Distribuciones y Análisis Univariado - Día 3

Change Status

in-progress 40 min

Learning Objectives

1. Identificar y caracterizar diferentes tipos de distribuciones en datos reales
2. Calcular e interpretar medidas de forma: asimetría y curtosis
3. Aplicar métodos estadísticos para detección sistemática de outliers

Actividades y Aprendizajes

Aprende todo sobre funciones y módulos en Python con ejemplos prácticos.

Task 1: Tipos Comunes de Distribuciones (12 minutos)

Las distribuciones de datos reales siguen **patrones reconocibles** que revelan procesos subyacentes. Comprender estos patrones permite seleccionar análisis apropiados y detectar anomalías significativas.

Distribución Normal (Gaussiana): El Patrón de Equilibrio

La distribución normal es la **distribución más famosa** de la estadística, caracterizada por su **curva de campana simétrica**.

Características:

Simétrica: Media = Mediana = Moda
Forma de campana: La mayoría de valores cerca del centro
Definida por dos parámetros: Media (μ) y desviación estándar (σ)
Regla empírica: 68% dentro 1σ, 95% dentro 2σ, 99.7% dentro 3σ

Casos de aplicación:

Mediciones físicas: Altura, peso, presión arterial
Errores de medición: En experimentos científicos
Variables económicas: En mercados eficientes a corto plazo

Pruebas de normalidad: Shapiro-Wilk, Kolmogorov-Smirnov, visual con Q-Q plots.

Distribuciones Sesgadas: Asimetría en Datos Reales

La mayoría de datos del mundo real exhiben **asimetría**, donde la distribución no es simétrica alrededor del centro.

Sesgo positivo (right-skewed): Cola larga hacia la derecha

Ejemplos: Ingresos, precios de vivienda, tiempos de respuesta
Media > Mediana > Moda
Interpretación: Pocos valores muy altos elevan la media

Sesgo negativo (left-skewed): Cola larga hacia la izquierda

Ejemplos: Tasas de finalización, edades de fallecimiento
Media < Mediana < Moda
Interpretación: Pocos valores muy bajos deprimen la media
Distribuciones Multimodales: Múltiples Poblaciones

Las distribuciones multimodales indican **mezclas de poblaciones** diferentes dentro del mismo dataset.

Causas comunes:

Segmentos demográficos: Jóvenes vs adultos vs seniors
Categorías de producto: Productos premium vs estándar
Períodos temporales: Comportamiento diferente por temporada

Detección: Histogramas con múltiples picos, análisis de clustering.

Task 2: Medidas de Forma: Asimetría y Curtosis (10 minutos)

La asimetría y curtosis cuantifican la **forma** de la distribución más allá de medidas de tendencia central y dispersión.

Asimetría (Skewness): Medida de Simetría

La asimetría mide qué tan **asimétrica** es la distribución alrededor de su media.

Fórmula: $\gamma = \Sigma[(x_i - \mu)/\sigma]^3 / n$

Interpretación:

γ = 0: Distribución perfectamente simétrica (normal)
γ > 0: Sesgo positivo (cola derecha más larga)
γ < 0: Sesgo negativo (cola izquierda más larga)

Rango típico: -3 a +3, valores extremos indican distribuciones muy sesgadas.

Implicaciones prácticas:

Sesgo positivo: La media sobreestima el valor típico
Sesgo negativo: La media subestima el valor típico
Elección de medidas: Usar mediana en lugar de media para medidas centrales
Curtosis (Kurtosis): Medida de "Colas Pesadas"

La curtosis mide la **"pesadez" de las colas** comparada con una distribución normal.

Fórmula: `κ = Σ[(x_i - μ)/σ]⁴ / n - 3`

Interpretación:

κ = 0: Curtosis normal (mesocúrtica)
κ > 0: Colas más pesadas (leptocúrtica) - más outliers extremos
κ < 0: Colas más ligeras (platicúrtica) - menos valores extremos

Ejemplos:

Distribución normal: κ = 0
Distribución t-Student: κ > 0 (colas pesadas)
Distribución uniforme: κ < 0 (colas ligeras)

Significado práctico:

Alta curtosis: Mayor probabilidad de valores extremos
Baja curtosis: Distribución más "aplastada", menos outliers

Task 3: Detección Sistemática de Outliers (8 minutos)

Los outliers son valores que **se desvían significativamente** del patrón general, requiriendo métodos tanto estadísticos como de dominio para su identificación.

Métodos Estadísticos para Outliers

Método del rango intercuartílico (IQR):

```
# Calcular límites
Q1 = df['variable'].quantile(0.25)
Q3 = df['variable'].quantile(0.75)
IQR = Q3 - Q1

limite_inferior = Q1 - 1.5 * IQR
limite_superior = Q3 + 1.5 * IQR

# Identificar outliers
outliers = df[(df['variable'] < limite_inferior) | (df['variable'] > limite_superior)]
```

Ventajas: Robusto, no asume distribución específica, fácil de interpretar.

Método Z-Score:

```
from scipy import stats

# Calcular z-scores
z_scores = stats.zscore(df['variable'])

# Outliers: |z| > 3
outliers = df[abs(z_scores) > 3]
```

Ventajas: Estándar estadístico, útil para distribuciones aproximadamente normales.

Estrategias de Manejo de Outliers

Eliminación: Solo cuando outliers son claramente errores de medición.

```
df_sin_outliers = df[abs(z_scores) <= 3]
```

Transformación: Aplicar logaritmos o otras transformaciones.

```
df['variable_log'] = np.log1p(df['variable'])
```

Winsorización: Reemplazar outliers con percentiles extremos.

```
from scipy.stats.mstats import winsorize

df['variable_winsorized'] = winsorize(df['variable'], limits=[0.05, 0.05])
```

Análisis separado: Analizar outliers como grupo separado.

```
outliers_group = df[abs(z_scores) > 3]
normal_group = df[abs(z_scores) <= 3]
```

Consideraciones de Dominio

¿Son errores o señales?:

Errores: Valores físicamente imposibles (edades negativas)
Señales: Valores extremos válidos (compras de lujo, eventos raros)

Contexto temporal: Un outlier en contexto histórico puede ser normal en contexto actual.

Segmentación: Lo que es outlier en un segmento puede ser normal en otro.