

## Data Analysis: A Bird's Eye View

A lecture for SCSIS Birkbeck  
research students by  
Prof. Boris Mirkin  
06/11/07 SCSIS Birkbeck

(in [My academic interests](#) on my  
website)

---

---

---

---

---

---

---

---

## Outline Data analysis: statistics, machine learning, data mining, knowledge discovery

- ⌘ Data and feature scales; pre-processing
- ⌘ Two main problems in data analysis
  - ☒ Summarisation
  - ☒ Correlation
- ⌘ Four approaches:
  - ☒ statistics
  - ☒ machine learning
  - ☒ data mining
  - ☒ knowledge discovery
- ⌘ **Examples** (through): Mean, Correlation coefficient
- ⌘ **Methods**: global, local, nature-inspired

---

---

---

---

---

---

---

---

## A Data Analysis Application Involves

- ☒ **A. Developing a data set**
- ☒ **B. Data pre-processing**
- ☒ **C. Applying a method**
- ☒ **D. Interpreting results**
- ☒ **E. Drawing conclusions**

**C** is the only item that is treated  
scientifically,

**B** and **D** sometimes relate to **C**

---

---

---

---

---

---

---

---

## A generic data format

**Table 0.1.** Companies characterized by mixed scale features; the first three companies making product A, the next three making product B, and the last two product C.

Company name	Income, \$mln	SharP \$	NSup	EC	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Industrial
Bayernart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Industrial
Bunchista	12.1	16.9	2	Yes	Industrial
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

---

---

---

---

---

---

---

---

## Feature scales

- ⌘ Quantitative (admits averaging)
- ⌘ Ordinal (admits ordering)
- ⌘ Nominal (partition)
- ⌘ Categorical (not quantitative)

### Terminology:

**Entity = Instance = Case = Object = Observation**

**Feature = Attribute = Character = Variable**

**Value = Score = Grade = Category = State**

---

---

---

---

---

---

---

---

## Two types of data analysis tasks

(Why TWO? - later)

### ⌘ Correlation

- ⌘ **Given:** (sets of) features **X** (input) and **Y** (target/output)
- ⌘ **Find:** association **D** between **X** and **Y**

### ⌘ Summarisation

- ⌘ **Given:** set of features **X**
- ⌘ **Find:** (a smaller set of factors) **F** representing **X**

---

---

---

---

---

---

---

---

## Data analysis objectives for the example

- ⌘ How to meaningfully map companies to the screen? (Summarization)
- ⌘ Would clustering of companies reflect the product? What features to be involved? (Summarization)
- ⌘ Can rules be derived to make an attribution of the product for an outside company? (Correlation)
- ⌘ The structural features and market related features: any relation? (Correlation.)

---

---

---

---

---

---

---

---

## First stage: Preprocessing to quantitative format

Code	Income	SharP	NSup	EC	Util	Indu	Retail
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1

---

---

---

---

---

---

---

---

## Second stage: Feature standardisation

$$\boxtimes Y_{ik} = (X_{ik} - A_k) / B_k$$

- $X$  - original data
- $Y$  - standardised data
- $i$  - entities
- $k$  - features
- $A_k$  - shift of the origin
- $B_k$  - rescaling factor

⊗ **Statistics:**  $A$  - mean,  $B$  - standard deviation

⊗ **SVM:**  $A$  - midrange,  $B$  - half-range

⊗ **Clustering:**  $A$  - mean,  $B$  - half-range (Mirkin 2005)

---

---

---

---

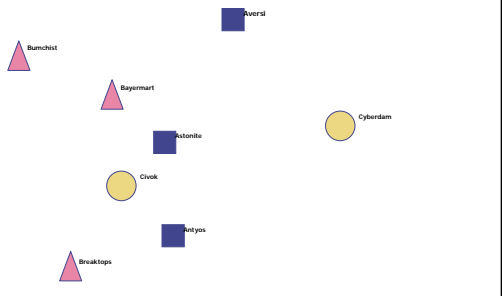
---

---

---

---

### Standardisation effect I: No normalization (B=1)



---

---

---

---

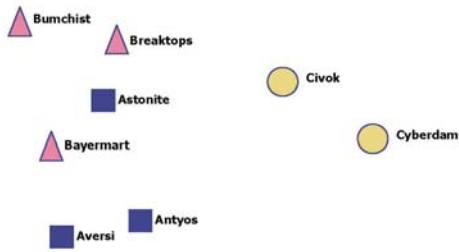
---

---

---

---

### Standardisation effect II: Z-Scoring



---

---

---

---

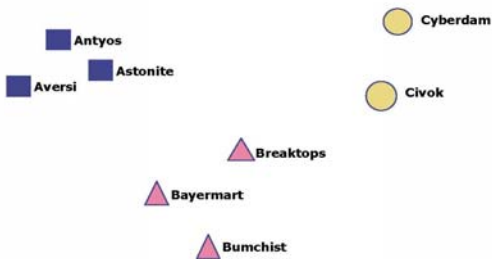
---

---

---

---

### Standardisation effect III: cluster-wise (Mirkin 2005)



Here: both visualization and clustering are ok

---

---

---

---

---

---

---

---

## Data mining perspective

⌘ **Goal:** Finding Patterns and Regularities within the Data (not quite a scientific perspective)

⌘ **Principles:**

☒ **Heuristics** (properties of methods to explore)

☒ **Data recovery**

⌘ **User-friendly:** No need to understand methods, only patterns matter

⌘ **Re-sampling based confidence:**

☒ Bootstrapping

☒ Cross-validation

---

---

---

---

---

---

---

---

## The mean in data mining perspective

⌘ **Heuristic:** just a central value

☒ Given real  $x_1, x_2, \dots, x_N$  compute

$$c = \sum_{i=1}^N x_i / N$$

⌘ **Data recovery:**

☒ Given real  $x_1, x_2, \dots, x_N$ , find  $c$  such that  $x_i = c + e_i$  ( $i=1, \dots, N$ ) to minimize the sum of squares,  $e_1^2 + e_2^2 \dots + e_N^2$ ?

---

---

---

---

---

---

---

---

## Correlation coefficient in data mining perspective

⌘ **Heuristic:** just **cosine** between z-scored feature vectors

☒ Given **z-scored** pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$  compute the average product

$$r = \sum_{i=1}^N x_i y_i / N$$

⌘ **Data recovery:**

☒ Given pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$  find  $a, b$  such that  $y_i = ax_i + b + e_i$  ( $i=1, \dots, N$ ) to minimize the sum of squares,  $e_1^2 + e_2^2 \dots + e_N^2$ .

☒ Property: **Residual y-variance =**  
**= (1-r<sup>2</sup>)\*y-variance**

---

---

---

---

---

---

---

---

## Validity of the mean with Boot-strapping I

### ⚡ Bootstrap:

- ☒ Trial: select randomly, with replacement,  $N$  entities, and compute the mean on this sample
- ☒ Consider distribution of the means over 500 or 5000 trials
- ☒ Find boundaries for the mean with a pre-specified confidence level, say 95%

---

---

---

---

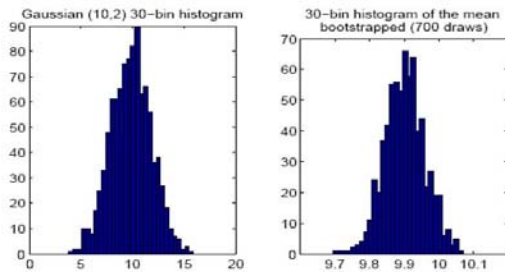
---

---

---

---

## Validity of the mean with Boot-strapping II



---

---

---

---

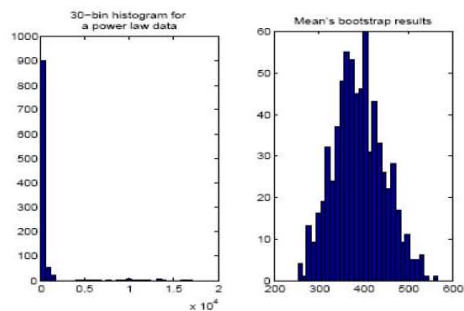
---

---

---

---

## Validity of the mean with Boot-strapping III



---

---

---

---

---

---

---

---

## Classical statistics perspective 1

- ⌘ Data from a probability distribution
- ⌘ Goal: Estimate its properties or parameters
- ⌘ The user must know the model and methods
- ⌘ Questions: “What regression curve is?”  
“How many clusters are out there?”,  
“How data should be pre-processed?”  
are well substantiated

---

---

---

---

---

---

---

---

## Mean

Given:  $x_1=1.2, x_2=1.3, x_3=1, x_4=1.1, x_5=1.4, x_6=1.2, \dots$

Model:  $\mathbf{x}=\mathbf{a}+\mathbf{e}, \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}), p(\mathbf{u}, \mathbf{0}, \boldsymbol{\sigma})=c^* \exp\{-\mathbf{u}^2/\boldsymbol{\sigma}^2\}$

Method: Max likelihood  $\prod_i \exp\{-(x_i-a)^2/\boldsymbol{\sigma}^2\} \rightarrow \max$   
Least Squares  $\sum_i (x_i-a)^2 \rightarrow \min$

Solution:  $\mathbf{a}$  the average

Statistical properties for testing statistical hypotheses

Data pre-processing:  $\mathbf{z}$ -scoring,  $\mathbf{z}=(\mathbf{x}-\mathbf{a})/\boldsymbol{\sigma}$

---

---

---

---

---

---

---

---

## Correlation coefficient

Given: number pairs  $\mathbf{w}_i=(x_i, y_i), i=1, \dots, N$

Model:  $\mathbf{w} = \mathbf{m} + \mathbf{e}, \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \text{density } p(\mathbf{u}, \mathbf{0}, \boldsymbol{\Sigma}) = c^* \exp\{-\mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}\}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x & \rho \\ \rho & \sigma_y \end{bmatrix}$$

Method: Max likelihood  $\prod_i \exp\{-(\mathbf{w}_i - \mathbf{m}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{w}_i - \mathbf{m}_i)\} \rightarrow \max$

Equivalently, Least Squares  $\sum_i (y_i - \mathbf{a}x_i - \mathbf{b})^2 \rightarrow \min \mathbf{a}, \mathbf{b}$

Solution:  $\rho = \mathbf{a}\boldsymbol{\sigma}_x / \boldsymbol{\sigma}_y$  or  
 $\rho$  = the average scalar product of  $\mathbf{x}$  and  $\mathbf{y}$ ,  $\mathbf{z}$ -scored

Statistical properties for testing statistical hypotheses

Data pre-processing:  $\mathbf{z}$ -scoring,  $\mathbf{z}=(\mathbf{var} - \mathbf{av}) / \mathbf{std}$

---

---

---

---

---

---

---

---

## Classical statistics perspective 2

⌘ **Great Distribution Based Principles: Hypothesis Testing & Statistical Confidence**

⌘ **Great Induction Principles:**

- ☒ Maximum Likelihood (Least-Squares, Least- Moduli)
- ☒ Minimum Description Length

⌘ **Troubles:** when data relate to a phenomenon of which not much is known. Response: Non-parametric concepts. Challenge: The existence of a distribution is questionable. Response: Test against disorder – non-satisfactory.

---

---

---

---

---

---

---

---

## Machine learning perspective

⌘ **Goal: Deriving Prediction Rule from Data**

⌘ **Principles:**

- ☒ Statistics
- ☒ Minimising structural risk (SVM)

⌘ **Re-sampling** (to substantiate methods)

- ☒ K-fold cross-validation
- ☒ Jack-knife

---

---

---

---

---

---

---

---

## Knowledge discovery perspective :

⌘ **Goal: Improving knowledge**

⌘ **What is Knowledge:** Structurally,

**Categories and Statements** relating them

- ☒ **Summarisation:** Generating new categories (factors, clusters)
- ☒ **Correlation:** Generating new relations (decision rules, regression)

---

---

---

---

---

---

---

---

## Methods

- ⌘ **Global:** Linear - PCA or Simple Combinatorial – MST (**Rarely work**)
- ⌘ **Local:** Hill-climbing (NN Back Propagation), Alternating Optimisation (Expectation-Maximisation), Neighbourhood Search Heuristics (**No deep minima**)
- ⌘ **Nature Inspired** - Population Driven
  - ⊗ Genetic Algorithms
  - ⊗ Evolutionary Algorithms
  - ⊗ Particle Swarm Algorithms

(**Great expectations**)

---

---

---

---

---

---

---

---

## Conclusion. Bird's Eye View of DA:

- ⌘ **Generic data types: feature table, network graph (as well as signal, image, video) – pre-processing is a must**
- ⌘ **Two major problems:**
  - ⊗ Summarisation
  - ⊗ Correlation
- ⌘ **Different perspectives and validation:**
  - ⊗ classical statistics
  - ⊗ machine learning
  - ⊗ data mining
  - ⊗ knowledge discovery
- ⌘ **Methods:**
  - ⊗ Global
  - ⊗ Local
  - ⊗ Nature Inspired

---

---

---

---

---

---

---

---