Fundamentos para al Análisis de Datos y la Investigación

Tema 2 - Análisis Exploratorio de Datos (1/2)

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

1 Introducción

2 Los datos

3 Análisis exploratorio

4 Análisis exploratorio gráfico univariado

5 Gráficos con ggplot2

1 Introducción

Contenidos:

- Los datos
- Análisis exploratorio de datos univariados
- Casos Prácticos con R

Contents

1 Introducción

2 Los datos

3 Análisis exploratorio

4 Análisis exploratorio gráfico univariado

5 Gráficos con ggplot2

2 Los datos

- *Cuantitativos o numéricos*: se expresan numéricamente. Pueden ser continuos (precios, renta, rendimientos, etc) o discretos (número de hijos, sectores económicos, etc).
- *Cualitativos, categóricos o factores*: expresan cualidades (color de ojos, grupos de edad) y para su tratamiento se codifican. Pueden ser de tipo nominal o bien ordinal (se puede establecer una ordenación).
- A su vez, los datos puede ser *univariantes o multivariantes*, si se mide más de una característica.
- Además, podemos trabajar con información *primaria* o *secundaria*, ya tratada.
- Otra división de los datos es datos de corte transversal, temporal o paneles de datos.

2 Estructura de Datos

- Trabajamos con *datos estructurados*
- Tipos de datos estructurados: numéricos y categóricos
- Datos rectangulares: marco de referencia básico en la ciencia de datos
- Estructuras de datos no estructurados (no rectangulares): datos espaciales, datos en forma de gráficos o redes, etc.

Contents

1 Introducción

2 Los datos

3 Análisis exploratorio

4 Análisis exploratorio gráfico univariado

5 Gráficos con ggplot2

3 Análisis exploratorio

- El análisis exploratorio de datos univariados (univariante) utiliza cuatro tipo de medidas-resumen:
  - Posición o localización
  - Dispersión
  - Forma
  - Concentración
- Partimos de un conjunto de datos cuantitativos $x_1, \ldots, x_n$.
- Nuestro objetivo será resumir la información contenida en este conjunto de datos.

3 Medidas de posición

- Medidas de posición o localización: son valores típicos o representativos de la variable en estudio.
- *Media aritmética*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Es la más usada. Sin embargo es muy sensible a valores atípicos y observaciones extremas.

- *Mediana*: considerando los datos ordenados de menor a mayor, la mediana es el valor que deja a izquierda y derecha el mismo número de observaciones (puede no ser única). Es menos sensible que la media a valores atípicos y valores extremos.

3 Medidas de posición

- *Cuantiles*: El cuantil de orden p (x_p) es el valor que deja a la izquierda un p por ciento de las observaciones, tomando los datos ordenados de menor a mayor.
- *Casos particulares*. Cuartiles: tres valores que dividen los datos en cuatro partes (Q_1: deja a la izquierda el 25 % de las observaciones; Q_2: mediana: deja a la izquierda el 50 % de las observaciones; Q_3: deja a la izquierda el 75 % de las observaciones. Deciles: nueve valores que dividen los datos en diez partes. Percentiles: 99 valores que dividen los datos en cien partes.
- Otras medidas de posición:
  Media geométrica: `psych:: geometric.mean(x)`;
  Media armónica: `psych:: harmonic.mean(x,na.rm=TRUE,zero=TRUE)`
- *Medias robustas*: media recortada al 10 %, (media aritmética excluyendo el 10 % de los mayores y el 10 % de los menores) `mean(x,trim=0.1)`

3 Medidas de dispersión y forma

- Las medidas de dispersión: miden la representatitividad de un promedio.
- *La varianza*:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  Media de los cuadrados de las distancias de los datos a la media. La *desviación típica* $S_X = \sqrt{S_X^2}$ tiene las mismas unidades que la variable. Ambas medidas de dispersión son muy sensibles a los valores extremos. No es posible comparar la dispersión de dos variables medidas en diferentes unidades.
- *Coeficiente de variación*: Es una medida de dispersión relativa, que permite la comparación de la dispersión de dos variables medidas en distintas unidades:

$$\mathrm{CV}(X) = \frac{S_X}{\bar{x}}$$

3 Medidas de dispersión y forma

- Otras medidas. *Rango* $R = x_{(n)} - x_{(1)}$
- *Recorrido intercuartílico* y *semi-intercuartílico*:

$$R_I = Q_3 - Q_1, \quad RSI = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

La principal diferencia entre ambas es que la segunda permite comparar la dispersión de dos variables independientemente de la escala.

3 Medidas de dispersión y forma

- Las medidas de forma incluyen los coeficientes de asimetría y de curtosis.
- *Coeficiente de asimetría de Pearson*:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S_X^3}$$

Si $g_1 = 0$, la distribución es simétrica; si $g_1 < 0$, la distribución es asimétrica negativa y si $g_1 > 0$, la distribución es asimétrica positiva.

- *Coeficiente de curtosis*: concentración de valores en torno a la media y mide el grado de apuntamiento de la distribución con respecto a la distribución normal:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{S_X^4} - 3$$

Si $g_2 = 0$ la distribución es mesocúrtica; si $g_2 < 0$, la distribución es platicúrtica; si $g_2 > 0$, la distribución es leptocúrtica.

3 Paquetes de R de estadística descriptiva

- En el módulo base están las funciones `summary(x)` y `fivenum(x)`
- Paquetes de estadística descriptiva
  - `libray(psych) :: describe(x)`
  - `library(moments) :: skewness(x); kurtosis(x)`
  - `library(pastecs) :: stat.desc(x)`

3 Medidas de concentración

- Las medidas de concentración estudian cómo se reparte el total de valores de una variable. Las medidas de concentracion más importantes son la curva de Lorenz y el índice de Gini.
- Se usan en el estudio de la desigualdad económica y como *métricas de evaluación y selección* en algunos métodos en CD.
- Disponemos de un conjunto de datos (x_i, n_i), donde $i = 1, 2, \ldots, k$ donde cada pareja representa un conjunto de renta, ordenados de menor a mayor $x_1 \leq x_2 \leq \cdots \leq x_n$.
- La *curva de Lorenz* se construye a partir de la poligonal de datos:

$$(p_i, q_i) = \left(\frac{n_1 + \cdots + n_i}{n}, \frac{x_1 n_1 + \cdots + x_i n_i}{\sum_{i=1}^{k} x_i n_i} \right)$$

con $i = 1, 2, \ldots, k$

3 Medidas de concentración

- El *índice de Gini* está definido por:

$$G = \frac{\sum_{i=1}^{k}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

donde $IG = 0$ significa equidistribución y $IG = 1$ máxima concentración (un sólo individuo se reparte el total de la renta). A partir de datos individuales, el índice de Gini se define de modo alternativo como:

$$G = \frac{\frac{1}{n^2}\sum_{i \neq j}|x_i - x_j|}{2\bar{x}}$$

- Paquetes de R para el cálculo de índices de concentración:
  - Paquete `ineq` `https://cran.r-project.org/web/packages/ineq/ineq.pdf`
  - Paquete `lawstat` `https://cran.r-project.org/web/packages/lawstat/lawstat.pdf`

Contents

1 Introducción

2 Los datos

3 Análisis exploratorio

4 Análisis exploratorio gráfico univariado

5 Gráficos con ggplot2

4 Análisis exploratorio gráfico

- Tenemos diversas alternativas que es necesario conocer
- Gráficos haciendo uso del **módulo base**. Se puede realizar cualquier tipo de gráfico
- Gráficos mediante paquete **lattice** (autor: Deepayan Sarkan). Permite realizar gráficos múltiples
- Gráficos mediante paquete **ggplot2** (autor: Hadley Wickham). Similar al anterior, pero sin duda el más extendido. Imprescindible para el data scientist. Se basa en la denominada *gramática de los datos*.

4 Análisis exploratorio gráfico

- Como primer aspecto a tener en cuenta, el *análisis exploratorio gráfico* es un *aspecto clave y complementario* al análisis exploratorio puramente numérico.
- Podemos realizar análisis gráfico en variables categóricas y en variables cuantitativas.
- Comenzaremos con el análisis gráfico en variables categóricas:
  - Las variables categóricas o factores se tabulan en R mediante `table(x)`
  - Para la representación gráfica de variables categóricas se usan diagramas de barras o diagramas circulares: `barplot(x)`, `pie(x)`.
  - En lenguaje R, en muchas ocasiones es necesario declarar a la variable categórica como factor.

4 Representaciones gráficas de variables numéricas

- Continuamos con el análisis gráfico en variables numéricas.
- Comenzamos con los *histogramas*, que en `R` se obtienen mediante `hist(x)`. Los histogramas permiten representar variables cuantitativas. Para ello, se divide el rango de los datos en intervalos y se cuentan los datos en cada intervalo. En `R` es posible, entre otras cosas, elegir el número de intervalos.
- Presentan varios inconvenientes. Entre ellos, el gráfico pierde información, depende del número de intervalos y puede ser difícil de interpretar.

4 Diagramas de caja

- Los *diagramas de caja o box-plot* se utilizan para representar gráficamente datos cuantitativos. Estos gráficos de basan en medidas robustas, como son los cuartiles
- Estos gráficos informan simultánemante sobre: Medidas de localización robustas; Medidas de dispersión robustas; Outliers de los datos.
- Los diagramas de caja se construyen a partir de los cuartiles y de los límites:

$$\begin{aligned} LI &= Q_1 - 1{,}5(Q_3 - Q_1) \\ LS &= Q_3 + 1{,}5(Q_3 - Q_1) \end{aligned}$$

- Los diagramas de cajas permiten comparar varias muestras fácilmente
- Permiten representar una variable (cuantitativa) según los niveles de una factor (por ejemplo, género)

4 Diagramas de caja

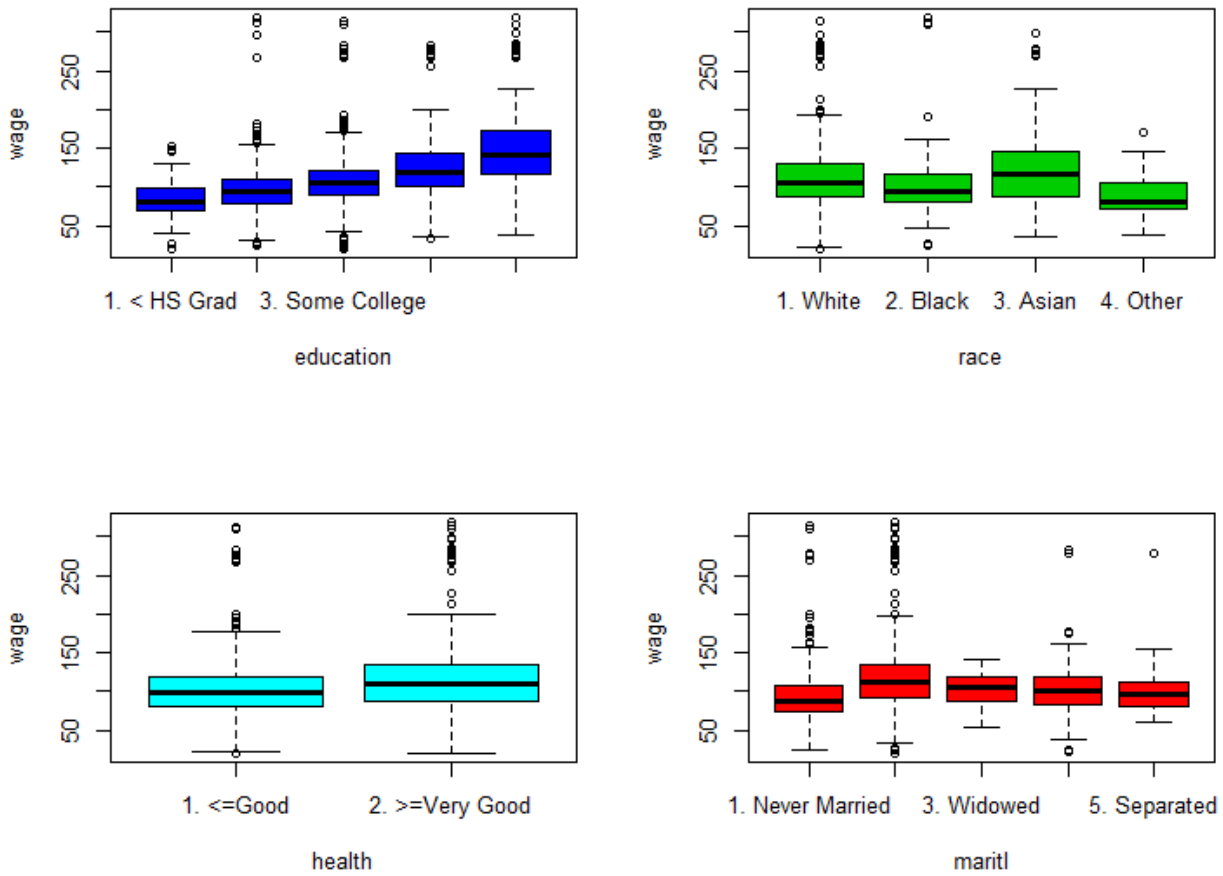

Figura: Ejemplos de box plot

4 Estimadores de la densidad tipo nucleo

- Con objeto de mejorar el histograma, se dispone de los *estimadores de la función de densidad tipo nucleo (kernel)*. Vienen dados por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

  donde $K(\cdot)$ es la función nucleo, n el número de datos y h el ancho de banda (bandwidth), que se obtiene minimizando el tipo de error MISE.

- Se obtiene por medio de la fórmula: $h = (4s_j^5/3n)^{1/5}$, donde s_j es la desviación típica de los datos.

4 Estimadores de la densidad tipo nucleo

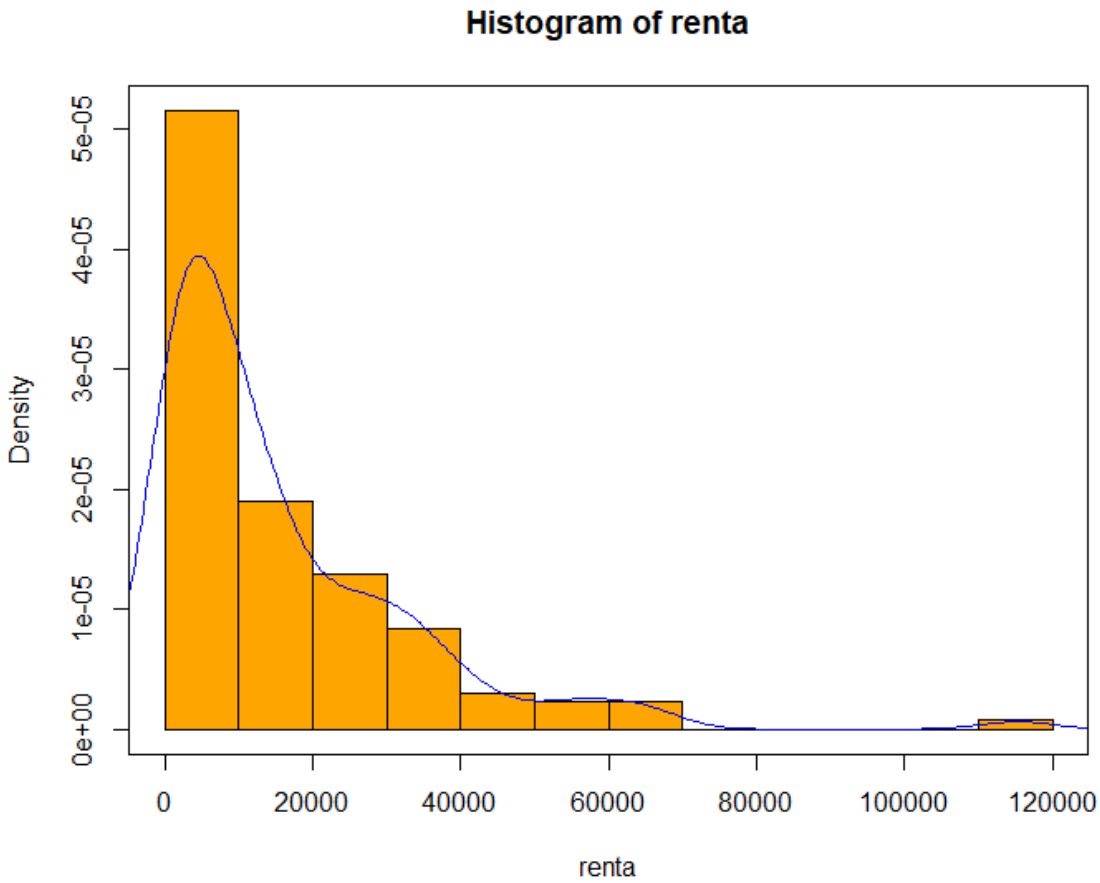

Figura: Histograma y estimación tipo nucleo de la variable renta

4 Estimadores de la densidad tipo nucleo

- R permite siete tipo de nucleos (kernel) diferentes. *Algunos de los más utilizados son*:
  - Kernel Gaussiano $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
  - Kernel Epanechnikov $K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$
  - Kernel uniforme $K(u) = \frac{1}{2}I(|u| \leq 1)$
  - Kernel triangular $K(u) = (1-|u|)I(|u| \leq 1)$
- Como es claro, el aspecto del gráfico depende de las elecciones de h y de la función nucleo

4 Violin plots

- Los *gráficos de violín* (violin plots) son otro tipo de gráfico para datos cuantitativos
- Son parecidos a los box-plots, pero además incluyen los diagramas de kernel en cada lado rotados
- Permiten captar datos de naturaleza bimodal.

4 Violin plots

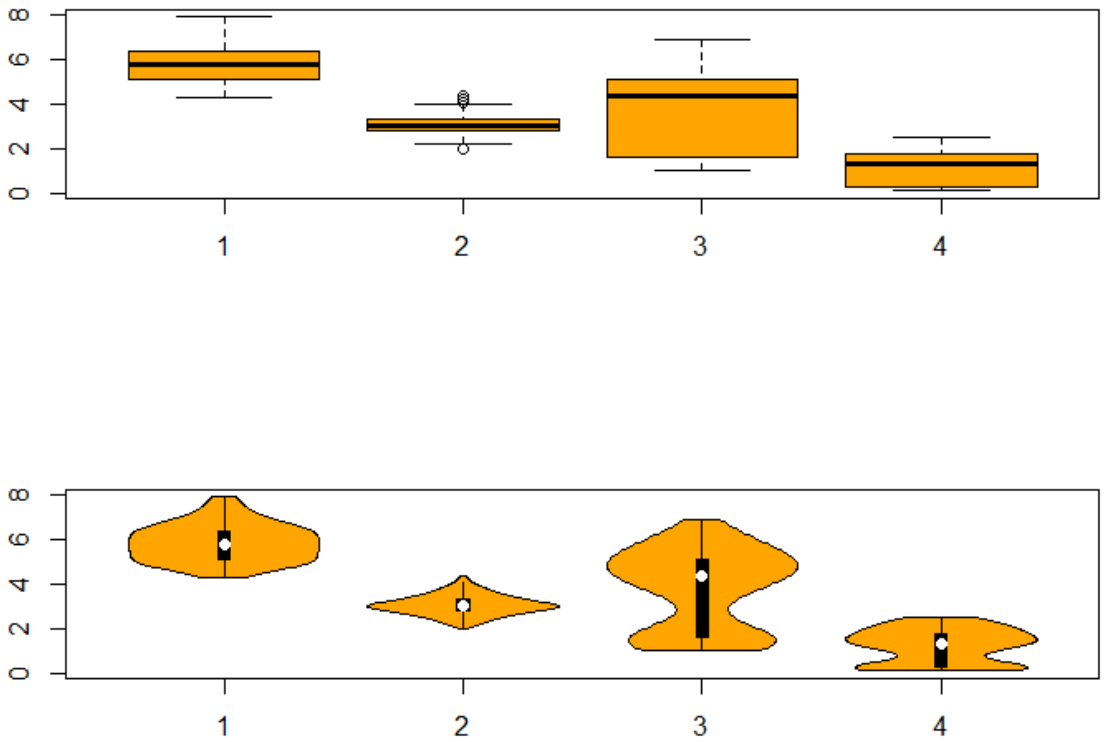

Figura: Ejemplos de violin plot

Contents

1 Introducción

2 Los datos

3 Análisis exploratorio

4 Análisis exploratorio gráfico univariado

5 Gráficos con ggplot2

5 Gráficos en ggplot2

- Proporciona gráficos profesionales: *grammar of graphics*
- La estructura de código es:

$$\textbf{ggplot}(\textbf{datos}, \textbf{aes}()) + \textbf{geom\_tipo}()$$

  de modo que se realiza la configuración a través de capas, donde cada capa representa una característica.
- Pasamos a detallar

5 Gráficos en ggplot2

Elementos de los gráficos ggplot2:

1. **Datos**: elemento más importante. Sólo se acepta estructuras **data.frame**
2. **Estética**: La estética del gráfico se refiere al color, ejes x e y, forma de los puntos, alpha (intensidad) etc. El argumento es **aes()**. La estética indicada al principio se hereda en el resto de las capas geom
3. **Capas**: las capas o geom son los verbos del ggplot2. Las capas se van superponiendo y se añaden con el signo+: geom_point(), geom_smooth() etc. Las capas xlab(), ylab() y ggtitle() se refieren a los títulos.
   Otros tipos de capas son: geom_point, geom_line, geom_histogram, geom_boxplot etc
4. **Temas**: La capa theme modifica los aspectos estéticos del gráfico que no tienen que ven con los datos: ejes, fondo, márgenes etc.
5. Se adjunta **cheat sheet** del ggplot2.

Fundamentos para al Análisis de Datos y la Investigación

Tema 2 - Análisis Exploratorio de Datos (1/2)

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad