

Probabilidad Básica y Distribuciones de Datos - Día 1

Change Status

in-progress 40 min

Learning Objectives

- 1 Comprender los fundamentos probabilísticos que explican el comportamiento de los datos
- 2 Identificar distribuciones de probabilidad apropiadas para diferentes contextos empresariales
- 3 Aplicar el teorema del límite central en análisis de muestras y estimación

Theory

Practice

Quiz

Evidence

<> **Ejercicio**: Análisis probabilístico de comportamiento de clientes en e-commerce

Ejercicio práctico para aplicar los conceptos aprendidos.

Análisis de distribución de compras por cliente:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Generar dataset de comportamiento de clientes
np.random.seed(42)
n_clientes = 10000

# Simular compras por cliente (distribución Poisson)
compras_por_cliente = np.random.poisson(lam=2.5, size=n_clientes)

# Simular valor de compra (distribución Log-normal)
valor_compra = np.random.lognormal(mean=4.5, sigma=0.8, size=n_clientes)

# Crear DataFrame
df = pd.DataFrame({
```




```
'cliente_id': range(1, n_clientes + 1),
'compras_anuales': compras_por_cliente,
'valor_promedio_compra': valor_compra,
'gasto_total_anual': compras_por_cliente * valor_compra
})

print("Análisis probabilístico de comportamiento de clientes")
print("=" * 55)
print(f"Total clientes analizados: {len(df),}")
print(f"Compras promedio por cliente/año: {df['compras_anuales'].mean():.2f}")
print(f"Valor promedio de compra: ${df['valor_promedio_compra'].mean():.2f}")
```

Análisis de distribuciones y probabilidades:

```
# Análisis de distribución de compras
print("\nDISTRIBUCIÓN DE COMPRAS ANUALES")
print("-" * 35)
print(df['compras_anuales'].value_counts().sort_index().head(10))

# Probabilidades usando distribución de Poisson
lambda_compras = df['compras_anuales'].mean()

print(f"\nProbabilidades (usando distribución de Poisson  $\lambda$ ={{lambda_compras:.2f}}):")
print(f"P(0 compras/año): {stats.poisson.pmf(0, lambda_compras):.4f}")
print(f"P(1 compra/año): {stats.poisson.pmf(1, lambda_compras):.4f}")
print(f"P(5+ compras/año): {1 - stats.poisson.cdf(4, lambda_compras):.4f}")

# Análisis de valor de compra
print(f"\nANÁLISIS DE VALOR DE COMPRA")
print("-" * 30)
print(f"Media: ${df['valor_promedio_compra'].mean():.2f}")
print(f"Mediana: ${df['valor_promedio_compra'].median():.2f}")
print(f"Desviación estándar: ${df['valor_promedio_compra'].std():.2f}")

# Test de normalidad para valor de compra
_, p_valor = stats.normaltest(df['valor_promedio_compra'])
print(f"Test de normalidad (p-valor): {p_valor:.4f}")
print(f"¿Es normal?: {'No' if p_valor < 0.05 else 'Sí'}")
```

Aplicación del teorema del límite central:



```
# Demostración del teorema del límite central
```

```
n_muestras = 1000
```

```
tamano_muestra = 50
```

```
medias_muestrales = []
```

```
for _ in range(n_muestras):
```

```
    muestra = np.random.choice(df['gasto_total_anual'], size=tamano_muestra, replace=True)
```

```
    medias_muestrales.append(np.mean(muestra))
```

```
# Análisis de la distribución de medias muestrales
```

```
medias_array = np.array(medias_muestrales)
```

```
media_muestral_global = np.mean(medias_array)
```

```
error_estandar = np.std(medias_array)
```

```
print(f"\nTEOREMA DEL LÍMITE CENTRAL")
```

```
print("-" * 30)
```

```
print(f"Media de medias muestrales: ${media_muestral_global:.2f}")
```

```
print(f"Media poblacional real: ${df['gasto_total_anual'].mean():.2f}")
```

```
print(f"Error estándar de medias: ${error_estandar:.2f}")
```

```
# Intervalo de confianza
```

```
z_score = 1.96 # 95% confianza
```

```
margen_error = z_score * error_estandar
```

```
ic_inferior = media_muestral_global - margen_error
```

```
ic_superior = media_muestral_global + margen_error
```

```
print(f"Intervalo de confianza 95%: (${ic_inferior:.2f}, ${ic_superior:.2f})")
```

```
print(f"Margen de error: ${margen_error:.2f}")
```

Visualización de distribuciones:

```
# Crear visualización de distribuciones
```

```
fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(15, 10))
```

```
# Distribución de compras (comparar empírica vs teórica)
```

```
ax1.hist(df['compras_anuales'], bins=range(0, 15), alpha=0.7, density=True,  
         color='#1f77b4', label='Datos empíricos')
```

```
x_poisson = range(0, 15)
```

```
y_poisson = [stats.poisson.pmf(x, lambda_compras) for x in x_poisson]
```

```
ax1.plot(x_poisson, y_poisson, 'ro-', label='Distribución Poisson', linewidth=2)
```

```
ax1.set_title('Distribución de Compras por Cliente')
```

```
ax1.set_xlabel('Número de compras anuales')
```




```
ax1.set_ylabel('Probabilidad')
ax1.legend()

# Distribución de valor de compra
ax2.hist(df['valor_promedio_compra'], bins=50, alpha=0.7, density=True,
         color='#ff7f0e', label='Datos')

# Ajuste de distribución Log-normal
params = stats.lognorm.fit(df['valor_promedio_compra'])
x_fit = np.linspace(0, df['valor_promedio_compra'].max(), 100)
y_fit = stats.lognorm.pdf(x_fit, *params)
ax2.plot(x_fit, y_fit, 'r-', linewidth=2, label='Ajuste log-normal')
ax2.set_title('Distribución de Valor de Compra')
ax2.set_xlabel('Valor ($)')
ax2.set_ylabel('Densidad')
ax2.legend()

# Distribución de medias muestrales (TLC)
ax3.hist(medias_array, bins=30, alpha=0.7, density=True, color='#2ca02c')

# Añadir curva normal teórica
x_normal = np.linspace(medias_array.min(), medias_array.max(), 100)
y_normal = stats.norm.pdf(x_normal, media_muestral_global, error_estandar)
ax3.plot(x_normal, y_normal, 'r-', linewidth=2, label='Distribución normal')
ax3.axvline(media_muestral_global, color='red', linestyle='--', alpha=0.7)
ax3.set_title('Teorema del Límite Central')
ax3.set_xlabel('Media muestral de gasto total')
ax3.set_ylabel('Densidad')
ax3.legend()

# Gráfico Q-Q para normalidad
stats.probplot(df['gasto_total_anual'], dist="norm", plot=ax4)
ax4.set_title('Gráfico Q-Q: Normalidad de Gasto Total')


plt.tight_layout()
plt.savefig('distribuciones_probabilidad_clientes.png', dpi=300, bbox_inches='tight')
print("\nVisualizaciones guardadas como 'distribuciones_probabilidad_clientes.png'")
```


Verificación: Explica cómo las distribuciones probabilísticas (Poisson para conteos, log-normal para valores) modelan mejor el comportamiento de clientes que las distribuciones normales, y cómo el teorema del límite central justifica el uso de estadística paramétrica en análisis de muestras.

Requerimientos:


Python con SciPy y NumPy
Pandas para manipulación de datos



 Dashboard

 Career Path

 Forms

 Profile

 Sign Out

