ITESO, Universidad
Jesuita de Guadalajara

# Identificación de datos

Dr. Gaddiel Desirena López

Primavera 2024

# Contenido

# Variables numéricas y variables categóricas

Variable categórica (cualitativa): Contienen un número finito de categorías o grupos distintos.

Variable numérica (cuantitativa): Los valores son números que suelen representar un control o una medición.

## Tipos de características

Categórica: El dominio es un conjunto de valores discretos.

Ordinal: El dominio es el conjunto de valores ordenados.

Numérica: El dominio es el conjunto de valores numéricos. La característica numérica puede ser continua o discreta. La característica numérica puede ser escalada: $u = 2v$.

# Variables numéricas y variables categóricas

Tabla 1: Datos demográficos

| Nombre | Edad | Sexo | Estudios |
|---|---|---|---|
| Fernando | 32 | Masculino | Maestría |
| Karen | 32 | Femenino | Maestría |
| Rosario | 58 | Femenino | Secundaria |
| Fernando | 59 | Masculino | Preparatoria |
| Carlos | 31 | Masculino | Doctorado |
| Marlene | 31 | Femenino | Mestría |
| Martín | 25 | Masculino | Licenciatura |

# Valores faltantes

Los datos faltantes no son raros en conjuntos de datos reales. De hecho, la probabilidad de que falte al menos un punto de datos aumenta a medida que aumenta el tamaño del conjunto de datos. Los datos faltantes pueden ocurrir de varias formas, algunas de las cuales incluyen las siguientes.

Fusión en la fuente de datos: un ejemplo sencillo suele ocurrir cuando dos conjuntos de datos se combinan mediante un identificador de muestra (ID). Si una ID está presente solo en el primer conjunto de datos, entonces los datos combinados contendrán valores faltantes para esa ID para todos los predictores en el segundo conjunto de datos.

# Valores faltantes

Eventos aleatorios: cualquier proceso de medición es vulnerable a eventos aleatorios que impiden la recopilación de datos. Por ejemplo, si una batería se agota o el dispositivo de recolección está dañado, las mediciones no se pueden recolectar y faltarán en los datos finales.

Fallos de medición: por ejemplo, las mediciones basadas en imágenes requieren que una imagen esté enfocada. Otro ejemplo de falla en la medición ocurre cuando un paciente en un estudio clínico pierde una visita médica programada. Las mediciones que se hubieran tomado para el paciente en esa visita faltarían en los datos finales.

## Tipos de valores faltantes

Deficiencias estructurales en los datos: se define como un componente faltante de un predictor que se omitió de los datos. Este tipo de falta es a menudo el más fácil de resolver una vez que se identifica el componente necesario.

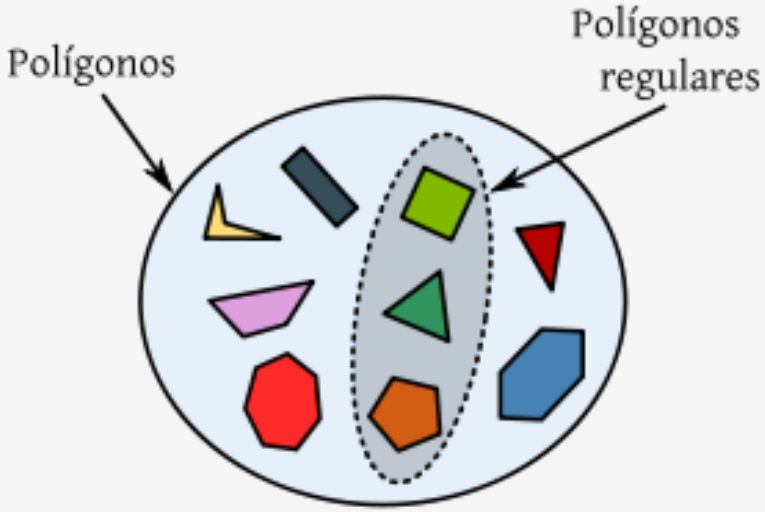
Por un caso específico o suceso no aleatorio: Este tipo de datos faltantes son los más difíciles de manejar.

# Valores faltantes

Sucesos aleatorios: Éste se subdivide en dos categorías:

- Datos perdidos completamente al azar: la probabilidad de que falte un resultado es igual para todos los puntos de datos (observados o no observados). En otras palabras, los valores perdidos son independientes de los datos. Esta es la mejor situación.
- Datos faltantes al azar: la probabilidad de que falten resultados no es igual para todos los puntos de datos (observados o no observados). En este escenario, la probabilidad de que falte un resultado depende de los datos observados pero no de los datos no observados.

# Cardinalidad de variables categóricas




La cardinalidad de un conjunto es la medida del "número de elementos en el conjunto". Por ejemplo, el conjunto $A = \{2, 4, 6\}$ contiene 3 elementos, y por tanto $A$ tiene cardinalidad 3. La cardinalidad de un conjunto $A$ usualmente se denota $|A|$ o como $\#A$.

## Características de color

Una manera de encontrar la cardinalidad de color es

# Cardinalidad de variables categóricas

## Características de color

Una manera de encontrar la cardinalidad de color es

- convertir la imagen (R,G,B) a un solo canal RGB.
- Cuantizar los datos obtenidos.

## Características de color

Una manera de encontrar la cardinalidad de color es

- convertir la imagen (R,G,B) a un solo canal RGB.
- Cuantizar los datos obtenidos.

Ej:
- color1=0x000000-0x000010
- color2=0x000010-0x000020
- ...

# Cardinalidad de variables categóricas

## Características de color

Una manera de encontrar la cardinalidad de color es

- convertir la imagen (R,G,B) a un solo canal RGB.
- Cuantizar los datos obtenidos.

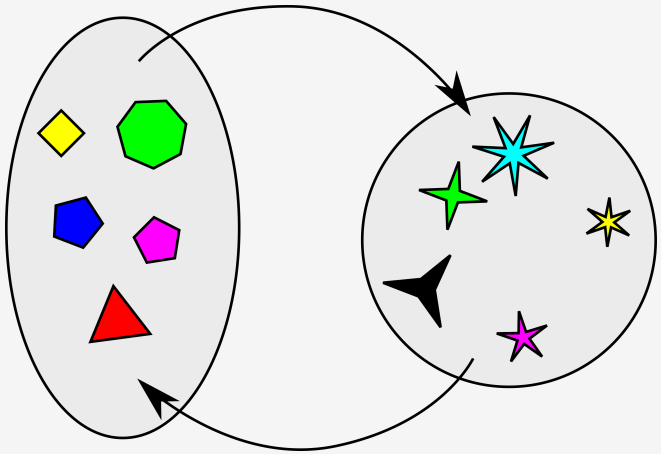
Ej:
- color1=0x000000-0x000010
- color2=0x000010-0x000020
- ...

## Características de textura

Una de las formas de modelar estas características es la matriz de co–ocurrencia en escala de grises definida como

$$M_{i,j} = \{\#[I_{i,j}] | I_{i,j} = I_{i+1,j+1}\}$$

# Relaciones lineales



- Cuantifica como se vinculan dos variables

  $$Y = f(X), \quad f : R^p \to R^q$$
  $$Y = WX, \quad W \in R^{q \times p}$$

- Este mapeo es independiente del orden de la variables:

  $$\{x_1, x_2, x_3, \dots\} \to \{y_1, y_2, y_3, \dots\}.$$
  $$\{x_5, x_1, x_6, \dots\} \to \{y_5, y_1, y_6, \dots\}.$$

- Nos permite reformular problemas no–lineales a lineales

  $$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 / x_2^2 + \dots$$
  con $x_1 / x_2^2 = x_3$.

# Relaciones lineales

Por ejemplo, si el conjunto de características $Y$ representa sensaciones de alimentos (picante/caliente, mentolado/fresco) y el conjunto $X$ representa color (rojo, verde y azul)

$$\begin{aligned} Y &= \{1, 2\} \\ X &= \{1, 2, 3\} \end{aligned}$$

La matriz que lo relaciona es

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.55 & 0.3 \end{bmatrix}$$

# Relaciones lineales

- Función exacta

$$f : X \to Y$$

Ejemplo: $X = [1.1,\ 2]$, $Y = [-0.9\ 0]$. Siendo $y = f(x) = mx + b$

  - Se sustituyen los pares de datos que se desean relacionar

| $X$ | | $Y$ | $y = mx + b$ |
|---|---|---|---|
| 1.1 | $\sim$ | -0.9 | $-0.9 = m(1.1) + b$ |
| 2 | $\sim$ | 0 | $0 = m(2) + b$ |

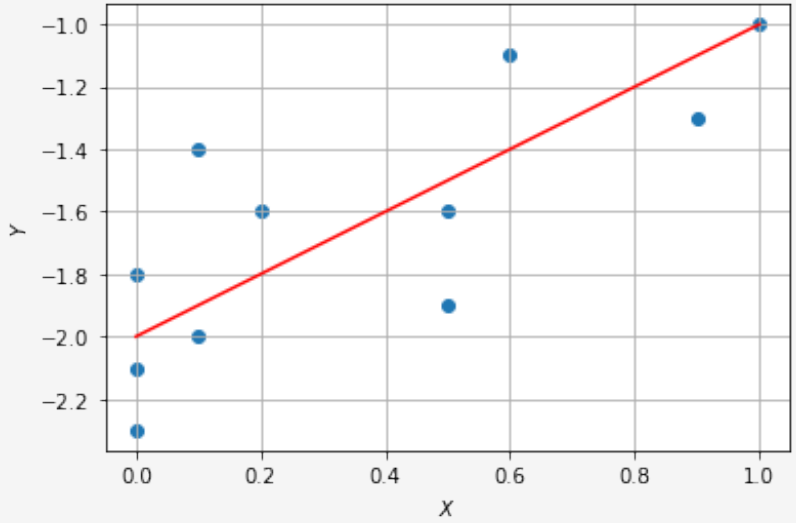
  - Se resuelve para $[m,\ b]^T$

$$\begin{bmatrix} 1.1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} -0.9 \\ 0 \end{bmatrix}$$

resultando $m = 1$ y $b = -2$, por lo que la función es $y = x - 2$.

# Relaciones lineales

Ahora bien, considere tener dos conjuntos de variables
Gráficamente




El objetivo sigue siendo encontrar los valores de $m$ y $b$ en la ecuación $y = mx + b$.

Entonces se busca la **mejor** línea que pase entre los puntos.
Criterio:

$$\frac{1}{N}\sum^{N}(mx + b - y)^2,$$

para cada $y \in Y$ y $x \in X$ con $N$ datos.

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | 4 |
| 2 | |
| 3 | |
| 4 | |

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | |
| 4 | |

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | 2 |
| 4 | |

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | 2 |
| 4 | 1 |

# Distribución de datos

Es la agrupación de datos en diferentes categorías indicando el número de observaciones:

$\{1, 1, 2, 1, 3, 2, 1, 4, 2, 3\}$

| Dato | Observaciones |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | 2 |
| 4 | 1 |

En un conjunto grande de datos, la representación en una tabla resulta poco útil. El objetivo, a partir de un conjunto de datos, es obtener un conjunto pequeño de números que resuman bien a éste a partir de **medidas de posición**, de **dispersión** y de **forma**.

## Medidas de posición o de tendencia central

Estas medidas centralizan la información, también se les conoce como promedios:

- Media aritmética $x = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$
- Media recortada ($\alpha$-trimmed)
- Media ponderada $x = \dfrac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n}$
- Media geométrica $\sqrt[n]{x_1 x_2 \cdots x_n}$
- Media armónica $x^{-1} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i^{-1}$
- Mediana
- Moda
- Cuantil o percentil

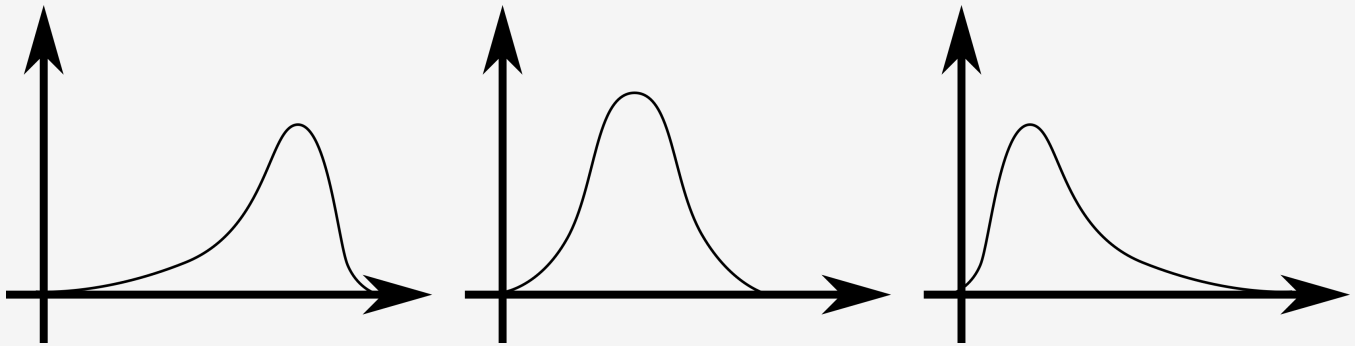
## Medidas de dispersión

Lo procedente es saber qué fiabilidad nos ofrecen esas pocas cantidades o números, es decir, cuánta variabilidad existe en el conjunto de datos. Si hay poca variabilidad, la información de los valores medios será muy precisa. Si existe mucha variabilidad, la información será menos precisa.

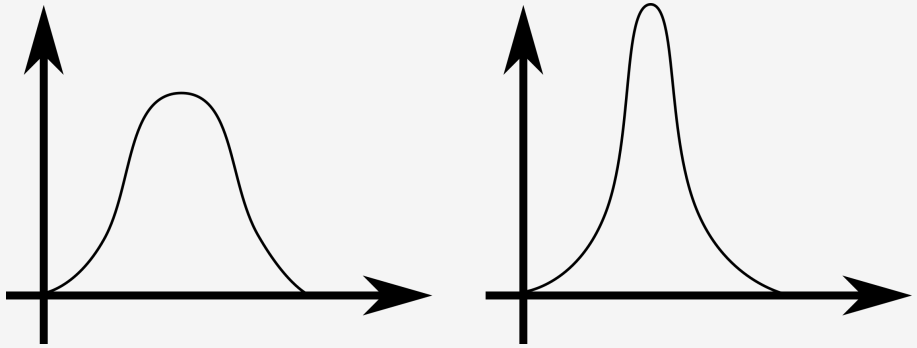
- Varianza y desviación estandar $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$
- Recorrido $\text{máx}(X) - \text{mín}(X)$
- Recorrido intercuartílico $(R_i)$ $Q_3 - Q_1$.
- Coeficiente de variación (Pearson) $cv = \frac{\sigma}{|x|}$

## Medidas de forma

- Simetría: Coeficiente de Fisher $g_1 = \frac{m_3}{\sigma^3}$, $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x - x_i)^3$



- Apuntamiento o Curtosis: $g_2 = \frac{m_4}{\sigma^4}$



## Diagramas de caja

Sirve para visualizar tanto la dispersión como la forma del conjunto de datos.

- A los datos que se encuentren a una distancia de $Q_1$ por la izquierda, o de $Q_3$ por la derecha, superior a 1.5 veces el recorrido intercuartílico $R_i = Q_3 - Q_1$, se le llaman **atípicos** de primer nivel.
- Cuando la distancia, por uno de los dos lados, es superior a $3R_i$, el valor atípico se denomina de segundo nivel, o **dato extremo**.

# Valores atípicos

Figura 1: Características de un diagrama de caja.

# Valores atípicos

Figura 1: Características de un diagrama de caja.

# Valores atípicos

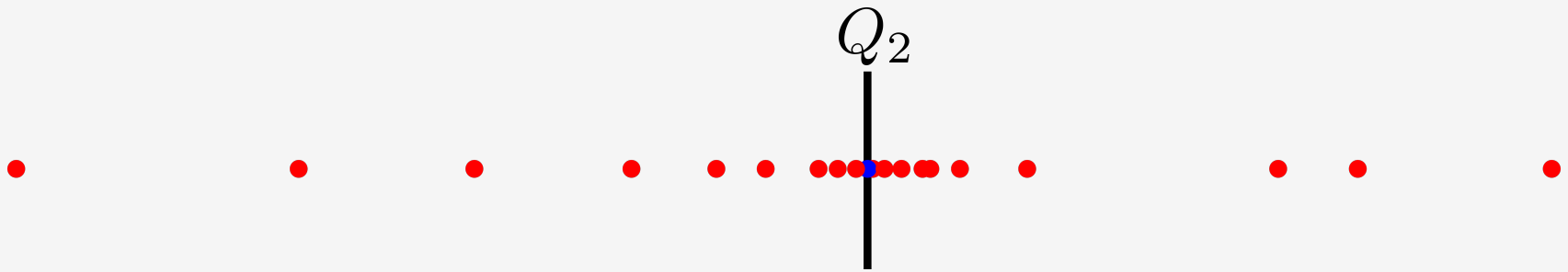



Figura 1: Características de un diagrama de caja.

# Valores atípicos

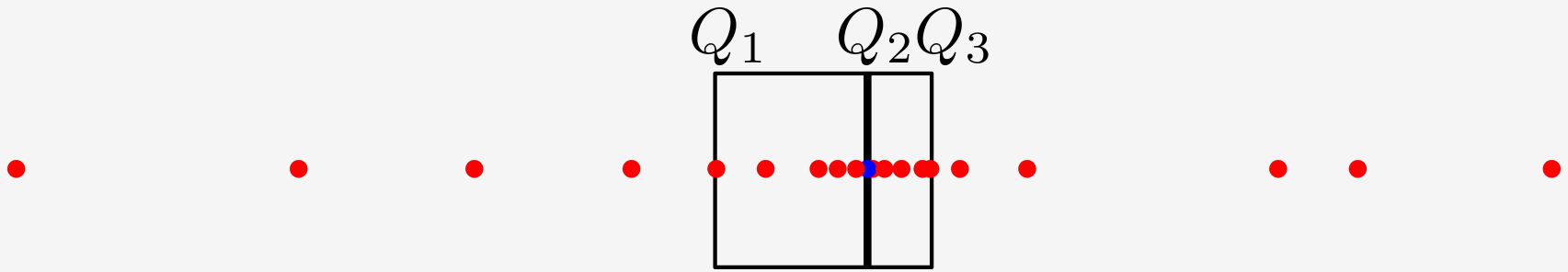



Figura 1: Características de un diagrama de caja.

# Valores atípicos

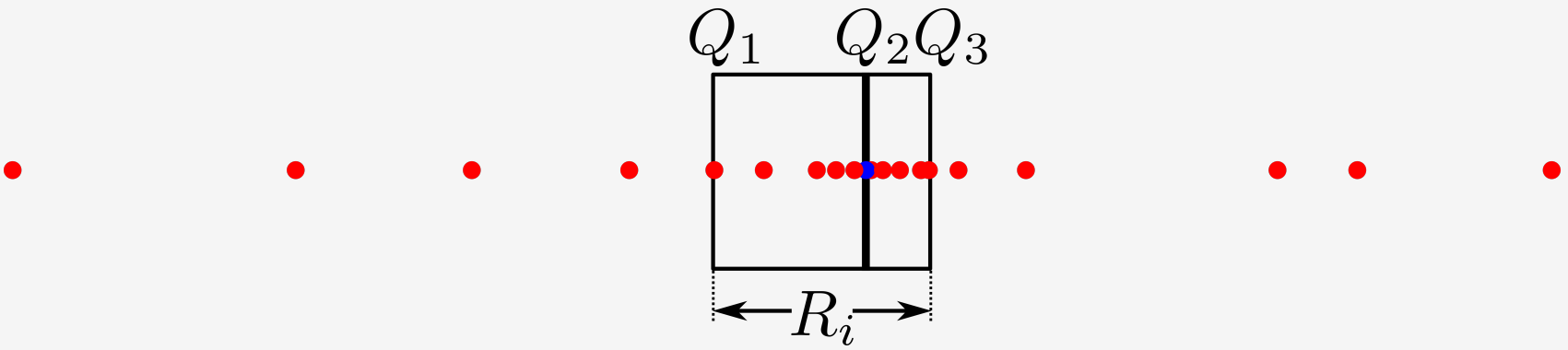



Figura 1: Características de un diagrama de caja.

# Valores atípicos

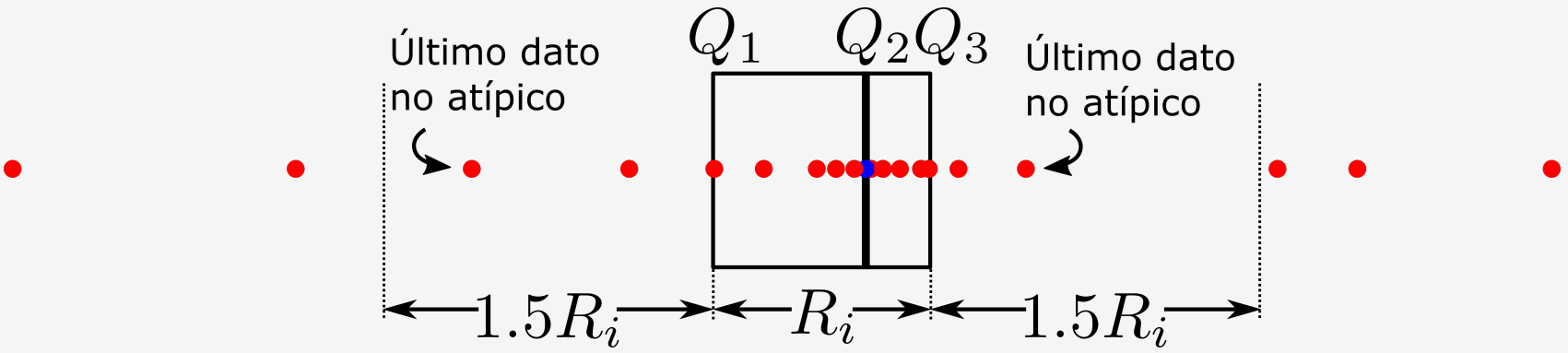



Figura 1: Características de un diagrama de caja.

# Valores atípicos

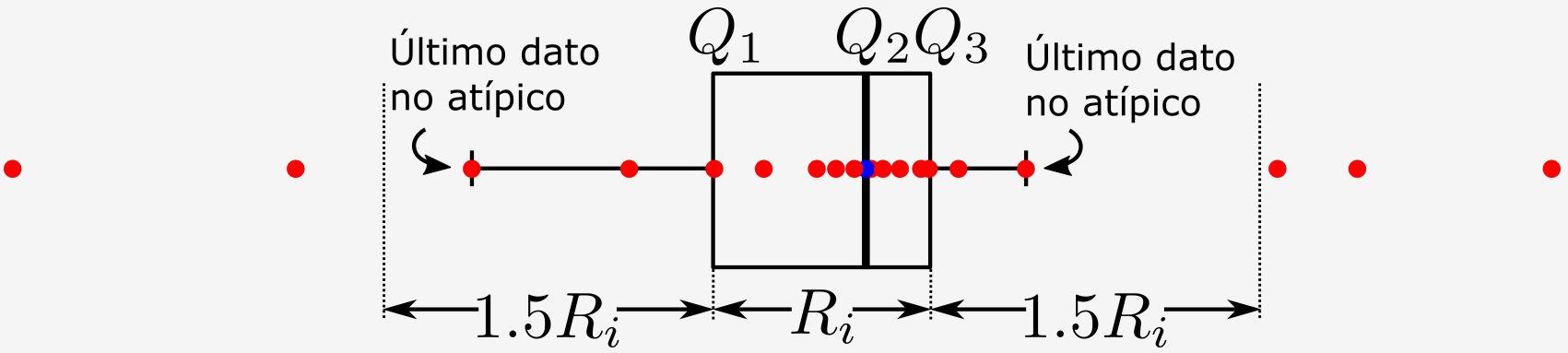



Figura 1: Características de un diagrama de caja.

# Valores atípicos

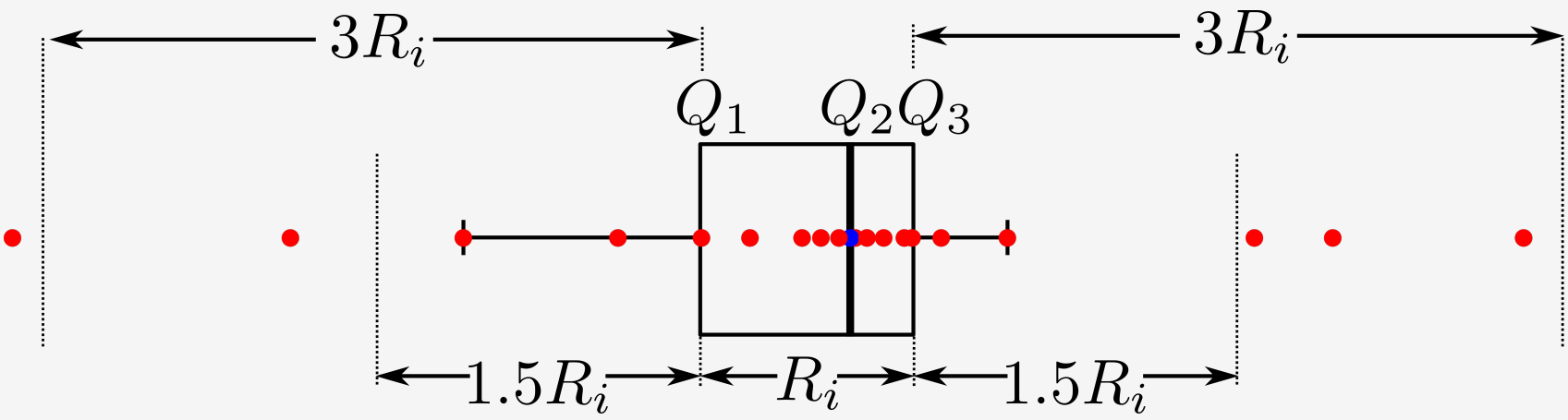



Figura 1: Características de un diagrama de caja.

# Valores atípicos

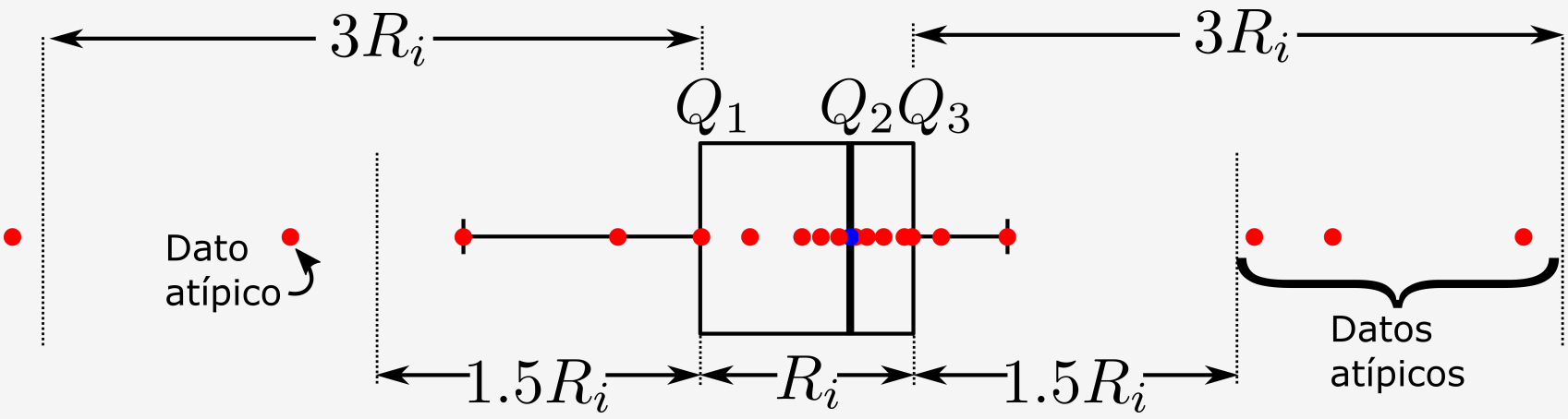



Figura 1: Características de un diagrama de caja.

# Valores atípicos

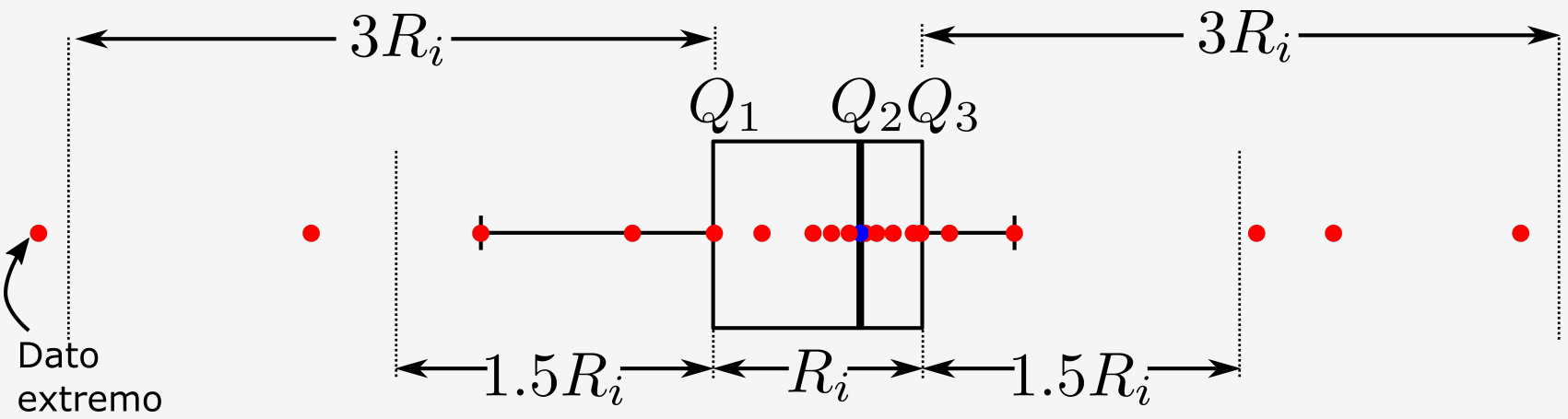



Figura 1: Características de un diagrama de caja.

# Valores atípicos

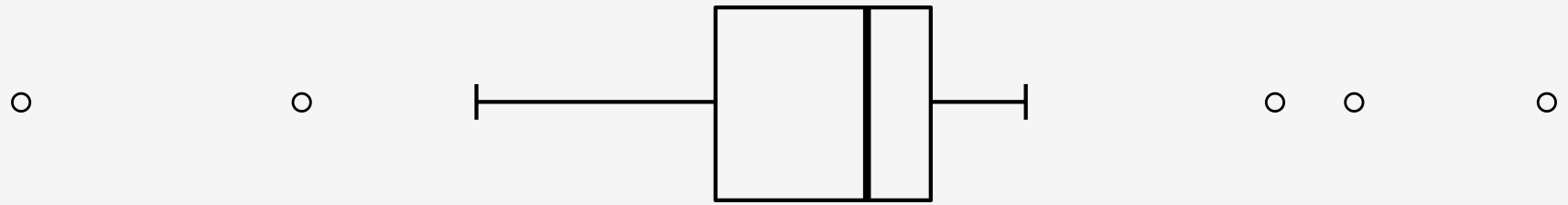


Figura 1: Características de un diagrama de caja.