Symbolic Data Analysis: Basic Statistics

Lynne Billard

Department of Statistics
University of Georgia
lynne@stat.uga.edu

COMPSTAT - August 2010

Text: Billard and Diday (2006):
Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.
(Equation, table, and figure numbers taken from text.)

Schweizer (1985): "Distributions are the numbers of the future"

Classical Data Value X:
- A single point in p-dimensional space

E.g., $X = 17$, $X = 2.1$, $X =$ blue

Symbolic Data Value Y:
- Hypercube or Cartesian product of distributions in p-dimensional space

I.e. $Y =$ list, interval, modal in structure

Modal data:
Histogram,
empirical distribution function,
probability distribution,
model, ...

Weights:
Relative frequencies
capacities,
credibilities,
necessities,
possibilities, ...

Multi-valued data, Lists

E.g., 1. **Bird Colors** - Y = Color (Table 2.5)

ω_u	Bird	Major Colors
ω_1	Magpie	{black, white}
ω_2	Kookaburra	{brown, black, white, blue}
ω_3	Galah	{pink, grey}
ω_4	Cardinal	{red, black}
ω_5	Goldfinch	{black, yellow}
ω_6	Quetzal	{red, green, white}
ω_7	Toucan	{ black, yellow, red, green}
ω_8	Rainbow Lorikeet	{blue, yellow, green, red, violet, orange}

Here, a magpie eg, can be a single bird,
or, a collection of birds, a species

Multi-valued data, Lists

E.g.,2. **Marital Status:** (Table 2.1a – Medical Dataset)

Y_5 = status, with possible values in $\mathcal{Y}_5 = \{S = \text{single}, M = \text{married}\}$

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
1	Boston	M	24	M	S	2	2	0	165
2	Boston	M	56	M	M	1	2	2	186
3	Chicago	D	48	M	M	1	3	2	175
4	El Paso	M	47	F	M	0	1	1	141
5	Byron	D	79	F	M	0	3	4	152
6	Concord	M	12	M	S	2	1	0	73
7	Atlanta	M	67	F	M	1	6	0	166
8	Boston	O	73	F	M	0	2	4	164
9	Lindfield	D	29	M	M	2	0	2	227
10	Lindfield	D	44	M	M	1	3	3	216
11	Boston	D	54	M	S	1	5	0	213
12	Chicago	M	12	F	S	2	2	0	75
13	Macon	M	73	F	M	0	3	1	152
14	Boston	D	48	M	M	0	2	4	206
15	Peoria	O	79	F	M	0	3	3	153

For a single individual/observation, Y_5 takes one value only (single or married);
e.g., for $i = 1$, Y_5 = single. .... Not a list.

If the observation unit is gender ($\equiv Y_4$), then for these 15 individuals:
For $u = 1$ (men), Y_5 = { single, married}, or {single, 3/8; married 5/8}
For $u = 2$ (women), Y_5 = { single, married}, or {single, 1/7; married 6/7}

Types of Data - Multi-valued data, Lists

A formal definition for a multi-valued variable is:

Definition 2.2: A multi-valued symbolic random variable Y is one whose possible value takes one or more values from the list of values in its domain $\mathcal{Y}$. The complete list of possible values in $\mathcal{Y}$ is finite, and values may be well-defined categorical or quantitative values.

where

Definition 2.1:A categorical variable is one whose values are names; also called qualitative variable. A quantitative variable is one whose values are subsets of the real line $\mathcal{R}^1$. Note however that sometimes qualitative values can be recoded into apparent quantitative values.

Interval-valued data

E.g., 1. Naturally occuring data: **Mushrooms** (from Table 3.3)

ω_u	Species	Pileus Cap Width	Stipe Length	Stipe Thickness	Edibility
ω_1	*arorae*	[3.0, 8.0]	[4.0, 9.0]	[0.50, 2.50]	U
ω_2	*arvenis*	[6.0, 21.0]	[4.0, 14.0]	[1.00, 3.50]	Y
ω_3	*benesi*	[4.0, 8.0]	[5.0, 11.0]	[1.00, 2.00]	Y
ω_4	*bernardii*	[7.0, 6.0]	[4.0, 7.0]	[3.00, 4.50]	Y
ω_5	*bisporus*	[5.0, 12.0]	[2.0, 5.0]	[1.50, 2.50]	Y
ω_6	*bitorquis*	[5.0, 15.0]	[4.0, 10.0]	[2.00, 4.00]	Y
ω_7	*califorinus*	[4.0, 11.0]	[3.0, 7.0]	[0.40, 1.00]	T
ω_8	*campestris*	[5.0, 10.0]	[3.0, 6.0]	[1.00, 2.00]	Y
ω_9	*comtulus*	[2.5, 4.0]	[3.0, 5.0]	[0.40, 0.70]	Y
...	...	...	...	...	...

Unlike the magpie, a unit is a species or collection of mushrooms

Interval-valued data: Credit-card Data (Table 2.3)

i	Name	Month	Food	Social	Travel	Gas	Clothes
1	Jon	February	23.65	14.56	218.02	16.79	45.61
2	Leigh	May	28.47	8.99	141.60	21.74	86.04
3	Leigh	July	30.86	9.55	193.14	24.26	95.68
4	Tom	July	24.13	15.97	190.40	35.71	20.02
5	Jon	April	23.40	11.61	179.38	23.73	48.89
6	Jon	November	23.11	16.71	178.78	20.55	47.96
7	Leigh	September	32.14	12.34	165.65	17.62	66.40
8	Leigh	August	25.92	20.78	201.18	32.97	70.96
9	Leigh	November	31.52	16.62	177.50	20.95	71.18
10	Jon	November	23.11	14.41	179.86	20.53	51.49
11	Jon	November	22.80	11.35	184.55	20.94	50.36
12	Leigh	September	32.83	13.93	158.65	17.04	69.41
13	Leigh	November	31.13	12.82	179.57	20.67	69.01
14	Tom	August	23.01	13.20	220.52	29.44	18.09
15	Jon	December	21.09	9.90	180.66	22.95	47.87
16	Leigh	August	30.90	13.29	202.22	32.29	68.71
17	Leigh	December	37.36	15.63	184.22	20.32	71.74
18	Tom	July	24.25	15.71	149.01	30.68	21.75
19	Tom	April	21.83	14.95	154.43	30.48	21.09
20	Jon	January	25.94	12.38	197.90	20.06	47.09
...	...	...	...	...	...	...	...

For a single individual/observation, e.g.,

once in February Jon spent $Y_1 = 23.65$ on food.

For the unit/category Jon, $\xi = [21.09, 25.94]$.

Likewise,

for the unit/cateogory Tom, $\xi = [29.44, 35.71]$ on gas ($\equiv Y_4$);

for the unit/category Leigh, $\xi = [\ 66.40, 95.68]$ on clothes ($\equiv Y_5$).

A formal definition of an interval-valued variable is:

Definition 2.3: An interval-valued symbolic random variable Y is one that takes values in an interval; i.e., $Y = \xi = [a, b] \subset \mathcal{R}^1$, with $a \leq b$, $a, b \subset \mathcal{R}^1$. The interval can be closed or open at either end, i.e., (a, b), $[a, b)$, or $(a, b]$.

Further, note that:

When the intervals emerge as the result of aggregating classical data, then the symbolic values a_{uj}, b_{uj} for the variable j in category ω_u , are given by

$$a_{uj} = \min_{i \in \Omega_u} x_{ij}, \quad b_{uj} = \max_{i \in \Omega_u} x_{ij},$$

where Ω_u is the set of $i \in \Omega_u$ values which make up category ω_u.

Modal-valued data:

Definition 2.4: Let a random variable Y take possible values $\{\eta_k;\ k = 1, 2, \ldots\}$ over a domain $\mathcal{Y}$. Then, a particular outcome is modal-valued if it takes the form

$$Y(\omega_u) \equiv Y(u) = \{\eta_k, \pi_k; k = 1, ..., s_u\}$$

for an observation ω_u, where π_k is a non-negative measure associated with η_k and where s_u is the number of values actually taken from $\mathcal{Y}$. The η_k may be categorical or quantitative in value and the domain $\mathcal{Y}$ can be finite or infinite in size.

Measures/Weights:
Relative frequencies
capacities, credibilities, necessities, possibilities, ...
(Definitions 2.7 - 2.10)

Modal multi-valued data:

E.g., 1. See marital status earlier

If the observation unit is gender ($\equiv Y_4$), then for these 15 individuals:

For $u = 1$ (men), $Y_5 = \{$ single, married$\}$, or {single, 3/8; married 5/8}

For $u = 2$ (women), $Y_5 = \{$ single, married$\}$, or {single, 1/7; married 6/7}

E.g., 2. Opinion poll: Y_1 = opinion on Q1 (from Example 2.13)

Possible opinions are:

$\mathcal{Y}$ = {Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree}

For Product 1, i.e., ω_1, or $u = 1$,

$Y(\omega_1)$ = {Strongly Agree, 0.3; Agree, 0.4; Neutral, 0.15; Disagree, 0.12; Strongly Disagree, 0.03}

For Product 2, i.e., ω_2, or $u = 2$,

$Y(\omega_2)$ = {Strongly Agree, 0.1; Agree, 0.2; Neutral, 0.5; Disagree, 0.15; Strongly Disagree, 0.15}

Quantitative Modal-valued data, i.e., Histograms: **Recall**

Definition 2.4: Let a random variable Y take possible values $\{\eta_k; \, k = 1, 2, \ldots\}$ over a domain $\mathcal{Y}$. Then, a particular outcome is modal-valued if it takes the form $Y(\omega_u) = \{\eta_k, \pi_k; \, k = 1, ..., s_u\}$ for an observation ω_u, where π_k is a non-negative measure associated with η_k and where s_u is the number of values actually taken from $\mathcal{Y}$. The η_k may be categorical or quantitative in value and the domain $\mathcal{Y}$ can be finite or infinite in size.

For histogram data, the possible values η_k are intervals, to give us:

Definition 2.6: Let Y be a quantitative random variable that can take values on a finite number of nonoverlapping intervals $\{[a_k, b_k), k = 1, 2, \ldots\}$ with $a_k \leq b_k$. Then, an outcome for observation ω_u for an histogram interval-valued random variable takes the form

$$Y(\omega_u) \equiv Y(u) = \{[a_{uk}, b_{uk}), p_{uk}; k = 1, \ldots, s_u\}$$

where $s_u < \infty$ is the finite number of intervals forming the support for the outcome $Y(\omega_u)$ for observation ω_u , and where p_{uk}, is the support for the particular subinterval $[a_{uk}, b_{uk}), k = 1, \ldots, s_u$, with $\sum_k p_{uk} = 1$. The intervals (a_k, b_k) can be open or closed at either end.

E.g., 1. (Our HMO medical dataset)

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
1	Boston	M	24	M	S	2	2	0	165
2	Boston	M	56	M	M	1	2	2	186
3	Chicago	D	48	M	M	1	3	2	175
4	El Paso	M	47	F	M	0	1	1	141
5	Byron	D	79	F	M	0	3	4	152
6	Concord	M	12	M	S	2	1	0	73
7	Atlanta	M	67	F	M	1	6	0	166
8	Boston	O	73	F	M	0	2	4	164
9	Lindfield	D	29	M	M	2	0	2	227
10	Lindfield	D	44	M	M	1	3	3	216
11	Boston	D	54	M	S	1	5	0	213
12	Chicago	M	12	F	S	2	2	0	75
13	Macon	M	73	F	M	0	3	1	152
14	Boston	D	48	M	M	0	2	4	206
15	Peoria	O	79	F	M	0	3	3	153

Take $Y_3 =$ Age

Y_3(men) $= [12, 56]$, Y_3(women) $= [12, 79]$; Intervals too wide?

Histogram is better –

Y_3(men) $= \{[12, 34), 3/8; [34, 56], 5/8\}$,

Y_3(women) $= \{[12, 40), 1/7; [40, 60), 1/7; [60, 80], 5/7\}$

E.g., 2. Cholesterol for Gender x Age categories (from Table 4.5)

ω_u	Concept Gender	Age	Frequency Histogram
ω_1	Female	20s	{[80, 100), .025; [100, 120), .075; [120, 135); .175; [135, 150), .250; [150, 165), .200; [165, 180), .162; [180, 200), .088; [200, 240), .025}
ω_2	Female	30s	{[80, 100), .013; [100, 120), .088; [120, 135), .154; [135, 150), .253; [150, 165), .210; [165, 180), .177; [180, 195), .066; [195, 210), .026; [210, 240), .013}
ω_3	Female	40s	{[95, 110), .012; [110, 125), .029; [125, 140), .113; [140, 155), .206; [155, 170), .235; [170, 185), .186; [185, 200), .148; [200, 215), .043; [215, 230), .020; [230, 245), .008}
ω_4	Female	50s	{[105, 120), .009; [120, 135), .026; [135, 150), .046; [150, 165), .105; [165, 180), .199; [180, 195), .248; [195, 210), .199; [210, 225), .100; [225, 240), .045; [240, 260), .023}
ω_5	Female	50s	{[115, 140), .012; [140, 160), .069; [160, 180), .206; [180, 200), .300; [200, 220), .255; [220, 240), .146; [240, 260), .012}
ω_6	Female	70s	{[120, 140), .017; [140, 160), .083; [160, 180), .206; [180, 200), .294; }
...	...	...	...
ω_{14}	Male	80+	{[155, 170), .067; [170, 185), .133; [185, 200), .200; [200, 215), .267; [215, 230), .200; [230, 245), .067; [245, 260), .066}

Note, the subintervals need not be of equal length

Some other types of data: Classical → Fuzzy → Symbolic data $Y_1 = \text{Height}$

Classical Data

Individual	Height	Weight	Hair
Sean	1.85	80	blonde
Kevin	1.60	45	blonde
Rob	1.35	30	black
Jack	1.95	90	black

$y \geq 0$, Classical Y

→

Fuzzy Data on $Y_1 = \text{Height}$

	Height				
Individual	Short	Average	Tall	Weight	Hair
Sean	0.00	0.50	0.50	80	blonde
Kevin	0.70	0.30	0.00	45	blonde
Rob	0.50	0.00	0.00	30	black
Jack	0.00	0.00	0.48	90	black

y = short, triangular distribution (1.50)
average, triangular distribution(1.80)
tall, triangular distribution (1.90)
for Fuzzy Y

→

Symbolic data for categories based on $Y_3 = \text{Hair Color}$

	Height				
Hair Color	Short	Average	Tall	Weight	
blonde	[0, 0.70]	[0.30, 0.50]	[0, 0.50]	[45, 80]	{Sean,Kevin}
black	[0, 0.50]	0	[0, 0.48]	[30, 90]	{Rob, Jack}

Some other types of data: Fuzzy, Imprecise, Conjunctive data

E.g.,1. Many medical measurements are imprecise
e.g., Y = Pulse rate, e.g., $Y = 64 + / - 2$, i.e., $Y = [62, 66]$
... there are many examples

E.g., 2. Need for confidentiality:
e.g., Y = household income, e.g., $Y = 110$ may become $Y = [107, 120]$

Types of Data - Conjunctive

Some other types of data: Fuzzy, Imprecise, Conjunctive data

E.g., 1. Y = bird color

Y(galah) = { gray, **pink**}

Written as a conjunctive (i.e., $\wedge$) value, we have

Y(galah) = {**gray** $\wedge$ **pink**}

E.g., 2. Y = color of sweet pea

Y(sweet pea) = {red, **purple**, red$\wedge$**purple**}

This is conjunctive and captures fact can have

all red, all **purple**, or red **and** **purple** sweet peas.

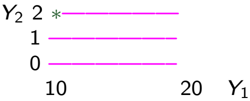
Rules

Logical dependency rules

E.g., $Y_1 = \text{age}, \quad Y_2 = \#\ \text{children}$

Classical: $Y_a = (10, 0), Y_b = (20, 2), Y_c = (18, 1)$

Aggregation $\longrightarrow$
Symbolic: $\xi = [10, 20] \times \{0, 1, 2\}$



I.e. ξ implies classical $Y_d = (10, 2)$ is possible

Need rule $\nu : \{\text{If}\, Y_1 < 15, \text{then}\, Y_2 = 0\}$

Logical dependency rules

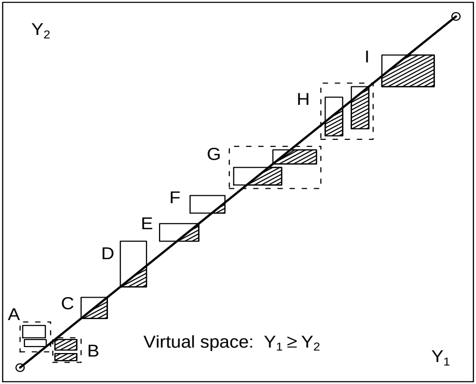
Interval data Eg, baseball, soccer, blood pressures,....

At Bats and Hits by Team

ω_u Team	Y_1 #At-Bats	Y_2 #Hits	Pattern	ω_u Team	Y_1 # At-Bats	Y_2 #Hits	Pattern
ω_1	(289, 538)	(75, 162)	B	ω_{11}	(212, 492)	(57, 151)	B
ω_2	(88, 422)	(49, 149)	I	ω_{12}	(177, 245)	(189, 238)	G
ω_3	(189, 223)	(201, 254)	F	ω_{13}	(342, 614)	(121, 206)	B
ω_4	(184, 476)	(46, 148)	B	ω_{14}	(120, 439)	(35, 102)	B
ω_5	(283, 447)	(86, 115)	B	ω_{15}	(80, 468)	(55, 115)	I
ω_6	(24, 26)	(133, 141)	A	ω_{16}	(75, 110)	(75, 110)	C
ω_7	(168, 445)	(37, 135)	B	ω_{17}	(116, 557)	(95, 163)	I
ω_8	(123, 148)	(137, 148)	E	ω_{18}	(197, 507)	(52, 53)	B
ω_9	(256, 510)	(78, 124)	B	ω_{19}	(167, 203)	(48, 232)	H
ω_{10}	(101, 126)	(101, 132)	D				

$\xi_2 : Y_2 = 149$ not possible when $Y_1 < 149$

Different patterns, for Rule $\nu : Y_1 \geq \alpha Y_2$



Virtual Descriptions, Rules:

Rules

- may be necessary for data coherence/integrity
- condition(s) underlying analysis
- data cleaning, and so on.

Need notion of virtual description space:

Virtual Descriptions: First,

Definition 3.1: Individual descriptions, denoted by $\mathbf{x}$, are those descriptions for which each D_j is a set of one value only, i.e., $\mathbf{x} = (x_1, \ldots, x_p) \equiv \mathbf{d} = (\{x_1\}, \ldots, \{x_p\})$, $\mathbf{x} \in \mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

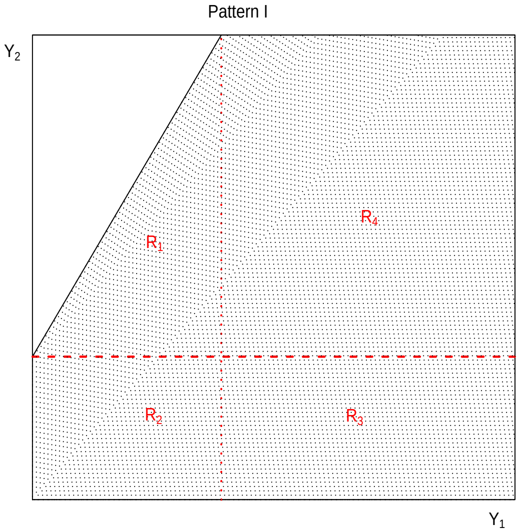
Definition 3.2: Virtual description of the description vector d is the set of individual description vectors x that satisfy all the (logical dependency) rules ν in $\mathcal{X}$. We write this as

$$vir(\mathbf{d}) = \{\mathbf{x} \in \mathcal{D}; \nu(x) = 1, \text{ for all } \nu \text{ in } \mathcal{V}_{\mathcal{X}}$$

where $\mathcal{V}_{\mathcal{X}}$ is the set of all rules ν operating on $\mathcal{X}$.

A rule ν: $[\mathbf{x} \in A] \Rightarrow [\mathbf{x} \in B]$ is a mapping of $\mathcal{X}$ onto $\{0, 1\}$ with $\nu(x) = 0(1)$ if the rule is not (is) true.
An $\mathbf{x}$ satisfies ν if and only if $\mathbf{x} \in A \cap B$ or $\mathbf{x} \neq A$.

E.g., Baseball dataset: Pattern **I**: Rule: $\nu : Y_1 \geq \alpha Y_2$



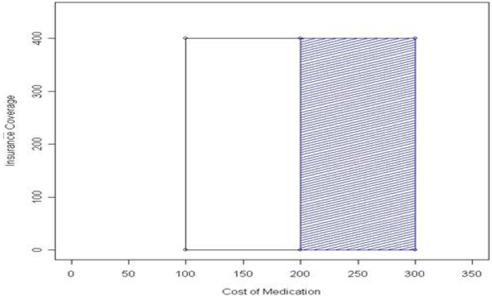
Virtual dataspace $V(u)$ is a histogram-valued variable

HMO medical-insurance dataset (from Example 3.6, 3.7)
Y_1 = Cost of medication, Y_2 =Insurance coverage

Suppose $\xi = ([100, 300], [0, 400])$
Then, D is all $\mathbf{x}$ in the rectangle $(100, 300) \times (0, 400)$

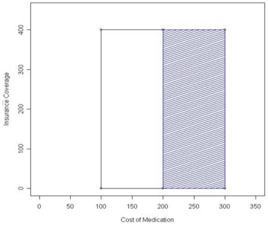
Consider the rule: ν_1 : If $Y_1 < 200$, then $Y_2 = 0$.

Then the virtual description space are those points $\mathbf{x}$ in the hypercube (rectangle) $(200, 300) \times (0, 400)$

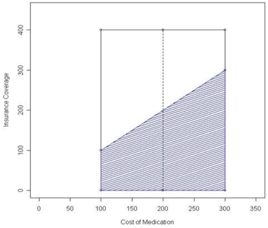


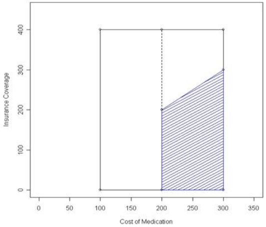
There can be many rules - $\nu = (\nu_1, \nu_2)$

ν_1 : If $Y_1 < 200$, then $Y_2 = 0$



ν_2 : $[Y_2 \leq Y_1]$





$\leftarrow \quad \nu = (\nu_1, \nu_2)$

I.e., you pay the first 200 costs, and your coverage cannot exceed your costs

Then the virtual description space are those points **x** in the hypercube bounded by the vertices (200, 0), (300, 0), (300, 300), (200, 200)

Rules

Cancer treatments:
Y_1 = presence of cancer, $\mathcal{Y}_1 = \{\text{No}=0, \text{Yes}=1\}$
Y_2 = # treatments, $\mathcal{Y}_2 = \{0,1,2,3\}$

Description space D consists of all possible $\mathbf{x} = (Y_{u1}, Y_{u2})$:

ω_u	Y_1	Y_2	D
ω_1	{0,1}	{2}	{(0,2),(1,2)}
ω_2	{0,1}	{0,1}	{(0,0), (0,1), (1,0), (1,1)}
ω_3	{0,1}	{3}	{(0,3),(1,3)}
ω_4	{0}	{1}	{(0,1)}
ω_5	{0}	{0,1}	{(0,0),(0,1)}
ω_6	{1}	{2,3}	{(1,2), (1,3)}

However, some values are not possible, e.g., $\mathbf{x} = (0, 1)$, since there would be no treatments if there is no cancer;
i.e, If $Y_1 = 0$, then $Y_2 = 0$.

This is a
Rule - ν : If $Y_1 \in \{0\}$, then $Y_2 \in \{0\}$

Rules

- may be necessary for data coherence/integrity,
  e.g., baseball example $\nu : Y_1 \leq Y_2$
- condition(s) underlying analysis,
  e.g., HMO medical example ν_1 : If $Y_1 < 200$, then $Y_2 = 0$, and/or $\nu_2 : Y_2 \leq Y_1$
- data cleaning,
  e.g., cancer example ν : If $Y_1 \in \{0\}$, then $Y_2 \in \{0\}$
- and so on.

Descriptive Statistics

Let $Y = (Y_1, \ldots, Y_p)$ have realization $\xi = (\xi_{u1}, \ldots, \xi_{up})$ with $\xi_{uj} = [a_{uj}, b_{uj}],\ \ j = 1 \ldots, p$, for observation $\omega_u \in E \quad$ (or, $u \in E = \{1, \ldots, m\}$)

Take $p = 1,\ Y(\omega_u) = [a_u, b_u], u \in E$.

Then, if we assume a uniform distribution across the intervals, we have for each ω_u,

$$\begin{aligned} P\{X = \xi | x \in vir(D_u)\} &= 0, \xi < a_u, \\ &= (\xi - a_u)/(b_u - a_u), a_u \leq \xi < b_u, \\ &= 1, \xi \geq b_u. \end{aligned}$$

The empirical distribution function $F(\xi)$ is the distribution of a mixture of m uniform distributions $\{Y(\omega_u), u = 1, \ldots, m\}$, with weights p_u. That is,

$$F(\xi) = \frac{1}{\sum p_u} \sum_{\xi \in Y(u)} p_u[(\xi - a_u)/(b_u - a_u) + |(u|\xi \geq b_u)|].$$

For equally weighted observations, $p_u = 1/m$.

Histograms:

Construct a histogram on g subintervals

$I_g = [\zeta_{g-1}, \zeta_g), g = 1, \ldots, r-1, \quad I_r = [\zeta_{r-1}, \zeta_r]$

which span $\quad I = [\min_{u \in E} a_u, \max_{u \in E} b_u]$.

Definitions 3.8: For the interval-valued variate Y, the observed frequency for the histogram subinterval $I_g = [\zeta_{g-1}, \zeta_g), g = 1, \ldots, r$, is

$$f_g = \sum_{u \in E} ||Y(u) \cap I_g|| / ||Y(u)|| \tag{3.20}$$

where $||A||$ is the length of the interval A; and the relative frequency is

$$p_g = f_g / m. \tag{3.21}$$

Then, the histogram for Y is the set of $\{(I_g, f_g), g = 1, \ldots, r\}$.

Blood Pressure Data

ω_u	Y_1 Pulse Rate	Y_2 Systolic Pressure	Y_3 Diastolic Pressure
ω_1	[44, 68]	[90, 110]	[50, 70]
ω_2	[60, 72]	[90, 130]	[70, 90]
ω_3	[56, 90]	[140, 180]	[90, 100]
ω_4	[70, 112]	[110, 142]	[80, 108]
ω_5	[54, 72]	[90, 100]	[50, 70]
ω_6	[70, 100]	[134, 142]	[80, 110]
ω_7	[72, 100]	[130, 160]	[76, 90]
ω_8	[76, 98]	[110, 190]	[70, 110]
ω_9	[86, 96]	[138, 180]	[90, 110]
ω_{10}	[86, 100]	[110, 150]	[78, 100]
ω_{11}	[53, 55]	[160, 190]	[205, 219]
ω_{12}	[50, 55]	[180, 200]	[110, 125]
ω_{13}	[73, 81]	[125, 138]	[78, 99]
ω_{14}	[60, 75]	[175, 194]	[90, 100]
ω_{15}	[42, 52]	[105, 115]	[70, 82]

Histogram of Y_1 = Pulse rate

Min $\{a_{u1}\} = 42$, Max $\{b_{u1}\} = 112$

Let I span $[40, 115]$, $r = 5$

$I_g : [40, 55), \ldots, [100, 115]$

$$f_g = \sum_{u \in E} ||Y(u) \cap I_g|| / ||Y(u)|| \qquad (3.20)$$

Observed frequency f_1 on $I_1 = [40, 55)$ is:

$$f_1 = (55-44)/(68-44) + 0 + 0 + 0 + (55-54)/(72-54) + 0 + 0 + 0 + 0 + 0$$
$$+ (55-53)/(55-53) + (55-50)/(55-50) + 0 + 0 + (52-42)/(52-42) = \mathbf{3.514}$$

The complete histogram for Y_1 = Pulse Rate is

g	Histogram Subinterval I_g	Observed Frequency f_g	Relative Frequency p_g
1	[40, 55)	**3.514**	0.234
2	[55, 70)	3.287	0.219
3	[70, 85)	3.783	0.252
4	[85, 100)	4.131	0. 275
5	[100, 115]	0.286	0.019

Notice frequency f_g for histogram sub-interval I_g is not an integer as in classical data

Blood Pressure Data

ω_u	Y_1 Pulse Rate	Y_2 Systolic Pressure	Y_3 Diastolic Pressure
ω_1	[44, 68]	[90, 110]	[50, 70]
ω_2	[60, 72]	[90, 130]	[70, 90]
ω_3	[56, 90]	[140, 180]	[90, 100]
ω_4	[70, 112]	[110, 142]	[80, 108]
ω_5	[54, 72]	[90, 100]	[50, 70]
ω_6	[70, 100]	[134, 142]	[80, 110]
ω_7	[72, 100]	[130, 160]	[76, 90]
ω_8	[76, 98]	[110, 190]	[70, 110]
ω_9	[86, 96]	[138, 180]	[90, 110]
ω_{10}	[86, 100]	[110, 150]	[78, 100]
ω_{11}	[53, 55]	[160, 190]	[205, 219]
ω_{12}	[50, 55]	[180, 200]	[110, 125]
ω_{13}	[73, 81]	[125, 138]	[78, 99]
ω_{14}	[60, 75]	[175, 194]	[90, 100]
ω_{15}	[42, 52]	[105, 115]	[70, 82]

However,
Y_3 = Diastolic Pressure
$<$ Systolic Pressure $= Y_2$
I.e., we need a Rule $\nu : Y_3 \leq Y_2$

Histogram of Y_1 = Pulse rate
Min $\{a_{u1}\} = 42$, Max $\{b_{u1}\} = 112$
Let I span $[40, 115]$, $r = 5$
$I_g : [40, 55), \ldots, [100, 115]$

$$f_g = \sum_{u \in E} ||Y(u) \cap I_g||/||Y(u)|| \qquad (3.20)$$

Observed frequency f_1 on $I_1 = [40, 55)$ is now:

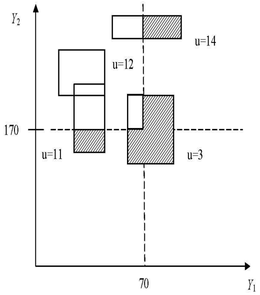
$$f_1 = (55-44)/(68-44) + 0 + 0 + 0 + (55-54)/(72-54) + 0 + 0 + 0 + 0 + 0$$
$$+ \underline{(55-53)/(55-53)} + (55-50)/(55-50) + 0 + 0 + (52-42)/(52-42) = \mathbf{2.514}$$

The application of some rules on original hypercube bounded by intervals produces a virtual description space based on histograms.

ω_u	Y_1 Pulse Rate	Y_2 Systolic Pressure	Y_3 Diastolic Pressure
ω_1	[44, 68]	[90, 110]	[50, 70]
ω_2	[60, 72]	[90, 130]	[70, 90]
ω_3	[56, 90]	[140, 180]	[90, 100]
ω_4	[70, 112]	[110, 142]	[80, 108]
ω_5	[54, 72]	[90, 100]	[50, 70]
ω_6	[70, 100]	[134, 142]	[80, 110]
ω_7	[72, 100]	[130, 160]	[76, 90]
ω_8	[76, 98]	[110, 190]	[70, 110]
ω_9	[86, 96]	[138, 180]	[90, 110]
ω_{10}	[86, 100]	[110, 150]	[78, 100]
ω_{11}	[53, 55]	[160, 190]	[205, 219]
ω_{12}	[50, 55]	[180, 200]	[110, 125]
ω_{13}	[73, 81]	[125, 138]	[78, 99]
ω_{14}	[60, 75]	[175, 194]	[90, 100]
ω_{15}	[42, 52]	[105, 115]	[70, 82]

E.g., the rule

$$\nu : A = \{Y_1 = 70\} \Rightarrow B = \{Y_2 = 170\}.$$



For histogram-valued data –
Realization for ω_u is $Y_u \equiv \xi_u = \{[a_{uk}, b_{uk}), p_{uk}; k = 1, \ldots, s_u\}$, $u = 1, \ldots, m$
Same idea as for intervals, iterating through each of the s_u data subintervals.

———

Histogram of Histogram-valued Data Algorithm (SAS Marco)

```
/* Variable v, Histogram interval [ha, hb], # Observations m */
%macro hist(datain=, dataout=, v =, ha =, hb =, s =);
data &dataout; set &datain end=last;
retain sum&v  m&v 0;
%do k = 1 %to &s;
  if a&v < &ha & b&v ≤ &ha then do; add=0; end;
  if a&v ≤ &ha & b&v > &ha & b&v < &hb then do;
     add=p&v * (b&v − &ha)/(b&v − a&v); end;
  if a&v > &ha & b&v < &hb then do; add=p&v; end;
  if a&v = &ha & b&v = &hb then do; add=p&v; end;
  if a&v > &ha & a&v < &hb & b&v ≥ &hb then do;
      add=p&v * (&hb − a&v)/(b&v − a&v); end;
  if a&v < &ha & b&v > &hb then do;
      add=p&v * (&hb − &ha)/(b&v − a&v); end;
  if a&v > &ha & a&v ≥ &hb then do; add=0; end;
  if a&v < &ha & b&v = &hb then do;
      add=p&v * (&hb − &ha)/(b&v − a&v); end;
  if a&v = &ha & &hb < b&v then do;
      add=p&v * (&hb − &ha)/(b&v − a&v); end;
sum&v=sum&v+add; output;
%end;
m&v = m&v + 1;
if last then do;
  prob&v=sum&v/m&v;
    file print;
    put " For Variable &v on Interval g&v = (&ha, &hb):";
    put " Frequency = " sum&v "Probability = " prob&v;
end; run;
%mend hist;

%hist(datain=one, dataout=two, v=2, ha=90, hb=110, s=10);
```

Histogram for Y_1 = Pulse Rate:

		No rule		Rule $\nu_1 : Y_3 \leq Y_2$		Rule $\nu_3 = (\nu_1, \nu_2)$	
g	Histogram Subinterval I_g	Observed Frequency f_g	Relative Frequency p_g	Observed Frequency f_g	Relative Frequency p_g	Observed Frequency f_g	Relative Frequency p_g
1	[40, 55)	3.514	0.234	**2.514**	**0.180**	*1.514*	*0.116*
2	[55, 70)	3.287	0.219	3.287	**0.235**	*2.552*	*0.196*
3	[70, 85)	3.783	0.252	3.783	**0.270**	*4.500*	*0.346*
4	[85, 100)	4.131	0. 275	4.131	**0.295**	*4.148*	*0.319*
5	[100, 115]	0.286	0.019	0.286	**0.020**	*0.286*	*0.022*
	m	15		14		13	

Rule $\nu_3 = (\nu_1, \nu_2)$ where
$\nu_1 : Y_3 \leq Y_2$
$\nu_2 : A = \{Y_1 = 70\} \Rightarrow \quad B = \{Y_2 = 170\}$.

Under rule ν_2, some observations are now histogram-valued

Histograms of Histograms:

Construct a histogram on g subintervals $I_g = [\zeta_{g-1}, \zeta_g)$, $g = 1, \ldots, r$, which span $I = [\min_{u \in E, k} a_{uk}, \max_{u \in E, k} b_{uk}]$.

Definitions 3.8: For the histogram-valued variate Y, the observed frequency for the histogram subinterval $I_g = [\zeta_{g-1}, \zeta_g), g = 1, \ldots, r$, is

$$O(g) = \sum_{u \in E} \pi(g; u), \quad \pi(g; u) = \sum_{k \in \mathcal{Y}(g)} p_{uk} ||Y(k; u) \cap I_g|| / ||Y(k; u)||$$

where $Y(k; u) = [a_{uk}, b_{uk})$ and $\mathcal{Y}(g)$ is the set of those intervals overlapping with $\{(I_g, f_g), g = 1, \ldots, r\}$, and where $||A||$ is the length of the interval A; and the relative frequency is

$$p_g = O(g)/m.$$

Then, the histogram for Y is the set $\{(I_g, O_g), g = 1, \ldots, r\}$;

Income by Age-Groups Y_1 = Income (Table 3.10)

Age	ω_u	
20s	ω_1	{[70, 84), .02; [84, 96), .06; [96, 108), .24; [108, 120), .30; [120, 132), .24; [132, 144), .06; [144, 160), .08}
30s	ω_2	{[100, 108), .02; [108, 116), .06; [116, 124), .40; [124, 132), .24; [132, 140), .24; [140, 150), .04}
40s	ω_3	{[110, 125), .04; [125, 135), .14; [135, 145), .20; [145, 155), .42; [155, 165), .14; [165, 175), .04; [175, 185), .02}
50s	ω_4	{[100, 114), .04; [114, 126), .06; [126, 138), .20; [138, 150), .26; [150, 162), .28; [162, 174), .12; [174, 190), .04}
60s	ω_5	{[125, 136), .04; [136, 144), .14; [144, 152), .38; [152, 160), .22; [160, 168), .16; [168, 180), .06}
70s	ω_6	{[135, 144), .04; [144, 150), .06; [150, 156), .24; [156, 162), .26; [162, 168), .22; [168, 174), .14; [174, 180), .04}
80s	ω_7	{(100, 120), .02; [120, 135), .12; [135, 150), .16; [150, 165), .24; [165, 180), .32; [180, 195), .10; [195, 210), .04}

$I = [70, 210]; r = 10;$ $I_g : [60, 75), \ldots, [195, 210]$ $m = 7$

For $g = 5$, contribution of ω_1 to $I_5 = [120, 135)$ is

$$\pi(5; 1) = 0 + 0 + 0 + 0 + (0.24)(132\text{-}120)/(132\text{-}120) + (.06)(135\text{-}132)/144\text{-}132) + 0 = \mathbf{0.2550}.$$

Likewise, for observations $\omega_u, u = 2, \ldots, 7,$

$\pi(5; 2) = .5300, \pi(5; 3) = .1533, \pi(5, 4) = .1800,$
$\pi(5, 5) = .0364, \pi(5, 6) = .0000, \pi(5, 7) = .1200$

Age	ω_u	
20s	ω_1	$\{[70, 84), .02; [84, 96), .06; [96, 108), .24; [108, 120), .30;$ $[120, 132), .24; [132, 144), .06; [144, 160), .08\}$
30s	ω_2	$\{[100, 108), .02; [108, 116), .06; [116, 124), .40; [124, 132), .24;$ $[132, 140), .24; [140, 150), .04\}$
40s	ω_3	$\{[110, 125), .04; [125, 135), .14; [135, 145), .20; [145, 155), .42;$ $[155, 165), .14; [165, 175), .04; [175, 185), .02\}$
50s	ω_4	$\{[100, 114), .04; [114, 126), .06; [126, 138), .20; [138, 150), .26;$ $[150, 162), .28; [162, 174), .12; [174, 190), .04\}$
60s	ω_5	$\{[125, 136), .04; [136, 144), .14; [144, 152), .38; [152, 160), .22;$ $[160, 168), .16; [168, 180), .06\}$
70s	ω_6	$\{[135, 144), .04; [144, 150), .06; [150, 156), .24; [156, 162), .26;$ $[162, 168), .22; [168, 174), .14; [174, 180), .04\}$
80s	ω_7	$\{(100, 120), .02; [120, 135), .12; [135, 150), .16; [150, 165), .24;$ $[165, 180), .32; [180, 195), .10; [195, 210), .04\}$

$I = [70, 210]; r = 10;$ $I_g : [60, 75), \ldots, [195, 210]$ $m = 7$

For $g = 5$, contributions of ω_u to $I_5 = [120, 135)$ are
$\pi(5;1) =$ **.2550**. $\pi(5;2) = .5300, \pi(5;3) = .1533, \pi(5,4) = .1800,$
$\pi(5,5) = .0364, \pi(5,6) = .0000, \pi(5,7) = .1200$

Hence, observed frequency for I_5 is:
$O(5) = \sum_{u \in E} \pi(5;u) = .2550 + .5300 + .1533 + .1800 + .0364 + .0000 + .1200 =$ **1.2747**.
Relative frequency is: $p_5 = O(5)/m = 1.2747/7 = 0.1821$.

Complete Histogram of Histogram Data is:

g	Histogram Subinterval I_g	Observed Frequency O_g	Relative Frequency p_g
1	[60, 75)	0.0071	0.0010
2	[75, 90)	0.0429	0.0061
3	[90, 105)	0.2418	0.0345
4	[105, 120)	0.7249	0.1036
5	**[120, 135)**	**1.2747**	**0.1821**
6	[135, 150)	1.6736	0.2391
7	[150, 165)	1.9750	0.2821
8	[165, 180)	0.8850	0.1264
9	[180, 195)	0.1350	0.0193
10	[195, 210]	0.0400	0.0057

Sample Mean, Sample Variance (Bertrand and Goupil (2000))

Definitions 3.9: For an interval-valued random variable Y, the symbolic sample mean is given by

$$\bar{Y} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u) \tag{3.22}$$

and the symbolic sample variance is given by

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2} [\sum_{u \in E} (b_u + a_u)]^2, \tag{3.23}$$

for observations $Y_u = [a_u, b_u), u = 1, \dots, m$

Let us study the sample variance more closely:

Rewrite S^2 as

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

When $m = 1$, this becomes (for $u = 1$)

$$S^2 = \frac{1}{3 \times 1}[(a_u - \bar{Y})^2 + (a_u - \bar{Y})(b_u - \bar{Y}) + (b_u - \bar{Y})^2]$$

Suppose $m = 1$ with $Y_u = [7, 13]$; then $\bar{Y} \equiv \bar{Y}_u = 10$ and $S^2 = 3 \neq 0$

In general, we have

$$S_u^2 = \frac{1}{3}[(a_u - \bar{Y}_u)^2 + (a_u - \bar{Y}_u)(b_u - \bar{Y}_u) + (b_u - \bar{Y}_u)^2], \quad \bar{Y}_u = (a_u + b_u)/2$$

This S_u^2 measures variation Within Observation u

___________-

Note, if $X \sim U(7, 13)$, then $\mathrm{Var}(X) = 3$

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

This is Total Variance $\equiv$ TotalSS$/m$

We can show
Total Variation = Within Variation + Between Variation

$$\text{WithinSS} = \frac{1}{3} \sum_{u=1}^{m} [(a_u - \bar{Y}_u)^2 + (a_u - \bar{Y}_u)(b_u - \bar{Y}_u) + (b_u - \bar{Y}_u)^2],$$

$$\text{BetweenSS} = \frac{1}{3} \sum_{u=1}^{m} [(a_u - \bar{Y})^2 + (a_u - \bar{Y})(b_u - \bar{Y}) + (b_u - \bar{Y})^2],$$

$$\bar{Y}_u = (a_u + b_u)/2, \quad \bar{Y} = \frac{1}{m} \sum_{u=1}^{m} (a_u + b_u)/2$$

For classical data, $Y_u = a_u = [a_u, a_u] = \bar{Y}_u$, $\Rightarrow$ WithinSS $= 0$.

Histogram and Sample Statistics for Y_1 = Pulse Rate:

g	Histogram Subinterval I_g	No rule: Observed Frequency f_g	No rule: Relative Frequency p_g	Rule $\nu_1 : Y_3 \leq Y_2$: Observed Frequency f_g	Rule $\nu_1 : Y_3 \leq Y_2$: Relative Frequency p_g	Rule $\nu_3 = (\nu_1, \nu_2)$: Observed Frequency f_g	Rule $\nu_3 = (\nu_1, \nu_2)$: Relative Frequency p_g
1	[40, 55)	3.514	0.234	**2.514**	**0.180**	*1.514*	*0.116*
2	[55, 70)	3.287	0.219	3.287	**0.235**	*2.552*	*0.196*
3	[70, 85)	3.783	0.252	3.783	**0.270**	*4.500*	*0.346*
4	[85, 100)	4.131	0. 275	4.131	**0.295**	*4.148*	*0.319*
5	[100, 115]	0.286	0.019	0.286	**0.020**	*0.286*	*0.022*
	m	15		14		13	
	$\bar{Y}_1$	72.433		73.750		76.050	
	S_1^2	272.501		265.938		239.648	
	S_1	16.508		16.308		15.481	
	Data Type	Interval		Interval		Histogram	

Rules $\nu_1 : Y_3 \leq Y_2 \quad \nu_2 : A = \{Y_1 = 70\} \Rightarrow \quad B = \{Y_2 = 170\}$.

Although rules deal only with Y_2 and Y_3, Y_1 is impacted

Two or more variables: Covariance between Y_1 and Y_2 - $Y_{u1} = [a_{u1}, b_{u1}]$ and $Y_{u2} = [a_{u2}, b_{u2}]$, $u = 1, \ldots, m$

We can show

$$\text{TotalSP} = \text{WithinSP} + \text{BetweenSP}$$

where

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

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

$$\bar{Y}_j = \frac{1}{m} \sum_{u=1}^{m} (a_{uj} + b_{uj}), \; j = 1, 2.$$

Hence,

$$Cov(Y_1, Y_2) = \frac{1}{6m} \sum_{u=1}^{m} [2(a_{u1} - \bar{Y}_1)(a_{u2} - \bar{Y}_2) + (a_{u1} - \bar{Y}_1)(b_{u2} - \bar{Y}_2) + (b_{u1} - \bar{Y}_1)(a_{u2} - \bar{Y}_2) + 2(b_{u1} - \bar{Y}_1)(b_{u2} - \bar{Y}_2)].$$

$Cov(Y_1, Y_2) = S_1^2$, and classical formula holds for classical (Y_1, Y_2)

Blood Pressure Data

ω_u	Y_1 Pulse Rate	Y_2 Systolic Pressure	Y_3 Diastolic Pressure
ω_1	[44, 68]	[90, 110]	[50, 70]
ω_2	[60, 72]	[90, 130]	[70, 90]
ω_3	[56, 90]	[140, 180]	[90, 100]
ω_4	[70, 112]	[110, 142]	[80, 108]
ω_5	[54, 72]	[90, 100]	[50, 70]
ω_6	[70, 100]	[134, 142]	[80, 110]
ω_7	[72, 100]	[130, 160]	[76, 90]
ω_8	[76, 98]	[110, 190]	[70, 110]
ω_9	[86, 96]	[138, 180]	[90, 110]
ω_{10}	[86, 100]	[110, 150]	[78, 100]
ω_{11}	[53, 55]	[160, 190]	[205, 219]
ω_{12}	[50, 55]	[180, 200]	[110, 125]
ω_{13}	[73, 81]	[125, 138]	[78, 99]
ω_{14}	[60, 75]	[175, 194]	[90, 100]
ω_{15}	[42, 52]	[105, 115]	[70, 82]

Take (Y_2, Y_3)
No rules,
$Cov(Y_2, Y_3) = 674.989$
$S_{Y_2} = 30.371$
$S_{Y_3} = 34.701$
$\rho(Y_2, Y_3) = Cov(Y_2, Y_3)/(S_{Y_2} S_{Y_3}) = .640$

Rule ν_1:
$Cov(Y_2, Y_3) = 411.469$
$S_{Y_2} = 29.843$
$S_{Y_3} = 15.914$
$\rho(Y_2, Y_3) = Cov(Y_2, Y_3)/(S_{Y_2} S_{Y_3}) = .866$

Rule $\nu_3 = (\nu_1, \nu_2)$: $Cov(Y_2, Y_3) = 308.647$
$S_{Y_2} == 26.801$
$S_{Y_3} = 14.001$
$\rho(Y_2, Y_3) = Cov(Y_2, Y_3)/(S_{Y_2} S_{Y_3}) = .823$

$\nu_1 : Y_3 \leq Y_2$,
$\nu_2 : A = \{Y_1 \leq 70\} \Rightarrow \quad B = \{Y_2 \leq 170\}$