TalentOps

Dashboard
Career Path
Forms
Profile

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

Actividades y Aprendizajes

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

Task 1: Conceptos Básicos de Probabilidad (10 minutos)

La probabilidad es el lenguaje matemático de la incertidumbre, permitiendo cuantificar qué tan seguros podemos estar sobre eventos futuros basados en evidencia histórica.

Probabilidad Frecuentista vs Bayesiana

Probabilidad frecuentista: Basada en frecuencias relativas de eventos repetidos.

Definición: P(A) = (número de veces que ocurre A) / (número total de ensayos)
Ejemplo: Probabilidad de que un cliente compre = (compras en período) / (total visitantes)
Aplicación: Análisis histórico de datos empresariales

Probabilidad bayesiana: Incorpora conocimiento previo con evidencia nueva.

Teorema de Bayes: P(A|B) = P(B|A) × P(A) / P(B)
Ejemplo: Probabilidad de que un cliente sea "premium" dado su comportamiento de compra
Aplicación: Modelos de scoring de riesgo, segmentación predictiva

us English

Sign Out

Toggle theme: Light Theme

TalentOps
Dashboard
Career Path
Forms
Profile

Eventos y Reglas de Probabilidad

Eventos mutuamente excluyentes: No pueden ocurrir simultáneamente.

P(A ∪ B) = P(A) + P(B)
Ejemplo: Un cliente compra producto A O producto B (no ambos)

Eventos independientes: La ocurrencia de uno no afecta al otro.

P(A ∩ B) = P(A) × P(B)
Ejemplo: Probabilidad de que dos clientes diferentes compren en el mismo día

Probabilidad condicional: Probabilidad de A dado que B ocurrió.

P(A|B) = P(A ∩ B) / P(B)
Ejemplo: Probabilidad de que un cliente repita compra dado que recibió descuento

Task 2: Distribuciones de Probabilidad Comunes (12 minutos)

Las distribuciones de probabilidad describen cómo se distribuyen los valores posibles de una variable aleatoria, permitiendo modelar y predecir comportamiento de datos.

Distribución Normal (Gaussiana): La Reina de las Distribuciones

La distribución normal es la más importante en estadística, describiendo muchos fenómenos naturales y sociales.

Características:

Forma: Curva de campana simétrica
Parámetros: Media (μ) y desviación estándar (σ)
Propiedad: ~68% de datos dentro de 1σ, ~95% dentro de 2σ, ~99.7% dentro de 3σ
Casos de aplicación: Errores de medición, alturas humanas, puntajes de exámenes

En análisis de datos:

```
from scipy import stats
import numpy as np

# Generar datos normales
datos_normales = np.random.normal(loc=100, scale=15, size=1000)

# Calcular probabilidades
prob_menor_85 = stats.norm.cdf(85, loc=100, scale=15)  # P(X < 85)
```

Sign Out
Toggle theme: Light Theme

```python
prob_entre_85_115 = stats.norm.cdf(115, 100, 15) - stats.norm.cdf(85, 100, 15)

print(f"P(X < 85): {prob_menor_85:.3f}")
print(f"P(85 < X < 115): {prob_entre_85_115:.3f}")
```

Distribución Binomial: Éxitos en Ensayos Independientes

Modela el número de éxitos en n ensayos independientes, cada uno con probabilidad p de éxito.

Características:

Parámetros: n (número de ensayos), p (probabilidad de éxito)
Casos: Compras exitosas, conversiones de ventas, respuestas a encuestas
Ejemplo: Número de clientes que compran en una semana de 100 visitantes

Aplicación práctica:

```python
# Probabilidad de exactamente k conversiones en n visitantes
n_visitantes = 100
p_conversion = 0.05  # 5% tasa de conversión

prob_exactly_3 = stats.binom.pmf(3, n_visitantes, p_conversion)
prob_at_least_5 = 1 - stats.binom.cdf(4, n_visitantes, p_conversion)

print(f"P(exactamente 3 conversiones): {prob_exactly_3:.4f}")
print(f"P(al menos 5 conversiones): {prob_at_least_5:.4f}")
```

Distribución de Poisson: Eventos en Intervalos de Tiempo

Modela el número de eventos que ocurren en un intervalo de tiempo o espacio fijo.

Características:

Parámetro: λ (tasa promedio de eventos por intervalo)
Casos: Llamadas a soporte, visitas a sitio web, defectos en producción
Ejemplo: Número de pedidos por hora en un e-commerce

Aplicación en análisis de demanda:

```python
# Tasa de pedidos por hora
lambda_pedidos = 12  # 12 pedidos por hora promedio
```

```python
prob_exactly_10 = stats.poisson.pmf(10, lambda_pedidos)
prob_more_than_15 = 1 - stats.poisson.cdf(15, lambda_pedidos)

print(f"P(exactamente 10 pedidos/hora): {prob_exactly_10:.4f}")
print(f"P(más de 15 pedidos/hora): {prob_more_than_15:.4f}")
```

Task 3: Teorema del Límite Central y Muestreo (8 minutos)

El teorema del límite central explica por qué la distribución normal es tan ubicua en estadística aplicada.

El Teorema Fundamental

Enunciado: La distribución de las medias muestrales se aproxima a una normal, sin importar la distribución original, cuando el tamaño de muestra es suficientemente grande ($n \geq 30$).

Implicaciones prácticas:

Justificación de pruebas t: Podemos usar distribuciones normales para inferencia
Estimación de intervalos: Construcción de intervalos de confianza
Tamaño de muestra: Para n grande, podemos asumir normalidad

Error estándar de la media: $\sigma/\sqrt{n}$ - cuánto varían las medias muestrales.

Aplicación en Análisis de Datos Empresariales

Estimación de parámetros poblacionales:

```python
# Estimación del ingreso promedio de clientes
ingresos_muestra = np.random.lognormal(10, 0.5, 100)  # Muestra de 100 clientes
media_muestra = np.mean(ingresos_muestra)
error_estandar = stats.sem(ingresos_muestra)  # Error estándar

# Intervalo de confianza al 95%
intervalo_confianza = stats.t.interval(0.95, len(ingresos_muestra)-1,
                                       loc=media_muestra, scale=error_estandar)

print(f"Media muestral: ${media_muestra:.2f}")
print(f"Error estándar: ${error_estandar:.2f}")
print(f"IC 95%: (${intervalo_confianza[0]:.2f}, ${intervalo_confianza[1]:.2f})")
```

Consideraciones de Muestreo

Sesgo de selección: Muestras no representan a la población. **Tamaño de muestra**: $n = (Z^2 \times \sigma^2) / E^2$ para precisión deseada. **Poder estadístico**: Capacidad de detectar diferencias reales.