Computational Statistics (2006) 21:187-210
DOI 10.1007/s00180-006-0259-6

Descriptive Statistics for Interval-valued Observations in the Presence of Rules

L. Billard[1] and E. Diday[2]

[1]Department of Statistics, University of Georgia, Athens, GA 30602-1952 USA
[2] CEREMADE, Universite de Paris 9 Dauphine, 75775 Paris Cedex 16 France

Summary

While symbolic data exist in their own right, contemporary datasets can be too large to analyse using traditional statistical methodologies. Aggregation of these large datasets into sets of more managable size perforce produce datasets whose entries are symbolic data. This paper studies the derivation of basic description statistics, in particular, histograms and mean and variances plus joint histograms for interval-valued datasets when logical dependency rules are present. Algorithms for calculating these histograms are also provided.

Keywords: Interval-valued data, logical dependency rules, univariate histograms, sample means, sample variances, joint histogram

1 Introduction

While symbolic data exist in their own right as small or large datasets, the advent of the modern computer has brought with it classical (and/or symbolic) datasets that are too large in size to be analysed using traditional statistical methodologies even with the computational assistance of those same computers that generated such data. Therefore, in order to elicit reasonable and appropriate analyses and conclusions from the data, it becomes necessary to aggregate the data in some meaningful manner first before analyses can proceed. How this aggregation occurs will depend on some of the underlying questions and/or answers being asked and/or sought. For example, suppose a dataset consists of the medical records for a country (say), and suppose that apart from the more-direct medically related variables, there are also demographic variables such as the individual's age, gender, town or residence, and so on. One basic question may relate to what happens across towns or residence sites, while another may be concerned with age $\times$ gender differences. Thus, these questions led to aggregations by towns, or by age $\times$ gender, respectively. The number of possible aggregations is limited only by the number of such basic questions. Whether the original data were classical or symbolic data, the aggregated values will now be as lists, and/or intervals, and/or modal values, regardless of the nature of the aggregation method adopted. For example, a list could be the types of cancer observed, $Y = \{\text{lung, colon, ...}\}$; an interval value could be the pulse rate, $Y = 64 \pm 1 = (63, 65)$; a modal value could be a histogram, $Y = \{(\text{red}, p_1), (\text{green}, p_2), \ldots\}$ with $\Sigma p_i = 1$. For a review of symbolic data, see Billard and Diday (2003) and for a more detailed description, see Bock and Diday (2000).

In this paper our focus will be on interval-valued data in the presence of rules, and in particular on obtaining basic descriptive statistics such as frequency histograms, joint frequency histograms and sample means and variances. Rules, so-called, can arise in two (or three) broadly defined ways. The first relates to underlying conditions that exist, be the data classical- or symbolic-valued. For example, interest may center on children, in which case any analysis conditions the data to contain only children; or, the variables Y_1 and Y_2 may be required to satisfy a condition that $Y_1 + Y_2 = \beta$ (say), or so on. In contrast, when aggregating data into symbolic-valued variables, the very action of aggregation may produce data that perforce engage the adoption of rule(s) to maintain data integrety. For example, suppose we have values for $Y_1 =$ age and $Y_2 =$ number of children, and suppose we have particular classical values $Y_a = (21, 2)$, $Y_b = (10, 0)$, $Y_c = (16, 1), \ldots$, where $Y = (Y_1, Y_2)$; and suppose further that the concept of interest, after appropriate aggregation, put these three individuals into the same category and produced the symbolic interval-valued observation $\xi = \{(10, 21), (0, 1, 2)\}$. As it stands, the value ξ implies that persons in the age interval (10, 21) years had (0, 1,

2) children, including the possibility that the 10-year-old had 1 (or 2) children. To maintain data integrity here, it is necessary to include a rule such as $\nu = \{\text{If } Y_1 < 14 \text{ (say), then } Y_2 = 0\}$. The need for this type of rule is unique to symbolic data. The precise nature of such rules could perforce vary with the description of the symbolic data value. A possible third type of rule is what would amount to a form of data cleaning; e.g., a "rule" such as age $= Y_1 > 0$, could be used to catch observed (classical, or symbolic) values of $Y_1 = -15$ (say, an obvious miskeying situation). In some circumstances, data cleaning rules are absorbed into either of the first two categories defined above. Data cleaning rules however do need to be present for datasets too large to be "eye-balled" for correctness.

Since classical data are but single points in p-dimensional space (where p is the number of variables), rules are relatively easy to manage. However, since symbolic values are p-dimensional hypercubes and/or Cartesian products of distributions in p-dimensional space, rules can and do create difficulties. We focus on rules for interval-valued data; the methodology can be extended to histogram-valued data reasonably easy conceptually (less easy computationally!)

Bertrand and Goupil (2000) derived formula for finding the univariate histogram and sample mean and variance for a single interval-valued variable Y without rules. They also developed the corresponding results for multi-valued (list) data with and without rules. To accommodate rules, their basic approach was to convert each actual possible symbolic data value into a so-called virtual data value where the virtual values were those that satisfied the given rule(s). Billard and Diday (2003), alluded to extending Bertrand and Goupil's virtual data idea to interval data with rules, but gave no details. Our aim here is to develop this concept further and also to extend it to finding joint histograms for (Y_1, Y_2) where Y_1 and Y_2 are each interval-valued variables and where rules exist. We develop our basic approach through rules applied to the interval-valued data of Table 1.

Therefore, in Section 2, we consider the nature of the virtual observation space in the presence of rules and show how the virtual observation values can be determined. Then, in Section 3, we use these virtual observations to obtain univariate histograms under a variety of specific rules. Calculating the sample mean and variance in the presence of rules is studied in Section 4. Derivation of a joint histogram for the bivariate $Y = (Y_1, Y_2)$ is considered in Section 5. The basic principles involved are discussed and summarized in Section 6. These form the nucleus of the methodology required to obtain basic statistics for interval-valued data in the presence of rules. In the course of these derivations, the need arises for calculating a histogram of histogram-valued data and an algorithm for calculating a joint histogram for interval-valued data; these algorithms are outlined in Section 7.

2 Observed and Virtual Symbolic Intervals

The data of Table 1 represent two random variables, viz., Y_1 = Number of At-Bats; and Y_2 = Number of Hits, for baseball players over a season. Players are aggregated by teams, so that the resulting team statistics are now intervals. The results shown in Table 1 are based on actual (Y_1, Y_2) statistics for a sample of players from a variety of baseball teams obtained from Vanessa and Vanessa (2004). Some additional results have been inserted for illustrative purposes.

We denote a particular realization of $Y = (Y_1, Y_2)$ by $\xi = (\xi_1, \xi_2)$ with $\xi_i = (a_i, b_i)$, $i = 1, 2$. Following Bertrand and Goupil (2000), we make the assumption that specific (point) values of Y_i are uniformly distributed across the interval (a_i, b_i). Further, ξ takes values in the p = 2-dimensional hypercube (i.e., rectangle) bounded by $(a_1, b_1) \times (a_2, b_2)$. We denote a specific observation by $\xi(u)$, which is bounded by the rectangle $R(u) = (a_{1u}, b_{1u}) \times (a_{2u}, b_{2u})$ for $u = 1, \ldots, n$, where n is the number of observations.

Table 1 - At-Bats and Hits by Team

u Team	Y_1 #At-Bats	Y_2 #Hits	Pattern	u Team	Y_1 # At-Bats	Y_2 #Hits	Pattern
1	(289, 538)	(75, 162)	B	11	(212, 492)	(57, 151)	B
2	(88, 422)	(49, 149)	I	12	(177, 245)	(189, 238)	G
3	(189, 223)	(201, 254)	F	13	(342, 614)	(121, 206)	B
4	(184, 476)	(46, 148)	B	14	(120, 439)	(35, 102)	B
5	(283, 447)	(86, 115)	B	15	(80, 468)	(55, 115)	I
6	(24, 26)	(133, 141)	A	16	(75, 110)	(75, 110)	C
7	(168, 445)	(37, 135)	B	17	(116, 557)	(95, 163)	I
8	(123, 148)	(137, 148)	E	18	(197, 507)	(52, 53)	B
9	(256, 510)	(78, 124)	B	19	(167, 203)	(48, 232)	H
10	(101, 126)	(101, 132)	D				

To examine these data more closely, we first make the logical deduction that the Number of At-Bats cannot be less than the Number of Hits, i.e., $Y_1 \geq Y_2$. Consider the second observation $\xi(2)$. Each of the $\xi_1(2)$ and $\xi_2(2)$ values is possible. The resulting rectangle $R(2)$ has vertices at (x_1, x_2) = (88, 49), (88, 149), (422, 49) and (422, 149). All (x_1, x_2) values contained in this rectangle appear as possible values. This includes the vertex value $(x_1, x_2) = (88, 149)$, i.e., the number of hits is 149 from 88 at-bats - clearly not a logical possibility. However, another player can have 149 hits from 422 at-bats for example, and so on. Here, the logical rule $\nu : Y_1 \geq Y_2$ implies that the actual apparent hypercube $R(u)$ has to be transformed to a virtual hypercube $V(u)$ containing only those values of $R(u)$ that satisfy the rule ν. In contrast, the observation $u = 6$, with $\xi(6) = \{(24, 26), (133, 141)\}$ would suggest that the ξ_1 and ξ_2 have been transposed. The logical rule here catches this, as part of a data cleaning process for example.

Formally, we adapt the definition of virtual data, from Bertrand and Goupil (2000), as follows.

Definition: The *virtual observation* space $V \equiv V(u)$ of an actual observation space $R \equiv R(u)$ consists of all possible values x in R which satisfy all the rules $\nu = (\nu_1, \nu_2 \ldots)$ operating on R. That is, for the observation u,

$$V(u) = \{x \in R(u), \quad \nu_i(x) = 1, \quad \text{for all } \nu_i \text{ in } \nu\} \tag{1}$$

where $\nu_i(x) = 1$ if the rule is true for the vector-value x and is 0 if the rule is not true for x. Let us denote the virtual observation by $\xi' = (\xi'_1, \ldots, \xi'_p)$ with $\xi'_i = (a'_i, b'_i)$, $i = 1, \ldots, p$.

To illustrate this further, suppose that for the Table 1 data, there is a logical rule

$$\nu : Y_2 \leq \alpha Y_1. \tag{2}$$

Setting $\alpha = 1.0$ allows for the removal of x values that are not logically possible; while setting $\alpha = 0.400$, say, is acknowledging that batting averages $(= Y_2/Y_1)$ above 0.400 are unlikely and therefore in this present sense also not logically possible. The impact of this rule ν on the observed rectangle R will produce a virtual hypercube V which has one of the eight patterns, denoted by $A, B, \ldots, I$, displayed in Figure 1. Those values which fall above the line $Y_2 = \alpha Y_1$ are not logically possible values and so are excluded from R to produce the virtual value V. The shaded regions correspond to the virtual values V. The conditions that apply that give $\nu(x) = 1$ for the respective patterns are given in Table 2.

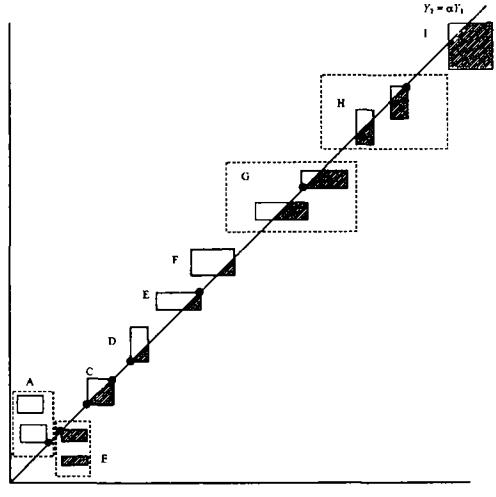
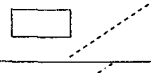
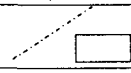
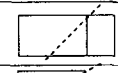
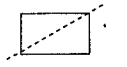
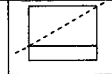
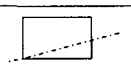
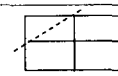
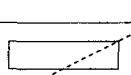


Figure 1 - Patterns for Virtual V - shaded regions

The pattern A corresponds to those observations $\xi(u)$ for which $V(u)$ is empty; e.g., $\xi(6)$ in Table 1. In this case, the underlying condition from equation (1) that generates $\nu(x) = 0$ is $\{\alpha b_1 < a_2\}$. The pattern B represents those observations that are unaffected by the rule ν, i.e., $V(u) = R(u)$. The condition that gives $\nu(x) = 1$ in equation (1) for these patterns translate, in terms of the (a_i, b_i) values, $i = 1, 2$, to $\{\alpha a_1 \geq b_2\}$.

The four patterns C, D, E, F are similar in that the virtual observation hypercube is a triangle, though they differ as to whether or not particular triangle vertices do or do not fall on the line $Y_2 = \alpha Y_1$. Therefore, the virtual ξ values differ accordingly. Thus, for pattern D, the virtual value for the original observation ξ is $\xi_1' = \xi_1 = (a_1, b_1)$ and $\xi_2' = (a_2, \alpha b_1)$. Notice that the virtual observation for Y_1 (alone) is unaffected by ν. In contrast, for pattern E, Y_2 is unaffected, $\xi_2' = \xi_2$, but the Y_1 values are affected giving the virtual value as $\xi_1' = (a_2/\alpha, b_1)$. For pattern F, both Y_1 and Y_2 values are affected by ν; whereas in pattern C, neither are. Table 3 displays these virtual values ξ_i', $i = 1, 2$.

Table 4.9 - Virtual Patterns and Conditions

Pattern	Pattern V	Conditions	Pattern	(*) Pattern V	Conditions
A		$\alpha b_1 \leq a_2$	F		$\begin{cases} \alpha a_1 < a_2 < \alpha b_1 \\ \alpha b_1 < b_2 \end{cases}$
B		$\alpha a_1 \leq b_2$	G		$\begin{cases} \alpha a_1 \leq a_2 \\ \alpha b_1 > b_2 \end{cases}$
C		$\begin{cases} \alpha a_1 = a_2 \\ \alpha b_1 = b_2 \end{cases}$	H		$\alpha a_1 > a_2$ $\alpha b_1 \leq b_2$
D		$\begin{cases} \alpha a_1 = a_2 \\ \alpha b_1 < b_2 \end{cases}$	I		$\begin{cases} \alpha a_1 > a_2 \\ \alpha b_1 > b_2 > \alpha a_1 \end{cases}$
E		$\begin{cases} \alpha a_1 < a_2 \\ \alpha b_1 = b_2 \end{cases}$			

(*) The dotted line represents the line $Y_2 = \alpha Y_1$

Also shown, in Table 3, are the apparent virtual values for the bivariate pair (ξ_1', ξ_2') for the C, D, E and F patterns. When calculating the histogram for Y_1 (or Y_2) alone, these virtual ξ_i' values are used in the usual manner using Bertrand and Goupil (2000) methodology. However, when calculating the joint histogram for (Y_1, Y_2), routine application of the methodology (see Billard and Diday, 2003) would in this case produce answers as though the hypercube (ξ_1', ξ_2') were the rectangle $(a_1', b_1') \times (b_2', b_2')$ with area $(b_1' - a_1')(a_2' - a_2')$, instead of the triangle whose vertices are $\{(a_1', b_1'), (b_1', a_2'), (a_2', b_2')\}$, and with area $|V| = (b_1' - a_1')(b_2' - a_2')/2$ where $|A|$ is the area of the region A. Clearly, this feature has to be accommodated, and is addressed further in Section 5. The corresponding areas $|V|$ for each pattern C, D, E, F are also displayed in Table 3.

Table 3 - Virtual Observation Space Values - by Pattern

Pattern	Virtual ξ_1'	Virtual ξ_2'	Apparent Virtual ξ'
A	ϕ	ϕ	ϕ
B	(a_1, b_1)	(a_2, b_2)	$\{(a_1, b_1),\ (a_2, b_2)\}$
C	(a_1, b_1)	(a_2, b_2)	$\{(a_1, b_1),\ (a_2, b_2)\}$
D	(a_1, b_1)	$(a_2, \alpha b_1)$	$\{(a_1, b_1),\ (a_2, \alpha b_1)\}$
E	$(a_2/\alpha, b_1)$	(a_2, b_2)	$\{(a_2/\alpha, b_1),\ (a_2, b_2)\}$
F	$(a_2/\alpha, b_1)$	$(a_2, \alpha b_1)$	$\{(a_2/\alpha, b_1),\ (a_2, \alpha b_1)\}$
G	$\{(a_2/\alpha, b_2/\alpha)p_1,$ $(b_2/\alpha, b_1)p_2\}$	(a_2, b_2)	$\{[(a_2/\alpha, b_2/\alpha), (a_2, b_2)]p_1,$ $[(b_2/\alpha, b_1), (a_2, b_2)]p_2\}$
	$p_i = \lvert R_i\rvert/\lvert V\rvert, \lvert V\rvert = \lvert R_1\rvert + \lvert R_2\rvert, \lvert R_1\rvert = (b_2 - a_2)^2/(2\alpha), \lvert R_2\rvert = (\alpha b_1 - b_2)(b_2 - a_2)/\alpha$		
H	(a_1, b_1)	$\{(a_2, \alpha a_1)p_1,$ $(\alpha a_1, \alpha b_1)p_2\}$	$\{[(a_1, b_1), (a_2, \alpha a_1)]p_1,$ $[(a_1, b_1), (\alpha a_1, \alpha b_1)]p_2\}$
	$p_1 = \lvert R_1\rvert/\lvert V\rvert, \lvert V\rvert = \lvert R_1\rvert + \lvert R_2\rvert, \lvert R_1\rvert = (b_1 - a_1)(\alpha a_1 - a_2), \lvert R_2\rvert = \alpha(b_1 - a_1)^2/2$		
I	$\{(a_1, b_2/\alpha)p_1,$ $(b_2/\alpha, b_1)p_2\}$	$\{(a_2, \alpha a_1)p_1^*,$ $(\alpha a_1, b_2)p_2^*\}$	$\{[(a_1, b_2/\alpha), (\alpha a_1, b_2)]p_1^{**},$ $[(a_1, b_2/\alpha), (a_2, \alpha a_1)]p_2^{**},$ $[(b_2/\alpha, b_1), (a_2, \alpha a_1)]p_3^{**},$ $[(b_2/\alpha, b_1), (\alpha a_1, b_2)]p_4^{**}\}$
	$p_1 = (\lvert R_1\rvert + \lvert R_2\rvert)/\lvert V\rvert$ $p_2 = (\lvert R_2\rvert + \lvert R_4\rvert)/\lvert V\rvert$	$p_1^* = (\lvert R_2\rvert + \lvert R_3\rvert)/\lvert V\rvert$ $p_2^* = (\lvert R_1\rvert + \lvert R_4\rvert)/\lvert V\rvert$	$p_i^{**} = \lvert R_i\rvert/\lvert V\rvert i = 1, \ldots, 4$
	$\lvert R_1\rvert = (b_2 - \alpha a_1)^2/(2\alpha), \lvert R_2\rvert = (b_2 - \alpha a_1)(\alpha a_1 - a_2)/\alpha, \lvert R_3\rvert = (\alpha b_1 - b_2)(\alpha a_1 - a_2)/\alpha,$ $\lvert R_4\rvert = (\alpha b_1 - b_2)(b_2 - \alpha a_1)/\alpha, \lvert V\rvert = \lvert R_1\rvert + \lvert R_2\rvert + \lvert R_3\rvert + \lvert R_4\rvert$		

The two patterns G and H have the common feature that their 4-sided (non-rectangular) hypercube can be viewed as the union of a triangle and a rectangle. For the pattern G, the virtual description for Y_1 (alone) is now a histogram-valued variable (and not the interval-valued observation of the original data); while for the pattern H, it is the variable Y_2 (considered alone) which has a histogram-valued virtual description. Thus, we can show that in pattern G, the virtual observation becomes

$$\xi_1' = \{(a_2/\alpha, b_2/\alpha)p_1, \quad (b_2/\alpha, b_1)p_2\} \tag{3}$$

where the relative frequencies p_i, $i = 1, 2$, are given by

$$p_i = |R_i|/|V| \tag{4}$$

with

$$|R_1| = (b_2 - a_2)^2/(2\alpha), \quad |R_2| = (\alpha b_1 - b_2)(b_2 - a_2)/\alpha \tag{5}$$

and

$$|V| = |R_1| + |R_2|; \tag{6}$$

and where the virtual description of Y_2 (alone) is unaffected, with $\xi_2' = \xi_2 = (a_2, b_2)$. These are displayed in Table 3 for both patterns G and H. Then, by using the methodology developed in Billard and Diday (2003) for obtaining a histogram of histograms, the respective (univariate) histograms can be obtained. Also, shown in Table 3 is the apparent virtual description of the bivariate pair (Y_1, Y_2). These too are now histogram-valued, rather than interval-valued, observations. However, again as cautioned above for the patterns C, D, E, F, care is required for the "triangle" pieces, viz., $R_1 \equiv [(a_2/\alpha, b_2/\alpha), (a_2, b_2)]$ in pattern G, and $R_2 \equiv [(a_1, b_1), (\alpha a_1, \alpha b_1)]$ in pattern H.

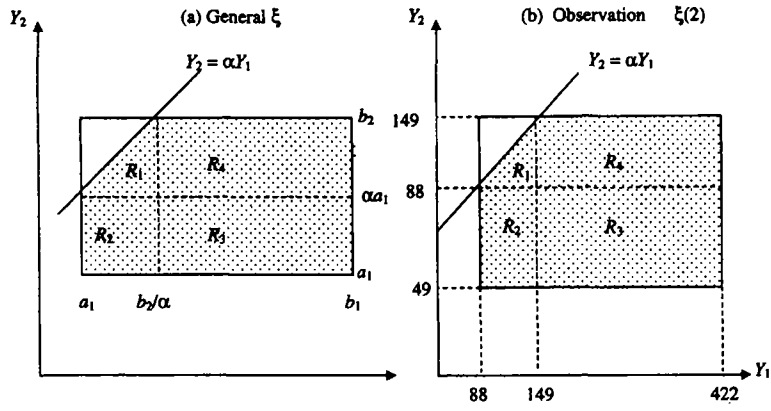


Figure 2 – Pattern I Detail: (a) General, (b) Observation $\xi(2)$

Finally, we consider the pattern I, reproduced in Figure 2a. In these cases, the virtual observation space V is a 5-sided hypercube which can be partitioned into the triangle R_1, and three different rectangles R_2, R_3, R_4 with respective vertices as indicated in Figure 2a. For data that follow this pattern, the virtual values of both the Y_1 and Y_2 variables (each considered alone) differ from the actual observed values; and in each case the virtual values become histogram-valued instead of the original integral-valued. It follows that for Y_1 (alone) the virtual observation is

$$\xi_1' = \{(a_1, b_2/\alpha)p_1, \quad (b_2\alpha, b_1)p_2\} \tag{7}$$

where

$$p_1 = (|R_1| + |R_2|)/|V|, \quad p_2 = (|R_3| + |R_4|)/|V| \tag{8}$$

with

$$|R_1| = (b_2 - a_1)^2/(2\alpha), \quad |R_2| = (b_2 - \alpha a_1)(\alpha a_1 - a_2)/\alpha,$$
$$|R_3| = (\alpha b_1 - b_2)(\alpha a_1 - a_2)/\alpha, \quad |R_4| = (\alpha b_1 - b_2)(b_2 - \alpha a_1)/\alpha \tag{9}$$

and

$$|V| = |R_1| + \cdots + |R_4|. \tag{10}$$

The virtual observation for Y_2 (alone) is

$$\xi_2' = \{(a_2, \alpha a_1)p_1^*, \quad (\alpha a_1, b_2)p_2^*\} \tag{11}$$

where

$$p_1^* = (|R_2| + |R_3|)/|V|, \quad p_2^* = (|R_1| + |R_4|)/|V|. \tag{12}$$

These values are summarized in Table 3. The table also shows the corresponding apparent virtual observation for (Y_1, Y_2) taken together as a bivariate pair. Here, we can show that the virtual value ξ' of ξ is

$$\begin{aligned}\xi' \;=\; & \{[(a_1, b_2/\alpha), (\alpha a_1, b_2)]p_1^{**}, [(a_1, b_2/\alpha), (a_2, \alpha a_1)]p_2^{**}, \\ & [(b_2/\alpha, b_1), (a_2, \alpha a_1)]p_3^{**}, [(b_2/\alpha, b_1), (\alpha a_1, b_2)]p_4^{**}\} \end{aligned} \quad (13)$$

where

$$p_i^{**} = |R_i|/|V|, \quad i = 1, \ldots, 4, \quad (14)$$

with $|R_i|$ and $|V|$ as given in equations (9) and (10). Again, the "triangle" piece ($R_1 \equiv [(a_1, b_2/\alpha), (\alpha a_1, b_2)]$) requires special care when calculating a joint histogram function.

3 Construction of Histograms

When, after application of the rule $\nu = (\nu_1, \nu_2, \ldots)$, the resulting virtual dataset consists entirely of interval-valued data, the histogram of the virtual dataset can be constructed by using the Bertrand and Goupil methodology which is available computationally in the SODAS software (and can be found on the web at www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm).

For comparative purposes, we first give the histogram for the baseball dataset of Table 1 when there are no rules. Suppose we build the histogram for Y_1 = Number of At-Bats on the $r_1 = 7$ intervals $[0, 100), \ldots, [600, 650]$; and suppose the histogram for Y_2 = Number of Hits is constructed on the $r_2 = 9$ intervals $[0, 50), [50, 75), \ldots, [200, 225), [225, 275]$. The resulting histograms are given in column (a) of Table 4 for Y_1 and Table 5 for Y_2, respectively.

Table 4 - Histogram for Y_1 = # At-Bats

g_1	Histogram Interval, I_{g_1}	(a) ν_0: No Rules	(b) ν_1 : $Y_1 \geq 120$	(c) ν_2 : $Y_2 \leq Y_1$	(d) ν_3 : $Y_2 \leq 0.35Y_1$	(e) ν_4 : (ν_1, ν_2)
1	$[0, 100)$	1.802	–	0.778	0.000	–
2	$[100, 200)$	5.043	4.687	4.490	1.491	4.163
3	$[200, 300)$	4.180	4.244	4.701	2.203	4.743
4	$[300, 400)$	4.099	4.162	4.132	3.919	4.174
5	$[400, 500)$	3.114	3.143	3.131	3.601	3.151
6	$[500, 600)$	0.711	0.713	0.716	1.287	0.717
7	$[600, 650]$	0.051	0.051	0.051	0.499	0.051
	n	19	17	18	13	17
	$\bar{X}$	268.079	295.500	283.921	370.504	297.263
	S	136.101	119.449	125.531	118.946	118.133

Suppose now interest is restricted to those situations with 120 or more at-

bats. This translates to the rule

$$\nu_1 : \{Y_1 \geq 120\}. \tag{15}$$

Under this rule, observation $\xi(6)$ and $\xi(16)$ are deleted entirely. Observations $u = 2, 10, 15$, and 17, are truncated; so that the virtual observation for $u = 2$, becomes $\xi'(2) = \{(120, 422), (49, 149)\}$; likewise, $\xi'(10)$, $\xi'(15)$, and $\xi'(17)$ can be found. After application of the rule ν_1, all virtual observations are integral-valued. Then, by building the relevant histograms on the same histogram intervals used in column (a), we obtain the frequencies of column (b) in Table 4 for Y_1 and Table 5 for Y_2, respectively. Comparing the two histograms of columns (a) and (b), we see the impact of this rule. For the variable Y_1, since this rule directly truncates Y_1 values, the two histogram intervals I_{g_1}, $g_1 = 1, 2$, are clearly affected. However, so are other histogram intervals affected (in contrast to the corresponding comparison for classical data when these latter intervals are not affected). Take, e.g., the histogram

Table 5 - Histogram for Y_2 = # Hits

g_2	Histogram Interval, I_{g_2}	(a) ν_0: No Rules	(b) ν_1 : $Y_2 \geq 120$	(c) ν_2 : $Y_2 \leq Y_1$	(d) ν_3 : $Y_2 \leq 0.35Y_1$	(e) ν_4 : (ν_1, ν_2)
1	[0, 50)	0.417	0.417	0.421	0.632	0.421
2	[50, 75)	2.784	2.784	2.854	3.342	2.834
3	[75, 100)	3.978	3.264	4.026	3.242	3.316
4	[100, 125)	4.233	3.947	4.458	2.907	4.149
5	[125, 150)	4.144	3.144	2.990	1.721	3.030
6	[150, 175)	0.770	0.770	0.771	0.668	0.771
7	[175, 200)	0.654	0.654	0.569	0.348	0.569
8	[200, 250)	1.169	1.169	1.631	0.125	1.631
9	[225, 275]	0.851	0.851	0.279	0.015	0.279
	n	19	17	18	13	17
	$\bar{X}$	119.684	120.265	116.455	97.795	117.871
	S	48.678	50.799	47.207	37.287	48.112

interval $I_{g_1} = I_3 = [200, 300)$ all of whose internal values are valid under ν_1. Take also, e.g., the contribution of the $u = 2$ observation to this I_3 interval. Then, the virtual data value $\xi_1'(2)$ contributes a portion equal to $(300 - 200)/(422 - 120) = 100/302$ to the frequency of I_3, while the original data $\xi_1(2)$ contributes the amount $(300 - 200)/(422 - 88) = 100/334$ ($\neq 100/302$) to the frequency in I_3. A comparison of columns (a) and (b) in Table 5 for the histogram for the Y_2 variable also reveals differences. This occurs even though the rule ν_1 does not involve Y_2 directly, and even though for every observation u the virtual $\xi_2'(u) = \xi_2(u)$. The impact of ν_1 on the histogram for Y_2 is a reflection of the $u = 6$ and $u = 16$ observations being deleted.

Suppose now we apply the rule of equation (2) with $\alpha = 1.0$, viz.,

$$\nu_2 : Y_1 \geq Y_2, \tag{16}$$

i.e., the number of hits cannot exceed the number of at-bats. Table 6, column (a) identifies the pattern of the virtual observation in the presence of this rule. Columns (b) and (c) give the virtual observation value for Y_1 and Y_2, respectively, for each case by utilizing Table 3. For example, clearly when $u = 1$, pattern B pertains. Hence, $\xi_1' = \xi_1$, $\xi_2' = \xi_2$; also, $\xi' = (\xi_1', \xi_2') = \xi$. The $u = 2$ observation under ν_2 reduces to a virtual observation space with pattern I (see Figure 2b). It is really verified that the areas $|R_i|$, $i = 1, \ldots, 4$, and $|V|$ are

$$|R_1| = 1860.5, \quad |R_2| = 2379, \quad |R_3| = 10647, \quad |R_4| = 16653, \quad |V| = 31539.5.$$

Table 6(i) - Virtual ξ_1' and ξ_2' under rule $\nu_2 : Y_1 \geq Y_2$

Team u	(a)	(b) $\xi_1' : Y_1 =$ # Hits	(c) $\xi_2' : Y_2 =$ # At-Bats
1	B	(289, 538)	(75, 162)
2	I	{(88, 149), 0.134; (149, 422), 0.866}	{(49, 88), 0.413; (88, 149), 0.587}
3	F	(201, 223)	(201, 223)
4	B	(184, 476)	(46, 148)
5	B	(283, 447)	(86, 115)
6	A	ϕ	ϕ
7	B	(168, 445)	(37, 135)
8	E	(137, 148)	(137, 148)
9	B	(256, 510)	(78, 124)
10	D	(101, 126)	(101, 126)
11	B	(212, 492)	(57, 151)
12	G	{(189, 238), 0.778; (238, 245), 0.222}	(189, 238)
13	B	(342, 614)	(121, 206)
14	B	(120, 439)	(35, 102)
15	I	{(80, 115), 0.066; (115, 468), 0.934}	{(55, 80), 0.428; (80, 115) 0.572}
16	C	(75, 110)	(75, 110)
17	I	{(116, 163), 0.072; (163, 557), 0.928}	{(95, 116), 0.321; (116, 163) 0.679}
18	B	(197, 507)	(52, 53)
19	H	(167, 203)	{(48, 167) 0.869; (167, 232), 0.131}

Hence, the virtual values for this observation become $\xi'(2) = (\xi_1'(2), \xi_2'(2))$ where for Y_1 (considered alone), by substitution into equations (7)-(9), we have

$$\xi_1'(2) = \{(88, 149), 0.134; \quad (149, 422), 0.866\};$$

for Y_2 considered alone, from equations (9)-(12),

$$\xi_2'(2) = \{(49, 88), 0.413; \quad (88, 149), 0.587\};$$

and that for (Y_1, Y_2) the virtual value is, from equations (9), (10) and (13),

$$\begin{aligned} \xi'(2) \quad = \quad & \{[(88, 149), (88, 149)], 0.059; \quad [(88, 149), (49, 88)], 0.075; \\ & [(149, 422), (49, 88)], 0.338; \quad [(149, 422), (88, 149)], 0.528\}. \end{aligned}$$

Under this rule, only the $u = 6$th observation fails entirely, as a pattern A virtual observation. However, only the nine observations corresponding to $u = 1, 4, 5, 7, 9, 11, 13, 14, 18$, are unaffected by this rule, to be identified as a pattern B value. The remaining eight observations are affected in various ways (with a variety of patterns occuring) but with all eight observations having some portion of the original $R(u)$ space eliminated as not being logically possible under ν_2. The virtual values for all the observations in the dataset of Table 1 after application of the rule ν_2 are displayed in Table 6 in columns (b), (c), and (d) for the variable Y_1, Y_2 and (Y_1, Y_2), respectively. Clearly, the virtual dataset contains histogram-valued observations. An algorithm for the determination of a histogram from histogram-valued observations is outlined in Section 7. Therefore, building our histograms for Y_1 (or Y_2) on the same histogram intervals as were used previously, we can obtain the histograms for Y_1 (and Y_2) as displayed in column (c) of Table 4 (and Table 5, respectively).

Table 6 (ii) - Virtual $\xi' = (\xi'_1, \xi'_2)$ under rule $\nu_2 : Y_1 \geq Y_2$

	(d) $\xi' = (\xi'_1, \xi'_2) : Y = (Y_1, Y_2)$
1	{(289, 538), (75, 162)}
2	{[(88, 149), (88, 149)], 0.059; [(88, 149), (49, 88)], 0.075; [(149, 422), (49, 88)], 0.338; [(149, 422), (88, 149)], 0.528}
3	{(201, 223), (201, 223)}
4	{(184, 476), (46, 148)}
5	{(283, 447), (86, 115)}
6	ϕ
7	{(168, 445), (37, 135)}
8	{(137, 148), (137, 148)}
9	{(256, 510), (78, 124)}
10	{101, 126), (101, 126)}
11	{(212, 492), (57, 151)}
12	{[(189, 238), (189, 238)], 0.778; [(238, 245), (189, 238)], 0.222}
13	{(342, 614), (121, 206)}
14	{(120, 439), (35, 102)}
15	{[(80, 115), (80, 115)], 0.027; [(80, 115), (55, 80)], 0.039; [(115, 468), (55, 80)], 0.389; [(115, 468), (80, 115)], 0.545}
16	{(75, 110), (75, 110)}
17	{[(116, 163), (116, 163)], 0.039; [(116, 163), (95, 116), 0.034; [163, 557), (95, 116)], 0.286; [(163, 557), (116, 163), 0.641}
18	{(197, 507), (52, 53)}
19	{[(167, 203), (48, 167)], 0.869; [(167, 203)(, 167, 232)], 0.131}

Column (d) of Table 4 and Table 5 give the corresponding histograms for Y_1 and Y_2, respectively, when in equation (2), $\alpha = 0.350$, i.e., under the rule

$$\nu_3 : Y_2 \leq 0.350 Y_1. \tag{17}$$

In this case, several more of the original observations have virtual values which follow pattern A, as would be expected; and the resulting histograms reflect this restriction. This is especially evident for the $I_{g_2} = I_8 = [200, 225)$ interval for the histogram for the number of hits Y_2. Under ν_3, the $u = 3$ and $u = 12$ observations are deleted by virtue of their becoming pattern A values in their virtual space. Yet, both of these observations, contributed nonzero

frequencies to this I_8 interval for the histograms of columns (a), (b) and (c) in Table 5. We can show that $\xi_2(3) = (201, 254)$ and $\xi_2(12) = (189, 238)$ contributed a frequency equal to 0.453 and 0.510, respectively, with a total contribution of 1.063 when there were no rules.

Finally, column (e), in Table 4 and Table 5 provides the histogram results for the set of rules

$$\nu_4 : (\nu_1, \nu_2) \equiv \{Y_1 \geq 120 \text{ and } Y_2 \leq Y_1\} \tag{18}$$

for Y_1 alone and Y_2 alone, respectively. The details are omitted.

4 Sample Means and Variances

Formulae for calculating the empirical mean and variance for interval-valued data were given by Bertrand and Goupil (2000) and for histogram-valued data by Billard and Diday (2003). We have seen how rules in effect transform the actual interval-valued data $R(u)$ into virtual data $V(u)$, $u = 1, \ldots, n$, with these virtual data also being interval-valued or histogram-valued values. Thus, use of the Bertrand-Goupil or Billard-Diday formula subsequently apply. For completeness, we provide here the formula for histogram-valued data.

Suppose our random variable Y has histogram values $\xi(u) = \{(a_{uj}, b_{uj}), p_{uj}; j = 1, \ldots, s_u\}$ with $\sum_j p_{uj} = 1$, where, for observation u, p_{uj} is the relative frequency (or probability) of taking values on the jth interval (a_{uj}, b_{uj}), $j = 1, \ldots, s_u$ where s_u is the total number of histogram-intervals. Note that when $s_u = 1$ and hence $p_{uj} = 1$ for all j, we have an interval-valued observation. Then, from Billard and Diday (2003), the sample mean is given by

$$\bar{Y} = \frac{1}{2n} \sum_{u=1}^{n} \left\{ \sum_{j=1}^{s_u} (a_{uj} + b_{uj}) p_{uj} \right\} \tag{19}$$

and the sample variance S^2 and standard deviation S are found from

$$S^2 = \frac{1}{3n} \sum_{u=1}^{n} \left\{ \sum_{j=1}^{s_u} (a_{uj}^2 + a_{uj} b_{uj} + b_{uj}^2) p_{uj} \right\} - \frac{1}{4n^2} \left\{ \sum_{u=1}^{n} \sum_{j=1}^{s_u} (a_{uj} + b_{uj}) p_{uj} \right\}^2 . \tag{20}$$

Therefore, by using equations (19) and (20) on the original data of Table 1, we obtain the $\bar{Y}$ and S values as shown in Table 4 for the Number of At-Bats Y_1, and in Table 5 for the Number of Hits Y_2. Likewise, under the rules $\nu_1, \ldots, \nu_4$, we can apply these equations (19) and (20) to the relevant virtual data to obtain the corresponding values for $\bar{Y}$ and S; these are also displayed in Table 4 and Table 5 for Y_1 and Y_2, respectively.

5 Joint Histograms

Principles underlying the univariate case apply to constructing histograms for $p \geq 2$ variables. For illustrative clarity, let us take $p = 2$ and let us construct the joint histogram for $Y = (Y_1, Y_2)$ on the histogram rectangles $R(g_1, g_2) = \{[ha_1, hb_1) \times [ha_2, hb_2)\}$, $g_1 = 1, \ldots, r_1$, $g_2 = 1, \ldots, r_2$. Then, when there are no rules present, we have from Billard and Diday (2003) that the frequency that observations lie in the rectangle $R_{g_1 g_2}$ is

$$O(g_1, g_2) = \sum_u \frac{|R(u) \cap R(g_1, g_2)|}{|R(u)|}. \tag{21}$$

The relative frequency is $p_{g_1 g_2} = O(g_1, g_2)/n$.

An algorithm for calculating these $p_{g_1 g_2}$ and $O(g_1, g_2)$ terms is given in Section 7. To illustrate this, we construct a joint histogram for $Y = (Y_1, Y_2)$ using the baseball data of Table 1. Suppose we take histogram intervals on Y_1 as $[0, 50), [50, 200), \ldots, [500, 650]$ and the histogram intervals on Y_2 as $[0, 75), [75, 125), \ldots, [225, 275]$. Thus, e.g., for $g_1 = 3$, $g_2 = 4$, we have the histogram rectangle $R(3, 4) = [200, 350) \times [175, 225)$. The observed frequencies are shown in Table 7. The corresponding relative frequencies $pg_1 g_2$ are plotted in Figure 3.

When rules are present we replace the actual observation $R(u)$ by its virtual observation $V(u)$. When the $V(u)$ values are themselves rectangles, then the same computational algorithm used for equation (21) pertains. It is often the case that this virtual space $V(u)$, itself a multi-sided hypercube, can be partitioned into components. The patterns G, H, I for the baseball data are examples of such partitioning. When these components are themselves rectangular, then again use of the basic joint histogram algorithm of Section 7 pertains.

Table 7 - Joint Histogram for (Y_1, Y_2) - No Rules (ν_0)

g_2	Y_2 \ g_1 / Y_1	1	2	3	4	5	
		[0, 50)	[50, 200)	[200, 350)	[350 – 500)	[500 – 650]	Freq Y_2
1	[0, 75)	0.000	0.544	1.473	1.161	0.022	3.201
2	[75, 125)	0.000	2.668	2.556	2.783	0.204	8.211
3	[125, 175)	1.000	1.686	0.749	1.095	0.384	4.914
4	[175, 225)	0.000	0.644	0.826	0.201	0.153	1.824
5	[225, 275]	0.000	0.302	0.549	0.000	0.000	0.850
Freq Y_1		1.000	5.844	6.153	5.240	0.763	19

Some of the virtual observations in the baseball example have components that are triangles. For these components, routine use of this basic algorithm treats these triangular pieces as though they are rectangles. For these components, appropriate adjustment has to be made. We omit the details.

Table 8 - Joint Histogram for (Y_1, Y_2) - Rule $\nu_2 : Y_2 \leq Y_1$

g_2	Y_2 \ g_1 / Y_1	1 [0, 50)	2 [50, 200)	3 [200, 350)	4 [350 − 500)	5 [500 − 650]	Freq Y_2
1	[0, 75)	0.000	0.599	1.487	1.167	0.023	3.275
2	[75, 125)	0.000	2.888	2.589	2.801	0.206	8.484
3	[125, 175)	0.000	1.501	0.767	1.105	0.386	3.761
4	[175, 225)	0.000	0.221	1.626	0.201	0.153	2.201
5	[225, 275]	0.000	0.059	0.220	0.000	0.000	0.279
Freq Y_1		0.000	5.268	6.690	5.274	0.768	18

Thus, to illustrate, we construct the joint histogram for the baseball data, when the rule ν_2 holds on the same histogram intervals as were used above in Table 6 and Figure 3. The observed frequencies are displayed in Table 8, and the relative frequencies are plotted in Figure 4.

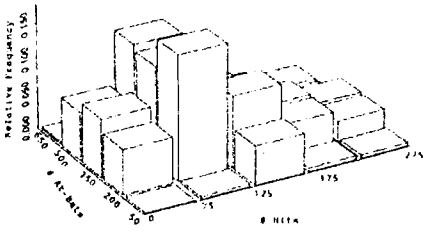


Figure 3 - Joint Histogram (Y_1, Y_2), No Rules

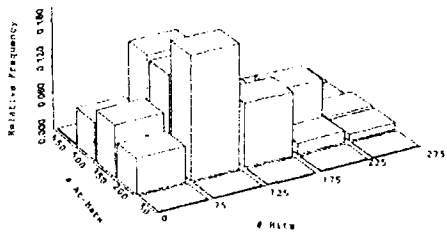


Figure 4 - Joint Histogram (Y_1, Y_2), under Rule ν_2

6 Basic Principles

Suppose we have observations on the p-dimensional variable $Y = (Y_1, \ldots, Y_p)$ with Y_j taking values on the interval $\xi_j = (a_j, b_j)$, $j = 1, \ldots, p$, for each observation $u = 1, \ldots, n$. Let $R(u)$ be the p-dimensional rectangle that represents the observation u. Let there be a set of rules $\nu = (\nu_1, \nu_2, \ldots)$.

The basic issue is to find that subspace $V(u)$ of $R(u)$ which represents those values of $R(u)$ for which the rule ν holds; i.e., those $x = (x_1, \ldots, x_p)$ such that $\nu_i(x) = 1$, for all rules ν_i; see equation (1). For some u, this $V(u) = \phi$, the empty set; for other u, this $V(u) = R(u)$, the original observation; and for others, this $V(u)$ is a nonempty p-dimensional rectangle $V(u) \equiv R^*(u)$ contained in $R(u)$, i.e., $V(u) \equiv R^*(u) \subseteq R(u)$. For these observations, the adjustment to the relevant calculation for the descriptive statistic of interest is routine.

Frequently, the virtual observation $V(u)$ is a p-dimensional non-rectangular hypercube. However, it is usually the case that this virtual space can be

partitioned into components each of which is itself a rectangle (or a shape such as a triangle which is clearly a half-rectangle). For example, pattern I observed for the baseball data under rule ν_2 (see Table 3) can be partitioned into components R_j, $j = 1, \ldots, 4$. Each component, R_j is a proportion, p_j, of the whole virtual space V, with $V = \bigcup_j R_j$, and $\Sigma p_j = 1$, for a given u. Each R_j component is then added to the dataset as though it were an "observation" but is an observation with probability weight p_j. This necessitates a probability weight of $p = 1$ for those observations u for which $V(u)$ is itself a p-dimensional rectangle. When all virtual components $\{R_j, j = 1, 2, \ldots\}$ are rectangles, then the direct use of the methodologies presented herein apply.

When (as in the baseball example) an R_j is a triangle, adjustment has to be made to ensure that the calculated area of the "triangle" is indeed that, and not the area of the corresponding rectangle. Components R_j that are not rectangles are different. In some instances, this non-rectanglar shape in and of itself is not a problem though calculating the probability might (but should not in general) be tricky. Situations that are otherwise difficult will be treated elsewhere. This present work assumes $V(u)$ can be partitioned into rectangular components (with appropriate adjustment for trianglar pieces).

7 Histogram Algorithms

7.1 Univariate Histograms of Histogram - Valued Data

An algorithm for calculating the histogram of a set of histogram-valued data is briefly outlined as follows. Suppose the random variable Y has realizations

$$\xi(u) = \{\xi_{uj} = [a_{uj}, b_{uj}), p_{uj};\ j = 1, \ldots, s_u\}$$

for each $u = 1, \ldots, n$, where p_{uj} is the observed relative frequency on the interval $[a_{uj}, b_{uj})$ with $\sum_j p_{uj} = 1$, and where s_u is the number of histogram intervals for the data value u. In the virtual descriptions of patterns G, H, I in Section 4, $s_u = 2$. Note that when $s_u = 1$, and hence $p_1 = 1$, the observation is interval-valued (as a special case of histogram-valued variables). Suppose we want to construct a histogram of these $\{\xi(u), u = 1, \ldots, n\}$ observations. Let there be r histogram intervals $I_g = [ha, hb), g = 1, \ldots, r-1$, and $I_r = [ha, hb]$, where clearly in I_1, $ha \leq \min_{u,j} a_{uj}$, and in I_r, $hb \geq \max_{u,j} b_{uj}$. Then, from Billard and Diday (2005), the observed frequency for the histogram interval I_g is given by

$$O(g) = \sum_u \sum_{j \in Z(g)} p_{uj} \frac{||\xi_{uj} \cap I_g||}{||\xi_{uj}||} \tag{22}$$

where, for each $u = 1, \ldots, n$, $Z(g)$ is the set of all ξ_{uj} intervals which overlap with I_g and where $||A||$ is the length of the interval A. The relative frequency is $p_g = O(g)/n$.

The basic algorithm for computing $O(g)$ from equation (22) essentially requires ascertaining the precise nature of the $(\xi_{uj} \cap I_g)$ term across all $j = 1, \ldots, s_u$ values, taking specific care of the exact endpoint values (a_{uj}, b_{uj}) and (ha, hb) and their relative relationships with each other. Once all possibilities have been identified, the process is reasonably straightforward. The algorithm is presented as a SAS macro; using SAS is not essential.

The algorithm itself is presented in Appendix A. This algorithm has in effect three components. The first (identified A) relates to the various initial commands to set up the computer program (such as options, titles, ...) including reading in the data. This version of the algorithm assumes data are inputted as

$$\xi = \{[aj, bj)pj;\ j = 1, \ldots, nsu\}$$

where $nsu = \max_u s_u$. Adjustments to the data to accommodate data where $s_u \neq s$ for all u (commonly the case) can also be made at this data manipulation stage (or the appropriate terms in the core macro of part B can be adjusted if preferred). For ease of presentation, we assume these are appropriately handled in the first stage A. The core macro utilizes generic terms for the maximum number of data-histogram intervals (nsu), the first and last histogram ha values ($first\_ha$, and $last\_ha$) and the histogram interval length ($hinc$). Thus, these are also set in Part A.

Part B is the core macro, here called "**hist**". Part B1 addresses the values of the frequencies to be added (the *add* term) for each data histogram entry and its relationship to the histogram interval (ha, hb). Part B2 adds these frequencies over all data values and calculates the relative frequencies. This part also includes a simple format for outputting the resulting frequencies and relative frequencies (here referred to as 'probabilities'); whatever format suits the reader should be substituted. This macro can then be invoked to calculate the $O(g)$ and p_g for a given g.

Rather than repeatedly invoking the 'hist' macro for each I_g, we can, alternatively, use part C which is a simple macro, called '**histall**', which calculates all the histogram frequencies over all I_g inside a simple do-loop routine. Then, this '**histall**' macro can be called once, to give $O(g)$ and p_g for all $g = 1, \ldots, r$. This is particularly useful when all histogram intervals I_g are of the same length.

Clearly, this is a basic algorithm to calculate $O(g)$ and p_g. Variations to accommodate different features (e.g., histograms with different I_g interval lengths) can be readily made.

7.2 Joint Histograms for Interval-Valued Data

Let the p-dimensional interval-valued observation be $Y = (Y_1, \ldots, Y_p)$ with Y_v taking values on the interval $(a_v, b_v), v = 1, \ldots, p$. We want to construct the joint histogram for the two variables (Y_i, Y_j); for illustrative clarity we take (Y_1, Y_2). We may rewrite equation (21) as

$$O(g_1, g_2) = \sum_u \left(\frac{b_1^* - a_1^*}{b_1 - a_1}\right)\left(\frac{b_2^* - a_2^*}{b_2 - a_2}\right) \tag{23}$$

where $R^* = \{(a_1^*, b_1^*) \times (a_2^*, b_2^*)\}$ is the rectangle which represents the intersection of the data rectangle $R(u)$ and the histogram rectangle $I(g_1, g_2)$. This R^* rectangle can be empty. We note that the interval (a_i^*, b_i^*), $i = 1, 2$, may or may not overlap with the relevant (ha, hb) interval, and that the various possibilities observed when calculating the histogram of histogram data in Section 7.1 pertain here also (see Appendix A); but they pertain for both the Y_1 and Y_2 dimensions. More specifically, $O(g_1, g_2)$ is the sum (over all observations) of cross-product terms, with each cross-product term equal to the product of one term from each of Y_1 and Y_2. A basic algorithm is given in Appendix B and proceeds as follows. Part A, as before relates to the relevant preliminary program statements, including the input of the data.

Calculation of the $O(g_1, g_2)$ of equation (23) consists of two parts, presented here as macros under Parts B and C, respectively. The macro of Part B, called '**hist**' (comparable to but different from the '**hist**' macro of Section 7.1) calculates the term $(b_v^* - a_v^*)/(b_v - a_v)$ for a single v value. This macro is written to allow for any specified v value (as shown in, e.g., the *a&v* term). This term is called *prod&k*. This macro will be invoked twice, once for each k value (e.g., $k = 1$ and $k = 2$), to give a product term value of prod1 and prod2 (for Y_1, and Y_2, respectively) for each observation u.

The second macro of Part C, called '**hist2**', reads in the calculated prod1 and prod2 terms, takes their product and sums these over all observations, i.e., it completes the calculation of equation (23). The cross-product and their summation is achieved via an IML routine, as shown. This particular macro calculates the observed joint frequency and the corresponding joint probability for a single histogram rectangle. There are also included simple format lines for printing these results. Thus, invoking the '**hist2**' macro will produce the joint histogram value for a given histogram rectangle. A third macro, along the lines of the '**histall**' macro shown in (Part C) of Section 7.1 (see Appendix A), could also be written to enable all histogram rectangles to be considered with one invocation. The details are omitted.

There is one final, but important, feature. Let us first consider standard rectangular ξ spaces; i.e., consider 2-dimensional interval-valued rectangular data $R(u)$ for all u such as when there are no rules, or for 2-dimensional virtual data $V(u)$ which are also rectangles. For these situations, the algo-

rithm as described thus far proceeds without any problem. However, when, as often occurs, the virtual observation $V(u)$ is the union of smaller rectangles each with some probability $p_i < 1$, then appropriate adjustment must be made. For example, in the baseball example, under the rule ν_2, we see from Table 6(ii) that the virtual observation for the $u = 2$ contains the rectangle $[(149, 422), (88, 149)]$ with probability $p = 0.528$ ($\neq 1$). In contrast, the virtual observation for $u = 1$ is the same as the actual observation, viz., the rectangle $[(289, 538), (75, 162)]$ with probability $p = 1$. We saw from Section 6 that in general a non-rectangular $V(u)$ can be decomposed into nonoverlapping rectangles $R_j(u)$, $j = 1, \ldots, k$, each with probability p_j, $\Sigma p_j = 1$. Therefore, in the data input and manipulation stage (of Part A), these rectangles and probabilities are calculated. We treat each of these $R_j(u)$ as though it were a whole "observation" but with probability $p = p_j$ (instead of the initially set $p = 1$ value). This is reflected in the '**hist**' macro by summing these probabilities to obtain the sample size n ($n\&v = n\&v + p\&v$, of line 4, instead of the more intuitive $n = n + 1$). It is also reflected in the '**hist2**' macro by taking the product $prod1 * prod2 * p1$ in the IML routine.

8 Conclusion

Rules can have many forms and can impact data in various ways. While our study herein focused on logical dependency rules on interval-valued data, other forms of data may require different types of rules. There can also be situations where the rule itself varies depending on the "value" of the symbolic data (interval-valued or not). For example, outlier values may induce their own dependency rules. In another situation, it may be that one variable is correlated with another variable (as is prevalent with medical and/or biologically based variables) with the resulting need for rules that are themselves observation-dependent. In a different direction, once histograms (for example) have been developed in the presence of rules, then other parametric distribution procedures (such as fitting, estimation, and so forth) can be developed. The field is wide-open for more research.

Appendix A - Histogram Algorithm

```
(A): program option (statements), as approprite;
data one;
title (statements);
input data (statements);
/* Data Histograms:
  Data(u) = {(a1, b1)p1, .... (as, bs)ps}, u= 1, ...n,
```

```
     where s=#data-histogram intervals, and
     n=Number of Observations */

%let nsu=  ;       /* #intervals Data histogram,
                      nsu >= max(su, u=1,...n)*/
%let first_ha=  ; /* First ha value */
%let last_ha=   ; /* Last ha value */
%let hinc=  ;      /* Width of Histogram interval (ha, hb)*/
%\end{quote}

(B1):/*Macro for Histogram on Histogram Interval (ha, hb) */
%macro hist(datain=, dataout=, ha= );
data &dataout;
set &datain end=last;
ha=&ha;
hb=&ha + &hinc;
retain add freq n 0;
%do j= 1 %to nsu;
/*Each j corresponds to each data histogram j piece */
  if a&j < ha & b&j <= ha then do;
   add=0;
  end;
  if a&j <= ha & b&j > ha & b&j < hb then do;
   add=p&j*(b&j-ha)/(b&j-a&j);
  end;
  if a&j > ha & b&j < hb then do;
   add=p&j;
  end;
  if a&j=ha & b&j=hb then do;
   add=p&j;
  end;
  if a&j > ha & a&j < hb & b&j >= hb then do;
   add=p&j*(hb-a&j)/(b&j-a&j);
  end;
  if a&j < ha & b&j > hb then do;
   add=p&j*(hb-ha)/(b&j-a&j);
  end;
  if a&j > ha & a&j >= hb then do;
   add=0;
  end;
  if a&j < ha & b&j = hb then do;
   add=p&j*(hb-ha)/(b&j-a&j);
  end;
  else if a&j=ha & hb < b&j then do;
```

```
   add=p&j*(hb-ha)/(b&j-a&j);
  end;

(B2): freq=freq+add;
 output;
%end;
 n = n + 1;
if last then do;
  prob=sum/n;
  file print;
  put "Interval g = ("ha "- " hb 3.0"):  Frequency = " freq 8.4 "
  Probability = " prob 6.4;
end;
run;
%mend hist;

(C): /*macro to calculate all histogram frequencies
        in one step */
%macro histall(datain=, dataout=);
data &dataout; set &datain;
%do h = &first_ha %to &last_ha %by &hinc;
  %hist (datain=one,dataout=two,ha=&h);
%end;
run;
%mend histall;

/* call macro to do complete histogram over
   all histogram intervals */
%histall(datain=one, dataout=two);

/* call macro to do one histogram interval */
%hist (datain=one,dataout=two,ha=125);
quit;
```

Appendix B - Joint Histogram Algorithm

```
(A): Program option statements, as appropriate;
data one;
input data statements;
/*Data intervals Data(u)={(av, bv), v=1,...,p}, u = 1,...,n;
  n = # observations.
```

```
  Set initial probabilities pv = 1; see text. */

(B): /*Macro for Histogram Calculation */
/* Variable v1, Histogram Interval = (ha1, hb1),
   and variable v2, Histogram Interval = (ha2, hb2);
   Joint Histogram Frequencies on
   Histogram Rectangle (ha1, hb1)x(ha2,hb2)*/

 %macro hist(datain=, dataout=, v=, ha=, hb=, k=);
data &dataout;
set &datain;
retain prod&k n&v 0;
 n&v = n&v + p&v;
 if a&v < &ha & b&v <= &ha then do;
   prod&k=0;
 end;
 if a&v <= &ha & b&v > &ha & b&v < &hb then do;
   prod&k=(b&v-&ha)/(b&v-a&v);
 end;
 if a&v > &ha & b&v < &hb then do;
   prod&k=1;
 end;
 if a&v=&ha & b&v=&hb then do;
   prod&k=1;
 end;
 if a&v > &ha & a&v < &hb & b&v >= &hb then do;
   prod&k=(&hb-a&v)/(b&v-a&v);
 end;
 if a&v < &ha & b&v > &hb then do;
   prod&k=(&hb-&ha)/(b&v-a&v);
 end;
 if a&v > &ha & a&v >= &hb then do;
   prod&k=0;
 end;
 if a&v < &ha & b&v = &hb then do;
   prod&k=(&hb-&ha)/(b&v-a&v);
 end;
 if a&v=&ha & &hb < b&v then do;
   prod&k=(&hb-&ha)/(b&v-a&v);
 end;
 output;
run;
%mend hist;
```

```
(C): /*Macro to find the 2-dim histogram frequencies*/
%macro hist2(datain2=, dataout2=, v1=, ha1=, hb1=, v2=,
        ha2=, hb2=);
data data2; set &datain2;
%hist(datain=&datain2, dataout=two1, v=&v1, ha= &ha1,
        hb= &hb1, k=1);
%hist(datain=&datain2, dataout=two2, v=&v2, ha= &ha2,
        hb= &hb2, k=2);
data &dataout2;
merge two1 two2;
proc iml;
use &dataout2;
read all var {prod1 prod2 p1} into pr;
c=J(nr,1,0);
c=pr[,1]#pr[,2]#pr[,3];
nob=sum(pr[,3]);
jointfreq=sum(c);
jointprob=jointfreq/nob;
print "Frequency for (&ha1,&hb1)x(&ha2,&hb2) = " jointfreq;
print "Probability for (&ha1,&hb1)x(&ha2,&hb2) = "
        jointprob;
print "Number of Observations = " nob;
%mend hist2;

/*Invoke macro */
%hist2(datain2=use,dataout2=two,v1=1,ha1=0,hb1=50,v2=2,
       ha2=0,hb2=75);

quit;
```

References

Bertrand, P. & Goupil, F. (2000), 'Descriptive statistics for symbolic data', *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday), Berlin, Springer-Verlag, pp 103–124.

Billard, L. & Diday, E. (2003), 'From the statistics of data to the statistics of knowledge: Symbolic data analysis', *Journal of the American Statistical Association* **98**, pp 470–487

Billard, L. & Diday, E. (2005), 'Histograms in symbolic data analysis', *Bulletin International Statistical Institute* (in press).

Bock, H. -H. & Diday, E. (2000), 'Analysis of Symbolic Data: Exploratory

Methods for Extracting Statistical Information from Complex Data', Berlin, Springer-Verlag.

Vanessa, A. & Vanessa, L. (2004), 'La meilleure équipe de Base-ball. CEREMADE, Université dé Paris 9, Dauphine.