Resumen del capítulo: Estadística descriptiva

Variables continuas y discretas

Las variables categóricas (cualitativas) toman sus valores de un conjunto limitado. Si se les asignan valores numéricos, esto es puramente simbólico y solo para facilitar el procesamiento (por ejemplo, rojo = 1, azul = 2).

Las variables cuantitativas (numéricas) toman valores numéricos. Existen dos formas:

- **Variables continuas** que pueden tomar cualquier valor (con cualquier grado de precisión) dentro de un rango (por ejemplo, cualquier valor entre 0 y 1)
- **Variables discretas** que no son continuas en ningún rango (por ejemplo, una variable que toma valores enteros entre 0 y 100)

Histogramas de frecuencia

Los histogramas funcionan bien para variables **discretas**. Para visualizar la frecuencia de las variables continuas necesitamos algo más.

Una forma de visualizar la distribución de variables continuas es dividir el conjunto de valores posibles en intervalos y contar el número de valores en cada intervalo.

Al trazar un histograma en pandas, puedes establecer el número de intervalos (contenedores) y establecer sus límites de forma explícita:

```
data.hist(bins=[value1, value2, value3, value4, ..., valueN])
```

No obstante, recuerda siempre que el éxito de este enfoque depende de qué tan bien elijas los límites del intervalo, una tarea que puede ser difícil incluso para analistas de datos experimentados.

Histogramas de densidad

Para superar las dificultades de crear un histograma para una variable continua, podemos usar una técnica un poco diferente que representa la frecuencia no como la altura de una columna, sino como su área (la longitud del intervalo multiplicada por la altura de la columna). Esta área es la **frecuencia** de la variable continua y la altura de la columna es la **densidad de frecuencia**. Un histograma que usa densidad de frecuencia se llama **histograma de densidad.**

Para estimar cuántos valores se encuentran en un intervalo particular, toma dos valores y encuentra el área total de los histogramas de densidad entre ellos. El número que obtengas será una estimación del numero de valores en ese intervalo.

También podemos mostrar la densidad de frecuencia de las variables continuas mediante líneas continuas. Se aplica el mismo principio: el área bajo la curva entre dos valores es proporcional a la frecuencia de los valores en un intervalo dado.

Medidas de posición

Podemos usar **medidas de posición**, como la mediana y la media, para estimar aproximadamente *dónde* se encuentra un dataset en el eje numérico.

La media y la mediana se denominan de manera más formal como **medida de posición algebraica** y **medida de posición estructural**, respectivamente.

¿Quién dispersó los datos?

Las medidas de posición no son suficientes si realmente quieres entender los datos. También necesitas saber cómo se **dispersan** o **están dispersos** los datos alrededor de estas medidas.

Para las medianas, la dispersión puede medirse en términos de cuartiles.

Varianza

La varianza es otra medida común de dispersión. Se puede calcular elevando al cuadrado la distancia media de los datos desde la media:

$$\sigma^2 = \frac{\sum (\mu - x_i)^2}{n}$$

donde la letra griega mu, μ, corresponde a la media aritmética de los datos.

$$\mu = \frac{\sum (x_i)}{n}$$

La librería **NumPy** en Python contiene una gran librería de funciones matemáticas de alto nivel. Así es como se importa:

```
import numpy as np
```

La varianza se calcula con el método **var()**:

```
import numpy as np
variance = np.var(x)
```

Desviación estándar

La varianza tiene un inconveniente: sus unidades de medida son cuadrados de las unidades originales de la variable. Para volver a las unidades de medida originales necesitamos sacar la raíz cuadrada de la varianza. El valor resultante se conoce como **desviación estándar**.

$$\sigma = \sqrt{\frac{\sum (\mu - x_i)^2}{n}}$$

La desviación estándar se puede encontrar con el método `std()` de NumPy:

```
import numpy as np
standard_deviation = np.std(x)
```

Recuerda también que si ya conoces la varianza, puedes usar el método `sqrt()` de NumPy para obtener la desviación estándar.

```
import numpy as np
variance = 2.9166666666666665
standard_deviation = np.sqrt(variance)
```

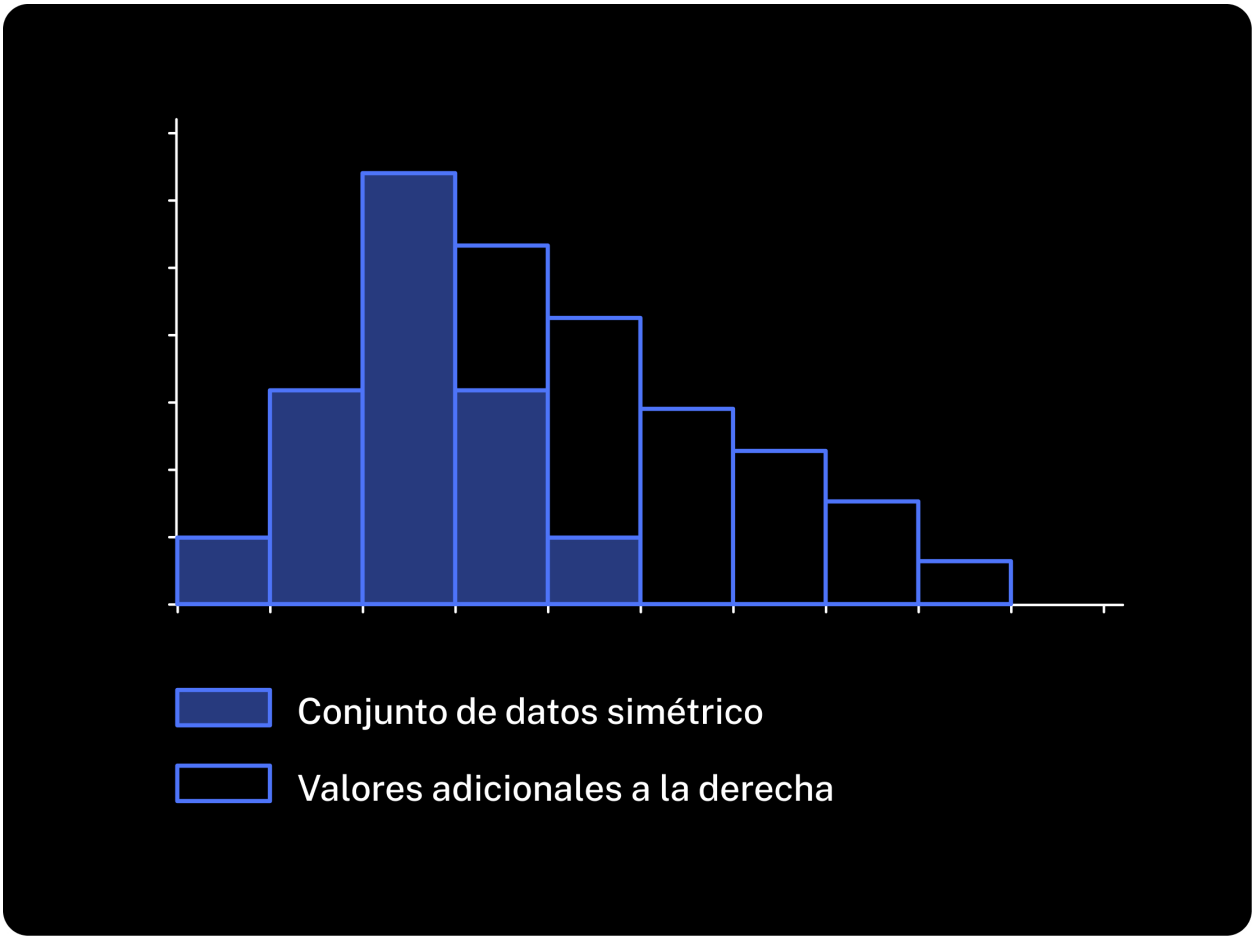
La regla de las tres desviaciones estándar, o la **regla de las tres sigma**, es válida para las distribuciones más utilizadas. Esta regla dice que casi todos los valores (aproximadamente el 99 %) se encuentran dentro de las tres desviaciones estándar de la media:

$(\mu - 3\sigma, \mu + 3\sigma)$

Esta regla no solo te ayuda a encontrar el intervalo en el que se encuentran la mayoría de los valores que te interesan, sino que también te ayuda a encontrar valores fuera de ese intervalo (**valores atípicos**).

Datos sesgados

En muchos casos, los datos se distribuyen normalmente y de forma simétrica. Pero los datasets también pueden ser asimétricos o "sesgados" en una dirección positiva o negativa. Es fácil reconocer la asimetría si observas un histograma. Puedes visualizar la asimetría como una cola de un lado al otro de la "masa" simétrica de los datos.



Se dice que una distribución de datos con valores adicionales a la derecha **sesga a la derecha**. Esto a menudo se denomina **asimetría positiva** porque hay valores adicionales a medida que te mueves a lo largo del eje en una dirección positiva.

Por el contrario, se dice que un dataset que se diferencia de uno simétrico en que tiene valores adicionales a la izquierda está **sesgado a la izquierda** o tiene **asimetría negativa**.

Los diagramas de caja también muestran la asimetría de una distribución.

Hay una manera de determinar la asimetría de un dataset sin trazar gráficos: simplemente compara la media y la mediana. Dado que la mediana no se ve afectada por los valores atípicos tanto como la media, la media es mayor que la mediana para datasets sesgados a la derecha, y viceversa para datasets sesgados a la izquierda.