Capítulo 2
Análisis Exploratorio de Datos

SEBASTIÁN MALDONADO

ASIGNATURA: IN3401

SEMESTRE OTOÑO, 2010

Introducción

- La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística.

- De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

Introducción(2)

El AED proporciona métodos para:

- Organizar, visualizar y preparar los datos
- detectar fallos en el diseño y recolección de datos
- tratamiento y evaluación de datos ausentes
- identificación de casos atípicos
- comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

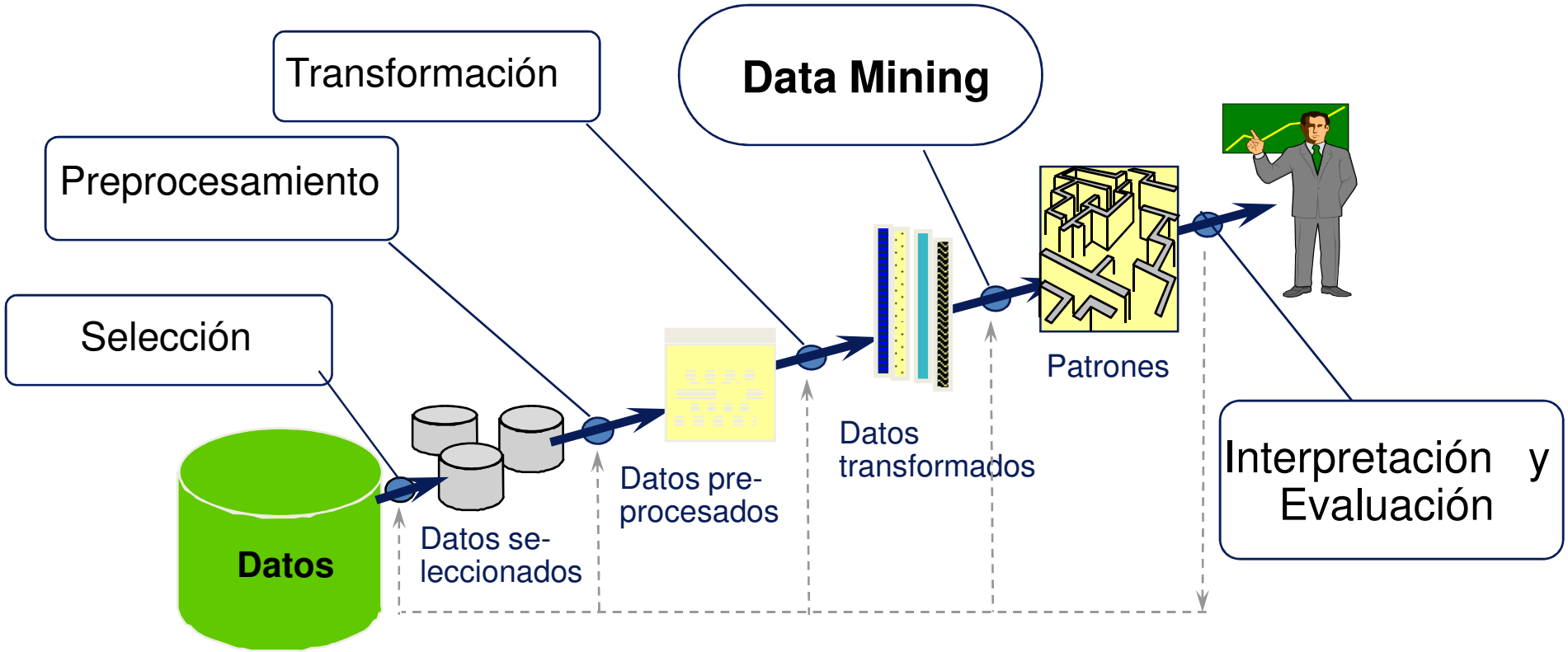
Temario del Capítulo

- Análisis de valores perdidos e inconsistencias
- Preparación y transformación de variables
- Análisis estadístico unidimensional
  - Supuestos univariados (normalidad)
- Análisis estadístico bidimensional
  - Supuestos de independencia
  - Análisis de varianza y tests no paramétricos
- Reducción de la dimensionalidad
  - Selección de Atributos
  - Análisis de Componentes Principales y Análisis Factorial

Proceso de KDD

Knowledge Discovery in Databases

"KDD es el proceso no-trivial de identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos"



Limpieza de datos

- Tipos de Datos perdidos (Taxonomía Clásica) [Little and Rubin, 1987]:
  - **Missing Completely at Random (MCAR):**
    - Los valores perdidos no se relacionan con las variables en la base de datos
  - **Missing at Random (MAR):**
    - Los valores perdidos se relacionan con los valores de las otras variables dentro de la base de datos.
  - **Not Missing at Random or Nonignorable (NMAR):**
    - Los valores perdidos dependen del valor de la variable.

Técnicas Populares de Tratamiento

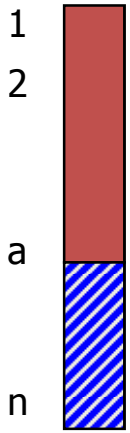
1. Eliminación de datos:
   - Eliminación de Casos (listwise or casewise deletion)
   - Eliminación de pares (o tuplas) de casos (pairwise data delection)

- Donde encontrarlo: La mayoría de paquetes estadísticos, SAS, SPSS, etc.

- Cuando Ocuparlo ➔ MCAR

Técnicas Populares de Imputation

2. Sustitución por la media (mediana y moda):



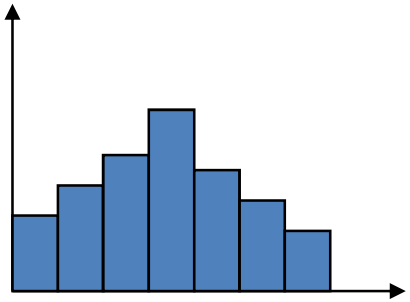
Observados

$y_1, y_2, \ldots, y_a$

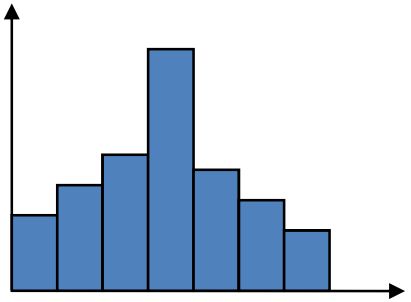
$y_{a+1}, \ldots, y_n$ Perdidos

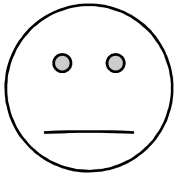
Reemplazar con $y_{obs} = \frac{1}{a}\sum_{i=1}^{a} y_i$

Antes



Después



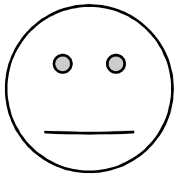


- Corrompe la distribución de Y

Técnicas Populares de Imputation

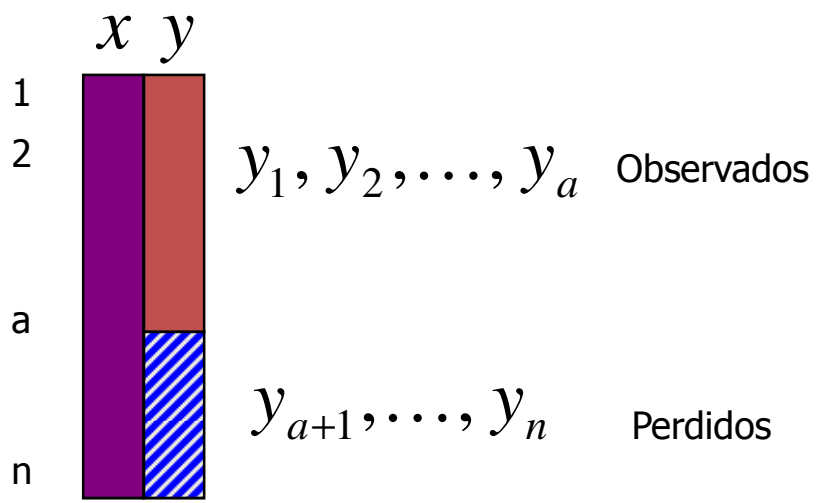
3. Simple Hot Deck:

- Reemplaza los valores perdidos con un valor aleatorio obtenido de la distribución de probabilidades de la variable.
- Preserva la distribución de la variable.
- Distorsiona las correlaciones y covarianzas.

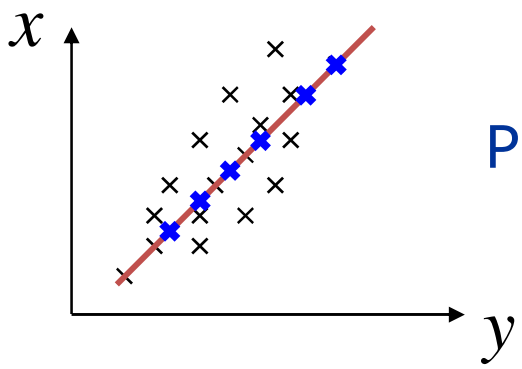


Técnicas Populares de Imputation

4. Métodos de Regresión:



- Reemplazar los valores perdidos con un valor obtenido a través de un modelo de regresión

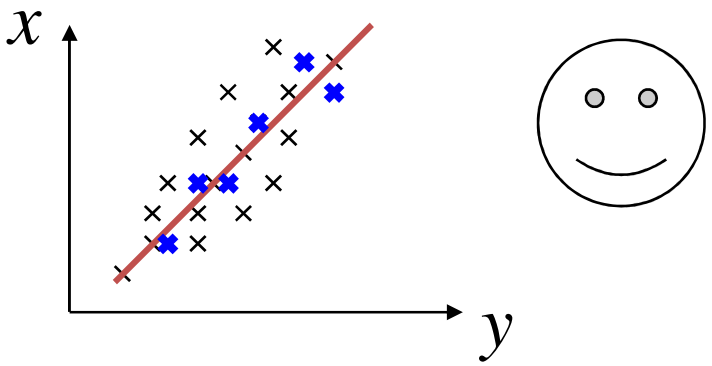


Problema: Esto aumenta las correlaciones

Técnicas Populares de Imputation

4. Métodos de Regresión:

Mejor idea: Reemplazar los valores perdidos con un valor obtenido a través de un modelo de regresión más los residuos de éste



- Se requiere un modelo
- Se asume que los datos perdidos no dependen de los valores de y
- Es difícil de ocupar cuando se tiene que todos los campos presentan valores perdidos.