SESIÓN CORRELACIÓN

CONTENIDOS:

- Graficando la correlación de variables: tablas de contingencia y gráfico scatterplot.
- Midiendo la correlación de variables con el indicador r-pearson.
- Causalidad v/s correlación.

GRAFICANDO LA CORRELACIÓN DE VARIABLES: TABLAS DE CONTINGENCIA Y GRÁFICO SCATTERPLOT

La correlación es una medida estadística que describe el grado de relación entre dos variables. En ciencia de datos, comprender la correlación es esencial para identificar patrones, predecir comportamientos y tomar decisiones informadas. Sin embargo, es fundamental recordar que correlación no implica causalidad, lo que significa que una relación entre dos variables no necesariamente indica que una causa a la otra.

Existen diferentes tipos de correlación:

- Correlación positiva: Cuando una variable aumenta, la otra también lo hace.
- Correlación negativa: Cuando una variable aumenta, la otra disminuye.
- Correlación nula: No hay una relación clara entre las variables.

Graficando la Correlación de Variables

1. Tablas de Contingencia

Una tabla de contingencia es una tabla de frecuencias utilizada para analizar la relación entre dos variables categóricas. Muestra cuántas veces ocurre cada combinación de categorías y permite evaluar si existe alguna asociación entre ellas.

¿Cuándo usar una tabla de contingencia?

- Para examinar la relación entre dos variables categóricas (por ejemplo, género y preferencia de producto).
- Para calcular probabilidades condicionales y frecuencias relativas.
- Para construir una tabla de chi-cuadrado y evaluar independencia entre variables.

Ejemplo en Python:

```python
import pandas as pd

# Crear un DataFrame con variables categóricas
df = pd.DataFrame({
    'Género': ['Masculino', 'Femenino', 'Femenino', 'Masculino'],
    'Preferencia': ['Deportes', 'Cine', 'Deportes', 'Cine']
})

# Crear tabla de contingencia
tabla_contingencia = pd.crosstab(df['Género'], df['Preferencia'])
print(tabla_contingencia)
```

Ilustración 1 Ejemplo en Python de tabla de contingencia

Tabla de Contingencia Resultante:

La función pd.crosstab() cuenta la frecuencia de cada combinación de valores entre las columnas Género y Preferencia. El resultado es el siguiente:

GÉNERO	CINE	DEPORTES
Femenino	1	1
Masculino	1	1

Ilustración 2 Tabla contingencia resultante

Interpretación de la tabla de contingencia:

Filas (Género):

- Femenino: Hay 1 persona de género femenino que prefiere Cine y 1 persona que prefiere Deportes.
- Masculino: Hay 1 persona de género masculino que prefiere Cine y 1 persona que prefiere Deportes.

Columnas (Preferencia):

- Cine: En total, 2 personas prefieren Cine (1 femenino y 1 masculino).
- Deportes: En total, 2 personas prefieren Deportes (1 femenino y 1 masculino).

2. Gráfico Scatterplot (Diagrama de Dispersión)

Un scatterplot (diagrama de dispersión) es una herramienta visual utilizada para examinar la relación entre dos variables numéricas. Cada punto en el gráfico representa una observación en el conjunto de datos.

¿Cuándo usar un scatterplot?

- Para visualizar si existe una correlación entre dos variables numéricas.
- Para detectar outliers o patrones en los datos.
- Para analizar la relación entre variables en problemas de regresión.

Ejemplo en Python

```python
import matplotlib.pyplot as plt
import numpy as np

# Datos ficticios de ventas (cantidad de productos vendidos y precio promedio)
np.random.seed(42)
cantidad_vendida = np.random.randint(10, 100, 50)  # Número de productos vendidos
precio_promedio = cantidad_vendida * np.random.uniform(0.8, 1.2, 50)  # Precio con variación aleatoria

# Crear el scatterplot
plt.figure(figsize=(8, 5))
plt.scatter(cantidad_vendida, precio_promedio, color='blue', alpha=0.5)
plt.xlabel('Cantidad Vendida')
plt.ylabel('Precio Promedio')
plt.title('Relación entre Cantidad Vendida y Precio Promedio')
plt.grid(True)
plt.show()
```

Ilustración 3 Ejemplo gráfico scatterplot

Interpretación del scatterplot:

- Si los puntos siguen una línea ascendente, hay una correlación positiva (cuando una variable aumenta, la otra también).
- Si los puntos forman una línea descendente, hay una correlación negativa.
- Si los puntos están dispersos sin una forma clara, no hay correlación.

MIDIENDO LA CORRELACIÓN DE VARIABLES CON EL INDICADOR R-PEARSON

El coeficiente de correlación de Pearson **(*r*)** es una medida estadística utilizada para cuantificar la relación lineal entre dos variables numéricas. Se usa comúnmente en análisis exploratorio de datos para determinar si existe una relación entre dos conjuntos de datos y en qué grado.

Coeficiente de Correlación de Pearson (r)

Mide la intensidad y dirección de la relación lineal entre dos variables.

Rango de valores: **$-1 \leq r \leq 1$**

Interpretación:

- r = 1: Correlación positiva perfecta (cuando una variable aumenta, la otra también).
- r = -1: Correlación negativa perfecta (cuando una variable aumenta, la otra disminuye).
- r = 0: No hay correlación lineal (las variables no tienen relación).

Cálculo del Coeficiente de Pearson

La fórmula del coeficiente de correlación de Pearson es:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Ilustración 4 Fórmula coeficiente de Pearson

Donde:

- Xi e Yi Son los valores de las variables.
- $\bar{X}$ e $\bar{Y}$ Son las medias de las variables X e Y
- ∑ indica la suma de los valores.

Ejemplo en Python:

Podemos calcular el coeficiente de correlación de Pearson en Python usando scipy.stats.pearsonr() o numpy.corrcoef().

Ejemplo 1: Usando SciPy

```python
import numpy as np
import scipy.stats as stats

# Datos ficticios: Ventas de un producto y presupuesto de marketing
presupuesto_marketing = np.array([1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500])
ventas = np.array([200, 270, 340, 410, 480, 550, 600, 660])

# Calcular el coeficiente de Pearson
coef, p_valor = stats.pearsonr(presupuesto_marketing, ventas)

print(f"Coeficiente de correlación de Pearson: {coef:.2f}")
print(f"P-valor: {p_valor:.4f}")
```

Ilustración 5 Ejemplo en Python coeficiente de Pearson con scipy.stats.pearsonr()

Ejemplo 2: Usando NumPy

```python
# Calcular correlación usando NumPy
corr_matrix = np.corrcoef(presupuesto_marketing, ventas)
print(f"Coeficiente de Pearson usando NumPy: {corr_matrix[0, 1]:.2f}")
```

Ilustración 6 Ejemplo en Python coeficiente de Pearson con NumPy

Interpretación del Coeficiente de Pearson

Los valores de **r** se interpretan según la siguiente escala:

Rango de r	Interpretación
0.8 ≤ r ≤ 1.0	Correlación positiva fuerte
0.5 r 0.8	Correlación positiva moderada
0.2 r 0.5	Correlación positiva débil
-0.2 r 0.2	Correlación despreciable o nula
-0.5 r -0.2	Correlación negativa débil
-0.8 r -0.5	Correlación negativa moderada
-1.0 r -0.8	Correlación negativa fuerte

Si el p-valor es menor a 0.05, la correlación es estadísticamente significativa.

Consideraciones Importantes al usar el Coeficiente de Pearson

1. Solo mide relaciones lineales: Si la relación entre las variables es no lineal, Pearson puede no ser la mejor medida.
2. Es sensible a outliers: Valores extremos pueden influir en el coeficiente y dar resultados engañosos.
3. No implica causalidad: Una alta correlación entre dos variables no significa que una cause la otra.
4. Es mejor complementarlo con visualización: Un gráfico de dispersión (scatterplot) puede ayudar a interpretar la relación entre las variables.

Ejemplo: Visualizar la Correlación con Scatterplot

```
import matplotlib.pyplot as plt

plt.scatter(presupuesto_marketing, ventas, color='blue', alpha=0.5)
plt.xlabel('Presupuesto de Marketing')
plt.ylabel('Ventas')
plt.title('Relación entre Presupuesto y Ventas')
plt.grid(True)
plt.show()
```

Ilustración 7 Ejemplo visualizar la correlación con Scatterplot

Explicación:

- Importa la librería Matplotlib, específicamente el módulo pyplot, que permite crear gráficos de manera sencilla. Se usa plt como alias para facilitar su uso en el código.
- Crea un gráfico de dispersión con los datos de presupuesto_marketing (eje X) y ventas (eje Y).
- color='blue' → Establece el color de los puntos en azul.
- alpha=0.5 → Define la transparencia de los puntos (0 es totalmente transparente, 1 es opaco).
- Etiqueta el eje X con el nombre "Presupuesto de Marketing", lo que indica que en este eje se representan los valores del presupuesto en publicidad.
- Etiqueta el eje Y con el nombre "Ventas", indicando que en este eje se representa la cantidad de ventas.
- plt.title() Asigna un título al gráfico para que el usuario pueda entender su propósito. En este caso, el título indica que el gráfico representa la relación entre el presupuesto en marketing y las ventas.
- plt.grid(true) Activa la cuadrícula en el gráfico para mejorar la lectura y visualización de los datos.
- plt.show() Muestra el gráfico en pantalla. Sin esta línea, el gráfico no aparecería.

CAUSALIDAD V/S CORRELACIÓN

En el análisis de datos, es común encontrar relaciones entre variables, pero correlación no implica causalidad. Es fundamental comprender la diferencia entre ambos conceptos para evitar interpretaciones erróneas y conclusiones inexactas en estudios y modelos predictivos.

¿Qué es Causalidad?

La causalidad ocurre cuando un cambio en una variable provoca directamente un cambio en otra. Para demostrar una relación causal, es necesario realizar pruebas rigurosas y descartar otros factores que puedan estar influyendo en la relación.

Ejemplo de Causalidad

- Un aumento en la dosis de un medicamento mejora la recuperación de pacientes.

Aquí, existe un vínculo directo y comprobado entre la causa (medicamento) y el efecto (recuperación).

¿Qué es Correlación?

La correlación mide la relación entre dos variables, pero no necesariamente implica que una causa el cambio de la otra. Puede ser positiva (cuando ambas variables aumentan o disminuyen juntas) o negativa (cuando una aumenta y la otra disminuye).

Ejemplo de Correlación

- Las ventas de helado y el número de personas en la playa aumentan al mismo tiempo.

Esto no significa que vender más helado haga que más personas vayan a la playa. Ambas variables están influenciadas por un tercer factor: el clima cálido.

Diferencias Clave entre Correlación y Causalidad

Característica	Correlación	Causalidad
Relación Directa	No necesariamente	Sí
Dirección del Efecto	No se puede determinar	Se establece mediante pruebas
Terceros Factores	Pueden influir	Se intentan descartar
Comprobación	Métodos estadísticos como Pearson	Experimentos y estudios longitudinales

Ejemplo de Confusión entre Correlación y Causalidad

Caso Falso: "Las personas felices viven más tiempo"

Podemos encontrar una correlación entre felicidad y longevidad, pero esto no prueba causalidad.

- Tal vez las personas felices tienen menos estrés, lo que mejora su salud.
- Tal vez las personas con buena salud son más felices y viven más tiempo.

Para demostrar causalidad, tendríamos que controlar otras variables (ejemplo: nivel socioeconómico, hábitos de salud, genética, etc.).

Cómo Establecer Causalidad en Ciencia de Datos

Para demostrar una relación causal, es necesario realizar pruebas y análisis rigurosos:

1. Experimentos Controlados
   - Se manipula una variable (grupo de prueba) y se compara con un grupo de control sin intervención.
   - Ejemplo: Probar un nuevo fármaco en pacientes y comparar con un placebo.
2. Análisis de Regresión
   - Se estudia el impacto de una variable en otra, controlando otros factores.
   - Ejemplo: Evaluar cómo el nivel de ingresos afecta el rendimiento académico, controlando el acceso a educación de calidad.
3. Estudios Longitudinales
   - Se observa a lo largo del tiempo cómo una variable afecta a otra.
   - Ejemplo: Analizar durante 10 años el impacto del ejercicio en la salud cardiovascular.
4. Pruebas de Intervención
   - Se introduce un cambio en una variable y se mide la respuesta en la otra.
   - Ejemplo: Aumentar el presupuesto en publicidad y medir si se incrementan las ventas.

La correlación es útil para identificar patrones, pero no implica causalidad. Para demostrar causalidad, se requieren experimentos, estudios longitudinales y modelos de regresión. En ciencia de datos, es crucial diferenciar entre ambas para evitar conclusiones erróneas y tomar decisiones informadas. Muchas veces, encontrar una correlación puede ser el primer paso para investigar una posible causalidad, pero nunca debe ser la única evidencia.