Indian Journal of Agricultural Sciences **84** (3): 424–7, March 2014/Short Communication
https://doi.org/10.56093/ijas.v84i3.38600

Descriptive statistics for symbolic interval-valued data

H K RANGANATH[1], PRAJNESHU[2] and HIMADRI GHOSH[3]

Indian Agricultural Statistics Research Institute, New Delhi 110012

Received: 18 August 2012; Revised accepted: 10 January 2014

Key words: Descriptive statistics, Histogram, Interval-valued data, Pan evaporation, Symbolic data

In statistics, data are usually formatted as single values. However, sometimes data are represented by intervals, histograms or even distributions. To deal with these kinds of data, concept of Symbolic data was introduced, which have an internal structure unlike classical data. Therefore, usual statistical methods cannot be readily applied for analysing Symbolic data. These types of data generally arise in two situations: Throughout data collection and processing. Some data collected are inherently Symbolic; some become Symbolic data after processing. An example of naturally collected Symbolic data is measurement of say, blood pressure where measuring device actually identifies an interval (even though value may be recorded as a single value) due to inherent continual changes in a person's blood pressure. In contrast, by recording changes throughout the day and from day to day, result is not a single value but a range of values, i.e. an interval. Another example of Symbolic data is Income level. Survey analysts know well that asking a person about his income directly usually does not elicit correct answer. Instead, a multiple choice question with different income levels, such as A[5000,10000), B[10000,20000), C[20000,50000), D[50000 and more) is usually utilized in a questionnaire. Another main reason to induce Symbolic data is that sometimes datasets are very large. With the introduction of modern computer science, massive size datasets are becoming routine, while methods of analysis are not so. Even performing a simple exploratory statistical analysis may involve a huge amount of computing power. To solve this problem, a lot of work has been done to obtain more efficient algorithms. While these improved algorithms help, they are still limited in usefulness. However, a simple way to solve this problem is to aggregate individual observations into groups of interest, especially when characteristics of groups are of higher interest to an analyst than those of individual observations. Thus, original dataset is summarized into one of a more manageable size while retaining as much interesting knowledge as possible. As an example, suppose a hospital has thousands of patient records. However, usually interest is cast in the characteristics and behaviours of certain groups, not those of a single person. Therefore, original data can be collapsed according to underlying scientific interest, which results in Symbolic data. A good description of Symbolic data is available in Billard and Diday (2003, 2006) and Diday and Noirhomme-Fraiture (2008). In this note, our purpose is to describe the methodology for obtaining descriptive statistics for Symbolic interval-valued data and also illustrate it on real data.

It may be noted that an interval cost of an item, say [20, 28] is different from an interval cost of [22, 26] even though both intervals have the same midpoint value of 24. A classical analysis using the same midpoint would lose the fact that these are two differently valued realizations with different internal variations. Suppose that Z_2 is some interval-valued random variable and that its realization for u^{th} observation is the interval $Z_2(u) = [a_u, b_u]$, where a_u and b_u are respectively the lower and upper limits of u^{th} observation. To construct a histogram, let $I=[\min_{u\in E} a_u, \max_{u\in E} b_u]$ be the interval that spans all observed values of random variable Z_2, where $E = \{1,2,\ldots,m\}$, m being the number of observations. Further, suppose that I is partitioned into r subintervals $I_g = [\xi_{g-1},\xi_g]$, g= 1,..., r-1, and $I_r = [\xi_{r-1},\xi_r]$. There is no hard and fast rule about the choice of appropriate value of r. It may be recalled that in the case of construction of histograms for point data, usually the number of classes is taken as neither too large nor too small, say between 5 and 10. For interval-data also, the same guidelines are followed. Then, histogram for Z_2 is a graphical representation of the frequency distribution $\{(I_g, f_g), g=1,..., r\}$, where f_g is the frequency that an arbitrary individual interval-valued observation Z_2 lies in the interval I_g and is given by

$$f_g = \sum_{u\in E} \left\| Z_2(u) \cap I_g \right\| \Big/ \left\| Z_2(u) \right\| , \; u=1, 2, \ldots, m. \quad \ldots (1)$$

Here $\left\| Z_2(u) \cap I_g \right\| = I_u(\xi)$ is the indicator function, which takes the values 1 or 0 according as ξ is or is not in

A part of the M Sc thesis of the first author submitted to the Indian Agricultural Research Institute, New Delhi, in 2012 (unpublished)

[1] Ph D Scholar (e mail: ranga533@gmail.com), [2]Principal Scientist and Head (e mail: prajnesh@iasri.res.in), [3]Senior Scientist (e mail: hghosh@gmail.com)

the interval I_g, $\|Z_2(u)\|$ is the length of interval I_g, and summation in (1) is only over those objects u for which $\xi \in I_g$. If some observation lies in the interval [a, b] and corresponding class-interval for construction of histogram is [r, s], then contribution of this observation towards f_g is computed as follows:

(i) If a < r and b ≥ s, then its contribution = (s-r)/(b-a)
(ii) If a < s and b > s, then its contribution=(s-a)/(b-a)
(iii) If a ≥ r and b ≤ s, then its contribution = 1
(iv) If b < s and b > r, then its contribution=(b-r)/(b-a)
(v) If r ≤ a and b < s, then its contribution = 1
(vi) Otherwise, its contribution = 0.

Thus, frequencies f_g are not necessarily integers. Further, symbolic sample mean and symbolic sample variance for an interval-valued variable $Z_2(u) = [a_u, b_u]$, are given by

$$\bar{Z}_2 = \sum_{u \in E} (b_u + a_u) / (2m), \qquad ...(2)$$

and

$$S^2 = \frac{1}{3m}\sum_{u \in E}(b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2}\left[\sum_{u \in E}(b_u + a_u)\right]^2 \qquad ...(3)$$

For calculation of bivariate histogram frequency of interval-valued random variables Z_1 and Z_2 having observations on the rectangle for each u∈E, the empirical joint probability density function for (Z_1, Z_2) is defined as

$$f(\xi_1,\xi_2) = \frac{1}{m}\sum_{u \in E} I_u(\xi_1,\xi_2) / \|Z'(u)\|, \quad u=1, 2, \ldots, m. \qquad ...(4)$$

Here $I_u(\xi_1,\xi_2)$ is the indicator function, which takes the values 1 or 0 according as (ξ_1,ξ_2) is or is not in the rectangle Z′(u) where ||Z′(u)|| denotes the area of this rectangle. Analogously with eq (4), joint histogram for Z_1 and Z_2 may be constructed by graphically plotting $\{R_{g_1g_2}, f_{g_1g_2}\}$ over the rectangle $R_{g_1g_2} = \{[\xi_{1,g_1-1}, \xi_{1g_1}) \times [\xi_{2,g_2-1}, \xi_{2g_2})\}$, where $f_{g_1g_2}$, not necessarily integers, are the frequencies of observations falling in the rectangle f_{g1g2} and are given by

$$f_{g_1g_2} = \sum_{u \in E} \left\| Z'(u) \cap R_{g_1g_2} \right\| \Big/ \left\| Z'(u) \right\|. \qquad ...(5)$$

The frequencies $f_{g_1g_2}$ are then computed along similar lines as for the univariate case. Thus, frequencies $f_{g_1g_2}$ are not necessarily integers.

Computer programs for computation of frequencies and construction of histograms for both univariate as well as bivariate situations were developed in SAS (2011) software package.

As an illustration, three symbolic random variables, viz. Pan evaporation, Temperature and Relative humidity are considered. Pan evaporation is an important climatic variable and is often used to estimate potential evaporation (Kirono *et al.* 2009) and reference evapotranspiration (Chen *et al.* 2005), as well as to forecast agricultural production (Wang *et al.* 2009). Changes in it will alter hydrological cycle in a region and so the changes of pan evaporation are of great significance for anticipating water balance and planning irrigation. Estimation of evaporation is important for design, planning and operation of water systems in agriculture. In India, water resources are scarce; estimation of this loss becomes more important in planning and management of irrigation practices. Weekly interval-valued data of Pan evaporation (Z_1) in mm, Temperature (Z_2) in °C and Relative humidity (Z_3) in percentage (%) were collected during year 2011 at the Meteorological Observatory, Division of Agricultural Physics, IARI, New Delhi.

As an example, the random variable pertaining to Temperature (Z_2) was considered. It was seen that, since minimum of lower limit in the data for Temperature was 10.8°C and maximum of upper limit was 44.5°C, interval span was I = [10.8, 44.5], m = 26, and E = {1,2,…,26}. Since the range, i.e. 44.5 - 10.8 is equal to 34.7, which is approximately 35, therefore I was partitioned into, say r =7 subintervals each of width 5, i e the class-intervals considered were: [10, 15), [15, 20), ..., [40, 45]. In order to construct the histogram, frequencies f_g given by eq (1) were required to be computed. The contribution to f_1 from the u^{th} observation was the fraction of observations in the interval [a,b], which overlapped the histogram class-interval I_1. For the interval I_1 = [10, 15], it was seen that there are only 3 observations, viz. u = 1,2,3 for which the lower limit of the intervals is less than the upper limit of histogram class-interval, i e 15. As an example, first observation (u=1), i e [10.8, 34.5], overlapped I_1 = [10, 15] by the fraction {(15-10.8) / (34.5-10.8)}, and so on. Therefore, from eq (1), frequency f_1 for interval-valued random variable Z_2 was obtained as follows:

$$f_1 = \left(\frac{15-10.8}{34.5-10.8}\right)_{u=1} + \left(\frac{15-14.1}{35.4-14.1}\right)_{u=2} + \left(\frac{15-11.7}{35-11.7}\right)_{u=3}$$
$$+ 0_{u=4} + 0_{u=5} + \ldots + 0_{u=26} = 0.361$$

Similarly, other frequencies could be computed. However, this was done easily in SAS (2011) software package by using computer program and the results are reported in the fourth column of Table 1. Further, I_g and f_g were computed for the remaining random variables of Relative humidity and Pan evaporation and the results are reported respectively

Table 1 Computation of frequencies

Pan evaporation (Z_1) (mm)		Temperature (Z_2) (in °C)		Relative humidity (Z_3) (in %)	
I_g	f_g	I_g	f_g	I_g	f_g
[1, 2)	0.494	[10, 15)	0.361	[10, 25)	1.120
[2, 3)	3.496	[15, 20)	1.294	[25, 40)	2.957
[3, 4)	4.960	[20, 25)	3.033	[40, 55)	3.864
[4, 5)	4.391	[25, 30)	8.526	[55, 70)	6.903
[5, 6)	4.166	[30, 35)	8.551	[70, 85)	7.373
[6, 7)	4.210	[35, 40)	3.312	[85, 100]	3.783
[7, 8)	2.497	[40, 45]	0.922		
[8, 9)	1.331				
[9, 10]	0.453				

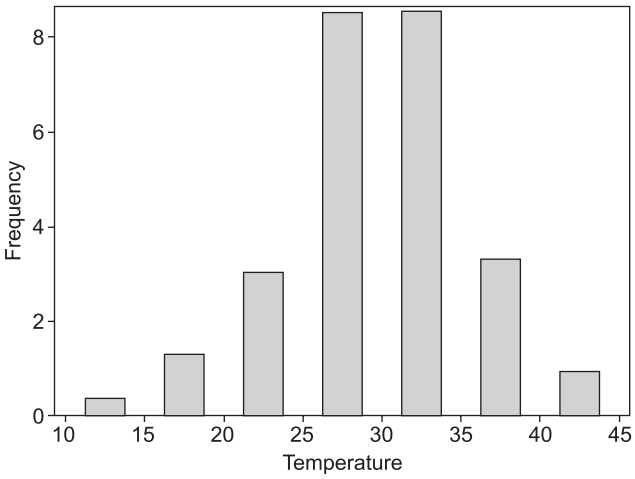

Fig 1 Histogram for temperature data

in second and sixth columns of Table 1. To save space, only the histogram for Temperature constructed, using computer program, is exhibited in Fig 1, which shows that its distribution is approximately normal.

For present data, symbolic sample mean and symbolic sample variance for interval-valued random variables were also computed using computer program and the same are reported in Table 2.

A joint histogram for say, (Z_1, Z_2) was subsequently constructed. As per Table 2, class-intervals for Z_1 and Z_2 were respectively taken as [1, 2), [2, 3), ..., [9, 10] and [10, 15), [15, 20) ,..., [40, 45]. Thus, the histogram rectangles were $R_{g_1g_2} = \{[\xi_{1,g_1-1}, \xi_{Ig_1}) \times [\xi_{2,g_2-1}, \xi_{2g_2})\}$, $g_1 = 1, 2, ..., 9$, $g_2 = 1, 2, ..., 7$, where the sides of these rectangles were the intervals [1, 2) × [10, 15), [1, 2) × [15, 20), ..., [1, 2) × [40, 45], ..., [8, 9] × [40, 45]. Then, observed frequencies $f_{g_1g_2}$ were calculated using eq (5). For example, using Table 3, when g_1 =3 and g_2 = 4, the rectangle $R_{3,4}$ = [3, 4) × [25, 30)

Table 2 Descriptive statistics

Statistic	Pan evaporation (Z_1) (mm)	Temperature (Z_2) (°C)	Relative humidity (Z_3) (%)
Sample Mean	5.06	29.66	63.60
Sample variance	33.64	396.09	3.49

Table 3 Computation of bivariate frequencies for Pan evaporation and Temperature data

g_1\	I_1\	Joint histogram (Z_1, Z_2): $f_{g_1g_2}$						
g_2	I_2	1	2	3	4	5	6	7
		[10,15)	[15,20)	[20,25)	[25,30)	[30,35)	[35,40)	[40,45]
1	[1, 2)	0.00	0.00	0.04	0.22	0.21	0.02	0.00
2	[2, 3)	0.00	0.00	0.21	1.53	1.51	0.25	0.00
3	[3, 4)	0.00	0.00	0.24	2.13	2.13	0.44	0.02
4	[4, 5)	0.09	0.19	0.39	1.55	1.59	0.55	0.04
5	[5, 6)	0.14	0.46	0.71	1.10	1.13	0.50	0.12
6	[6, 7)	0.13	0.55	0.84	0.98	0.98	0.55	0.20
7	[7, 8)	0.00	0.09	0.39	0.59	0.59	0.57	0.26
8	[8, 9)	0.00	0.00	0.17	0.32	0.32	0.32	0.20
9	[9, 10]	0.00	0.00	0.06	0.10	0.10	0.10	0.08

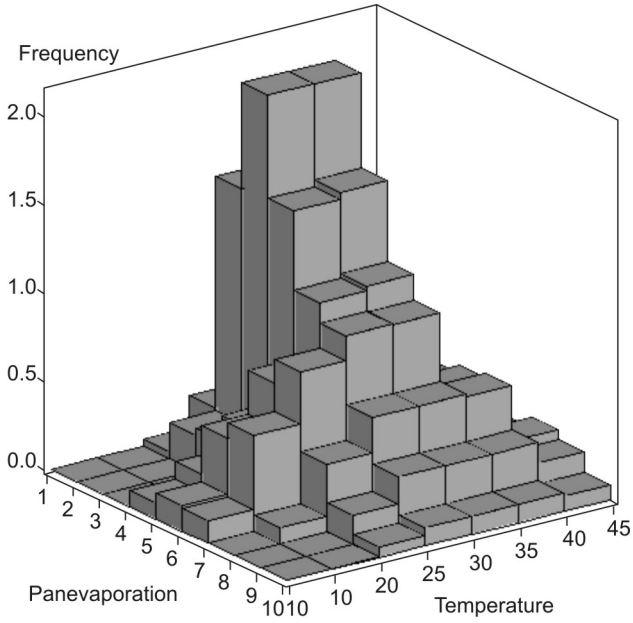

Fig 2 Bivariate histogram for Pan evaporation and Temperature data

Table 4 Pan evaporation, temperature and relative humidity data

Observation number (u)	Pan evaporation (Z_1) (in mm)	Temperature (Z_2) (in 0C)	Relative humidity (Z_3) (in %)
1	[4.0, 6.2]	[10.8, 34.5]	[24, 90]
2	[4.5, 7.2]	[14.1, 35.4]	[21, 90]
3	[5.5, 7.0]	[11.7, 35.0]	[10, 78]
4	[4.5, 7.0]	[15.6, 35.4]	[20, 83]
5	[5.2, 7.0]	[15.0, 36.2]	[17, 83]
6	[5.9, 8.0]	[16.5, 40.5]	[14, 76]
7	[5.5, 8.2]	[22.4, 41.5]	[18, 66]
8	[5.3, 9.4]	[21.7, 44.0]	[18, 77]
9	[6.8, 9.5]	[24.0, 44.5]	[19, 68]
10	[4.0, 8.1]	[19.7, 42.0]	[21, 84]
11	[5.7, 8.7]	[21.9, 41.3]	[33, 78]
12	[5.1, 9.8]	[19.6, 43.6]	[26, 84]
13	[3.2, 9.0]	[23.4, 42.2]	[36, 86]
14	[4.0, 6.0]	[23.2, 38.0]	[48, 95]
15	[1.7, 5.2]	[24.5, 38.2]	[61, 98]
16	[2.5, 5.1]	[25.7, 38.1]	[44, 92]
17	[2.0, 3.9]	[24.6, 35.8]	[61, 97]
18	[2.9, 4.8]	[24.4, 37.5]	[50, 94]
19	[2.6, 5.1]	[25.0, 34.5]	[61, 95]
20	[2.0, 6.0]	[26.3, 38.2]	[56, 92]
21	[2.2, 4.2]	[24.4, 34.4]	[62, 98]
22	[1.6, 3.9]	[24.0, 34.2]	[59, 98]
23	[1.9, 4.0]	[24.0, 35.2]	[57, 94]
24	[3.0, 5.4]	[25.7, 35.5]	[60, 91]
25	[1.6, 4.2]	[23.8, 35.0]	[65, 98]
26	[1.9, 4.9]	[24.2, 35.4]	[63, 97]

was obtained. So

$$f_{34} = 0_{u=1} + 0_{u=2} + \ldots + \left(\frac{4-3.2}{9-3.2}\right)\left(\frac{30-25}{42.20-23.40}\right)_{u=13} + 0_{u=14} + \left(\frac{4-3}{5.2-1.7}\right)\left(\frac{30-25}{38.2-24.5}\right)_{u=15} + \ldots + \left(\frac{4-3}{4.9-1.9}\right)\left(\frac{30-25}{35.4-24.2}\right)_{u=26} = 2.13$$

Similarly, all other frequencies f_{g1g2} could be computed. However, this was done easily in SAS (2011) software package by using computer program and the results are reported in the fourth column of Table 2. The bivariate histogram for Pan evaporation (Z_1) and Temperature (Z_2) was then constructed, using computer program and the same is exhibited in Fig 2, which shows that the joint distribution is approximately bivariate normal.

SUMMARY

It is, by now, well recognized that real data are in intervals and not in points. Unfortunately, classical statistical theory is not capable of handling data in intervals. Here, methodology for drawing univariate and bivariate histograms and computation of descriptive statistics, like sample mean and sample variance for Symbolic interval-valued data is discussed. It is hoped that researchers would start employing this type of 'Symbolic data analysis' to their datasets.

REFERENCES

Billard L and Diday E. 2003. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association* **98**: 470–87.

Billard L and Diday E. 2006. Descriptive statistics for interval-valued observations in the presence of rules. *Computational Statistics and Data Analysis* **21**: 187–210.

Chen D, Gao G, Xu C Y, Guo J and Ren G. 2005. Comparison of the Thornthwaite method and pan data with the standard Penman-Monteith estimates of reference evapotranspiration in China. *Climate Research* **28**: 123–32.

Diday E and Noirhomme-Fraiture M. 2008. *Symbolic Data Analysis and the SODAS Software*. John Wiley, England.

Kirono D G C, Jones R N and Cleugh H A. 2009. Pan-evaporation measurements and Morton-point potential evaporation estimates in Australia: Are their trends the same? *International Journal of Climatology* **29**: 711–8.

SAS. 2011. SAS/IML User's Guide 9.2. SAS Institute, North Carolina.

Wang Z Y, Liu Z X, Zhang Z X and Liu X B. 2009. Subsurface drip irrigation scheduling for cucumber (*Cucumis sativus* L.) grown in solar greenhouse based on 20 cm standard pan evaporation in Northeast China. *Scientia Horticulturae* **123**: 51–7.