Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

CIMAT

Visualización univariada

- Objetivo: estudiar la **distribución de una sola variable**.
- Preguntas típicas:
  - ¿Dónde se concentran los datos? (medidas de tendencia central).
  - ¿Qué tan dispersos están? (varianza, RIQ).
  - ¿Existen asimetrías o colas largas?
  - ¿Hay valores atípicos visibles?
- Herramientas principales:
  1. Histogramas.
  2. Estimaciones de densidad (kernel).
  3. Boxplots.

Histogramas

Definición

Dada una partición del rango de datos en intervalos I_j, el histograma estima:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}, \quad h = \text{ancho de clase.}$$

- Aproximación empírica de la densidad subyacente.
- El ancho de clase h afecta el balance:
  - h pequeño $\rightarrow$ gráfico ruidoso (alta varianza).
  - h grande $\rightarrow$ gráfico sobre-suavizado (alto sesgo).
- Reglas prácticas: $h \approx 2 \times IQR(x)\, n^{-1/3}$ (Friedman–Diaconis).

Referencias: Freedman & Diaconis (1981); Scott (1979).

Propiedades del histograma

- **No negatividad:**
$$\hat{f}_h(x) \geq 0 \quad \forall x.$$
- **Normalización:**
$$\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = 1.$$
- **Consistencia:** Si $h \to 0$ y $nh \to \infty$, entonces $\hat{f}_h(x) \to f(x)$ en probabilidad.
- **Sesgo–varianza:**
  - h muy grande $\Rightarrow$ bajo varianza, alto sesgo (sobre-suavizado).
  - h muy pequeño $\Rightarrow$ baja sesgo, alta varianza (ruidoso).

Referencias: Freedman & Diaconis (1981); Scott (1979); Silverman (1986).

Estimación de densidad kernel (KDE)

Definición

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

donde $K(\cdot)$ es una función kernel (p. ej., gaussiano).

- Suaviza la distribución en lugar de usar cortes discretos.
- El parámetro h (bandwidth) controla el grado de suavizamiento.
- Comparación:
  - Histograma: dependiente de cortes arbitrarios.
  - KDE: continuidad, mejor para detectar multimodalidad.

Referencia: Silverman (1986).

Propiedades de la estimación de densidad kernel

Definición

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

donde $K(\cdot)$ es una función kernel y $h > 0$ es el parámetro de suavizamiento.

- **No negatividad:** $\hat{f}_h(x) \geq 0$ si $K \geq 0$.
- **Normalización:**
$$\int_{-\infty}^{\infty} \hat{f}_h(x)\, dx = 1.$$
- **Consistencia:** Si $h \to 0$ y $nh \to \infty$, entonces $\hat{f}_h(x) \to f(x)$ en probabilidad.
- **Elección de h:** controla el balance sesgo–varianza.
  - h pequeño $\Rightarrow$ captura detalles finos, pero alta varianza.
  - h grande $\Rightarrow$ suaviza más, pero introduce sesgo.
- **Kernel:** típicamente K es gaussiano, uniforme, Epanechnikov, etc.

Referencias: Silverman (1986); Wand & Jones (1995).

Boxplots

Definición

El boxplot resume la distribución con:

- Mediana Q_2.
- Rango intercuartílico $RIQ = Q_3 - Q_1$.
- "Bigotes": $[Q_1 - 1.5\,RIQ,\ Q_3 + 1.5\,RIQ]$.
- Puntos fuera de los bigotes $\rightarrow$ posibles outliers.

- Conecta visualización con medidas robustas (mediana, cuartiles).
- Permite comparar distribuciones de distintas escalas en paralelo.

Referencias: Tukey (1977); Rousseeuw & Leroy (1987).

Boxplot: Propiedades

- **Medida de robustez:** basado en cuartiles, es poco sensible a valores extremos (a diferencia de la media y desviación estándar).
- **Identificación de outliers:** utiliza el rango intercuartílico (IQR) para definir umbrales:

$$\text{Límite inferior} = Q_1 - 1.5 \times IQR, \quad \text{Límite superior} = Q_3 + 1.5 \times IQR.$$

- **Comparación de distribuciones:** facilita la visualización de simetría, sesgo y dispersión entre varios grupos.
- **Información resumida:** muestra mediana, cuartiles y posibles observaciones atípicas en una sola gráfica.
- **Invarianza:** es invariante a transformaciones afines positivas (traslación y escala).

Referencias: Tukey (1977); Rousseeuw & Leroy (1987); McGill, Tukey & Larsen (1978).

Visualización bivariada

- Objetivo: analizar la **relación entre dos variables**.
- Preguntas típicas:
  - ¿Existe asociación lineal o no lineal?
  - ¿Cuál es la dirección y magnitud de la correlación?
  - ¿Existen subgrupos o patrones ocultos?
- Herramienta principal: **scatterplot** (diagrama de dispersión).
- Complemento: medidas estadísticas como $\mathrm{Cov}(X, Y)$ y $\rho(X, Y)$.

Scatterplot

- Cada punto representa un par (x_i, y_i).
- Permite visualizar:
  - Tendencia general (lineal o curvilínea).
  - Concentración o dispersión de puntos.
  - Subgrupos según categorías (usando color o símbolo).
- Relación con estadística:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{s_x s_y}.$$

Referencia: Anscombe (1973); Cleveland (1993).

Visualización multivariada

- Objetivo: explorar relaciones simultáneas entre varias variables.
- Preguntas clave:
  - ¿Qué pares de variables muestran mayor asociación?
  - ¿Existen patrones de agrupamiento o subpoblaciones?
  - ¿Qué tan redundantes son algunas variables?
- Herramientas principales:
  1. Matriz de diagramas de dispersión (pairplot).
  2. Mapas de calor de correlación.

Pairplot (matriz de dispersión)

- Construye un conjunto de scatterplots para todas las combinaciones de variables.
- Permite detectar:
  - Correlaciones lineales y no lineales.
  - Distribuciones marginales (en la diagonal).
  - Agrupamientos o patrones según categorías (colores o símbolos).
- Conexión estadística:
  - Cada celda resume la relación bivariada $\leftrightarrow$ covarianza/correlación.
  - La diagonal resume la distribución univariada $\leftrightarrow$ histograma o KDE.

Referencia: Tukey (1977); Cleveland (1993).

Mapa de calor de correlación

- Resume visualmente la matriz de correlaciones:

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{s_{X_i} s_{X_j}}, \quad -1 \leq \rho_{ij} \leq 1.$$

- Colores representan magnitud y signo de la correlación.
- Útil para:
  - Identificar variables redundantes.
  - Detectar bloques de alta asociación (posibles subestructuras).
  - Guiar selección de variables en modelos posteriores.
- Precaución: la correlación mide relación **lineal**, no capta dependencias no lineales.

Referencia: Friendly (2002), *Corrgrams: Exploratory displays for correlation matrices*.

Limitaciones y buenas prácticas

- **Sobreposición**: scatterplots en alta dimensión son difíciles de leer $\rightarrow$ usar transparencia o submuestreo.
- **Percepción del color**: elegir paletas perceptualmente uniformes para heatmaps.
- **Escala**: siempre revisar si las variables fueron estandarizadas antes de interpretar correlaciones.
- **Complementariedad**: usar pairplots para detectar patrones y heatmaps para resumir toda la estructura.

Referencias: Cleveland (1993); Wilkinson (2005).

Referencias: Escalamiento y Normalización

- T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- G. James, D. Witten, T. Hastie, R. Tibshirani (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.
- C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer.
- I. T. Jolliffe, J. Cadima (2016). "Principal component analysis: a review and recent developments". *Phil. Trans. Royal Society A*, 374(2065), 20150202.
- M. Kuhn, K. Johnson (2013). *Applied Predictive Modeling*. Springer.

Referencias: Escaladores robustos

- P. J. Rousseeuw, A. M. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley.
- P. J. Huber, E. M. Ronchetti (2009). *Robust Statistics* (2nd ed.). Wiley.
- G. E. P. Box, D. R. Cox (1964). "An Analysis of Transformations". *Journal of the Royal Statistical Society B*, 26(2), 211–252.
- I. K. Yeo, R. A. Johnson (2000). "A New Family of Power Transformations to Improve Normality or Symmetry". *Biometrika*, 87(4), 954–959.

Referencias: Visualización univariada

- J. W. Tukey (1977). *Exploratory Data Analysis*. Addison–Wesley.
- D. Freedman, P. Diaconis (1981). "On the histogram as a density estimator: L^2 theory". *Zeitschrift f. Wahrscheinlichkeitstheorie*, 57, 453–476.
- D. W. Scott (1979). "On optimal and data-based histograms". *Biometrika*, 66(3), 605–610.
- B. W. Silverman (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.
- R. McGill, J. W. Tukey, W. A. Larsen (1978). "Variations of Boxplots". *The American Statistician*, 32(1), 12–16.

Referencias: Visualización bivariada y multivariada

- F. J. Anscombe (1973). "Graphs in Statistical Analysis". *The American Statistician*, 27(1), 17–21.
- W. S. Cleveland (1993). *Visualizing Data*. Hobart Press.
- L. Wilkinson (2005). *The Grammar of Graphics* (2nd ed.). Springer.
- M. Friendly (2002). "Corrgrams: Exploratory displays for correlation matrices". *The American Statistician*, 56(4), 316–324.
- J. VanderPlas (2016). *Python Data Science Handbook*. O'Reilly.