Descriptive Statistics for Symbolic Data

Paula Brito

Fac. Economia & LIAAD-INESC TEC, Universidade do Porto

ECI 2015 - Buenos Aires
T3: Symbolic Data Analysis:
Taking Variability in Data into Account

Outline

Outline

1. A new framework
2. Interval-valued variables
3. Histogram-valued variables

Descriptive Statistics for Symbolic Variables

No unique and straightforward definitions !

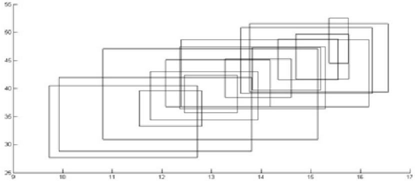
What is the variance of a set of interval observations ?

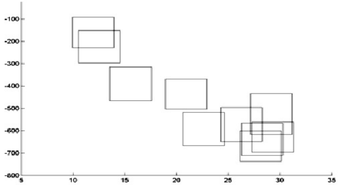
How de we measure correlation ?

- Measures based on interval parameters
- Measures based on distributional assumptions
- Measures based on distances

Outline

Biplots for Interval variables





Descriptive Statistics for Interval Variables

First option :

Using the dispersion of the interval centers

The mean value and the dispersion of all interval midpoints are given by

$$\overline{Y_j} = \frac{1}{n}\sum_{i=1}^{n}\frac{l_{ij}+u_{ij}}{2}$$

$$S^2_{Y_k} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_{ij}+u_{ij}}{2} - \overline{Y_j}\right)^2$$

Descriptive Statistics for Interval Variables

Second option :

Using the dispersion of the interval boundaries.

The mean value and the dispersion of all interval midpoints are given by

$$\overline{Y_j} = \frac{1}{n} \sum_{i=1}^{n} \frac{l_{ij} + u_{ij}}{2}$$

$$S^2_{Y_k} = \frac{1}{n} \sum_{i=1}^{n} \frac{(l_{ij} - \overline{Y_j})^2 + (u_{ij} - \overline{Y_j})^2}{2}$$

Descriptive Statistics for Interval Variables

Under the assumption that the observed $Y_j(s_i)$ and $Y_{j'}(s_i)$ values, $i = 1, \ldots, n$, are **uniformly distributed** across each interval $I_{ik} = [l_{ik}, u_{ik}]$, $k = j, j'$, we have

$$E(Y_{ik}) = (l_{ik} + u_{ik})/2 = c_{ik} \text{ and } Var(Y_{ik}) = (u_{ik} - l_{ik})^2/12$$

- **symbolic sample mean** :

$$\overline{Y_k} = \frac{1}{2n} \sum_{i=1}^{n} (l_{ik} + u_{ik}) = \frac{1}{n} \sum_{i=1}^{n} c_{ik}$$

- **symbolic sample variance** :

$$S^2_{Y_k} = \frac{1}{3n} \sum_{i=1}^{n} [(l_{ik} - \overline{Y_k})^2 + (l_{ik} - \overline{Y_k})(u_{ik} - \overline{Y_k}) + (u_{ik} - \overline{Y_k})^2]$$
$$= \frac{1}{3n} \sum_{i=1}^{n} (l^2_{ik} + l_{ik} u_{ik} + u^2_{ik}) - \overline{Y_k}^2$$

Bertrand and Goupil's (2000)
obtained from the empirical density function for an interval variable

Descriptive Statistics for Interval Variables

For the symbolic covariance three definitions were proposed :

- $Cov_1(Y_j, Y_{j'}) = \frac{1}{4n} \sum_{i=1}^{n} (l_{ij} + u_{ij})(l_{ij'} + u_{ij'}) - \overline{Y_j}.\overline{Y_{j'}}$

  Billard & Diday (2003)
  obtained from the empirical joint density function

- $Cov_2(Y_j, Y_{j'}) = \frac{1}{3n} \sum_{i=1}^{n} G_j G_{j'} \left[Q_j, Q_{j'}\right]^{1/2}$

  with $Q_k = (l_{ik} - \overline{Y_k})^2 + (l_{ik} - \overline{Y_k})(u_{ik} - \overline{Y_k}) + (u_{ik} - \overline{Y_k})^2$,

  $$G_k = \begin{cases} -1 & \text{if} \quad c_{ik} \leq \overline{Y_k} \\ 1 & \text{if} \quad c_{ik} > \overline{Y_k} \end{cases}$$

  Billard & Diday (2006)
  incorporating more accurately both between and within interval variations into the overall covariance

Descriptive Statistics for Interval Variables

- $$Cov_3(Y_j, Y_{j'}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{(u_{ij}-l_{ij})(u_{ij'}-l_{ij'})}{12}}_{\text{WithinSP}} + \underbrace{+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_{ij}+u_{ij}}{2}-\overline{Y_j}\right)\left(\frac{l_{ij'}+u_{ij'}}{2}-\overline{Y_{j'}}\right)}_{\text{BetweenSP}}$$
$$= \frac{1}{6n}\sum_{i=1}^{n}[2(l_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + (l_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}}) + (u_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + 2(u_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}})]$$

Billard (2008)
considering a decomposition into
Within observations Sum of Products (WithinSP) and
Between observations Sum of Products (BetweenSP)

Interval-valued variables: Distance measures

Many measures proposed in the litterature

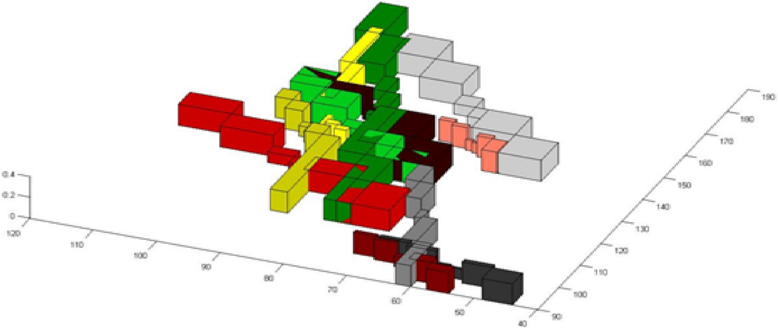
Hausdorff distance :
$d_H(I_i, I_j) = \max\{\{|l_i - l_j|, |u_i - u_j|\}$

Euclidean distance :
$d_2(I_i, I_j) = \sqrt{(l_i - l_j)^2 + (u_i - u_j)^2}$

City-Block distance :
$d_1(I_i, I_j) = |l_i - l_j| + |u_i - u_j|.$

Outline

Biplots for Histogram variables



Descriptive Statistics for Histogram Variables

Assumming an Uniform distributon within each sub-interval of $Y_k(s_i)$, $i = 1, \ldots, n$, $I_{ik\ell} = [l_{ik\ell}, u_{ik\ell}]$, $\ell = 1, \ldots K_j$, $k = j, j'$ we have

- **symbolic sample mean** :

$$\overline{Y_k} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} ((l_{ik\ell} + u_{ik\ell}) p_{ik\ell})$$

- **symbolic sample variance** :

$$S^2_{Y_k} == \frac{1}{3n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} ((l^2_{ik} + l_{ik} u_{ik} + u^2_{ik}) p_{ik\ell}) - \overline{Y_k}^2$$

Billard and Diday (2003)

Descriptive Statistics for Histogram Variables

And for the symbolic covariance three definitions :

$$Cov_1(Y_j, Y_{j'}) = \frac{1}{4n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} p_{ij\ell} p_{ij'\ell} (l_{ij} + u_{ij})(l_{ij'} + u_{ij'}) - \overline{Y_j}.\overline{Y_{j'}}$$

Billard & Diday (2003)
obtained from the empirical joint density function

Correlation Between Symbolic Variables

As in the classic variables: the **correlation coefficient** is defined as :

$$r_{Y_j Y_{j'}} = \frac{Cov(Y_j, Y_{j'})}{S_{Y_j} S_{Y_{j'}}}$$

where

$Cov(Y_j, Y_{j'})$ is the covariance function between Y_j and $Y_{j'}$

S_{Y_j}, $S_{Y_{j'}}$ the symbolic standard deviation of the variables Y_j and $Y_{j'}$, respectively.

In the particular case of interval variables the descriptive statistics depend on the assumed distribution within each interval.

Results already obtained for other distributions, e.g., the triangular distribution.