Hoja informativa: Estadística descriptiva

Práctica

```
# Creación de histogramas con límites de intervalo establecidos

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

# Importación de una librería de funciones matemáticas de alto nivel

import numpy as np

# Encontrar dispersión

import numpy as np
np.var(x)

# Sacar la raíz cuadrada

import numpy as np
np.sqrt(x)
```

Teoría

Variables cuantitativas (numéricas) toman valores numéricos.

Variables cualitativas (categóricas) toman valores no numéricos.

Una **variable continua** es una variable cuantitativa que puede tomar cualquier valor numérico (con cualquier grado de precisión) en algún rango (por ejemplo, cualquier valor entre 0 y 1).

Una **variable discreta** es cualquier variable que no es continua en ningún rango (por ejemplo, una variable que toma los valores enteros de 0 a 100).

Densidad de frecuencia: un valor equivalente a la altura de una columna de histograma cuya área refleja la frecuencia relativa de una variable continua.

Histograma de densidad: un histograma que usa la densidad de frecuencia.

Las métricas de posición te ayudan a calcular aproximadamente dónde se encuentra el conjunto de datos en el eje numérico.

Métrica algebraica de posición: la media, a menudo representada por la letra griega mu, μ.

Métrica estructural de posición: la mediana.

La **varianza** se usa para medir qué tan "dispersos" están los datos de la media. Se calcula tomando la distancia promedio elevada al cuadrado de la media de todos los puntos en el conjunto de datos:

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

La **desviación estándar** es la raíz cuadrada de la varianza y se representa con la letra griega sigma, σ. Se calcula con la ecuación siguiente:

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

Regla de las tres sigma: casi todos los valores (99,7%) se encuentran dentro del intervalo:

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

Sesgo es una medida de la asimetría de un dataset.

Los datos con **asimetría positiva (sesgo a la derecha)** tienen una media mayor que la mediana. Los datos tendrán más valores superiores a la media que inferiores.

Los datos con **asimetría negativa (sesgo a la izquierda)** tienen una media que es menor que la mediana. Los datos tendrán más valores inferiores a la media que superiores.