



Estadística Descriptiva Básica - Día 2

Change Status

in-progress

 40 min

 Learning Objectives

- 1

Calcular e interpretar medidas de tendencia central (media, mediana, moda)
- 2

Comprender y aplicar medidas de dispersión (varianza, desviación estándar, rango)
- 3

Utilizar estadísticos de posición (percentiles, cuartiles) para análisis de distribución

Theory

Practice

Quiz

Evidence

 Actividades y Aprendizajes

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

Task 1: Medidas de Tendencia Central (12 minutos)

Las medidas de tendencia central representan el **punto de equilibrio** de una distribución de datos, respondiendo a la pregunta fundamental: "¿Dónde se concentra la mayoría de los valores?". Cada medida ofrece una perspectiva diferente sobre el "centro" de los datos.

La Media Aritmética: El Centro de Gravedad

La media aritmética es el **promedio clásico** que todos conocemos, pero su interpretación requiere cuidado especial en análisis de datos.

Fórmula: $media = \frac{\sum(x_i)}{n}$ (suma de todos los valores dividida por el número de valores)

Fortalezas:

- Propiedad matemática:** Minimiza la suma de errores cuadráticos
- Sensibilidad a cambios:** Refleja inmediatamente cualquier modificación en los datos
- Base para estadística avanzada:** Fundamento de varianza, regresión, etc.

Limitaciones críticas:

Sensibilidad a outliers: Un solo valor extremo puede distorsionar completamente la media

No robusta: En distribuciones sesgadas, no representa el "típico"

Engañosa con datos faltantes: No indica si la media es representativa

Ejemplo práctico: En ingresos de empleados, si 99 ganan \$50,000 y 1 gana \$5,000,000, la media será \$100,000 pero la mediana \$50,000 refleja mejor la realidad.

La Mediana: El Valor Central Robusto

La mediana es el **punto medio** cuando los datos se ordenan, ofreciendo una medida robusta que resiste outliers.

Cálculo: Para datos ordenados, el valor en la posición $(n+1)/2$

Datos pares: Promedio de los dos valores centrales

Datos impares: Valor exactamente en el medio

Ventajas:

Robusta: Insensible a valores extremos

Intuitiva: El "valor típico" en distribuciones sesgadas

Siempre existe: Definida para cualquier distribución

Casos de uso:

Ingresos: Mejor que media para distribuciones con ricos extremos

Tiempos de respuesta: No afectado por timeouts ocasionales

Precios de vivienda: Resistente a mansiones excepcionales

La Moda: Los Valores Más Frecuentes

La moda representa los **valores más comunes** en el dataset, revelando preferencias naturales o categorías dominantes.

Tipos de moda:

Unimodal: Una sola categoría dominante

Bimodal: Dos categorías con frecuencias similares

Multimodal: Múltiples categorías frecuentes

Sin moda: Todos los valores igualmente frecuentes

Aplicaciones prácticas:

Segmentación de mercado: Productos más vendidos por categoría

Análisis de comportamiento: Acciones más frecuentes de usuarios

Detección de patrones: Categorías dominantes en datos categóricos

Task 2: Medidas de Dispersión y Variabilidad (10 minutos)

Mientras las medidas de tendencia central nos dicen **dónde** se concentran los datos, las medidas de dispersión explican **cuán dispersos o concentrados** están alrededor de ese centro.

El Rango: La Extensión Total

El rango es la **medida más simple** de dispersión: la distancia entre el valor más pequeño y el más grande.

Fórmula: rango = máximo - mínimo

Interpretación:

Rango amplio: Datos muy dispersos, mucha variabilidad

Rango estrecho: Datos concentrados alrededor del centro

Limitaciones:

Sensibilidad a outliers: Un solo valor extremo distorsiona el rango

No informativo sobre distribución: Dos datasets pueden tener mismo rango pero formas muy diferentes

La Varianza y Desviación Estándar: Medidas Estadísticas de Dispersión

La varianza mide la **dispersión promedio** de los datos alrededor de la media, cuantificando qué tan "extendidos" están.

Fórmula de varianza: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$ (para población) o $\frac{\sum (x_i - \mu)^2}{(n-1)}$ (para muestra)

Desviación estándar: Raíz cuadrada de la varianza, en las mismas unidades que los datos originales.

Interpretación práctica:

Desviación estándar pequeña: Datos concentrados alrededor de la media

Desviación estándar grande: Datos muy dispersos

Regla empírica: ~68% de datos están dentro de 1 DE, ~95% dentro de 2 DE, ~99.7% dentro de 3 DE

Coefficiente de variación: $CV = \frac{\sigma}{\mu}$ - medida relativa de dispersión, útil para comparar variables con unidades diferentes.

Task 3: Estadísticos de Posición y Percentiles (8 minutos)

Los percentiles dividen la distribución en **porciones iguales**, permitiendo análisis granular de la posición relativa de valores específicos.

Cuartiles y Percentiles Básicos

Percentiles: Dividen la distribución en 100 partes iguales

- P25 (Q1):** 25% de datos por debajo
- P50 (Mediana):** 50% de datos por debajo
- P75 (Q3):** 75% de datos por debajo

Rango intercuartílico (IQR): $IQR = Q3 - Q1$ - medida robusta de dispersión que ignora outliers.

Aplicaciones en Análisis de Datos

Identificación de outliers: Método IQR

- Límite inferior:** $Q1 - 1.5 \times IQR$
- Límite superior:** $Q3 + 1.5 \times IQR$
- Valores fuera de estos límites se consideran outliers

Análisis de distribución:

- Distribución simétrica:** $Q2 \approx (Q1 + Q3)/2$
- Sesgo positivo:** $Q2 > (Q1 + Q3)/2$
- Sesgo negativo:** $Q2 < (Q1 + Q3)/2$

Comparaciones relativas:

- P90:** Valor por encima del cual están el 10% superior
- P10:** Valor por debajo del cual están el 10% inferior