Business Analytics & Data Science

Día 5: Inferencia Estadística

EAE Business School Barcelona

6 de febrero de 2026

Plan del Día 5

Primera Parte (9:00-11:00)

1. De muestra a población: el problema de inferencia
2. Intervalos de confianza
3. Introducción a tests de hipótesis

Segunda Parte (11:30-13:30)

1. T-tests en práctica
2. Test chi-cuadrado
3. Patrón de análisis estadístico

Recap: Estadística Descriptiva

Ayer aprendimos:

- Media, mediana, desviación estándar
- Distribuciones de probabilidad
- Distribución normal: μ, σ, regla 68-95-99.7

Hoy: Usamos estadística para **tomar decisiones** basadas en datos

Primera Parte: De Muestra a Población

El Problema de Inferencia

Pregunta de negocio: ¿Cuál es el precio medio de pisos en Barcelona?

Problema: No tenemos acceso a los datos de todos los pisos (población completa)

Solución: Medimos una **muestra** e **inferimos** sobre la población

Inferencia estadística = hacer afirmaciones sobre la población basándonos en una muestra

Problema: Hasta qué punto podemos usar la muestra para sacar conclusiones acerca de la población?

Población vs Muestra

Población = todos los elementos de interés

- Ej: Todos los pisos en Barcelona (~500k)

Muestra = subconjunto de la población que medimos

- Ej: 1000 pisos que encontramos en el dataset

Parámetro = característica de la población (μ, σ)
Estadístico = característica de la muestra ($\overline{x}$, s)

Objetivo: Usar estadísticos muestrales para estimar parámetros poblacionales

Distribución Muestral

Si tomamos muchas muestras, cada una tendría una media diferente.

Distribución muestral = distribución de las medias muestrales

Teorema Central del Límite:

- Con muestras grandes (n > 30), la distribución muestral de la media es aproximadamente normal
- Incluso si los datos originales NO son normales

Error estándar (SE)

Si el tamaño de la muestra es n, entonces

$$SE = \frac{\sigma}{\sqrt{n}}.$$

Intervalos de Confianza

Intervalo de confianza (IC) = rango donde creemos que está el parámetro poblacional

Fórmula básica para la media: para un intervalo de confianza del 95%:

$$IC = \overline{x} \pm 1.96 \cdot SE = \overline{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Interpretación correcta:

"Si obtenemos repetidas muestras, el 95% de los intervalos contendrían el verdadero valor poblacional"

Interpretación INCORRECTA:

"Hay 95% de probabilidad de que μ esté en este intervalo"

Calcular IC en Python (1/2)

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

# Datos de muestra
precios = df['price'].dropna()
n = len(precios)
mean = precios.mean()
se = precios.std() / np.sqrt(n)
```

Calcular IC en Python (2/2)

```
# IC 95% (usando t-distribution)
confidence = 0.95
ic = stats.t.interval(confidence, df=n-1, loc=mean, scale=se)

print(f"Media: {mean:.0f}€")
print(f"IC 95%: [{ic[0]:.0f}, {ic[1]:.0f}]")
```

Ahora al notebook

Ejercicio Intervalos de Confianza (30 minutos)

Tests de Hipótesis: Concepto

Pregunta: ¿El precio medio de los pisos en el Eixample es diferente que en Sants?

Enfoque de test de hipótesis:

1. **H_0 (hipótesis nula)**: No hay diferencia (μ_Eixample = μ_Sants)
2. **H_1 (hipótesis alternativa)**: Sí hay diferencia (μ_Eixample ≠ μ_Sants)
3. **Recolectar datos** y calcular un estadístico
4. **Calcular p-valor**: ¿Qué tan probable es ver estos datos si H_0 es cierto?
5. **Decidir**: Si p-valor < 0.05 → rechazamos H_0

P-valor: Interpretación

P-valor = probabilidad de observar datos tan extremos (o más) si H_0 es cierto

Regla común: $\alpha = 0.05$ (5%)

- $p < 0.05$ → Rechazamos H_0 → Resultado "estadísticamente significativo"
- $p \geq 0.05$ → No rechazamos H_0 → No hay evidencia suficiente

Interpretación CORRECTA:
"Si no hubiera diferencia real, veríamos datos como estos solo el 5% de las veces"

Interpretación INCORRECTA:
"Hay 5% de probabilidad de que H_0 sea cierto"

Errores en Tests de Hipótesis

	H_0 es verdad	H_0 es falsa
Rechazamos H_0	❌ Error Tipo I (falso positivo)	✅ Correcto
No rechazamos H_0	✅ Correcto	❌ Error Tipo II (falso negativo)

Error Tipo I: Decir que hay efecto cuando no lo hay (controlado por α)

Error Tipo II: No detectar un efecto que existe (relacionado con poder estadístico)

Trade-off: Reducir uno aumenta el otro

Segunda Parte: Tests Estadísticos en Práctica

T-test: Una Muestra

Pregunta: ¿El precio medio es diferente de 300k€?

H_0: $\mu = 300000$

H_1: $\mu \neq 300000$

```python
from scipy import stats

# T-test de una muestra
t_stat, p_value = stats.ttest_1samp(df['price'].dropna(), 300000)

print(f"t-statistic: {t_stat:.3f}")
print(f"p-value: {p_value:.4f}")

if p_value < 0.05:
    print("Rechazamos H₀: el precio medio es diferente de 300k")
else:
    print("No rechazamos H₀")
```

T-test: Dos Muestras Independientes

Pregunta: ¿El precio medio de pisos con terraza es diferente que sin terraza?

H_0: μ_con_terraza = μ_sin_terraza
H_1: μ_con_terraza ≠ μ_sin_terraza

```python
# Separar grupos
con_terraza = df[df['terrace'] == 1]['price'].dropna()
sin_terraza = df[df['terrace'] == 0]['price'].dropna()

# T-test de dos muestras
t_stat, p_value = stats.ttest_ind(con_terraza, sin_terraza)

print(f"Media con terraza: {con_terraza.mean():.0f}€")
print(f"Media sin terraza: {sin_terraza.mean():.0f}€")
print(f"p-value: {p_value:.4f}")
```

Visualizar Comparaciones

Antes de hacer el test, **siempre visualizar**:

```
import plotly.express as px

# Preparar datos
df_viz = df[df['terrace'].isin([0, 1])].copy()
df_viz['tiene_terraza'] = df_viz['terrace'].map({0: 'No', 1: 'Sí'})
```

```
# Box plot para comparar distribuciones
fig = px.box(df_viz, x='tiene_terraza', y='price',
             title='Precio según Terraza',
             labels={'tiene_terraza': 'Tiene Terraza', 'price': 'Precio (€)'}
fig.show()
```

Visualizar primero ayuda a entender el efecto antes del test

Test Chi-Cuadrado: Tabla de Contingencia

Pregunta: ¿Hay asociación entre barrio y tipo de propiedad?

Para variables categóricas → Chi-cuadrado (χ^2)

```
# Crear tabla de contingencia
tabla = pd.crosstab(df['neighborhood'], df['type'])
print(tabla)
```

Test Chi-Cuadrado: Realizar Test

```python
# Test chi-cuadrado
chi2, p_value, dof, expected = stats.chi2_contingency(tabla)

print(f"χ² statistic: {chi2:.2f}")
print(f"p-value: {p_value:.4f}")

if p_value < 0.05:
    print("Hay asociación significativa entre barrio y tipo")
```

Supuestos de los Tests

T-test:

- Datos aproximadamente normales (o $n > 30$)
- Varianzas homogéneas (para dos muestras)
- Muestras independientes

Chi-cuadrado:

- Frecuencias esperadas > 5 en cada celda
- Observaciones independientes

Si no se cumplen: estos tests no sirven y hay que aplicar otros

Patrón de Análisis Estadístico

1. Pregunta de negocio

"¿Los pisos con parking son más caros?"

2. Explorar datos

- Visualizar con box plots
- Calcular estadísticas descriptivas

3. Seleccionar y realizar test

- T-test de dos muestras
- Verificar supuestos

4. Interpretar resultado

- p-valor y significancia
- Tamaño del efecto (diferencia práctica vs estadística)
- **Escribir conclusión en lenguaje de negocio**

Significancia vs Importancia Práctica

Significancia estadística ($p < 0.05$):

- "Hay evidencia de diferencia"
- Con muestras grandes, diferencias pequeñas pueden ser significativas

Importancia práctica:

- "La diferencia es suficientemente grande para importar"
- Diferencia de 1000€ en precio puede ser estadísticamente significativa pero no prácticamente importante

Siempre reportar ambos: p-valor Y magnitud del efecto

Buenas Prácticas

✅ **Visualizar antes de testear**
✅ **Verificar supuestos del test**
✅ **Reportar IC además de p-valores**
✅ **Considerar importancia práctica, no solo estadística**
✅ **Preregistrar hipótesis cuando sea posible** (evita p-hacking)

❌ **No buscar p-valores hasta encontrar $p < 0.05$**
❌ **No hacer múltiples tests sin corrección**
❌ **No interpretar "no significativo" como "no hay efecto"**

🔬 Ahora al notebook

Ejercicio Tests de Hipótesis (45 minutos)

Gracias!