

Business Analytics & Data Science

Día 4: Estadística Descriptiva y Probabilidad

EAE Business School Barcelona
5 de febrero de 2026

Plan del Día 4

Primera Parte (9:00-11:00)

1. Medidas de tendencia central
2. Medidas de dispersión
3. Distribuciones de datos
4. Correlación entre variables

Segunda Parte (11:30-13:30)

1. Fundamentos de probabilidad
2. Distribución normal
3. Muestreo y aleatoriedad
4. Aplicaciones prácticas

Reaso: ¿Dónde Estamos?

Día 1: Python básico

Día 2: Fundamentos de pandas

Día 3: Limpieza de datos y visualización

Hoy: Estadística descriptiva y probabilidad

¿Por qué es importante?

- Entender qué nos dicen los datos
- Cuantificar incertidumbre
- Base para Machine Learning (todo ML es probabilístico)
- **La estadística es el lenguaje del análisis de datos**

Primera Parte: Estadística Descriptiva

¿Qué Es la Estadística Descriptiva?

Objetivo: Resumir y describir características de un conjunto de datos

Dos grandes categorías:

1. **Tendencia central:** ¿Dónde está el "centro" de los datos?
2. **Dispersión:** ¿Qué tan "dispersos" están los datos?

Analogía:

- Tendencia central = ¿Cuál es el precio típico?
- Dispersión = ¿Los precios son similares o muy variables?

Medidas de Tendencia Central

Media (promedio): Suma de todos los valores / cantidad

```
df["price"].mean()
```

Mediana: Valor central cuando ordenamos los datos

```
df["price"].median()
```

Moda: Valor más frecuente

```
df["price"].mode()
```

Media vs Mediana: ¿Cuándo Usar Cada Una?

Ejemplo: Salarios en una empresa

- Empleados: 25k, 28k, 30k, 32k, 35k, **500k** (CEO)

Media: $(25 + 28 + 30 + 32 + 35 + 500) / 6 = \mathbf{108k}$

Mediana: $(30 + 32) / 2 = \mathbf{31k}$

- **¿Cuál representa mejor el salario "típico"?** La mediana
- **Mediana:** Es más robusta contra outliers y siempre es un elemento de la muestra

Medidas de Dispersión

¿Por qué importan?

Dos datasets pueden tener la misma media pero ser muy diferentes:

Dataset A: 10, 10, 10, 10, 10 → Media = 10, Dispersión = 0

Dataset B: 0, 5, 10, 15, 20 → Media = 10, Dispersión = alta

Rango

Rango: Diferencia entre máximo y mínimo

```
rango = df["price"].max() - df["price"].min()
```

Ventaja: Muy fácil de calcular e interpretar

Desventaja: Sensible a outliers extremos

Ejemplo: Precios entre 150k y 2M → Rango = 1.85M

(Pero si 99% están entre 200k-400k, el rango no es muy informativo)

Varianza

$$\text{Var} = \mathbb{E}[(x - \bar{x})^2]$$

O, para una columna en pandas:

```
df["price"].var() = ((df["price"] - df["price"].mean()) ** 2).mean()
```

Desviación Estándar

Desviación estándar (σ): Dispersión promedio respecto a la media, se obtiene como la raíz cuadrada de la Varianza:

$$\sigma = \sqrt{\text{Var}}$$

```
df["price"].std()
```

Ventaja: tiene las mismas unidades que los datos de la muestra

Interpretación intuitiva:

- σ pequeña → Datos concentrados cerca de la media
- σ grande → Datos dispersos, lejos de la media

Regla práctica (para distribuciones normales):

- ~68% de datos están dentro de $\mu \pm 1\sigma$
- ~95% de datos están dentro de $\mu \pm 2\sigma$
- ~99.7% de datos están dentro de $\mu \pm 3\sigma$

Cuartiles y Percentiles

Cuartiles: Dividen los datos en 4 partes iguales

- **Q1 (25%)**: 25% de datos están por debajo
- **Q2 (50%)**: La mediana
- **Q3 (75%)**: 75% de datos están por debajo

```
df["price"].quantile(0.25) # Q1  
df["price"].quantile(0.50) # Q2 (mediana)  
df["price"].quantile(0.75) # Q3
```

Rango intercuartílico (IQR): $Q3 - Q1$

- Contiene el 50% central de los datos
- Útil para detectar outliers

Regla estándar: Un valor es outlier si:

- $\text{Valor} < Q1 - 1.5 \times \text{IQR}$
- $\text{Valor} > Q3 + 1.5 \times \text{IQR}$

```
Q1 = df["price"].quantile(0.25)
Q3 = df["price"].quantile(0.75)
IQR = Q3 - Q1

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

outliers = df[(df["price"] < limite_inferior) |
              (df["price"] > limite_superior)]
```

Esto es lo que usa un box plot para mostrar outliers



Ahora al
notebook

Ejercicio: Estadísticas Descriptivas
(20 minutos)

Distribuciones de Datos

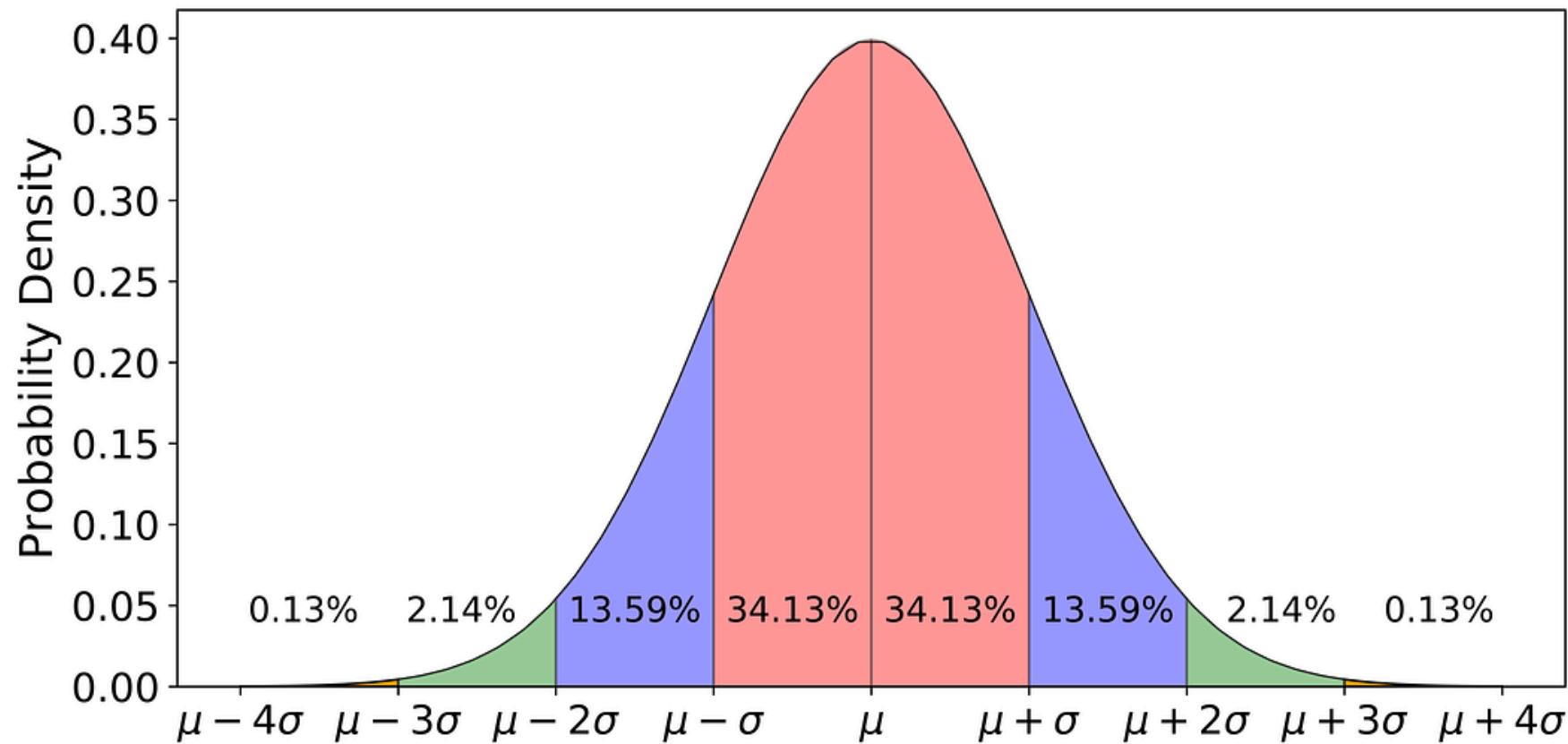
Distribución: Cómo se reparten los valores de una variable

Tipos comunes:

- **Normal (gaussiana):** Campana simétrica
- **Asimétrica derecha:** Cola larga hacia la derecha (precios, salarios)
- **Asimétrica izquierda:** Cola larga hacia la izquierda
- **Uniforme:** Todos los valores igual de probables
- **Bimodal:** Dos "picos"

Visualizar con histograma es esencial

Normal Distribution



La Distribución Normal

Características:

- Forma de campana simétrica
- Media = Mediana = Moda (todas en el centro)
- Definida por μ (media) y σ (desviación estándar)

¿Por qué es importante?

- Muchos fenómenos naturales siguen distribución normal
- Base de muchos tests estadísticos

Ejemplos: Altura de personas, errores de medición, rendimiento de estudiantes

Correlación Entre Variables

Correlación: Mide la relación lineal entre dos variables

```
df["price"].corr(df["sqrmtrs"])
```

Rango: -1 a +1

- **+1:** Correlación positiva perfecta (cuando X sube, Y sube)
- **0:** Sin correlación lineal
- **-1:** Correlación negativa perfecta (cuando X sube, Y baja)

Interpretación práctica:

- $|\rho| > 0.7$: Correlación fuerte
- $0.4 < |\rho| < 0.7$: Correlación moderada
- $|\rho| < 0.4$: Correlación débil

Matriz de Correlación

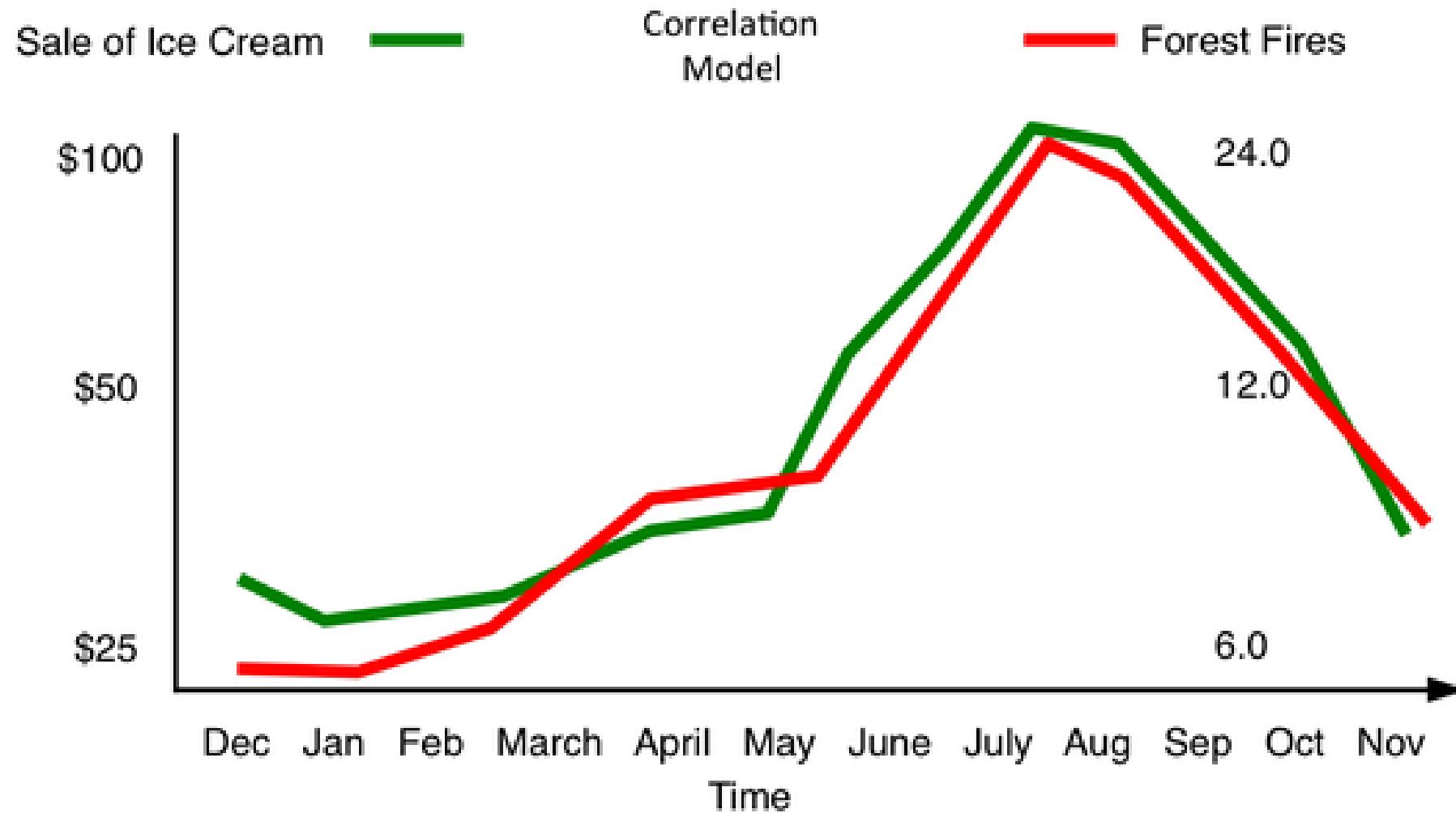
```
# Correlaciones entre todas las variables numéricas  
correlation_matrix = df.corr()  
print(correlation_matrix)
```

Útil para: Detectar relaciones entre variables antes de modelar

Visualizar Matriz de Correlación

```
import plotly.express as px

fig = px.imshow(correlation_matrix,
                 text_auto=True,
                 title="Matriz de Correlación")
fig.show()
```



Ejemplo clásico:

- Existe una fuerte correlación entre venta de helados e incendios forestales
- ¿Comer helado causa incendios?
- **NO**, hay una variable oculta que condiciona a ambas
- Temperatura (más calor → más helados Y más incendios)



Correlación NO Implica Causalidad

Causalidad requiere un análisis profundo

Para ejemplos de correlaciones alocadas, visitad [Spurious Correlations](#)



Ahora al
notebook

Ejercicio: Correlaciones (15
minutos)



Descanso de 30 minutos

Nos vemos a las 11:30 para probabilidad.

Segunda Parte: Fundamentos de Probabilidad

¿Qué Es la Probabilidad?

Definición: Medida de la incertidumbre de que ocurra un evento

Rango: 0 a 1

- $P = 0$: Imposible
- $P = 1$: Seguro
- $0 < P < 1$: Incierto

Ejemplo:

- Probabilidad de sacar cara en moneda justa: $P = 0.5$
- Probabilidad de que llueva mañana: $P = 0.3$

Reglas Básicas de Probabilidad (1/2)

1. Regla de la Suma (eventos mutuamente excluyentes)

$$P(A \cup B) = P(A) + P(B)$$

Ejemplo: Probabilidad de sacar 1 o 2 en dado = $1/6 + 1/6 = 1/3$

2. Regla del Producto (eventos independientes)

$$P(A \cap B) = P(A) \cdot P(B)$$

Ejemplo: Probabilidad de sacar dos caras seguidas es

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Reglas Básicas de Probabilidad (2/2)

3. Probabilidad Complementaria

$$P(\tilde{A}) = 1 - P(A)$$

Ejemplo: Si $P(\text{lluvia}) = 0.3$, entonces $P(\text{no lluvia}) = 0.7$

Eventos Independientes vs Dependientes

Independientes: Un evento no afecta al otro

- Lanzar dos monedas
- Resultados de dos tiradas de dado

Dependientes: Un evento afecta al otro

- Sacar dos cartas de baraja sin reemplazo
- Probabilidad de cancelación dada la temporada

Si A y B son independientes:

$$P(A \cap B) = P(A) \times P(B)$$

Probabilidad Condicional

$P(A|B)$: Probabilidad de A dado que B ya ocurrió

Fórmula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ejemplo real:

- $P(\text{cancelación}) = 0.20$
- $P(\text{cancelación} | \text{depósito_no_reembolsable}) = 0.05$

Interpretación: El depósito reduce la probabilidad de cancelación

Variable aleatoria

Es una cantidad X cuyo valor depende de un fenómeno aleatorio.

Ejemplos:

- La altura de una persona escogida al azar (contínua)
- 0 si un lanzamiento de moneda es cara, 1 si es cruz (discreta)
- El valor del resultado de lanzar un dado (discreta)
- La temperatura media que hará mañana (contínua)

Distribución de Probabilidad

Distribución de X es la ley que gobierna las probabilidades de los valores que puede tomar X

Ejemplo: Lanzar un dado justo

- $P(1) = 1/6$
- $P(2) = 1/6$
- ...
- $P(6) = 1/6$

Distribución uniforme: Todos los resultados son igual de probables

El histograma muestra la distribución empírica

La Distribución Normal

Es una distribución continua cuyos valores dependen sólo de dos parámetros:

- μ : Media (centro)
- σ : Desviación estándar (dispersión)

Se dice que X sigue una distribución normal con parámetros μ y σ y se escribe:

$$X \sim N(\mu, \sigma)$$

Ejemplo: Altura de mujeres españolas $\sim N(163, 6)$

- Media: 163 cm
- Desviación estándar: 6 cm

La distribución $N(0, 1)$ juega un papel central en la teoría de la probabilidad debido al **Teorema Central del Límite**

Si $X \sim N(\mu, \sigma)$, entonces

$$X \sim N(0, 1) * \sigma + \mu$$

Todas las distribuciones normales se pueden obtener a partir de $N(0, 1)$.

Regla Empírica (68-95-99.7)

Para distribución normal:

- **68%** de datos dentro de $\mu \pm 1\sigma$
- **95%** de datos dentro de $\mu \pm 2\sigma$
- **99.7%** de datos dentro de $\mu \pm 3\sigma$

Uso práctico:

- Detectar outliers (valores fuera de $\mu \pm 3\sigma$ son muy raros)
- Estimar intervalos de confianza
- Entender qué tan "extremo" es un valor

Estandarización (Z-Score)

Z-score: Cuántas desviaciones estándar está un valor de la media

Fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Interpretación:

- $z = 0$: Valor = media
- $z = 1$: Valor está 1 desviación estándar por encima de la media
- $z = -2$: Valor está 2 desviaciones estándar por debajo

```
from scipy import stats  
z_scores = stats.zscore(df["price"])
```

Uso: Detectar outliers ($|z| > 3$ es muy raro)

Muestreo y Aleatoriedad

Población: Todos los elementos de interés

Muestra: Subconjunto de la población

¿Por qué muestrear?

- No podemos medir toda la población (caro, imposible)
- Muestra representativa nos da información sobre población

Muestreo aleatorio: Cada elemento tiene igual probabilidad de ser seleccionado

```
# Muestra aleatoria de 100 propiedades  
muestra = df.sample(n=100, random_state=42)
```

Random Seed: Reproducibilidad

Operaciones aleatorias dan resultados diferentes cada vez

```
df.sample(5) # Cada vez diferente
```

Si fijamos el random seed, los resultados siempre serán iguales

```
df.sample(5, random_state=42) # Siempre igual
```

Usar `random_state` garantiza reproducibilidad a cambio de destruir la aleatoriedad.

Valor típico: 42 (por tradición - referencia a Guía del Autoestopista Galáctico)



Ahora al
notebook

Ejercicio: Probabilidad y
Distribuciones (30 minutos)

Aplicaciones Prácticas

1. Control de Calidad

- Si $\mu = 100\text{g}$, $\sigma = 2\text{g}$, productos fuera de $\mu \pm 3\sigma$ son defectuosos

2. Pricing

- Entender distribución de precios para estrategia competitiva

3. Riesgo

- Probabilidad de eventos extremos (pérdidas, cancelaciones)

4. A/B Testing

- ¿La diferencia observada es real o por azar?

5. Machine Learning

- Todos los modelos producen probabilidades
- Entender incertidumbre de predicciones

Gracias!