Análisis Exploratorio de Datos

DEPARTAMENTO DE ESTADÍSTICA, INFORMÁTICA Y MATEMÁTICAS – UPNA

Introducción

El análisis exploratorio de datos es el primer paso que se debe dar a la hora de analizar los datos de un experimento

Objetivos

- Detección de errores
- Comprobación de asunciones
- Selección preliminar de modelos adecuados
- Determinar relaciones entre variables de entrada
- Determinar relaciones “bastas” entre variables de entrada y de salida

Categorías

El análisis exploratorio de datos se puede clasificar en 4 grandes subgrupos, que atienden a dos categorías

- Análisis gráfico – Análisis no gráfico
- Análisis univariable – análisis multivariable (normalmente 2 variables)

Habitualmente es muy buena opción analizar el par de variables por separado antes de hacer el análisis de ambas a la vez

Análisis de una variable no gráfico

El objetivo principal es determinar la distribución de la muestra y hacer algunas conclusiones tentativas sobre qué distribución(es) de población son compatibles con la muestra

Además, también se pretende detectar outliers

Análisis de una variable no gráfico

Cuando las variables son categóricas se pueden hacer pocos estudios

- Cuenta de apariciones de cada valor
- Porcentaje de apariciones de cada valor

Análisis de una variable no gráfico

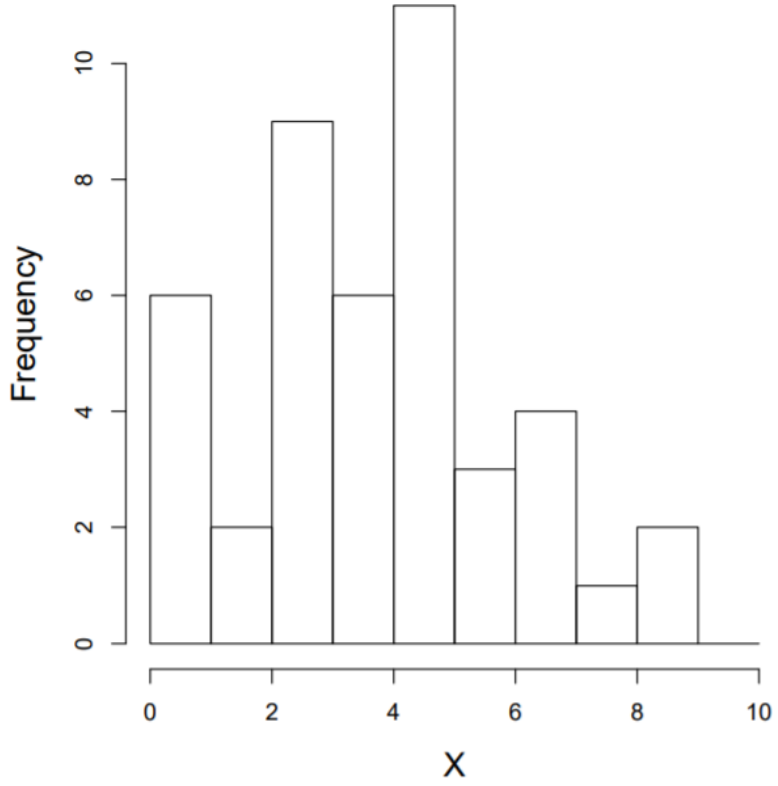
Cuando las variables son cuantitativas se estudia:

- Tendencia central
  - Media
  - mediana si quiero evitar los outliers
- Dispersión
  - varianza o desviación estándar
  - Rango intercuartil (Q3-Q1)
- Oblicuidad
  - Mide la asimetría
- Curtosis
  - Como de “empicado” es frente a una distribución gaussiana

Análisis de una variable gráfico

Histograma

- Ayuda a saber la tendencia central, dispersión, forma y outliers



Análisis de una variable gráfico

Gráficos stem-and-leaf

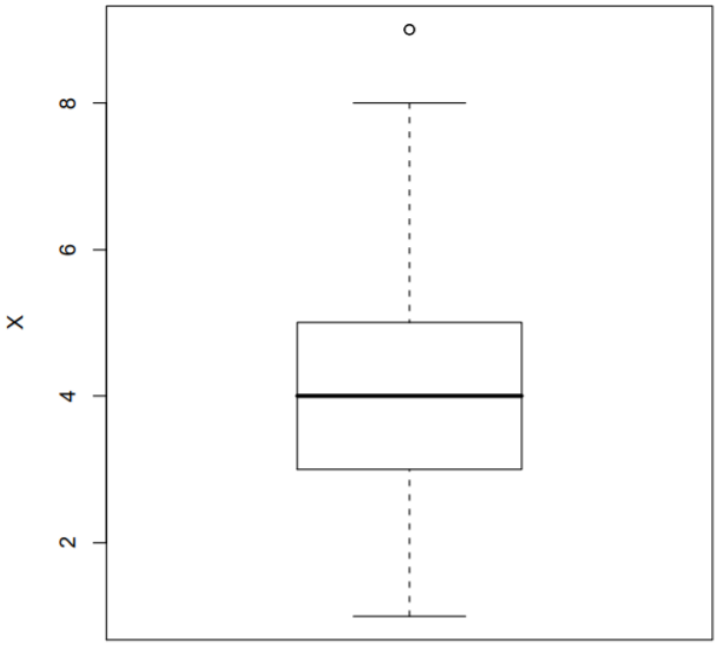
- Muestra todos los valores y la forma

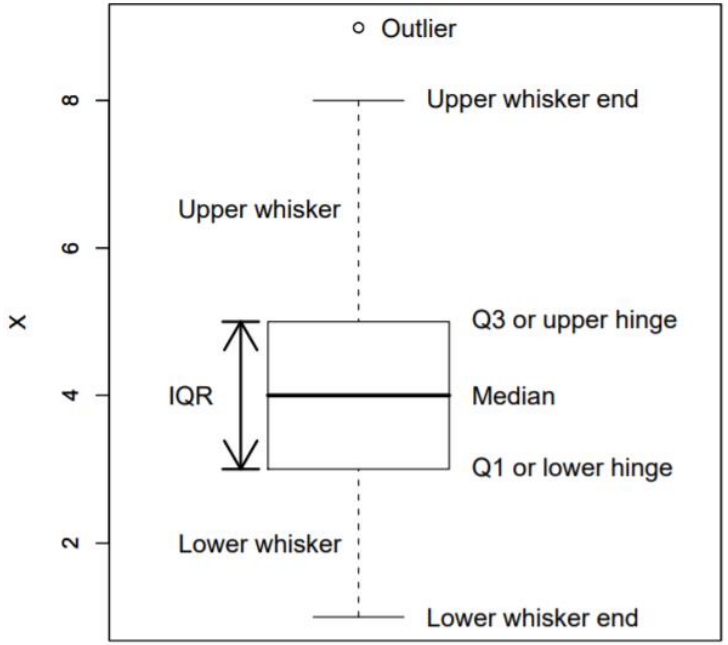
```
The decimal place is at the "|".
1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

Análisis de una variable gráfico

Boxplot

- Es una medida robusta de los valores de los datos y su dispersión. Además, añade información sobre la simetría y los outliers

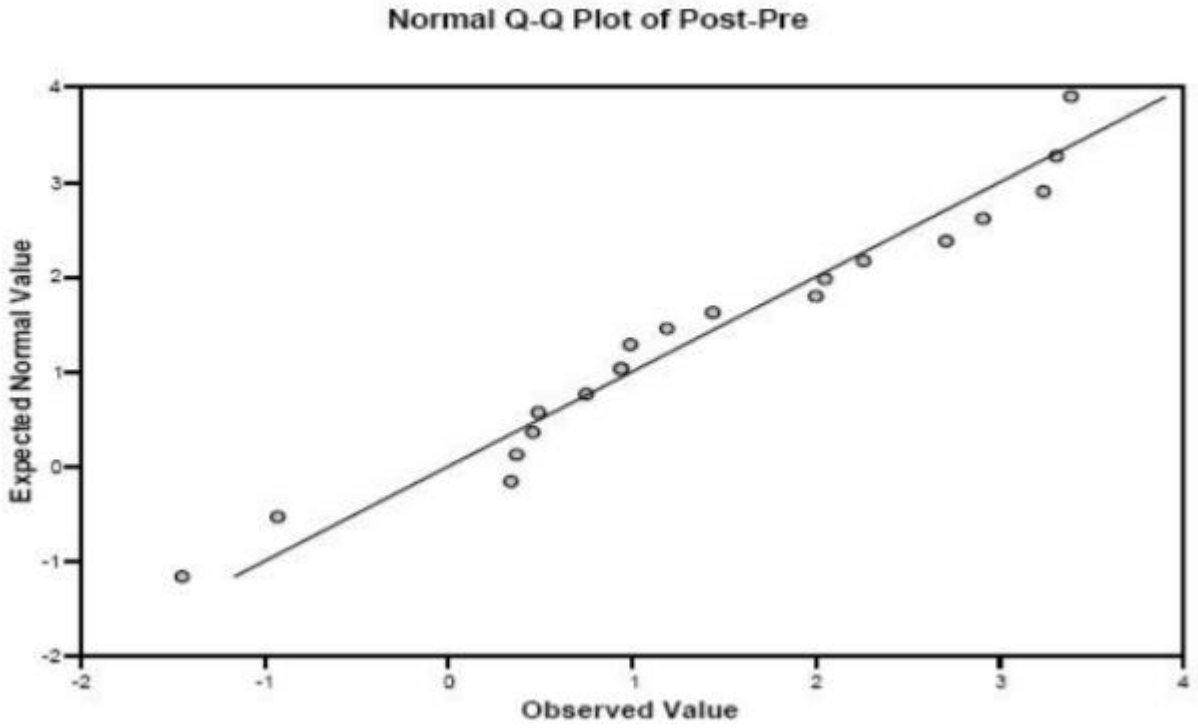




Análisis de una variable gráfico

Gráfico cuantil-normal

- Miden cuánto se parece una muestra a una distribución teórica (poblacional)



Análisis de varias variables no gráfico

Tabla de contingencia

- Es la técnica básica (funciona para cuantitativas y cualitativas) . Consiste en hacer recuento de apariciones de parejas de datos

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Análisis de varias variables no gráfico

Correlación y covarianza

- Para variables cuantitativas
- Covarianza: cuánto y en qué dirección debemos esperar que una variable cambie cuando cambia la otra
  - Difícil de interpretar los números resultantes (sólo positivos o negativos)

- Correlación (entre -1 y 1) – Coeficiente de correlación de Pearson
  - Independiente de la escala de las variables (covarianza entre desviaciones estándar)

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - -1 correlación negativa lineal perfecta
  - 1 correlación positiva lineal perfecta
  - 0 las variables no están correlacionadas

Análisis de varias variables no gráfico

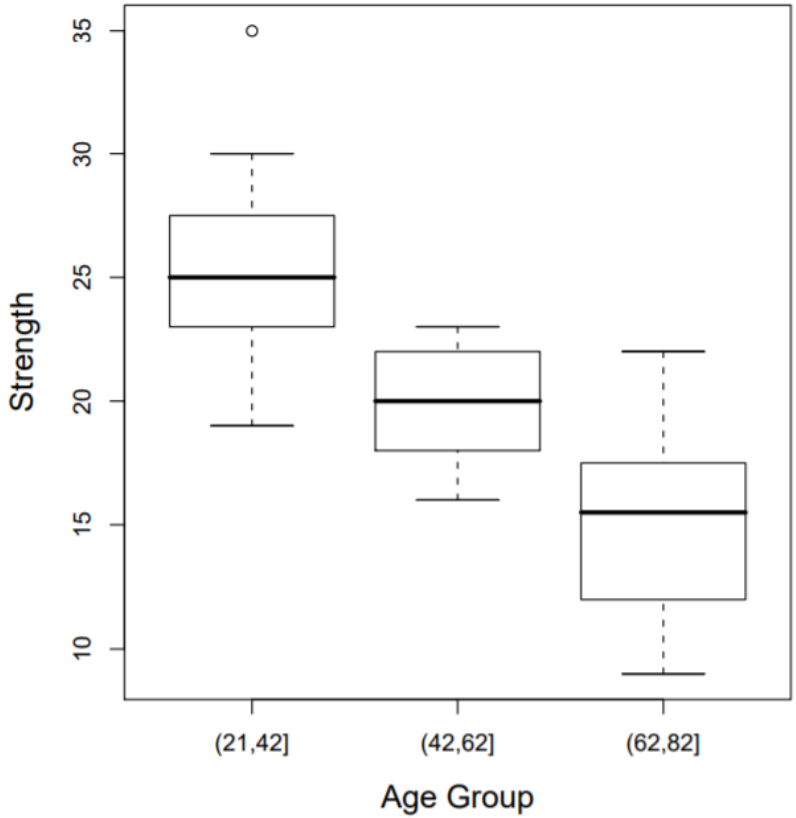
Matrices de correlación y covarianza

- Si tengo varios atributos, calculo la correlación (o covarianza) entre cada par de variables

Análisis de varias variables gráfico

Gráficos individuales para cada categoría de una variable

- Si una variable es categórica y otra numérica, para cada una de las categorías hago un estudio de los datos de la variable numérica y los comparo uno junto a otro



Análisis de varias variables gráfico

Gráficos de dispersión

- Si las dos variables son numéricas muestro una frente a la otra
- Si una variable es medida y la otra es de salida, la de salida debe ir en el eje y