2

Symbolic Data: Basics

In this chapter, we describe what symbolic data are, how they may arise, and their different formulations. Some data are naturally symbolic in format, while others arise as a result of aggregating much larger data sets according to some scientific question(s) that generated the data sets in the first place. Thus, section 2.2.1 describes non-modal multi-valued or lists of categorical data, with modal multi-valued data in section 2.2.2; lists or multi-valued data can also be called simply categorical data. Section 2.2.3 considers interval-valued data, with modal interval data more commonly known as histogram-valued data in section 2.2.4. We begin, in section 2.1, by considering the distinctions and similarities between individuals, classes, and observations. How the data arise, such as by aggregation, is discussed in section 2.3. Basic descriptive statistics are presented in section 2.4. Except when necessary for clarification purposes, we will write "interval-valued data" as "interval data" for simplicity; likewise, for the other types of symbolic data.

It is important to remember that symbolic data, like classical data, are just different manifestations of sub-spaces of the p-dimensional space $\mathbb{R}^p$ always dealing with the same random variables. A classical datum is a point in $\mathbb{R}^p$, whereas a symbolic value is a hypercube or a Cartesian product of distributions in $\mathbb{R}^p$. Thus, for example, the $p = 2$-dimensional random variable (Y_1, Y_2) measuring height and weight (say) can take a classical value $Y_1 = 68$ inches and $Y_2 = 70$ kg, or it may take a symbolic value with $Y_1 = [65, 70]$ and $Y_2 = [70, 74]$ interval values which form a rectangle or a hypercube in the plane. That is, the random variable is itself unchanged, but the realizations of that random variable differ depending on the format. However, it is also important to recognize that since classical values are special cases of symbolic values, then regardless of analytical technique, classical analyses and symbolic analyses should produce the same results when applied to those classical values.

Clustering Methodology for Symbolic Data, First Edition. Lynne Billard and Edwin Diday.

2.1 Individuals, Classes, Observations, and Descriptions

In classical statistics, we talk about having a random sample of n observations $Y_1, \ldots, Y_n$ as outcomes for a random variable Y. More precisely, we say Y_i is the observed value for individual i, $i = 1, \ldots, n$. A particular observed value may be $Y = 3$, say. We could equivalently say the description of the ith individual is $Y_i = 3$. Usually, we think of an individual as just that, a single individual. For example, our data set of n individuals may record the height Y of individuals, Bryson, Grayson, Ethan, Coco, Winston, Daisy, and so on. The "individual" could also be an inanimate object such as a particular car model with Y describing its capacity, or some other measure relating to cars. On the other hand, the "individual" may represent a class of individuals. For example, the data set consisting of n individuals may be n classes of car models, Ford, Renault, Honda, Volkswagen, Nova, Volvo, … , with Y recording the car's speed over a prescribed course, etc. However individuals may be defined, the realization of Y for that individual is a single point value from its domain $\mathcal{Y}$.

If the random variable Y takes quantitative values, then the domain (also called the range or observation space) is $\mathcal{Y}$ taking values on the real line $\mathbb{R}$, or a subset of $\mathbb{R}$ such as $\mathbb{R}^+$ if Y can only take non-negative or zero values. When Y takes qualitative values, then a classically valued observation takes one of two possible values such as {Yes, No} or coded to $\mathcal{Y} = \{0, 1\}$, for example. Typically, if there are several categories of possible values, e.g., bird colors with domain $\mathcal{Y} = \{\text{red, blue, green, white,}\ldots\}$, a classical analysis will include a different random variable for each category and then record the presence (Yes) or absence (No) of each category. When there are p random variables, then the domain of $\mathbf{Y} = (Y_1, \ldots, Y_p)$ is $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_p$.

In contrast, when the data are symbolic-valued, the observations $Y_1, \ldots, Y_m$ are typically realizations that emerge after aggregating observed values for the random variable Y across some specified class or category of interest (see section 2.3). Thus, for example, observations may refer now to m classes, or categories, of age×income, or to m species of dogs, and so on. Thus, the class Boston (say) has a June temperature range of [58°F, 75°F]. In the language of symbolic analysis, the individuals are ground-level or order-one individuals and the aggregations – classes – are order-two individuals or "objects" (see, e.g., Diday (1987, 2016), Bock and Diday (2000a,b), Billard and Diday (2006a), or Diday and Noirhomme-Fraiture (2008)).

On the other hand, suppose Gracie's pulse rate Y is the interval $Y = [62, 66]$. Gracie is a single individual and a classical value for her pulse rate might be $Y = 63$. However, this interval of values would result from the collection, or aggregation, of Gracie's classical pulse rate values over some specified time period. In the language of symbolic data, this interval represents the pulse rate of the class "Gracie". However, this interval may be the result of aggregating the classical point values of all individuals named "Gracie" in some larger data base. That is, some symbolic realizations may relate to one single individual, e.g., Gracie, whose pulse rate may be measured as [62, 66] over time, or to a set of all those Gracies of interest. The context should make it clear which situation prevails.

In this book, symbolic realizations for the observation u can refer interchangeably to the description Y_u of classes or categories or "individuals" u, $u = 1, \ldots, m$, that is, simply, u will be the unit (which is itself a class, category, individual, or observation) that is described by Y_u. Furthermore, in the language of symbolic data, the realization of Y_u is referred to as the "description" d of Y_u, $d(Y(u))$. For simplicity, we write simply Y_u, $u = 1, \ldots, m$.

2.2 Types of Symbolic Data

2.2.1 Multi-valued or Lists of Categorical Data

We have a random variable Y whose realization is the set of values $\{Y_1, \ldots, Y_{s'}\}$ from the set of possible values or categories $\mathcal{Y} = \{Y_1, \ldots, Y_s\}$, where s and s' with $s' \leq s$ are finite. Typically, for a symbolic realization, $s' > 1$, whereas for a classical realization, $s' = 1$. This realization is called a list (of s' categories from $\mathcal{Y}$) or a multi-valued realization, or even a multi-categorical realization, of Y. Formally, we have the following definition.

Definition 2.1 Let the p-dimensional random variable $\mathbf{Y} = (Y_1, \ldots, Y_p)$ take possible values from the list of possible values in its domain $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_p$ with $\mathcal{Y}_j = \{Y_{j1}, \ldots, Y_{js_j}\}, j = 1, \ldots, p$. In a random sample of size m, the realization $\mathbf{Y}_u$ is a **list** or **multi-valued** observation whenever

$$\mathbf{Y}_u = (\{Y_{ujk_j};\ k_j = 1, \ldots, s_{uj}\},\ j = 1, \ldots, p), \qquad u = 1, \ldots, m. \tag{2.2.1}$$

□

Notice that, in general, the number of categories s_{uj} in the actual realization differs across realizations (i.e., $s_{uj} \neq s_j$), $u = 1, \ldots, m$, and across variables Y_j (i.e., $s_{uj} \neq s_u$), $j = 1, \ldots, p$.

Example 2.1 Table 2.1 (second column) shows the list of major utilities used in a set of seven regions. Here, the domain for the random variable $Y =$ utility is $\mathcal{Y} = \{$coal, oil, wood, electricity, gas, …, (possible utilities), … $\}$. For example, for the fifth region ($u = 5$), $Y_5 = \{$gas, oil, other$\}$. The third region, $u = 3$, has a single utility (coal), i.e., the utility usage for this region is a classical realization. Thus, we write $Y_3 = \{$coal$\}$. If a region were uniquely identified by its utility usage, and we were to try to identify a region $u = 7$ (say) by a single usage, such as electricity, then it could be mistaken for region six, which is quite a

Table 2.1 List or multi-valued data: regional utilities (Example 2.1)

Region u	**Major utility**	**Cost**
1	{electricity, coal, wood, gas}	[190, 230]
2	{electricity, oil, coal}	[21.5, 25.5]
3	{coal}	[40, 53]
4	{other}	15.5
5	{gas, oil, other}	[25, 30]
6	{electricity}	46.0
7	{electricity, coal}	[37, 43]

different region and is clearly not the same as region seven. That is, a region cannot in general be described by a single classical value, but only by the list of types of utilities used in that region, i.e., as a symbolic list or multi-valued value. □

While list data mostly take qualitative values that are verbal descriptions of an outcome, such as the types of utility usages in Example 2.1, quantitative values such as coded values $\{1, 2, \dots\}$ may be the recorded value. These are not necessarily the same as ordered categorical values such as $\mathcal{Y} = \{$small, medium, large$\}$. Indeed, a feature of categorical values is that there is no prescribed ordering of the listed realizations. For example, for the seventh region ($u = 7$) in Table 2.1, the description {electricity, coal} is exactly the same description as {coal, electricity}, i.e., the same region. This feature does not carry over to quantitative values such as histograms (see section 2.2.4).

2.2.2 Modal Multi-valued Data

Modal lists or modal multi-valued data (sometimes called modal categorical data) are just list or multi-valued data but with each realized category occurring with some specified weight such as an associated probability. Examples of non-probabilistic weights include the concepts of capacities, possibilities, and necessities (see Billard and Diday, 2006a, Chapter 2; see also Definitions 2.6–2.9 in section 2.2.5). In this section and throughout most of this book, it is assumed that the weights are probabilities; suitable adjustment for other weights is left to the reader.

Definition 2.2 Let the p-dimensional random variable $\mathbf{Y} = (Y_1, \dots, Y_p)$ take possible values from the list of possible values in its domain $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_p$ with $\mathcal{Y}_j = \{Y_{j1}, \dots, Y_{js_j}\}$, $j = 1, \dots, p$. In a random sample of size m, the realization $\mathbf{Y}_u$ is a **modal list**, or **modal multi-valued** observation whenever

$$\mathbf{Y}_u = (\{Y_{ujk_j}, p_{ujk_j};\ k_j = 1, \dots, s_{uj}\},\ j = 1, \dots, p),\ \sum_{k_j=1}^{s_{uj}} p_{ujk_j} = 1,\ u = 1, \dots, m. \tag{2.2.2}$$

□

Without loss of generality, we can write the number of categories from $\mathcal{Y}_j$ for the random variable Y_j, as $s_{uj} = s_j$, for all $u = 1, \dots, m$, by simply giving unrealized categories ($Y_{jk'}$, say) the probability $p_{ujk'} = 0$. Furthermore, the non-modal multi-valued realization of Eq. (2.2.1) can be written as a modal multi-valued observation of Eq. (2.2.2) by assuming actual realized categories from $\mathcal{Y}_j$ occur with equal probability, i.e., $p_{ujk_j} = 1/s_{uj}$ for $k_j = 1, \dots, s_{uj}$, and unrealized categories occur with probability zero, for each $j = 1, \dots, p$.

Example 2.2 A study of ascertaining deaths attributable to smoking, for $m = 8$ regions, produced the data of Table 2.2. Here, Y = cause of death from smoking, with domain $\mathcal{Y}$ = {death from smoking, death from lung cancer, death from respiratory diseases induced by smoking} or simply $\mathcal{Y}$ = {smoking, lung cancer, respiratory}. Thus, for region $u = 1$, smoking caused lung cancer deaths in 18.4% of the smoking population, 18.8% of all smoking deaths were from respiratory diseases, and 62.8% of smoking deaths were from other smoking-related causes, i.e., $(p_{11}, p_{12}, p_{13}) = (0.628, 0.184, 0.188)$, where for simplicity we have dropped the $j = 1 = p$ subscript. Data for region $u = 7$ are limited to $p_{71} = 0.648$. However, we know that $p_{72} + p_{73} = 0.352$. Hence, we can assume that $p_{72} = p_{73} = 0.176$. On the other hand, for region $u = 8$, the categories {lung cancer, respiratory} did not occur, thus the associated probabilities are $p_{82} = p_{83} = 0.0$. □

Table 2.2 Modal multi-valued data: smoking deaths (p_{uk}) (Example 2.2)

Region	Proportion p_{uk} of smoking deaths attributable to:		
u	Smoking	Lung cancer	Respiratory
1	0.628	0.184	0.188
2	0.623	0.202	0.175
3	0.650	0.197	0.153
4	0.626	0.209	0.165
5	0.690	0.160	0.150
6	0.631	0.204	0.165
7	0.648		
8	1.000		

2.2.3 Interval Data

A classical realization for quantitative data takes a point value on the real line $\mathbb{R}$. An interval-valued realization takes values from a subset of $\mathbb{R}$. This is formally defined as follows.

Definition 2.3 Let the p-dimensional random variable $\mathbf{Y} = (Y_1, \dots, Y_p)$ take quantitative values from the space $\mathbb{R}^p$. A random sample of m realizations takes **interval** values when

$$\mathbf{Y}_u = ([a_{u1}, b_{u1}], \dots, [a_{up}, b_{up}]), \qquad u = 1, \dots, m, \tag{2.2.3}$$

where $a_{uj} \le b_{uj}$, $j = 1, \dots, p$, $u = 1, \dots, m$, and where the intervals may be open or closed at either end (i.e., $[a, b)$, $(a, b]$, $[a, b]$, or (a, b)). □

Example 2.3 Table 2.1 (right-hand column) gives the cost (in \$) of the regional utility usage of Example 2.1. Thus, for example, the cost in the first region, $u = 1$, ranges from 190 to 230. Clearly, a particular household in this region has its own cost, 199, say. However, not all households in this region have the same costs, as illustrated by these values. On the other hand, the recorded costs for households in region $u = 6$ is the classical value 46.0 (or the interval $[46, 46]$). □

There are numerous examples of naturally occurring symbolic data sets. One such scenario exists in the next example.

Example 2.4 The data of Table 2.3 show the minimum and maximum temperatures recorded at six weather stations in China for the months of January (Y_1) and July (Y_2) in 1988. Also shown is the elevation (Y_3), which is a classical value. The complete data set is extracted from <http://dss.ucar.edu/datasets/ds578.5> and is a multivariate time series with temperatures (in °C) for many

Table 2.3 Interval data: weather stations (Example 2.4)

Station	Y_1 = January	Y_2 = July	Y_3 = Elevation
u	$[a_{u1}, b_{u1}]$	$[a_{u2}, b_{u2}]$	a_{u3}
1	[−18.4, −7.5]	[17.0, 26.5]	4.82
2	[−23.4, −15.5]	[12.9, 23.0]	14.78
3	[−8.4, 9.0]	[10.8, 23.2]	73.16
4	[10.0, 17.7]	[24.2, 33.8]	2.38
5	[11.5, 17.7]	[25.8, 33.5]	1.44
6	[11.8, 19.2]	[25.6, 32.6]	0.02

stations for all months over the years 1974–1988 (see also Billard, 2014). Thus, we see that, in July 1988, station $u = 3$ enjoyed temperatures from a low of 10.8°C to a high of 23.2°C, i.e., $Y_{32} = [a_{32}, b_{32}] = [10.8, 23.2]$. □

2.2.4 Histogram Data

Histogram data usually result from the aggregation of several values of quantitative random variables into a number of sub-intervals. More formally, we have the following definition.

Definition 2.4 Let the p-dimensional random variable $\mathbf{Y} = (Y_1, \dots, Y_p)$ take quantitative values from the space $\mathbb{R}^p$. A random sample of m realizations takes **histogram** values when, for $u = 1, \dots, m$,

$$\mathbf{Y}_u = (\{[a_{ujk_j}, b_{ujk_j}), p_{ujk_j};\ k_j = 1, \dots, s_{uj}\}, \quad j = 1, \dots, p), \ \sum_{k_j=1}^{s_{uj}} p_{ujk} = 1, \tag{2.2.4}$$

where $a_{ujk_j} \leq b_{ujk_j}, j = 1, \dots, p, u = 1, \dots, m$. □

Usually, histogram sub-intervals are closed at the left end and open at the right end except for the last sub-interval, which is closed at both ends. Furthermore, note that the number of histogram sub-intervals s_{uj} differs across u and across j. For the special case that $s_{uj} = 1$, and hence $p_{uj1} = 1$ for all $u = 1, \dots, m$, the histogram is an interval.

Example 2.5 Table 2.4 shows a histogram-valued data set of $m = 10$ observations. Here, the random variable is Y = flight time for airlines to fly from several departure cities into one particular hub city airport. There were approximately 50000 flights recorded. Rather than a single flight, interest was on performance for specific carriers. Accordingly, the aggregated values by airline carrier were obtained and the histograms of those values were calculated in the usual way. Notice that the number of histogram sub-intervals s_u varies across $u = 1, \dots, m$; also, the sub-intervals $[a_{uk}, b_{uk})$ can differ for $u = 1, \dots, m$, reflecting, in this case, different flight distances depending on flight routes and the like. Thus, for example, $Y_7 = \{[10, 50), 0.117; [50, 90), 0.476; [90,130), 0.236; [130,170], 0.171\}$ (data extracted from Falduti and Taibaly, 2004). □

In the context of symbolic data methodology, the starting data are already in a histogram format. All data, including histogram data, can themselves be aggregated to form histograms (see section 2.4.4).

Table 2.4 Histogram data: flight times (Example 2.5)

Airline u	Y = Flight time $\{[a_{uk}, b_{uk}), p_{uk};\ k = 1 \ldots, s_u\}$
1	{[40, 100), 0.082; [100, 180), 0.530;[180, 240), 0.172; [240, 320), 0.118; [320, 380], 0.098}
2	{[40, 90), 0.171; [90, 140), 0.285; [140, 190), 0.351; [190, 240), 0.022; [240, 290], 0.171}
3	{[35, 70), 0.128; [70, 135), 0.114; [135, 195), 0.424; [195, 255], 0.334}
4	{[20, 40), 0.060; [40, 60), 0.458; [60, 80), 0.259; [80, 100), 0.117; [100, 120], 0.106}
5	{[200, 250), 0.164; [250, 300), 0.395; [300, 350), 0.340; [350, 400], 0.101}
6	{[25, 50), 0.280; [50, 75), 0.301; [75, 100), 0.250; [100, 125], 0.169}
7	{[10, 50), 0.117; [50, 90), 0.476; [90, 130), 0.236; [130, 170], 0.171}
8	{[10, 50), 0.069; [50, 90), 0.368; [90, 130), 0.514; [130, 170], 0.049}
9	{[20, 35), 0.066; [35, 50), 0.337; [50, 65), 0.281; [65, 80), 0.208; [80, 95], 0.108}
10	{[20, 40), 0.198; [40, 60), 0.474; [60, 80), 0.144; [80, 100), 0.131; [100, 120], 0.053}

2.2.5 Other Types of Symbolic Data

A so-called mixed data set is one in which not all of the p variables take the same format. Instead, some may be interval data, some histograms, some lists, etc.

Example 2.6 To illustrate a mixed-valued data set, consider the data of Table 2.5 for a random sample of joggers from each group of $m = 10$ body types. Joggers were timed to run a specific distance. The pulse rates (Y_1) of joggers at the end of their run were measured and are shown as interval values

Table 2.5 Mixed-valued data: joggers (Example 2.6)

Group u	Y_1 = Pulse rate	Y_2 = Running time
1	[73, 114]	{[5.3, 6.2), 0.3; [6.2, 7.1), 0.5; [7.1, 8.3], 0.2}
2	[70, 100]	{[5.5, 6.9), 0.4; [6.7, 8.0), 0.4; [8.0, 9.0], 0.2}
3	[69, 91]	{[5.1, 6.6), 0.4; [6.6, 7.4), 0.4; [7.4, 7.8], 0.2}
4	[59, 89]	{[3.7, 5.8), 0.6; [5.8, 6.3], 0.4}
5	[61, 87]	{[4.5, 5.9), 0.4; [5.9, 6.2], 0.6}
6	[69, 95]	{[4.1, 6.1), 0.5; [6.1, 6.9], 0.5}
7	[65, 78]	{[2.4, 4.8), 0.3; [4.8, 5.7), 0.5; [5.7, 6.2], 0.2}
8	[58, 83]	{[2.1, 5.4), 0.2; [5.4, 6.0), 0.5; [6.0, 6.9], 0.3}
9	[79, 103]	{[4.8, 6.5), 0.3; [6.5, 7.4); 0.5; [7.4, 8.2], 0.2}
10	[40, 60]	{[3.2, 4.1), 0.6; [4.1, 6.7], 0.4}

for each group. For the first group, the pulse rates fell across the interval $Y_{11} = [73, 114]$. These intervals are themselves simple histograms with $s_{uj} = 1$ for all $u = 1, \ldots, 10$.

In addition, the histogram of running times (Y_2) for each group was calculated, as shown. Thus, for the first group ($u = 1$), 30% of the joggers took 5.3 to 6.2 time units to run the course, 50% took 6.2 to 7.1, and 20% took 7.1 to 8.3 units of time to complete the run, i.e., we have $Y_{12} = \{[5.3, 6.2), 0.3; [6.2, 7.1), 0.5; [7.1, 8.3], 0.2\}$. On the other hand, half of those in group $u = 6$ ran the distance in under 6.1 time units and half took more than 6.1 units of time. □

Other types of symbolic data include probability density functions or cumulative distributions, as in the observations in Table 2.6(a), or models such as the time series models for the observations in Table 2.6(b).

The modal multi-valued data of section 2.2.2 and the histogram data of section 2.2.4 use probabilities as the weights of the categories and the histogram sub-intervals; see Eqs. (2.2.2) and (2.2.4), respectively. While these weights are the most common seen by statistical analysts, there are other possible weights. First, let us define a more general weighted modal type of observation. We take the number of variables to be $p = 1$; generalization to $p \geq 1$ follows readily.

Definition 2.5 Let the random sample of size m be realizations of the random variable Y taking values from its domain $\mathcal{Y} = \{\eta_1, \ldots, \eta_S\}$. Then, Y_u are **modal-valued** observations when they can be written in the form

$$Y_u = \{\eta_{uk}, \pi_{uk};\ k = 1, \ldots, s_u\}, \qquad u = 1, \ldots, m, \tag{2.2.5}$$

where π_{uk} is the weight associated with the category η_{uk}. □

Table 2.6 Some other types of symbolic data

	u	Description of u
(a)	1	Distributed as a normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
	2	Distributed as a normal $N(0, \sigma^2)$
	3	Distributed as exponential (β)
	.	…
(b)	5	Follows an AR(1) time-series model
	6	Follows a MA(q) time-series model
	7	Is a first-order Markov chain
	.	…

Thus, for a modal list or multi-valued observation of Definition 2.2, the category $Y_{uk} \equiv \eta_{uk}$ and the probability $p_{uk} \equiv \pi_{uk}$; $k = 1, \ldots, s_u$. Likewise, for a histogram observation of Definition 2.4, the sub-interval $[a_{uk}, b_{uk}) \equiv \eta_{uk}$ occurs with relative frequency p_{uk}, which corresponds to the weight π_{uk}, $k = 1, \ldots, s_u$. Note, however, that in Definition 2.5 the condition $\sum_{k=1}^{s_u} \pi_{uk} = 1$ does not necessarily hold, unlike pure modal multi-valued and histogram observations (see Eqs. (2.2.2) and (2.2.4), respectively). Thus, in these two cases, the weights π_k are probabilities or relative frequencies. The following definitions relate to situations when the weights do not necessarily sum to one. As before, s_u can differ from observation to observation.

Definition 2.6 Let the random variable Y take values in its domain $\mathcal{Y} = \{\eta_1, \ldots, \eta_S\}$. The **capacity** of the category η_k is the probability that at least one observation from the set of observations $\Omega = (Y_1, \ldots, Y_m)$ includes the category η_k, $k = 1, \ldots, S$. □

Definition 2.7 Let the random variable Y take values in its domain $\mathcal{Y} = \{\eta_1, \ldots, \eta_S\}$. The **credibility** of the category η_k is the probability that all observations in the set of observations $\Omega = (Y_1, \ldots, Y_m)$ include the category η_k, $k = 1, \ldots, S$. □

Definition 2.8 Let the random variable Y take values in its domain $\mathcal{Y} = \{\eta_1, \ldots, \eta_S\}$. Let C_1 and C_2 be two subsets of $\Omega = (Y_1, \ldots, Y_m)$. A **possibility** measure is the mapping π from Ω to $[0, 1]$ with $\pi(\Omega) = 1$ and $\pi(\phi) = 0$ where ϕ is the empty set, such that for all subsets C_1 and C_2, $\pi(C_1 \cup C_2) = \max\{\pi(C_1), \pi(C_2)\}$. □

Definition 2.9 Let the random variable Y take values in its domain $\mathcal{Y} = \{\eta_1, \ldots, \eta_S\}$. Let C be a subset of the set of observations $\Omega = (Y_1, \ldots, Y_m)$. A **necessity** measure of C, $N(C)$, satisfies $N(C) = 1 - \pi(C^c)$, where π is the possibility measure of Definition 2.8 and C^c is the complement of the subset C. □

Example 2.7 Consider the random variable Y = utility of Example 2.1 with realizations shown in Table 2.1. Then, the capacity that at least one region uses the utility η = coal is 4/7, while the credibility that every region uses both coal and electricity is 3/7. □

Example 2.8 Suppose a random variable Y = number of bedrooms in a house takes values $y = \{2, 3, 4\}$ with possibilities $\pi(y) = 0.3, 0.4, 0.5$, respectively. Let C_1 and C_2 be the subsets that there are two and three bedrooms, respectively. Then, the possibility $\pi(C_1 \cup C_2) = \max\{\pi(C_1), \pi(C_2)\} = \max\{0.3, 0.4\} = 0.4$.

Now suppose C is the set of three bedrooms, i.e., $C = \{3\}$. Then the necessity of C is $N(C) = 1 - \pi(C^c) = 1 - \max\{\pi(2), \pi(4)\} = 1 - \max\{.3, .5\} = 0.5$. □

More examples for these cases can be found in Diday (1995) and Billard and Diday (2006a). This book will restrict attention to modal list or multi-valued data and histogram data cases. However, many of the methodologies in the remainder of the book apply equally to any weights $\pi_{uk}, k = 1, \dots, s_u,\;\; u = 1, \dots, m,$ including those for capacities, credibilities, possibilities, and necessities.

More theoretical aspects of symbolic data and concepts along with some philosophical aspects can be found in Billard and Diday (2006a, Chapter 2).

2.3 How do Symbolic Data Arise?

Symbolic data arise in a myriad of ways. One frequent source results when aggregating larger data sets according to some criteria, with the criteria usually driven by specific operational or scientific questions of interest.

For example, a medical data set may consist of millions of observations recording a slew of medical information for each individual for every visit to a healthcare facility since the year 1990 (say). There would be records of demographic variables (such as age, gender, weight, height, …), geographical information (such as street, city, county, state, country of residence, etc.), basic medical tests results (such as pulse rate, blood pressure, cholesterol level, glucose, hemoglobin, hematocrit, …), specific aliments (such as whether or not the patient has diabetes, a heart condition and if so what, i.e. mitral value syndrome, congestive heart failure, arrhythmia, diverticulitis, myelitis, etc.). There would be information as to whether the patient had a heart attack (and the prognosis) or cancer symptoms (such as lung cancer, lymphoma, brain tumor, etc.). For given aliments, data would be recorded indicating when and what levels of treatments were applied and how often, and so on. The list of possible symptoms is endless. The pieces of information would in analytic terms be the variables (for which the number p is also large), while the information for each individual for each visit to the healthcare facility would be an observation (where the number of observations n in the data set can be extremely large). Trying to analyze this data set by traditional classical methods is likely to be too difficult to manage.

It is unlikely that the user of this data set, whether s/he be a medical insurer or researcher or maybe even the patient him/herself, is particularly interested in the data for a particular visit to the care provider on some specific date. Rather, interest would more likely center on a particular disease (angina, say), or respiratory diseases in a particular location (Lagos, say), and so on. Or, the focus may be on age × gender classes of patients, such as 26-year-old men or 35-year-old women, or maybe children (aged 17 years and under) with leukemia, again the list is endless. In other words, the interest is on characteristics between different groups of individuals (also called classes or categories, but these categories should not be confused with the categories that make up the lists or multi-valued types of data of sections 2.2.1 and 2.2.2).

However, when the researcher looks at the accumulated data for a specific group, 50-year-old men with angina living in the New England district (say), it is unlikely all such individuals weigh the same (or have the same pulse rate, or the same blood pressure measurement, etc.). Rather, thyroid measurements may take values along the lines of, e.g., $2.44, 2.17, 1.79, 3.23, 3.59, 1.67, \ldots$. These values could be aggregated into an interval to give $[1.67, 3.59]$ or they could be aggregated as a histogram realization (especially if there are many values being aggregated). In general, aggregating all the observations which satisfy a given group/class/category will perforce give realizations that are symbolic valued. In other words, these aggregations produce the so-called second-level observations of Diday (1987). As we shall see in section 2.4, taking the average of these values for use in a (necessarily) classical methodology will give an answer certainly, but also most likely that answer will not be correct.

Instead of a medical insurer's database, an automobile insurer would aggregate various entities (such as pay-outs) depending on specific classes, e.g., age $\times$ gender of drivers or type of car (Volvo, Renault, Chevrolet, …), including car type by age and gender, or maybe categories of drivers (such as drivers of red convertibles). Statistical agencies publish their census results according to groups or categories of households. For example, salary data are published as ranges such as \$40,000–50,000, i.e., the interval $[40, 50]$ in 1000s of \$.

Let us illustrate this approach more concretely through the following example.

Example 2.9 Suppose a demographer had before her a massively large data set of hundreds of thousands of observations along the lines of Table 2.7. The data set contains, for each household, the county in which the household is located (coded to $c = 1, 2, \ldots$), along with the recorded variables: Y_1 = weekly income (in \$) with domain $\mathcal{Y}_1 = \mathbb{R}^+ = Y_1 \geq 0$, Y_2 = age of the head of household (in years) with domain being positive integers, Y_3 = children under the age of 18 who live at home with domain $\mathcal{Y}_3 = \{\text{yes, no}\}$, Y_4 = house tenure with domain $\mathcal{Y}_4 = \{\text{owner occupied, renter occupied}\}$, Y_5 = type of energy used in the home with domain $\mathcal{Y}_5 = \{\text{gas, electric, wood, oil, other}\}$, and Y_6 = driving distance to work with $\mathcal{Y}_6 = \mathbb{R}^+ = Y_6 \geq 0$. Data for the first 51 households are shown.

Suppose interest is in the energy usage Y_5 within each county. Aggregating these household data for energy across the entire county produces the histograms displayed in Table 2.8. Thus, for example, in the first county 45.83% ($p_{151} = 0.4583$) of the households use gas, while 37.5% ($p_{152} = 0.375$) use electric energy, and so on. This realization could also be written as

Table 2.7 Household information (Example 2.9)

County u	Income Y_1	Age Y_2	Child Y_3	Tenure Y_4	Energy Y_5	Distance Y_6	County u	Income Y_1	Age Y_2	Child Y_3	Tenure Y_4	Energy Y_5	Distance Y_6
1	801	44	Yes	Owner	Gas	10	1	804	41	Yes	Owner	Gas	11
2	752	20	No	Owner	Gas	6	1	797	43	Yes	Owner	Gas	1
1	901	31	No	Owner	Electric	9	2	750	52	No	Owner	Electric	4
1	802	47	Yes	Owner	Gas	15	2	748	32	No	Owner	Electric	7
1	901	42	No	Renter	Electric	12	1	907	42	No	Renter	Electric	12
2	750	43	No	Owner	Electric	4	2	748	64	No	Renter	Wood	6
1	798	42	Yes	Owner	Gas	8	1	799	22	No	Owner	Gas	8
1	901	43	Yes	Renter	Electric	9	2	754	51	No	Owner	Electric	3
2	748	30	No	Owner	Gas	4	1	799	24	No	Owner	Gas	11
2	747	32	No	Owner	Electric	6	1	897	35	No	Renter	Electric	10
1	901	77	No	Renter	Other	10	2	749	38	No	Owner	Electric	5
2	751	66	No	Renter	Oil	6	1	802	39	Yes	Owner	Gas	8
1	899	45	No	Renter	Wood	8	2	751	37	No	Owner	Electric	5
1	897	48	No	Renter	Wood	9	2	747	29	No	Owner	Electric	6

(Continued)

Table 2.7 (Continued)

County u	Income Y_1	Age Y_2	Child Y_3	Tenure Y_4	Energy Y_5	Distance Y_6	County u	Income Y_1	Age Y_2	Child Y_3	Tenure Y_4	Energy Y_5	Distance Y_6
1	800	39	Yes	Owner	Gas	8	2	749	27	No	Owner	Gas	6
1	899	32	No	Owner	Electric	12	2	749	60	No	Renter	Wood	4
2	703	46	Yes	Owner	Gas	4	2	750	33	No	Renter	Gas	4
2	753	48	No	Owner	Electric	5	1	800	21	No	Owner	Electric	6
1	898	32	Yes	Renter	Electric	10	2	747	36	No	Owner	Electric	4
2	748	43	No	Owner	Electric	5	1	802	40	Yes	Owner	Gas	10
1	896	29	No	Owner	Electric	11	2	699	43	Yes	Owner	Gas	5
1	802	39	Yes	Owner	Gas	11	2	696	39	Yes	Owner	Gas	5
2	746	37	No	Owner	Electric	4	2	747	24	No	Renter	Gas	6
2	752	54	No	Owner	Electric	6	2	752	67	No	Owner	Wood	5
1	901	59	No	Renter	Oil	11	2	749	50	No	Owner	Electric	4
3	674	41	Yes	Renter	Gas	10	…	…	…	…	…	…	…

Table 2.8 Aggregated households (Example 2.9)

Class	Y_3 = Children	Y_5 = Energy
County 1	{no, 0.5417; yes, 0.4583}	{gas, 0.4583; electric, 0.375; wood, 0.0833; oil, 0.0417; other, 0.0417}
Owner	{no, 0.4000; yes, 0.6000}	{gas, 0.7333; electric, 0.2667}
Renter	{no, 0.7778; yes, 0.2222}	{electric, 0.5556; wood, 0.2222; oil, 0.111; other, 0.1111}
County 2	{no, 0.8846; yes, 0.1154}	{gas, 0.3077; electric, 0.5385; wood, 0.1154; oil, 0.0385}
Owner	{no, 0.8571; yes, 0.1429}	{gas, 0.2857; electric, 0.6667; wood, 0.0476}
Renter	{no, 1.000}	{gas, 0.4000; wood, 0.4000; oil, 0.2000}

Y_{15} = {gas, 0.4583; electric, 0.3750; wood, 0.0833; oil, 0.0417; other, 0.0417}. Of those who are home owners, 68.75% use gas and 56.25% use electricity, with no households using any other category of energy. These values are obtained in this case by aggregating across the class of county × tenure. The aggregated energy usage values for both counties as well as those for all county × tenure classes are shown in Table 2.8.

This table also shows the aggregated values for Y_3 which indicate whether or not there are children under the age of 18 years living at home. Aggregation by county shows that, for counties $u = 1$ and $u = 2$, respectively, Y_3 takes values $Y_{13} = \{\text{no}, 0.5417; \text{yes}, 0.4583\}$ and $Y_{23} = \{\text{no}, 0.8846; \text{yes}, 0.1154\}$, respectively. We can also show that $Y_{14} = \{\text{owner}, 0.625; \text{renter}, 0.375\}$ and $Y_{24} = \{\text{owner}, 0.8077; \text{renter}, 0.1923\}$, for home tenure Y_4.

Both Y_3 and Y_5 are modal multi-valued observations. Had the aggregated household values been simply identified by categories only, then we would have non-modal multi-valued data, e.g., energy Y_5 for owner occupied households in the first county may have been recorded simply as {gas, electric}. In this case, any subsequent analysis would assume that gas and electric occurred with equal probability, to give the realization $\{\text{gas}, 0.5; \text{electric}, 0.5; \text{wood}, 0; \text{oil}, 0; \text{other}, 0\}$.

Let us now consider the quantitative variable Y_6 = driving distance to work. The histograms obtained by aggregating across all households for each county are shown in Table 2.9. Notice in particular that the numbers of histogram sub-intervals s_{u6} differ for each county u: here, $s_{16} = 3$, and $s_{26} = 2$. Notice also that within each histogram, the sub-intervals are not necessarily of equal length: here, e.g., for county $u = 2$, $[a_{261}, b_{261}) = [3, 5)$, whereas $[a_{262}, b_{262}) = [6, 7]$. Across counties, histograms do not necessarily have the same sub-intervals: here, e.g., $[a_{161}, b_{161}) = [1, 5)$, whereas $[a_{261}, b_{261}) = [3, 5)$.

Table 2.9 Aggregated households (Example 2.9)

Class	Y_6 = Distance to work
County 1	$\{[1,5), 0.042; [6,10), 0.583; [11,15], 0.375\}$
Owner	$\{[1,5), 0.0625; [6,8), 0.3125; [9,12), 0.5625, [13,15], 0.0625\}$
Renter	$\{[8,9), 0.375; [10,12], 0.625\}$
County 2	$\{[3,5), 0.654; [6,7], 0.346\}$
Owner	$\{[3,5), 0.714; [6,7], 0.286\}$
Renter	$[4,6]$

The corresponding histograms for county × tenure classes are also shown in Table 2.9. We see that for renters in the second county, this distance is aggregated to give the interval $Y_6 = [4, 6]$, a special case of a histogram. □

Most symbolic data sets will arise from these types of aggregations usually of large data sets but it can be aggregation of smaller data sets. A different situation can arise from some particular scientific question, regardless of the size of the data set. We illustrate this via a question regarding hospitalizations of cardiac patients, described more fully in Quantin *et al.* (2011).

Example 2.10 Cardiologists had long suspected that the survival rate of patients who presented with acute myocardial infarction (AMI) depended on whether or not the patients first went to a cardiology unit and the types of hospital units to which patients were subsequently moved. However, analyses of the raw classical data failed to show this as an important factor to survival. In the Quantin *et al.* (2011) study, patient pathways were established covering a variety of possible pathways. For example, one pathway consisted of admission to one unit (such as intensive care, or cardiology, etc.) before being sent home, while another pathway consisted of admission to an intensive care unit at one hospital, then being moved to a cardiology unit at the same or a different hospital, and then sent home. Each patient could be identified as having followed a specific pathway over the course of treatment, thus the class/group/category was the "pathway." The recorded observed values for a vast array of medical variables were aggregated across those patients within each pathway.

As a simple case, let the data of Table 2.10 be the observed values for Y_1 = age and Y_2 = smoker for eight patients admitted to three different hospitals. The domain of Y_1 is $\mathbb{R}^+$. The smoking multi-valued variable records if the patient does not smoke, is a light smoker, or is a heavy smoker. Suppose the domain

Table 2.10 Cardiac patients (Example 2.8)

Patient	Hospital	Age	Smoker
Patient 1	Hospital 1	74	Heavy
Patient 2	Hospital 1	78	Light
Patient 3	Hospital 2	69	No
Patient 4	Hospital 2	73	Heavy
Patient 5	Hospital 2	80	Light
Patient 6	Hospital 1	70	Heavy
Patient 7	Hospital 1	82	Heavy
Patient 8	Hospital 3	76	No
…	…	…	…

Table 2.11 Hospital pathways (Example 2.8)

Pathway	Age	Smoker
Hospital 1	[70, 82]	{light,1/4; heavy, 3/4}
Hospital 2	[69, 80]	{no, light, heavy}
Hospital 3	[76, 76]	{no}
…	…	…

for Y_2 is written as $\mathcal{Y}_2 = \{\text{no, light, heavy}\}$. Let a pathway be described as a one-step pathway corresponding to a particular hospital, as shown. Thus, for example, four patients collectively constitute the pathway corresponding to the class "Hospital 1"; likewise for the pathways "Hospital 2" and "Hospital 3". Then, observations by pathways are the symbolic data obtained by aggregating classical values for patients who make up a pathway. The age values were aggregated into intervals and the smoking values were aggregated into list values, as shown in Table 2.11. The aggregation of the single patient in the "Hospital 3" pathway ($u = 3$) is the classically valued observation $Y_3 = (Y_{31}, Y_{32}) = ([76, 76], \{no\})$.

The analysis of the Quantin *et al.* (2011) study, based on the pathways symbolic data, showed that pathways were not only important but were in fact the most important factor affecting survival rates, thus corroborating what the cardiologists felt all along. □

There are numerous other situations which perforce are described by symbolic data. Species data are examples of naturally occurring symbolic data. Data with minimum and maximum values, such as the temperature data of

Table 2.4, also occur as a somewhat natural way to record measurements of interest. Many stockmarket values are reported as high and low values daily (or weekly, monthly, annually). Pulse rates may more accurately be recorded as 64 ± 2, i.e., $[62, 66]$ rather than the midpoint value of 64; blood pressure values are notorious for "bouncing around", so that a given value of say 73 for diastolic blood pressure may more accurately be $[70, 80]$. Sensitive census data, such as age, may be given as $[30, 40]$, and so on. There are countless examples.

A question that can arise after aggregation has occurred deals with the handling of outlier values. For example, suppose data aggregated into intervals produced an interval with specific values $\{9, 25, 26, 26.4, 27, 28.1, 29, 30\}$. Or, better yet, suppose there were many many observations between 25 and 30 along with the single value 9. In mathematical terms, our interval, after aggregation, can be formally written as [a,b], where

$$a = \min_{i \in \mathcal{X}} \{x_i\}, \; b = \max_{i \in \mathcal{X}} \{x_i\}, \tag{2.3.1}$$

where $\mathcal{X}$ is the set of all x_i values aggregated into the interval $[a, b]$. In this case, we obtain the interval $[9, 30]$. However, intuitively, we conclude that the value 9 is an outlier and really does not belong to the aggregations in the interval $[25, 30]$. Suppose instead of the value 9, we had a value 21, which, from Eq. (2.3.1), gives the interval $[21, 30]$. Now, it may not be at all clear if the value 21 is an outlier or if it truly belongs to the interval of aggregated values. Since most analyses involving interval data assume that observations within an interval are uniformly spread across that interval, the question becomes one of testing for uniformity across those intervals. Stéphan (1998), Stéphan et al. (2000), and Cariou and Billard (2015) have developed tests of uniformity, gap tests and distance tests, to help address this issue. They also give some reduction algorithms to achieve the deletion of genuine outliers.

2.4 Descriptive Statistics

In this section, basic descriptive statistics, such as sample means, sample variances and covariances, and histograms, for the differing types of symbolic data are briefly described. For quantitative data, these definitions implicitly assume that within each interval, or sub-interval for histogram observations, observations are uniformly spread across that interval. Expressions for the sample mean and sample variance for interval data were first derived by Bertrand and Goupil (2000). Adjustments for non-uniformity can be made. For list multi-valued data, the sample mean and sample variance given herein are simply the respective classical values for the probabilities associated with each of the corresponding categories in the variable domain.

2.4.1 Sample Means

Definition 2.10 Let Y_u, $u = 1, \ldots, m$, be a random sample of size m, with Y_u taking modal list multi-valued values $Y_u = \{Y_{uk}, p_{uk}; k = 1, \ldots, s\}$ from the domain $\mathcal{Y} = \{Y_1, \ldots, Y_s\}$ (as defined in Definition 2.2). Then, the **sample mean** for **list, multi-valued** observations is given by

$$\bar{Y} = \{Y_k, \bar{p}_k; k = 1, \ldots, s\}, \qquad \bar{p}_k = \frac{1}{m}\sum_{u=1}^{m} p_{uk}, \tag{2.4.1}$$

where, without loss of generality, it is assumed that the number of categories from $\mathcal{Y}$ contained in Y_u is $s_u = s$ for all observations by suitably setting $p_{uk} = 0$ where appropriate, and where, for non-modal list observations, it is assumed that each category that occurs has the probability $p_{uk} = 1/m$ and those that do not occur have probability $p_{uk} = 0$ (see section 2.2.2). □

Example 2.11 Consider the deaths attributable to smoking data of Table 2.2. It is easy to show, by applying Eq. (2.4.1), that

$$\bar{Y} = \{\text{smoking}, 0.687; \text{lung cancer}, 0.167; \text{respiratory}, 0.146\},$$

where we have assumed in observation $u = 7$ that the latter two categories have occurred with equal probability, i.e., $p_{72} = p_{72} = 0.176$ (see Example 2.2). □

Definition 2.11 Let Y_u, $u = 1, \ldots, m$, be a random sample of size m, with $Y_u = [a_u, b_u]$ taking interval values (as defined in Definition 2.3). Then, the **interval sample mean** $\bar{Y}$ is given by

$$\bar{Y} = \frac{1}{2m}\sum_{u=1}^{m}(a_u + b_u). \tag{2.4.2}$$

□

Definition 2.12 Let Y_u, $u = 1, \ldots, m$, be a random sample of size m, with Y_u taking histogram values (as defined in Definition 2.4), $Y_u = \{[a_{uk}, b_{uk}), p_{uk}; k = 1, \ldots, s_u\}$, $u = 1, \ldots, m$. Then, the **histogram sample mean** $\bar{Y}$ is

$$\bar{Y} = \frac{1}{2m}\sum_{u=1}^{m}\sum_{k=1}^{s_u}(a_{uk} + b_{uk})p_{uk}. \tag{2.4.3}$$

□

Example 2.12 Take the joggers data of Table 2.5 and Example 2.6. Consider pulse rate Y_1. Applying Eq. (2.4.2) gives

$$\bar{Y}_1 = [(73 + 114) + \cdots + (40 + 60)]/(2 \times 10) = 77.150.$$

Likewise, for the histogram values for running time Y_2, from Eq. (2.4.3), we have

$$\bar{Y}_2 = [\{(5.3+6.2)\times 0.3 + \cdots + (7.1+8.3)\times 0.2\} + \ldots \\ + \{3.2+4.1)\times 0.6 + (4.1+6.7)\times 0.4\}]/(2\times 10) = 5.866.$$

□

2.4.2 Sample Variances

Definition 2.13 Let Y_u, $u = 1,\ldots,m$, be a random sample of size m, with Y_u taking modal list or multi-valued values $Y_u = \{Y_{uk}, p_{uk}, k = 1,\ldots,s\}$ from the domain $\mathcal{Y} = \{Y_1,\ldots,Y_s\}$. Then, the **sample variance** for **list, multi-valued** data is given by

$$S^2 = \{Y_k, S_k^2; k = 1,\ldots,s\}, \qquad S_k^2 = \frac{1}{m-1}\sum_{u=1}^{m}(p_{uk} - \bar{p}_k)^2, \tag{2.4.4}$$

where $\bar{p}_k$ is given in Eq. (2.4.1) and where, as in Definition 2.10, without loss of generality, we assume all possible categories from $\mathcal{Y}$ occur with some probabilities being zero as appropriate. □

Example 2.13 For the smoking deaths data of Table 2.2, by applying Eq. (2.4.4) and using the sample mean $\bar{p} = (0.687, 0.167, 0.146)$ from Example 2.11, we can show that the sample variance S^2 and standard deviation S are, respectively, calculated as

$$S^2 = \{\text{smoking}, 0.0165; \text{lung cancer}, 0.0048; \text{respiratory}, 0.0037\},$$
$$S = \{\text{smoking}, 0.128; \text{lung cancer}, 0.069; \text{respiratory}, 0.061\}.$$

□

Definition 2.14 Let Y_u, $u = 1,\ldots,m$, be a random sample of size m, with $Y_u = [a_u, b_u]$ taking interval values with sample mean $\bar{Y}$ as defined in Eq. (2.4.2). Then, the **interval sample variance** S^2 is given by

$$S^2 = \frac{1}{3m}\sum_{u=1}^{m}[(a_u - \bar{Y})^2 + (a_u - \bar{Y})(b_u - \bar{Y}) + (b_u - \bar{Y})^2]. \tag{2.4.5}$$

□

Let us consider Eq. (2.4.5) more carefully. For these observations, it can be shown that the total sum of squares (SS), Total SS, i.e., mS^2, can be written as

$$mS^2 = \sum_{u=1}^{m}[(a_u + b_u)/2 - \bar{Y}]^2 + \sum_{u=1}^{m}[(a_u - \bar{Y}_u)^2 + (a_u - \bar{Y}_u)(b_u - \bar{Y}_u) \\ + (b_u - \bar{Y}_u)^2]/3, \tag{2.4.6}$$

where $\bar{Y}$ is the overall mean of Eq. (2.4.2), and where the sample mean of the observation Y_u is

$$\bar{Y}_u = (a_u + b_u)/2, \ u = 1, \dots, m. \tag{2.4.7}$$

The term inside the second summation in Eq. (2.4.6) equals S^2 given in Eq. (2.4.5) when $m = 1$. That is, this is a measure of the internal variation, the internal variance, of the single observation Y_u. When summed over all such observations, $u = 1, \dots, m$, we obtain the internal variation of all m observations; we call this the Within SS. To illustrate, suppose we have a single observation $Y = [7, 13]$. Then, substituting into Eq. (2.4.5), we obtain the sample variance as $S^2 = 3 \neq 0$, i.e., interval observations each contain internal variation. The first term in Eq. (2.4.6) is the variation of the interval midpoints across all observations, i.e., the Between SS.

Hence, we can write

$$\text{Total SS} = \text{Between SS} + \text{Within SS} \tag{2.4.8}$$

where

$$\text{Between SS} = \sum_{u=1}^{m} [(a_u + b_u)/2 - \bar{Y}]^2, \tag{2.4.9}$$

$$\text{Within SS} = \sum_{u=1}^{m} [(a_u - \bar{Y}_u)^2 + (a_u - \bar{Y}_u)(b_u - \bar{Y}_u) + (b_u - \bar{Y}_u)^2]/3. \tag{2.4.10}$$

By assuming that values across an interval are uniformly spread across the interval, we see that the Within SS can also be obtained from

$$\text{Within SS} = \sum_{u=1}^{m} (b_u - a_u)^2/12.$$

Therefore, researchers, who upon aggregation of sets of classical data restrict their analyses to the average of the symbolic observation (such as interval means) are discarding important information; they are ignoring the internal variations (i.e., the Within SS) inherent to their data.

When the data are classically valued, with $Y_u = a_u \equiv [a_u, a_u]$, then $\bar{Y}_u = a_u$ and hence the Within SS of Eq. (2.4.10) is zero and the Between SS of Eq. (2.4.9) is the same as the Total SS for classical data. Hence, the sample variance of Eq. (2.4.5) for interval data reduces to its classical counterpart for classical point data, as it should.

Definition 2.15 Let Y_u, $u = 1, \dots, m$, be a random sample of size m, with Y_u taking histogram values, $Y_u = \{[a_{uk}, b_{uk}), p_{uk}; k = 1, \dots, s_u\}$, $u = 1, \dots, m$, and let the sample mean $\bar{Y}$ be as defined in Eq. (2.4.3). Then, the **histogram sample variance** S^2 is

$$S^2 = \frac{1}{3m}\sum_{u=1}^{m}\sum_{k=1}^{s_u}\{[(a_{uk}-\bar{Y})^2 + (a_{uk}-\bar{Y})(b_{uk}-\bar{Y}) + (b_{uk}-\bar{Y})^2]p_{uk}\}. \tag{2.4.11}$$

□

As for intervals, we can show that the total variation for histogram data consists of two parts, as in Eq. (2.4.8), where now its components are given, respectively, by

$$\text{Between SS} = \sum_{u=1}^{m}\sum_{k=1}^{s_u}\{[(a_{uk}+b_{uk})/2 - \bar{Y}]^2 p_{uk}\}, \tag{2.4.12}$$

$$\begin{aligned}\text{Within SS} = \sum_{u=1}^{m}\sum_{k=1}^{s_u}\{[(a_{uk}-\bar{Y}_u)^2 + (a_{uk}-\bar{Y}_u)(b_{uk}-\bar{Y}_u)\\ + (b_{uk}-\bar{Y}_u)^2]p_{uk}\}/3\end{aligned} \tag{2.4.13}$$

with

$$\bar{Y}_u = \sum_{k=1}^{s_u} p_{uk}(a_{uk}+b_{uk})/2,\ u = 1,\dots,m. \tag{2.4.14}$$

It is readily seen that for the special case of interval data, where now $s_u = 1$ and hence $p_{u1} = 1$ for all $u = 1,\dots,m$, the histogram sample variance of Eq. (2.4.11) reduces to the interval sample variance of Eq. (2.4.5).

Example 2.14 Consider the joggers data of Table 2.5. From Example 2.12, we know that the sample means are, respectively, $\bar{Y}_1 = 77.150$ for pulse rate and $\bar{Y}_2 = 5.866$ for running time. Then applying Eqs. (2.4.5) and (2.4.11), respectively, to the interval data for pulse rates and the histogram data for running times, we can show that the sample variances and hence the sample standard deviations are, respectively,

$$S_1^2 = 197.611,\quad S_1 = 14.057;\quad S_2^2 = 1.458,\quad S_2 = 1.207.$$

□

2.4.3 Sample Covariance and Correlation

When the number of variables $p \geq 2$, it is of interest to obtain measures of how these variables depend on each other. One such measure is the covariance. We note that for modal data it is necessary to know the corresponding probabilities for the pairs of each cross-sub-intervals in order to calculate the covariances. This is not an issue for interval data since there is only one possible cross-interval/rectangle for each observation.

Definition 2.16 Let $\mathbf{Y}_u = (Y_1, Y_2)$, $u = 1, \dots, m$, be a random sample of interval data with $Y_{uj} = [a_{uj}, b_{uj}]$, $j = 1, 2$. Then, the **sample covariance** between Y_1 and Y_2, S_{12}, is given by

$$S_{12} = \frac{1}{6m} \sum_{u=1}^{m} [2(a_{u1} - \bar{Y}_1)(a_{u2} - \bar{Y}_2) + (a_{u1} - \bar{Y}_1)(b_{u2} - \bar{Y}_2) + (b_{u1} - \bar{Y}_1)(a_{u2} - \bar{Y}_2) + 2(b_{u1} - \bar{Y}_1)(b_{u2} - \bar{Y}_2)]. \quad (2.4.15)$$

□

As for the variance, we can show that the sum of products (SP) satisfies

$$mS_{12} = \text{Total SP} = \text{Between SP} + \text{Within SP} \quad (2.4.16)$$

where

$$\text{Between SP} = \sum_{u=1}^{m} [(a_{u1} + b_{u1})/2 - \bar{Y}_1][(a_{u2} + b_{u2})/2 - \bar{Y}_2], \quad (2.4.17)$$

$$\text{Within SP} = \frac{1}{6} \sum_{u=1}^{m} [2(a_{u1} - \bar{Y}_{u1})(a_{u2} - \bar{Y}_{u2}) + (a_{u1} - \bar{Y}_{u1})(b_{u2} - \bar{Y}_{u2}) + (b_{u1} - \bar{Y}_{u1})(a_{u2} - \bar{Y}_{u2}) + 2(b_{u1} - \bar{Y}_{u1})(b_{u2} - \bar{Y}_{u2})] \quad (2.4.18)$$

$$= \sum_{u=1}^{m} (b_{u1} - a_{u1})(b_{u2} - a_{u2})/12,$$

$$\bar{Y}_{uj} = \frac{1}{2}(a_{uj} + b_{uj}),\ j = 1, 2, \quad (2.4.19)$$

with $\bar{Y}_j, j = 1, 2$, obtained from Eq. (2.4.2).

Example 2.15 Consider the $m = 6$ minimum and maximum temperature observations for the variables Y_1 = January and Y_2 = July of Table 2.3 (and Example 2.4). From Eq. (2.4.2), we calculate the sample means $\bar{Y}_1 = -0.40$ and $\bar{Y}_2 = 23.09$. Then, from Eq. (2.4.15), we have

$$\begin{aligned} S_{12} = \frac{1}{6 \times 6} & [2(-18.4 - (-0.4))(17.0 - 23.09) + (-18.4 - (-0.4))(26.5 - 23.09) \\ & + (-7.5 - (-0.04))(17.0 - 23.09) + 2((-7.5 - (-0.04))(26.5 - 23.09)] \\ & + \dots \\ & + [2(11.8 - (-0.4))(25.6 - 23.09) + (11.8 - (-0.4))(32.6 - 23.09) \\ & + (19.2 - (-0.04))(25.6 - 23.09) + 2((19.2 - (-0.04))(32.6 - 23.09)] \\ = 69.197. & \end{aligned}$$

We can also calculate the respective standard deviations, $S_1 = 14.469$ and $S_2 = 6.038$, from Eq. (2.4.5). Hence, the correlation coefficient (see Definition 2.18 and Eq. (2.4.24)) is

$$Corr(Y_1, Y_2) = \frac{S_{12}}{S_1 S_2} = \frac{69.197}{14.469 \times 6.038} = 0.792.$$

□

Definition 2.17 Let Y_{uj}, $u = 1, \dots, m, j = 1, 2$, be a random sample of size m, with Y_u taking joint histogram values, $(Y_{u1}, Y_{u2}) = \{[a_{u1k_1}, b_{u1k_1}), [a_{u2k_2}, b_{u2k_2}), p_{uk_1k_2}; k_j = 1, \dots, s_{uj}, j = 1, 2\}, u = 1, \dots, m$, where $p_{uk_1k_2}$ is the relative frequency associated with the rectangle $[a_{u1k_1}, b_{u1k_1}) \times [a_{u2k_2}, b_{u2k_2})$, and let the sample means $\bar{Y}_j$ be as defined in Eq. (2.4.3) for $j = 1, 2$. Then, the **histogram sample covariance**, $\mathrm{Cov}(Y_1, Y_2) = S_{12}$, is

$$\begin{aligned} S_{12} = \frac{1}{6m} \sum_{u=1}^{m} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} [&2(a_{u1k_1} - \bar{Y}_1)(a_{u2k_2} - \bar{Y}_2) + (a_{u1k_1} - \bar{Y}_1)(b_{u2k_2} - \bar{Y}_2) \\ &+ (b_{u1k_1} - \bar{Y}_1)(a_{u2k_2} - \bar{Y}_2) + 2(b_{u1k_1} - \bar{Y}_1)(b_{u2k_2} - \bar{Y}_2)] p_{uk_1k_2}. \end{aligned} \tag{2.4.20}$$

□

As for the variance, we can show that Eq. (2.4.16) holds where now

$$\begin{aligned} \text{Between SP} = \sum_{u=1}^{m} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} &[(a_{u1k_1} + b_{u1k_1})/2 - \bar{Y}_1] \\ &\times [(a_{u2k_2} + b_{u2k_2})/2 - \bar{Y}_2] p_{uk_1k_2}, \end{aligned} \tag{2.4.21}$$

$$\begin{aligned} \text{Within SP} &= \frac{1}{6} \sum_{u=1}^{m} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} [2(a_{u1k_1} - \bar{Y}_{u1})(a_{u2k_2} - \bar{Y}_{u2}) \\ &\quad + (a_{u1k_1} - \bar{Y}_{u1})(b_{u2k_2} - \bar{Y}_{u2}) + (b_{u1k_1} - \bar{Y}_{u1})(a_{u2k_2} - \bar{Y}_{u2}) \\ &\quad + 2(b_{u1k_1} - \bar{Y}_{u1})(b_{u2k_2} - \bar{Y}_{u2})] p_{uk_1k_2} \qquad (2.4.22) \\ &= \sum_{u=1}^{m} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} (b_{u1k_1} - a_{u1k_1})(b_{u2k_2} - a_{u2k_2}) p_{uk_1k_2}/12, \end{aligned}$$

$$\bar{Y}_{uj} = \frac{1}{2} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} (a_{ujk_j} + b_{ujk_j}) p_{uk_1k_2},\ j = 1, 2, \tag{2.4.23}$$

with $\bar{Y}_j, j = 1, 2$, obtained from Eq. (2.4.3).

Definition 2.18 The Pearson (1895) product-moment **correlation coefficient** between two variables Y_1 and Y_2, $r_{sym}(Y_1, Y_2)$, for symbolic-valued observations is given by

$$r_{sym}(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{S_{Y_1} S_{Y_2}}. \tag{2.4.24}$$

□

Example 2.16 Table 2.12 gives the joint histogram observations for the random variables Y_1 = flight time (AirTime) and Y_2 = arrival delay time (ArrDelay) in minutes for airlines traveling into a major airport hub. The original values were aggregated across airline carriers into the histograms shown in these tables. (The data of Table 2.4 and Example 2.5 deal only with the single variable Y_1= flight time. Here, we need the joint probabilities $p_{uk_1k_2}$.)

Let us take the first $m = 5$ airlines only. Then, from Eq. (2.4.3), we have the sample means as $\bar{Y}_1 = 36.448$ and $\bar{Y}_2 = 3.384$. From Eq. (2.4.20), the covariance function, $\text{Cov}(Y_1, Y_2)$, is calculated as

$$\begin{aligned} S_{12} = \frac{1}{6 \times 5} \{ & [2(25 - 36.448)(-40 - 3.384) + (25 - 36.448)(-20 - 3.384) \\ & + (50 - 36.448)(-40 - 3.384) + 2(50 - 36.448)(-20 - 3.384)]0.0246 \\ & + \dots \\ & + [2(100 - 36.448)(35 - 3.384) + (100 - 36.448)(60 - 3.384) \\ & + (120 - 36.448)(35 - 3.384) + 2(120 - 36.448)(60 - 3.384)]0.0056\} \\ = {} & 119.524. \end{aligned}$$

Likewise, from Eq. (2.4.11), the sample variances for Y_1 and Y_2 are, respectively, $S_1^2 = 1166.4738$ and $S_s^2 = 280.9856$; hence, the standard deviations are, respectively, $S_1 = 34.154$ and $S_2 = 16.763$. Therefore, the sample correlation function, $\text{Corr}(Y_1, Y_2)$, is

$$\text{Corr}(Y_1, Y_2) = \frac{S_{12}}{S_1 S_2} = \frac{119.524}{34.154 \times 16.763} = 0.209.$$

Similarly, covariances and hence correlation functions for the variable pairs (Y_1, Y_3) and (Y_2, Y_3), where Y_3 = departure delay time (DepDelay) in minutes, can be obtained from Table 2.13 and Table 2.14, respectively, and are left to the reader. □

2.4.4 Histograms

Brief descriptions of the construction of a histogram based on interval data and on histogram data, respectively, are presented here. More complete details and examples can be found in Billard and Diday (2006a).

Table 2.12 Airlines joint histogram (Y_1, Y_2) (Example 2.16). Y_1 = flight time in minutes, Y_2 = arrival delay time in minutes

$u = 1$			$u = 2$			$u = 3$			$u = 4$			$u = 5$		
$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$
[25, 50)	[−40, −20)	0.0246	[10, 50)	[−30, −10)	0.0113	[10, 50)	[−50, −20)	0.0143	[20, 35)	[−35, −15)	0.0062	[20, 40)	[−30, −15)	0.0808
	[−20, 0)	0.1068		[−10, 10)	0.0676		[−20, 0)	0.0297		[−15, 10)	0.0412		[−15, 5)	0.0874
	[0, 25)	0.0867		[10, 30)	0.0218		[0, 30)	0.0132		[10, 35)	0.0075		[5, 35)	0.0220
	[25, 50)	0.0293		[30, 50]	0.0166		[30, 80]	0.0116		[35, 60]	0.0106		[35, 60]	0.0080
	[50, 75]	0.0328	[50, 90)	[−30, −10)	0.0689	[50, 90)	[−50, −20)	0.0388	[35, 50)	[−35, −15)	0.0301	[40, 60)	[−30, −15)	0.0714
[50, 75)	[−40, −20)	0.0215		[−10, 10)	0.2293		[−20, 0)	0.1725		[−15, 10)	0.1950		[−15, 5)	0.2836
	[−20, 0)	0.1013		[10, 30)	0.0976		[0, 30)	0.1047		[10, 35)	0.0674		[5, 35)	0.0933
	[0, 25)	0.0921		[30, 50]	0.0802		[30, 80]	0.0521		[35, 60]	0.0443		[35, 60]	0.0255
	[25, 50)	0.0398	[90,130)	[−30, −10)	0.0336	[90,130)	[−50, −20)	0.0535	[50, 65)	[−35, −15)	0.0182	[60, 80)	[−30, −15)	0.0172
	[50, 75]	0.0463		[−10, 10)	0.1011		[−20, 0)	0.1943		[−15, 10)	0.1503		[−15, 5)	0.0747
[75,100)	[−40, −20)	0.0070		[10, 30)	0.0562		[0, 30)	0.1726		[10, 35)	0.0700		[5, 35)	0.0390
	[−20, 0)	0.0677		[30, 50]	0.0449		[30, 80]	0.0941		[35, 60]	0.0430		[35, 60]	0.0130
	[0, 25)	0.0925	[130,170]	[−30, −10)	0.0344	[130,170]	[−50, −20)	0.0023	[65, 80)	[−35, −15)	0.0306	[80,100)	[−30, −15)	0.0288
	[25, 50)	0.0377		[−10, 10)	0.0711		[−20, 0)	0.0097		[−15, 10)	0.1126		[−15, 5)	0.0626
	[50, 75]	0.0449		[10, 30)	0.0418		[0, 30)	0.0186		[10, 35)	0.0381		[5, 35)	0.0314
[100,125]	[−40, −20)	0.0123		[30, 50]	0.0235		[30, 80]	0.0181		[35, 60]	0.0270		[35, 60]	0.0083
	[−20, 0)	0.0420							[80, 95]	[−35, −15)	0.0027	[100,120]	[−30, −15)	0.0045
	[0, 25)	0.0558								[−15, 10)	0.0585		[−15, 5)	0.0210
	[25, 50)	0.0258								[10, 35)	0.0301		[5, 35)	0.0217
	[50, 75]	0.0330								[35, 60]	0.0164		[35, 60]	0.0057

(Continued)

Table 2.12 (Continued)

$u=6$			$u=7$			$u=8$			$u=9$			$u=10$		
$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$
[40,100)	[−50, −20)	0.0079	[40, 90)	[−40, −20)	0.0011	[35, 70)	[−35, −15)	0.0152	[20, 40)	[−45, −20)	0.0113	[200,250)	[−50, −30)	0.0120
	[−20, 10)	0.0421		[−20, 0)	0.0675		[−15, 5)	0.0675		[−20, 0)	0.0249		[−30, 0)	0.0737
	[10, 40)	0.0147		[0, 20)	0.0660		[5, 25)	0.0262		[0, 30)	0.0117		[0, 30)	0.0556
	[40, 80)	0.0095		[20, 40]	0.0363		[25, 60]	0.0193		[30, 60]	0.0120		[30, 60]	0.0226
	[80,130]	0.0076	[90,140)	[−40, −20)	0.0119	[70,135)	[−35, −15)	0.0179	[40, 60)	[−45, −20)	0.0652	[250,300)	[−50, −30)	0.0226
[100,180)	[−50, −20)	0.0293		[−20, 0)	0.1259		[−15, 5)	0.0537		[−20, 0)	0.2258		[−30, 0)	0.1368
	[−20, 10)	0.2793		[0, 20)	0.1008		[5, 25)	0.0207		[0, 30)	0.0964		[0, 30)	0.1729
	[10, 40)	0.1216		[20, 40]	0.0463		[25, 60]	0.0220		[30, 60]	0.0709		[30, 60]	0.0632
	[40, 80)	0.0568	[140,190)	[−40, −20)	0.0197	[135,195)	[−35, −15)	0.0758	[60, 80)	[−45, −20)	0.0180	[300,350)	[−50, −30)	0.0135
	[80,130]	0.0435		[−20, 0	0.1173		[−15, 5)	0.1860		[−20, 0)	0.1096		[−30, 0)	0.1398
[180,240)	[−50, −20)	0.0095		[0, 20)	0.1251		[5, 25)	0.0978		[0, 30)	0.0791		[0, 30)	0.1474
	[−20, 10)	0.0798		[20, 40]	0.0886		[25, 60]	0.0647		[30, 60]	0.0520		[30, 60]	0.0391
	[10, 40)	0.0473	[190,240)	[−40, −20)	0.0019	[195,255]	[−35, −15)	0.0468	[80,100)	[−45, −20)	0.0050	[350,400]	[−30, 0)	0.0195
	[40, 80)	0.0189		[−20, 0)	0.0038		[−15, 5)	0.1708		[−20, 0)	0.0239		[0, 30)	0.0481
	[80,130]	0.0161		[0, 20)	0.0048		[5, 25)	0.0826		[0, 30)	0.0472		[30, 60]	0.0331
[240,320)	[−50, −20)	0.0202		[20, 40]	0.0119		[25, 60]	0.0331		[30, 60]	0.0413			
	[−20, 10)	0.0656	[240,290]	[−40, −20)	0.0314				[100,120]	[−45, −20)	0.0050			
	[10, 40)	0.0211		[−20, 0)	0.0868					[−20, 0)	0.0195			
	[40, 80)	0.0065		[0, 20)	0.0400					[0, 30)	0.0378			
	[80,130]	0.0046		[20, 40]	0.0130					[30, 60]	0.0435			
[320,380]	[−50, −20)	0.0082												
	[−20, 10)	0.0446												
	[10, 40)	0.0314												
	[40, 80)	0.0077												
	[80,130]	0.0063												

Table 2.13 Airlines joint histogram (Y_1, Y_3) (Example 2.16). Y_1 = flight time in minutes, Y_3 = departure delay time in minutes

$u = 1$			$u = 2$			$u = 3$			$u = 4$			$u = 5$		
$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$
[25, 50)	[−20, −5)	0.0835	[10, 50)	[−20, −5)	0.0227	[10, 50)	[−30, −5)	0.0165	[20, 35)	[−18, −5)	0.0106	[20, 40)	[−20, −5)	0.0636
	[−5, 0)	0.0767		[−5, 0)	0.0532		[−5, 0)	0.0114		[−5, 0)	0.0319		[−5, 0)	0.0924
	[0, 10)	0.0381		[0, 20)	0.0292		[0, 15)	0.0242		[0, 20)	0.0111		[0, 60]	0.0423
	[10, 50)	0.0547		[20, 50]	0.0122		[15, 50)	0.0089		[20,120]	0.0120	[40, 60)	[−20, −5)	0.0978
	[50,150]	0.0273	[50, 90)	[−20, −5)	0.1238		[50,150]	0.0079	[35, 50)	[−18, −5)	0.0550		[−5, 0)	0.2361
[50, 75)	[−20, −5)	0.0703		[−5, 0)	0.1844	[50, 90)	[−30, −5)	0.0410		[−5, 0)	0.1543		[0, 60]	0.1399
	[−5, 0)	0.0789		[0, 20)	0.1024		[−5, 0)	0.0896		[0, 20)	0.0718	[60, 80)	[−20, −5)	0.0312
	[0, 10)	0.0459		[20, 50]	0.0654		[0, 15)	0.1572		[20,120]	0.0559		[−5, 0)	0.0527
	[10, 50)	0.0681	[90,130)	[−20, −5)	0.0436		[15, 50)	0.0454	[50, 65)	[−18, −5)	0.0417		[0, 60]	0.0600
	[50,150]	0.0377		[−5, 0)	0.0889		[50,150]	0.0347		[−5, 0)	0.1246	[80,100)	[−20, −5)	0.0255
[75,100)	[−20, −5)	0.0525		[0, 20)	0.0628	[90,130)	[−30, −5)	0.0583		[0, 20)	0.0705		[−5, 0)	0.0546
	[−5, 0)	0.0550		[20, 50]	0.0405		[−5, 0)	0.0896		[20,120]	0.0448		[0, 60]	0.0510
	[0, 10)	0.0472	[130,170]	[−20, −5)	0.0471		[0, 15)	0.2377	[65, 80)	[−18, −5)	0.0470	[100,120]	[−20, −5)	0.0116
	[10, 50)	0.0595		[−5, 0)	0.0615		[15, 50)	0.0753		[−5, 0)	0.0971		[−5, 0)	0.0201
	[50,150]	0.0357		[0, 20)	0.0436		[50,150]	0.0535		[0, 20)	0.0368		[0, 60]	0.0213
[100,125]	[−20, −5)	0.0451		[20, 50]	0.0187	[130,170]	[−30, −5)	0.0066		[20,120]	0.0275			
	[−5, 0)	0.0377					[−5, 0)	0.0068	[80, 95]	[−18, −5)	0.0230			
	[0, 10)	0.0297					[0, 15)	0.0181		[−5, 0)	0.0505			
	[10, 50)	0.0316					[15, 50)	0.0088		[0, 20)	0.0230			
	[50,150]	0.0248					[50,150]	0.0084		[20,120]	0.0111			

(Continued)

Table 2.13 (Continued)

$u=6$			$u=7$			$u=8$			$u=9$			$u=10$		
$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$
[40,100)	[−20, −5)	0.0115	[40, 90)	[−20, −5)	0.0214	[35, 70)	[−15, −5)	0.0496	[20, 40)	[−20, −5)	0.0072	[200,250)	[−16, −5)	0.0135
	[−5, 0)	0.0227		[−5, 0)	0.0394		[−5, 0)	0.0551		[−5, 0)	0.0110		[−5, 0)	0.0526
	[0, 10)	0.0151		[0, 10)	0.0499		[0, 40]	0.0234		[0, 10)	0.0154		[0, 10)	0.0511
	[10, 50)	0.0188		[10, 50)	0.0534	[70,135)	[−15, −5)	0.0413		[10, 50)	0.0186		[10, 30]	0.0466
	[50,150]	0.0136		[50,100]	0.0069		[−5, 0)	0.0427		[50,100]	0.0076	[250,300)	[−16, −5)	0.0662
[100,180)	[−20, −5)	0.0714	[90,140)	[−20, −5)	0.0252		[0, 40]	0.0303	[40, 60)	[−20, −5)	0.0976		[−5, 0)	0.1368
	[−5, 0)	[0.1640		[−5, 0)	0.0648	[135,195)	[−15, −5)	0.0661		[−5, 0)	0.1411		[0, 10)	0.1143
	[0, 10)	0.1195		[0, 10)	0.0884		[−5, 0)	0.1887		[0, 10)	0.0951		[10, 30]	0.0782
	[10, 50)	0.1121		[10, 50)	0.0905		[0, 40]	0.1694		[10, 50)	0.0784	[300,350)	[−16, −5)	0.0541
	[50,150]	0.0634		[50,100]	0.0159	[195,255]	[−15, −5)	0.1019		[50,100]	0.0460		[−5, 0)	0.0887
[180,240)	[−20, −5)	0.0194	[140,190)	[−20, −5)	0.0358		[−5, 0)	0.1639	[60, 80)	[−20, −5)	0.0611		[0, 10)	0.1158
	[−5, 0)	0.0435		[−5, 0)	0.0897		[0, 40]	0.0675		[−5, 0)	0.0721		[10, 30]	0.0812
	[0, 10)	0.0468		[0, 10)	0.1167					[0, 10)	0.0444	[350,400]	[−16, −5)	0.0256
	[10, 50)	0.0377		[10, 50)	0.0947					[10, 50)	0.0450		[−5, 0)	0.0361
	[50,150]	0.0241		[50,100]	0.0138					[50,100]	0.0359		[0, 10)	0.0271
[240,320)	[−20, −5)	0.0216	[190,240)	[−20, −5)	0.0019				[80,100)	[−20, −5)	0.0148		[10, 30]	0.0120
	[−5, 0)	0.0432		[−5, 0)	0.0044					[−5, 0)	0.0202			
	[0, 10)	0.0278		[0, 10)	0.0090					[0, 10)	0.0302			
	[10, 50)	0.0170		[10, 50)	0.0063					[10, 50)	0.0265			
	[50,150]	0.0084		[50,100]	0.0008					[50,100]	0.0258			
[320,380]	[−20, −5)	0.0104	[240,290]	[−20, −5)	0.0289				[100,120]	[−20, −5)	0.0192			
	[−5, 0)	0.0292		[−5, 0)	0.0633					[−5, 0)	0.0183			
	[0, 10)	0.0295		[0, 10)	0.0413					[0, 10)	0.0195			
	[10, 50)	0.0210		[10, 50)	0.0314					[10, 50)	0.0214			
	[50,150]	0.0082		[50,100]	0.0063					[50,100]	0.0274			

Table 2.14 Airlines joint histogram (Y_2, Y_3) (Example 2.16). Y_2 = arrival delay time in minutes, Y_3 = departure delay time in minutes

$u = 1$			$u = 2$			$u = 3$			$u = 4$			$u = 5$		
$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$
[−40, −20)	[−20, −5)	0.0484	[−30, −10)	[−20, −5)	0.0667	[−50, −20)	[−30, −5)	0.0469	[−35, −15)	[−18, −5)	0.0434	[−30, −15)	[−20, −5)	0.1011
	[−5, 0)	0.0139		[−5, 0)	0.0693		[−5, 0)	0.0343		[−5, 0]	0.0443		[−5, 0)	0.0912
	[0, 10]	0.0031		[0, 20]	0.0122		[0, 15)	0.0275	[−15, 10)	[−18, −5)	0.1157		[0, 60]	0.0104
[−20, 0)	[−20, −5)	0.1349	[−10, 10)	[−20, −5)	0.1369		[15, 50]	0.0002		[−5, 0)	0.3409	[−15, 5)	[−20, −5)	0.1101
	[−5, 0)	0.1277		[−5, 0)	0.2345	[−20, 0)	[−30, −5)	0.0606		[0, 20)	0.1006		[−5, 0)	0.2935
	[0, 10)	0.0476		[0, 20)	0.0968		[−5, 0)	0.1156		[20,120]	0.0004		[0, 60]	0.1257
	[10, 50]	0.0076		[20, 50]	0.0009		[0, 15)	0.2191	[10, 35)	[−18, −5)	0.0155	[5, 35)	[−20, −5)	0.0163
[0, 25)	[−20, −5)	0.0601	[10, 30)	[−20, −5)	0.0283		[15, 50]	0.0107		[−5, 0)	0.0665		[−5, 0)	0.0640
	[−5, 0)	0.0898		[−5, 0)	0.0702	[0, 30)	[−30, −5)	0.0129		[0, 20)	0.0962		[0, 60]	0.1271
	[0, 10)	0.0865		[0, 20)	0.0920		[−5, 0)	0.0447		[20,120]	0.0350	[35, 60]	[−20, −5)	0.0021
	[10, 50)	0.0904		[20, 50]	0.0270		[0, 15)	0.1714	[35, 60]	[−18, −5)	0.0027		[−5, 0)	0.0071
	[50,150]	0.0002	[30, 50]	[−20, −5)	0.0052		[15, 50)	0.0787		[−5, 0)	0.0066		[0, 60]	0.0513
[25, 50)	[−20, −5)	0.0064		[−5, 0)	0.0139		[50,150]	0.0014		[0, 20)	0.0164			
	[−5, 0)	0.0129		[0, 20)	0.0371	[30, 80]	[−30, −5)	0.0020		[20,120]	0.1157			
	[0, 10)	0.0197		[20, 50]	0.1090		[−5, 0)	0.0029						
	[10, 50)	0.0824					[0, 15)	0.0191						
	[50,150]	0.0113					[15, 50)	0.0488						
[50, 75]	[−20, −5)	0.0016					[50,150]	0.1030						
	[−5, 0)	0.0039												
	[0, 10)	0.0041												
	[10, 50)	0.0334												
	[50,150]	0.1140												

(Continued)

Table 2.14 (Continued)

$u = 6$			$u = 7$			$u = 8$			$u = 9$			$u = 10$		
$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$
[−50, −20)	[−20, −5)	0.0267	[−40, −20)	[−20, −5)	0.0220	[−35, −15)	[−15, −5)	0.0689	[−45, −20)	[−20, −5)	0.0652	[−50, −30)	[−16, −5)	0.0135
	[−5, 0)	0.0331		[−5, 0)	0.0314		[−5, 0)	0.0813		[−5, 0)	0.0318		[−5, 0)	0.0165
	[0, 10)	0.0145		[0, 10)	0.0107		[0, 40]	0.0055		[0, 10]	0.0076		[0, 10)	0.0165
	[10, 50]	0.0008		[10, 50]	0.0019	[−15, 5)	[−15, −5)	0.1377	[−20, 0)	[−20, −5)	0.1090		[10, 30]	0.0015
[−20, 10)	[−20, −5)	0.0904	[−20, 0)	[−20, −5)	0.0691		[−5, 0)	0.2562		[−5, 0)	0.1729	[−30, 0)	[−16, −5)	0.0917
	[−5, 0)	0.2129		[−5, 0)	0.1601		[0, 40]	0.0840		[0, 10)	0.1043		[−5, 0)	0.1398
	[0, 10)	0.1545		[0, 10)	0.1446	[5, 25)	[−15, −5)	0.0510		[10, 50]	0.0176		[0, 10)	0.1023
	[10, 50)	0.0533		[10, 50]	0.0275		[−5, 0)	0.0978	[0, 30)	[−20, −5)	0.0230		[10, 30]	0.0361
	[50,150]	0.0003	[0, 20)	[−20, −5)	0.0199		[0, 40]	0.0785		[−5, 0)	0.0523	[0, 30)	[−16, −5)	0.0496
[10, 40)	[−20, −5)	0.0147		[−5, 0)	0.0627	[25, 60]	[−15, −5)	0.0014		[0, 10)	0.0828		[−5, 0)	0.1353
	[−5, 0)	0.0486		[0, 10)	0.1224		[−5, 0)	0.0152		[10, 50)	0.1128		[0, 10)	0.1564
	[0, 10)	0.0604		[10, 50)	0.1310		[0, 40]	0.1226		[50,100]	0.0013		[10, 30]	0.0827
	[10, 50)	0.1069		[50,100]	0.0006				[30, 60]	[−20, −5)	0.0028	[30, 60]	[−16, −5)	0.0045
	[50,150]	0.0055	[20, 40]	[−20, −5)	0.0023					[−5, 0)	0.0057		[−5, 0)	0.0226
[40, 80)	[−20, −5)	0.0025		[−5, 0)	0.0075					[0, 10)	0.0101		[0, 10)	0.0331
	[−5, 0)	0.0068		[0, 10)	0.0275					[10, 50)	0.0595		[10, 30]	0.0977
	[0, 10)	0.0074		[10, 50)	0.1157					[50,100]	0.1414			
	[10, 50)	0.0378		[50,100]	0.0430									
	[50,150]	0.0448												
[80,130]	[−20, −5)	0.0002												
	[−5, 0)	0.0013												
	[0, 10)	0.0019												
	[10, 50)	0.0077												
	[50,150]	0.0670												

Definition 2.19 Let Y_u, $u = 1, \dots, m$, be a random sample of interval observations with $Y_u = [a_u, b_u] \equiv Z(u)$. Then the **histogram** of these data is the set $\{(p_g, I_g), g = 1, \dots, r\}$, where $I_g = [a_{hg}, b_{hg})$, $g = 1, \dots, r-1$, and $[a_{hr}, b_{hr}]$ are the histogram sub-intervals, and p_g is the relative frequency for the sub-interval I_g, with

$$p_g = f_g/m, \quad f_g = \sum_{u=1}^{m} \frac{||Z(u) \cap I_g||}{||Z(u)||}, \tag{2.4.25}$$

where $||A||$ is the length of A. □

Example 2.17 Consider the Y_1 = pulse rate interval data of Table 2.5 considered in Example 2.6. Then, a histogram of these data is, from Eq. (2.4.25),

$$\{[40,55), 0.075; [55,70), 0.191; [70,85), 0.451; [85,100), 0.236; [100,115], 0.047\}.$$

□

Definition 2.20 Let Y_u, $u = 1, \dots, m$, be a random sample of histogram observations with $Y_u = \{[a_{uk}, b_{uk}), p_{uk}; k = 1, \dots, s_u\}$. Then the **histogram** of these data is the set $\{(p_g, I_g), g = 1, \dots, r\}$, where $I_g = [a_{hg}, b_{hg}), g = 1, \dots, r-1$, and $[a_{hr}, b_{hr}]$ are the histogram sub-intervals, and p_g is the relative frequency for the sub-interval I_g, with

$$p_g = f_g/m, \quad f_g = \sum_{u=1}^{m} \sum_{k=1}^{s_u} \frac{||Z(k;u) \cap I_g||}{||Z(k;u)||} p_{uk}, \tag{2.4.26}$$

where $Z(k;u) = [a_{uk}, b_{uk})$ is the kth, $k = 1, \dots, s_u$, sub-interval of the observation Y_u. □

Example 2.18 Consider the Y_2 = running time histogram data of Table 2.5 considered in Example 2.6. We can show that, from Eq. (2.4.26), a histogram of these histogram observations is

$$\{[1.5, 3.5), 0.042; [3.5, 5.5), 0.281; [5.5, 7.5), 0.596; [7.5, 9.5], 0.081\}.$$

□

2.5 Other Issues

There are very few theoretical results underpinning the methodologies pertaining to symbolic data. Some theory justifying the weights associated with modal valued observations, such as capacities, credibilities, necessities, possibilities, and probabilities (briefly described in Definitions 2.6–2.9), can be found in Diday (1995) and Diday and Emilion (2003). These concepts include the union and interception probabilities of Chapter 4, and are the only choice which gives Galois field sets. Their results embody *Choquet* (1954) capacities. Other Galois field theory supporting classification and clustering ideas can be found in Brito and Polaillon (2005).

The descriptive statistics described in section 2.4 are empirically based and are usually moment estimators for the underlying means, variances, and covariances. Le-Rademacher and Billard (2011) have shown that these estimators for the mean and variance of interval data in Eqs. (2.4.2) and (2.4.5), respectively, are the maximum likelihood estimators under reasonable distributional assumptions; likewise, Xu (2010) has shown the moment estimator for the covariance in Eq. (2.4.15) is also the maximum likelihood estimator. These derivations involve separating out the overall distribution from the internal distribution within the intervals, and then invoking conditional moment theory in conjunction with standard maximum likelihood theory. The current work assumed the overall distribution to be a normal distribution with the internal variations following appropriately defined conjugate distributions. Implicit in the formulation of these estimators for the mean, variance, and covariance is the assumption that the points inside a given interval are uniformly spread across the intervals. Clearly, this uniformity assumption can be changed. Le-Rademacher and Billard (2011), Billard (2008), Xu (2010), and Billard et al. (2016) discuss how these changes can be effected, illustrating with an internal triangular distribution. There is a lot of foundational work that still needs to be done here.

By and large, however, methodologies and statistics seem to be intuitively correct when they correspond to their classical counterparts. However, to date, they are not generally rigorously justified theoretically. One governing validity criterion is that, crucially and most importantly, methods developed for symbolic data must produce the corresponding classical results when applied to the special case of classical data.

Exercises

2.1 Show that the histogram data of Table 2.4 for Y = flight time have the sample statistics $\bar{Y} = 122.216$, $S_Y^2 = 7309.588$ and $S_Y = 85.486$.

2.2 Refer to Example 2.16 and use the data of Tables 2.12–2.14 for all $m = 10$ airlines for Y_1 = AirTime, Y_2 = ArrDelay, and Y_3 = DepDelay. Show that the sample statistics are $\bar{Y}_1 = 122.216$, $\bar{Y}_2 = 6.960$, $\bar{Y}_3 = 9.733$, $S_{Y_1} = 85.496$, $S_{Y_2} = 25.367$, $S_{Y_3} = 26.712$, $\mathrm{Cov}(Y_1, Y_2) = 110.250$, $\mathrm{Cov}(Y_1, Y_3) = -61.523$, $\mathrm{Cov}(Y_2, Y_3) = 466.482$, $\mathrm{Corr}(Y_1, Y_2) = 0.051$, $\mathrm{Corr}(Y_1, Y_3) = -0.027$, $\mathrm{Corr}(Y_2, Y_3) = 0.688$.

2.3 Consider the airline data with joint distributions in Tables 2.15–2.17.

(a) Using these tables, calculate the sample statistics $\bar{Y}_j$, S_j, $\mathrm{Cov}(Y_{j_1}, Y_{j_2})$ and $\mathrm{Corr}(Y_{j_1}, Y_{j_2})$, $j, j_1, j_2 = 1, 2, 3$.

(b) How do the statistics of (a) differ from those of Exercise 2.2, if at all? If there are differences, what do they tell us about different aggregations?

Appendix

Table 2.15 Airlines joint histogram (Y_1, Y_2) (Exercise 2.5.3). Y_1 = flight time in minutes, Y_2 = arrival delay time in minutes

$u=1$			$u=2$			$u=3$			$u=4$			$u=5$		
$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$
[25, 50)	[−40, −20)	0.0246	[10, 50)	[−30, −10)	0.0113	[10, 50)	[−50, −30)	0.0032	[20, 35)	[−35, −15)	0.0062	[20, 40)	[−30, −15)	0.0808
	[−20, 0)	0.1068		[−10, 10)	0.0676		[−30, 0)	0.0408		[−15, 10)	0.0412		[−15, 10)	0.0940
	[0, 25)	0.0867		[10, 30)	0.0218		[0, 30)	0.0132		[10, 35)	0.0075		[10, 35)	0.0154
	[25, 50)	0.0293		[30, 50]	0.0166		[30, 60]	0.0116		[35, 60]	0.0106		[35, 60]	0.0080
	[50, 75]	0.0328	[50, 90)	[−30, −10)	0.0689	[50, 90)	[−50, −30)	0.0088	[35, 50)	[−35, −15)	0.0301	[40, 60)	[−30, −15)	0.0714
[50, 75)	[−40, −20)	0.0215		[−10, 10)	0.2293		[−30, 0)	0.2025		[−15, 10)	0.1950		[−15, 10)	0.3155
	[−20, 0)	0.1013		[10, 30)	0.0976		[0, 30)	0.1047		[10, 35)	0.0674		[10, 35)	0.0614
	[0, 25)	0.0921		[30, 50]	0.0802		[30, 60]	0.0521		[35, 60]	0.0443		[35, 60]	0.0255
	[25, 50)	0.0398	[90,130)	[−30, −10)	0.0336	[90,130)	[−50, −30)	0.0157	[50, 65)	[−35, −15)	0.0182	[60, 80)	[−30, −15)	0.0172
	[50, 75]	0.0463		[−10, 10)	0.1011		[−30, 0)	0.2320		[−15, 10)	0.1503		[−15, 10)	0.0846
[75,100)	[−40, −20)	0.0070		[10, 30)	0.0562		[0, 30)	0.1726		[10, 35)	0.0700		[10, 35)	0.0291
	[−20, 0)	0.0677		[30, 50]	0.0449		[30, 60]	0.0941		[35, 60]	0.0430		[35, 60]	0.0130
	[0, 25)	0.0925	[130,170]	[−30, −10)	0.0344	[130,170]	[−50, −30)	0.0005	[65, 80)	[−35, −15)	0.0306	[80,100)	[−30, −15)	0.0288
	[25, 50)	0.0377		[−10, 10)	0.0711		[−30, 0)	0.0114		[−15, 10)	0.1126		[−15, 10)	0.0709
	[50, 75]	0.0449		[10, 30)	0.0418		[0, 30)	0.0186		[10, 35)	0.0381		[10, 35)	0.0232
[100,125]	[−40, −20)	0.0123		[30, 50]	0.0235		[30, 60]	0.0181		[35, 60]	0.0270		[35, 60]	0.0083
	[−20, 0)	0.0420							[80, 95]	[−35, −15)	0.0027	[100,120]	[−30, −15)	0.0045
	[0, 25)	0.0558								[−15, 10)	0.0585		[−15, 10)	0.0279
	[25, 50)	0.0258								[10, 35)	0.0301		[10, 35)	0.0149
	[50, 75]	0.0330								[35, 60]	0.0164		[35, 60]	0.0057

(Continued)

Table 2.15 (Continued)

$u=6$			$u=7$			$u=8$			$u=9$			$u=10$		
$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$	$[a_{u1k}, b_{u1k})$	$[a_{u2k}, b_{u2k})$	$p_{uk_1k_2}$
[40,100)	[−50, −20)	0.0079	[40, 90)	[−40, −20)	0.0011	[35, 70)	[−35, −15)	0.0193	[20, 40)	[−50, −30)	0.0022	[200,250)	[−50, −30)	0.0120
	[−20, 10)	0.0421		[−20, 0)	0.0675		[−15, 5)	0.0634		[−30, 0)	0.0340		[−30, 0)	0.0737
	[10, 40)	0.0147		[0, 20)	0.0660		[5, 25)	0.0262		[0, 30)	0.0117		[0, 30)	0.0556
	[40, 70)	0.0074		[20, 40]	0.0363		[25, 45)	0.0096		[30, 60]	0.0120		[30, 60]	0.0226
	[70,100)	0.0046	[90,140)	[−40, −20)	0.0119		[45, 65]	0.0096	[40, 60)	[−50, −30)	0.0110	[250,300)	[−50, −30)	0.0226
	[100,130]	0.0050		[−20, 0)	0.1259	[70,135)	[−35, −15)	0.0179		[−30, 0)	0.2800		[−30, 0)	0.1368
[100,180)	[−50, −20)	0.0293		[0, 20)	0.1008		[−15, 5)	0.0537		[0, 30)	0.0964		[0, 30)	0.1729
	[−20, 10)	0.2793		[20, 40]	0.0463		[5, 25)	0.0207		[30, 60]	0.0709		[30, 60]	0.0632
	[10, 40)	0.1216	[140,190)	[−40, −20)	0.0197		[25, 45)	0.0138	[60, 80)	[−50, −30)	0.0013	[300,350)	[−50, −30)	0.0135
	[40, 70)	0.0462		[−20, 0)	0.1173		[45, 65]	0.0083		[−30, 0)	0.1263		[−30, 0)	0.1398
	[70,100)	0.0238		[0, 20)	0.1251	[135,195)	[−35, −15)	0.0813		[0, 30)	0.0791		[0, 30)	0.1474
	[100,130]	0.0303		[20, 40]	0.0886		[−15, 5)	0.1804		[30, 60]	0.0520		[30, 60]	0.0391
[180,240)	[−50, −20)	0.0095	[190,240)	[−40, −20)	0.0019		[5, 25)	0.0978	[80,100)	[−50, −30)	0.0003	[350,400]	[−50, −30)	0.0000
	[−20, 10)	0.0798		[−20, 0)	0.0038		[25, 45)	0.0234		[−30, 0)	0.0287		[−30, 0)	0.0195
	[10, 40)	0.0473		[0, 20)	0.0048		[45, 65]	0.0413		[0, 30)	0.0472		[0, 30)	0.0481
	[40, 70)	0.0156		[20, 40]	0.0119	[195,255)	[−35, −15)	0.0565		[30, 60]	0.0413		[30, 60]	0.0331
	[70,100)	0.0079	[240,290]	[−40, −20)	0.0314		[−15, 5)	0.1612	[100,120]	[−50, −30)	0.0013			
	[100,130]	0.0115		[−20, 0)	0.0868		[5, 25)	0.0826		[−30, 0)	0.0233			
[240,320)	[−50, −20)	0.0202		[0, 20)	0.0400		[25, 45)	0.0207		[0, 30)	0.0378			
	[−20, 10)	0.0656		[20, 40]	0.0130		[45, 65]	0.0124		[30, 60]	0.0435			
	[10, 40)	0.0211												
	[40, 70)	0.0062												
	[70,100)	0.0013												
	[100,130]	0.0036												
[320,380]	[−50, −20)	0.0082												
	[−20, 10)	0.0446												
	[10, 40)	0.0314												
	[40, 70)	0.0065												
	[70,100)	0.0030												
	[100,130]	0.0046												

Table 2.16 Airlines joint histogram (Y_1, Y_3) (Exercise 2.5.3). Y_1 = flight time in minutes, Y_3 = departure delay time in minutes

$u=1$			$u=2$			$u=3$			$u=4$			$u=5$		
$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$
[25, 50)	[−20, −5)	0.0835	[10, 50)	[−20, −5)	0.0227	[10, 50)	[−30, −5)	0.0165	[20, 35)	[−18, −5)	0.0106	[20, 40)	[−20, −5)	0.0636
	[−5, 0)	0.0767		[−5, 0)	0.0532		[−5, 0)	0.0114		[−5, 0)	0.0319		[−5, 0)	0.0924
	[0, 10)	0.0381		[0, 20)	0.0292		[0, 10)	0.0224		[0, 20)	0.0111		[0, 60]	0.0423
	[10, 50)	0.0547		[20, 50]	0.0122		[10, 50)	0.0107		[20,120]	0.0120	[40, 60)	[−20, −5)	0.0978
	[50,150]	0.0273	[50, 90)	[−20, −5)	0.1238		[50,150]	0.0079	[35, 50)	[−18, −5)	0.0550		[−5, 0)	0.2361
[50, 75)	[−20, −5)	0.0703		[−5, 0)	0.1844	[50, 90)	[−30, −5)	0.0410		[−5, 0)	0.1543		[0, 60]	0.1399
	[−5, 0)	0.0789		[0, 20)	0.1024		[−5, 0)	0.0896		[0, 20)	0.0718	[60, 80)	[−20, −5)	0.0312
	[0, 10)	0.0459		[20, 50]	0.0654		[0, 10)	0.1447		[20,120]	0.0559		[−5, 0)	0.0527
	[10, 50)	0.0681	[90,130)	[−20, −5)	0.0436		[10, 50)	0.0580	[50, 65)	[−18, −5)	0.0417		[0, 60]	0.0600
	[50,150]	0.0377		[−5, 0)	0.0889		[50,150]	0.0347		[−5, 0)	0.1246	[80,100)	[−20, −5)	0.0255
[75,100)	[−20, −5)	0.0525		[0, 20)	0.0628	[90,130)	[−30, −5)	0.0583		[0, 20)	0.0705		[−5, 0)	0.0546
	[−5, 0)	0.0550		[20, 50]	0.0405		[−5, 0)	0.0896		[20,120]	0.0448		[0, 60]	0.0510
	[0, 10)	0.0472	[130,170]	[−20, −5)	0.0471		[0, 10)	0.2165	[65, 80)	[−18, −5)	0.0470	[100,120]	[−20, −5)	0.0116
	[10, 50)	0.0595		[−5, 0)	0.0615		[10, 50)	0.0966		[−5, 0)	0.0971		[−5, 0]	0.0201
	[50,150]	0.0357		[0, 20)	0.0436		[50,150]	0.0535		[0, 20)	0.0368		[0, 60]	0.0213
[100,125]	[−20, −5)	0.0451		[20, 50]	0.0187	[130,170]	[−30, −5)	0.0066		[20,120]	0.0275			
	[−5, 0)	0.0377					[−5, 0)	0.0068	[80, 95]	[−18, −5)	0.0230			
	[0, 10)	0.0297					[0, 10)	0.0161		[−5, 0)	0.0505			
	[10, 50)	0.0316					[10, 50)	0.0107		[0, 20)	0.0230			
	[50,150]	0.0248					[50,150]	0.0084		[20,120]	0.0111			

(Continued)

Table 2.16 (Continued)

$u = 6$			$u = 7$			$u = 8$			$u = 9$			$u = 10$		
$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$	$[a_{u1k}, b_{u1k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_1k_3}$
[40,100)	[−18, −5)	0.0115	[40, 90)	[−20, −5)	0.0214	[35, 70)	[−15, −5)	0.0496	[20, 40)	[−20, −5)	0.0072	[200,250)	[−16, −5)	0.0135
	[−5, 0)	0.0227		[−5, 0)	0.0394		[−5, 0)	0.0551		[−5, 0)	0.0110		[−5, 0)	0.0526
	[0, 10)	0.0151		[0, 10)	0.0499		[0, 40]	0.0234		[0, 10)	0.0154		[0, 10)	0.0511
	[10, 50)	0.0188		[10, 50)	0.0534	[70,135)	[−15, −5)	0.0413		[10, 50)	0.0186		[10, 30]	0.0466
	[50,150]	0.0136		[50,100]	0.0069		[−5, 0)	0.0427		[50,100]	0.0076	[250,300)	[−16, −5)	0.0662
[100,180)	[−18, −5)	0.0714	[90,140)	[−20, −5)	0.0252		[0, 40]	0.0303	[40, 60)	[−20, −5)	0.0976		[−5, 0)	0.1368
	[−5, 0)	0.1640		[−5, 0)	0.0648	[135,195)	[−15, −5)	0.0661		[−5, 0)	0.1411		[0, 10)	0.1143
	[0, 10)	0.1195		[0, 10)	0.0884		[−5, 0)	0.1887		[0, 10)	0.0951		[10, 30]	0.0782
	[10, 50)	0.1121		[10, 50)	0.0905		[0, 40]	0.1694		[10, 50)	0.0784	[300,350)	[−16, −5)	0.0541
	[50,150]	0.0634		[50,100]	0.0159	[195,255]	[−15, −5)	0.1019		[50,100]	0.0460		[−5, 0)	0.0887
[180,240)	[−18, −5)	0.0194	[140,190)	[−20, −5)	0.0358		[−5, 0)	0.1639	[60, 80)	[−20, −5)	0.0611		[0, 10)	0.1158
	[−5, 0)	0.0435		[−5, 0)	0.0897		[0, 40]	0.0675		[−5, 0)	0.0721		[10, 30]	0.0812
	[0, 10)	0.0468		[0, 10)	0.1167					[0, 10)	0.0444	[350,400]	[−16, −5)	0.0256
	[10, 50)	0.0377		[10, 50)	0.0947					[10, 50)	0.0450		[−5, 0)	0.0361
	[50,150]	0.0241		[50,100]	0.0138					[50,100]	0.0359		[0, 10)	0.0271
[240,320)	[−18, −5)	0.0216	[190,240)	[−20, −5)	0.0019				[80,100)	[−20, −5)	0.0148		[10, 30]	0.0120
	[−5, 0)	0.0432		[−5, 0)	0.0044					[−5, 0)	0.0202			
	[0, 10)	0.0278		[0, 10)	0.0090					[0, 10)	0.0302			
	[10, 50)	0.0170		[10, 50)	0.0063					[10, 50)	0.0265			
	[50,150]	0.0084		[50,100]	0.0008					[50,100]	0.0258			
[320,380]	[−18, −5)	0.0104	[240,290]	[−20, −5)	0.0289				[100,120]	[−20, −5)	0.0192			
	[−5, 0)	0.0292		[−5, 0)	0.0633					[−5, 0)	0.0183			
	[0, 10)	0.0295		[0, 10)	0.0413					[0, 10)	0.0195			
	[10, 50)	0.0210		[10, 50)	0.0314					[10, 50)	0.0214			
	[50,150]	0.0082		[50,100]	0.0063					[50,100]	0.0274			

Table 2.17 Airlines joint histogram (Y_2, Y_3) (Exercise 2.5.3). Y_2 = arrival delay time in minutes, Y_3 = departure delay time in minutes

$u = 1$			$u = 2$			$u = 3$			$u = 4$			$u = 5$		
$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$
[−40, −20)	[−20, −8)	0.0277	[−30, −5)	[−20, −5)	0.1116	[−50, −20)	[−30, −5)	0.0469	[−35, −15)	[−18, −5)	0.0434	[−30, −15)	[−20, −5)	0.1011
	[−8, −1)	0.0328		[−5, 0)	0.1373		[−5, 0)	0.0343		[−5, 0]	0.0443		[−5, 0)	0.0912
	[−1, 15]	0.0049		[0, 20]	0.0275		[0, 15)	0.0275	[−15, 10)	[−18, −5)	0.1157		[0, 60]	0.0104
[−20, 0)	[−20, −8)	0.0595	[−5, 10)	[−20, −5)	0.0920		[15, 50]	0.0002		[−5, 0)	0.3409	[−15, 5)	[−20, −5)	0.1101
	[−8, −1)	0.1837		[−5, 0)	0.1665	[−20, 0)	[−30, −5)	0.0606		[0, 20)	0.1006		[−5, 0)	0.2935
	[−1, 15)	0.0718		[0, 20)	0.0815		[−5, 0)	0.1156		[20,100]	0.0004		[0, 60]	0.1257
	[15, 60]	0.0029		[20, 50]	0.0009		[0, 15)	0.2191	[10, 35)	[−18, −5)	0.0155	[5, 35)	[−20, −5)	0.0163
[0, 20)	[−20, −8)	0.0273	[10, 30)	[−20, −5)	0.0283		[15, 50]	0.0107		[−5, 0)	0.0665		[−5, 0)	0.0640
	[−8, −1)	0.0960		[−5, 0)	0.0702	[0, 30)	[−30, −5)	0.0129		[0, 20)	0.0962		[0, 60]	0.1271
	[−1, 15)	0.1171		[0, 20)	0.0920		[−5, 0)	0.0447		[20,100]	0.0350	[35, 70]	[−20, −5)	0.0021
	[15, 60]	0.0457		[20, 50]	0.0270		[0, 15)	0.1714	[35, 85]	[−18, −5)	0.0027		[−5, 0)	0.0071
[20, 50)	[−20, −8)	0.0037	[30, 50]	[−20, −5)	0.0052		[15, 50)	0.0787		[−5, 0)	0.0066		[0, 60]	0.0513
	[−8, −1)	0.0207		[−5, 0)	0.0139		[50,150]	0.0014		[0, 20)	0.0164			
	[−1, 15)	0.0459		[0, 20)	0.0371	[30, 80)	[−30, −5)	0.0020		[20,100]	0.1157			
	[15, 60)	0.1007		[20, 50]	0.1090		[−5, 0)	0.0025						
	[60,150]	0.0027					[0, 15)	0.0168						
[50, 75]	[−20, −8)	0.0006					[15, 50)	0.0463						
	[−8, −1)	0.0037					[50,150]	0.0504						
	[−1, 15)	0.0076				[80,150]	[−5, 0)	0.0004						
	[15, 60)	0.0515					[0, 15)	0.0023						
	[60,150]	0.0937					[15, 50)	0.0025						
							[50,150]	0.0526						

(Continued)

Table 2.17 (Continued)

$u = 6$			$u = 7$			$u = 8$			$u = 9$			$u = 10$		
$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$	$[a_{u2k}, b_{u2k})$	$[a_{u3k}, b_{u3k})$	$p_{uk_2k_3}$
[−50, −20)	[−18, −5)	0.0267	[−40, −20)	[−20, −5)	0.0220	[−35, −15)	[−15, −5)	0.0702	[−45, −20)	[−20, −5)	0.0652	[−50, −30)	[−16, −5)	0.0135
	[−5, 0)	0.0331		[−5, 0)	0.0314		[−5, 0)	0.0829		[−5, 0)	0.0318		[−5, 0)	0.0165
	[0, 10)	0.0145		[0, 10)	0.0107		[0, 40]	0.0056		[0, 10]	0.0076		[0, 10)	0.0165
	[10, 50]	0.0008		[10, 50]	0.0019	[−15, 5)	[−15, −5)	0.1320	[−20, 0)	[−20, −5)	0.1090		[10, 30]	0.0015
[−20, 10)	[−18, −5)	0.0904	[−20, 0)	[−20, −5)	0.0691		[−5, 0)	0.2528		[−5, 0)	0.1729	[−30, 0)	[−16, −5)	0.0917
	[−5, 0)	0.2129		[−5, 0)	0.1601		[0, 40]	0.0829		[0, 10)	0.1043		[−5, 0)	0.1398
	[0, 10)	0.1545		[0, 10)	0.1446	[5, 25)	[−15, −5)	0.0520		[10, 50]	0.0176		[0, 10)	0.1023
	[10, 50)	0.0533		[10, 50]	0.0275		[−5, 0)	0.0997	[0, 30)	[−20, −5)	0.0230		[10, 30]	0.0361
	[50,150]	0.0003	[0, 20)	[−20, −5)	0.0199		[0, 40]	0.0801		[−5, 0)	0.0523	[0, 30)	[−16, −5)	0.0496
[10, 40)	[−18, −5)	0.0147		[−5, 0)	0.0627	[25, 60]	[−15, −5)	0.0014		[0, 10)	0.0828		[−5, 0)	0.1353
	[−5, 0)	0.0486		[0, 10)	0.1224		[−5, 0)	0.0154		[10, 50)	0.1128		[0, 10)	0.1564
	[0, 10)	0.0604		[10, 50)	0.1310		[0, 40]	0.1250		[50,100]	0.0013		[10, 30]	0.0827
	[10, 50)	0.1069		[50,100]	0.0006				[30,100]	[−20, −5)	0.0028	[30, 60]	[−16, −5)	0.0045
	[50,150]	0.0055	[20, 40]	[−20, −5)	0.0023					[−5, 0)	0.0057		[−5, 0)	0.0226
[40, 80)	[−18, −5)	0.0025		[−5, 0)	0.0075					[0, 10)	0.0101		[0, 10)	0.0331
	[−5, 0)	0.0068		[0, 10)	0.0275					[10, 50)	0.0595		[10, 30]	0.0977
	[0, 10)	0.0074		[10, 50)	0.1157					[50,100]	0.1414			
	[10, 50)	0.0378		[50,100]	0.0430									
	[50,150]	0.0448												
[80,150]	[−18, −5)	0.0002												
	[−5, 0)	0.0013												
	[0, 10)	0.0019												
	[10, 50)	0.0077												
	[50,150]	0.0670												