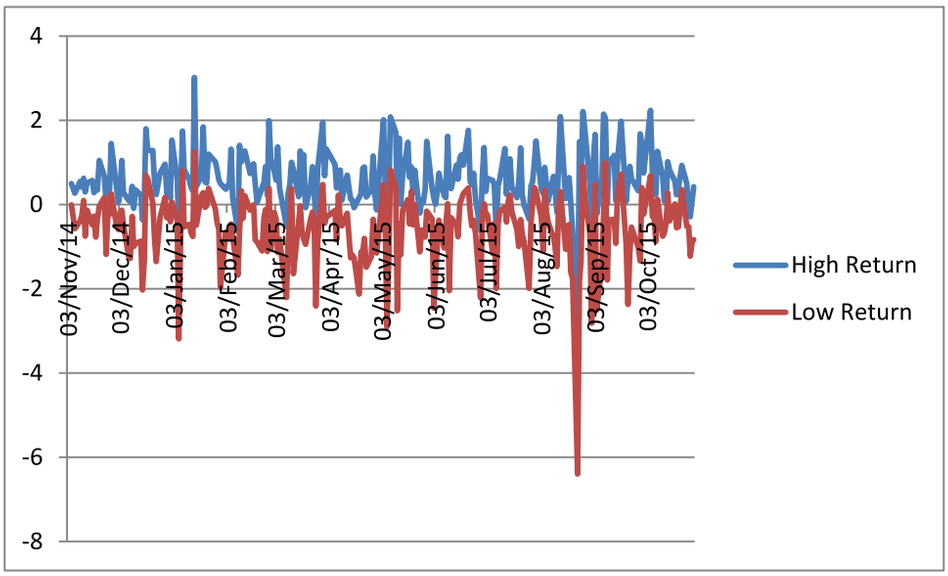
As a second example we consider CNX NIFTY data for the period 3rd November, 2014 to 30th October, 2015. Let H(t), L(t) and C(t) denote the High, Low and Closing values of NIFTY for day t. These are used to compute the variables High and Low Returns of day t (which are denoted as HR(t) and LR(t) respectively) as $HR(t) = \frac{H(t)-C(t-1)}{C(t-1)}\times 100$ and $LR(t) = \frac{L(t)-C(t-1)}{C(t-1)}\times 100$. The data is described graphically in Figure 2 below:

As in the earlier case we compute the mean interval by all the three methods and also the measure of inaccuracy (L). The results are given in Table 1 below:

Table 1: The Mean Intervals Obtained by the Three Methods and Their Measure of Inaccuracy

	$[\tilde{a},\tilde{b}]$	$[a',b']$	$[\hat{a},\hat{b}]$
Obtained Interval	[-0.59, 0.62]	[-0.47, 0.51]	[-0.65, 0.68]
Measure of Inaccuracy (L)	1.058	1.036	1.082

Fig 2: High and Low Returns of CNX NIFTY during the period 3rd November, 2014 to 30th October, 2015



We observe that there is substantial difference among the three intervals possibly due to the presence of outliers in the dataset. Since the interval [-0.47, 0.51] yields the least value of the inaccuracy measure (L) we use this interval for representing the variation of returns on CNX NIFTY within a day. An investor investing in CNX NIFTY would be advised to expect a variation in return between -0.47% to +0.51% during the course of a normal day.

References

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining*, 4(2), 149-156.

Le-Rademacher, J., & Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference*, 141, 1593-1602.