S12 - P1 (Sesión 23): Estadística descriptiva con NumPy, Pandas y Matplotlib

Departamento de Física.

Corodinadora: C Loyola
Profesores C Femenías / F Bugini / D Basantes

Primer Semestre 2025

Universidad Andrés Bello
Departamento de Física y Astronomía

Universidad Andrés Bello®

python powered

Resumen – Semana 12, Sesión 1 (Sesión 23)

Introducción y Repaso

Objetivos de la Sesión

Medidas Descriptivas con NumPy

Resumen con pandas

Visualización de Datos

Ejercicios Prácticos

Cierre

Introducción y Repaso

- Hasta la Semana 10 trabajamos con:
  - **NumPy**: arrays, álgebra lineal, aleatoriedad.
  - **Matplotlib**: subplots, histogramas, 3D.
- Hoy re-iniciamos la **unidad de Estadística**:
  - Medidas descriptivas (media, mediana, varianza).
  - Uso de **pandas** para tablas de datos.
  - Visualización: Ej: histogramas, boxplots, violin plots.
- Tarea V se publicará al final de la próxima clase.

Objetivos de la Sesión

Objetivos de la Sesión $\in$ Objetivos

- Comprender **estadística descriptiva** con **NumPy**.
- Manejar **DataFrames** de **pandas** para resumen rápido.
- Generar **gráficos estadísticos** en Matplotlib.
- Preparar bases para la **Tarea V** (análisis de datos reales).

Medidas Descriptivas con NumPy

Medidas Descriptivas con NumPy ∈ NumPy: estadísticas básicas

```python
import numpy as np
data = np.random.normal(loc=0, scale=1, size=1_000)

print("Media :", data.mean())
print("Mediana :", np.median(data))
print("Desvación estándar :", data.std(ddof=1))
print("Percentiles 25/75 :", np.percentile(data, [25, 75]))
```

- `ddof=1` usa varianza muestral (i.e. divide por $n - 1$).
- `np.percentile` devuelve los cuantiles 25% y 75% de la muestra.

Un percentil es simplemente el valor que deja por debajo de sí a un cierto porcentaje de los datos ordenados. Por ejemplo, el percentil 90 (P90) marca el punto por debajo del cual cae el 90% de las observaciones.

Resumen con pandas

Resumen con pandas $\in$ pandas: `describe()` y más

```python
import pandas as pd

df = pd.DataFrame({"x": data})
print(df.describe())
```

- `describe()` produce conteo, media, std, min, max y cuartiles.
- $\rightarrow$ Ideal para inspección rápida de datasets.

Visualización de Datos

Visualización de Datos ∈ Histogramas y Boxplots

```python
import matplotlib.pyplot as plt

fig, axs = plt.subplots(1, 2, figsize=(8,4))

axs[0].hist(df["x"], bins=30, alpha=0.8)
axs[0].set_title("Histograma")

axs[1].boxplot(df["x"], vert=False, patch_artist=True)
axs[1].set_title("Boxplot")

plt.tight_layout(); plt.show()
```

- El boxplot muestra mediana, cuartiles y outliers.
- `patch_artist=True` permite colorear la caja.

Ejercicios Prácticos

Enunciado

- Archivo csv: `https://gitarra.cl/lectures/gfiles/-/raw/main/pcfi161/S12/alturas.csv`
- Cargar el archivo `alturas.csv` (columna *height_cm*).
- Usar **NumPy** y **pandas** para calcular: media, mediana, desvío y cuartiles.
- Graficar histograma y boxplot en la misma figura.
- Comentar si los datos muestran *sesgo* o *asimetría*.

Ejercicios Prácticos $\in$ Ejercicio 2: Correlación Masa-Altura (opcional)

Pasos

1. Archivo csv: `https://gitarra.cl/lectures/gfiles/-/raw/main/pcfi161/S12/masas_alturas.csv`
2. Cargar `masas_alturas.csv` con columnas `mass_kg`, `height_cm`.
3. Calcular coeficiente de correlación ρ con `np.corrcoef`.
4. Graficar scatter y ajustar recta (`np.polyfit`).

Cierre

Cierre $\in$ Próxima Sesión

- **Sesión 24 (en dos días)**: Reforzamiento + **Tarea V**.
- Recuerden estudiar para que les vaya muy bien!.