Estadística aplicada

Correlación, regresión y métodos de muestreo

Clase 04

CONQUER BLOCKS

¿Qué es la estadística?

¿Qué es la estadística?

La estadística es una **rama de las matemáticas** que se encarga de **recolectar, analizar, interpretar, presentar y organizar datos**. Es una herramienta fundamental en muchas disciplinas, como la economía, la biología, la ingeniería, la psicología, la sociología, **la inteligencia artificial** y muchas otras, ya que **permite tomar decisiones informadas** basadas en datos y tendencias.

Descriptiva

Esta área se enfoca en **describir y resumir un conjunto de datos**. Utiliza **medidas** como la media, la mediana, la moda, la desviación estándar, y **gráficos** como histogramas y diagramas de dispersión para representar la información de manera comprensible.

Inferencial - Regresión

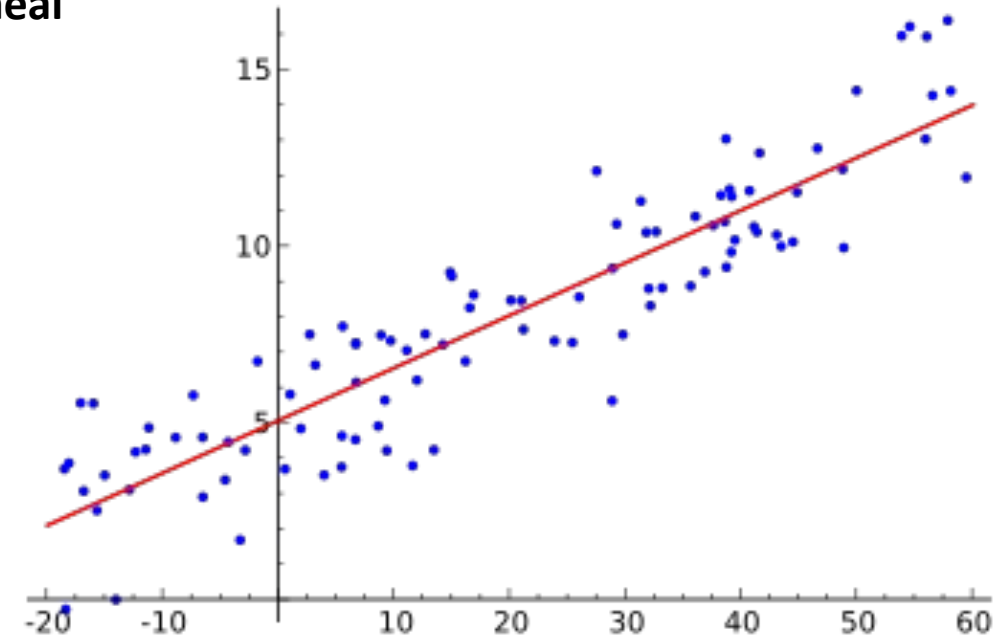
Esta área se centra en hacer inferencias o **predicciones** sobre una población **a partir de una muestra de datos.** Utiliza técnicas como la estimación de intervalos de confianza, pruebas de hipótesis, análisis de regresión, y análisis de varianza **para llegar a conclusiones sobre la población de interés**.

Correlación y regresión

¿Qué es la regresión?

La regresión es un método estadístico utilizado para **comprender la relación entre dos o más variables**. En la regresión lineal simple, se usa una variable independiente para predecir el valor de una variable dependiente. La relación entre las variables se modela ajustando una ecuación lineal a los datos observados.

Lineal



$$y = a + bx$$

Donde:
y es la variable dependiente,
x es la variable independiente,
a es la intersección en el eje Y,
b es la pendiente de la línea.

Correlación y regresión

¿Para qué sirve o se utiliza la regresión?

Predicción: La regresión se usa para **pronosticar** el valor de una variable dependiente **basándose en una o más variables independientes**. *Por ejemplo, se puede utilizar para predecir los precios **de una vivienda**: Usando la regresión lineal, se puede predecir el precio de una casa (variable dependiente) basado en características como el tamaño de la casa, el número de habitaciones, la ubicación, etc. (variables independientes).*

Identificación de relaciones: Nos ayuda a identificar y cuantificar la **relación entre variables**. *Por ejemplo, saber si existe **relación entre la publicidad y las ventas**: Una empresa puede usar la regresión para entender cómo sus gastos en publicidad (variable independiente) afectan sus ventas (variable dependiente). mantener la calidad del producto.*

Análisis de tendencias: Nos ayuda a analizar y entender **tendencias** en datos históricos. *Por ejemplo, **Análisis de tendencias en temperatura**: Los climatólogos pueden usar la regresión para analizar datos históricos de temperaturas y así identificar tendencias de calentamiento o enfriamiento global.*

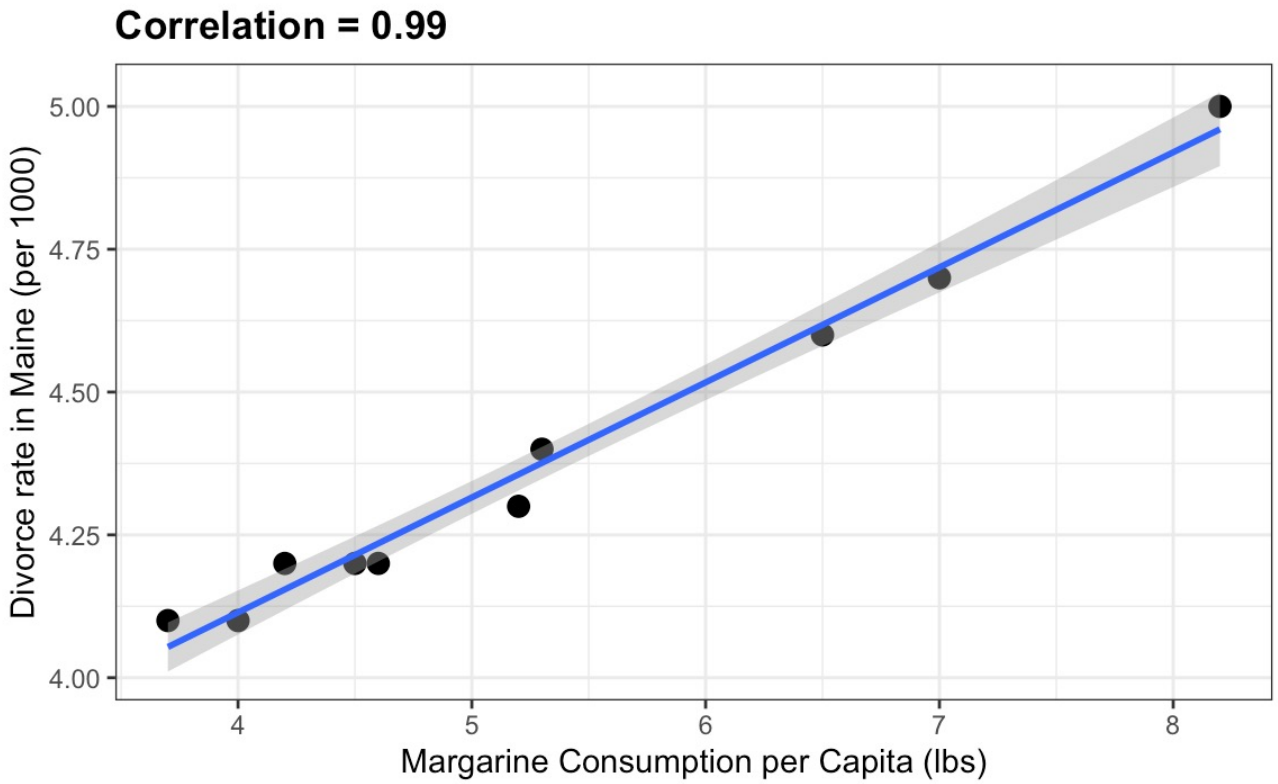
Optimización: La regresión se emplea para encontrar los **valores óptimos** de variables que **maximicen o minimicen algún resultado**. *Por ejemplo, en la Optimización de campañas de marketing: Una empresa puede usar la regresión para determinar **la combinación óptima de canales de marketing** (por ejemplo, televisión, radio, internet) **que maximice las ventas** o minimice los costes.*

Correlación y regresión

¿Qué es la correlación?

La correlación mide la **fuerza** y la dirección de la **relación lineal entre dos variables**

La correlación NO implica causalidad



Correlación y regresión

¿Cómo se pude medir la correlación entre variables?

Coeficiente de correlación de Pearson

Mide la relación lineal entre dos variables. Sus valores van de -1 a 1, donde:
•1 indica una correlación positiva perfecta,
•-1 indica una correlación negativa perfecta,
•0 indica que no hay correlación lineal.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

donde

- n es el tamaño de la muestra.
- x_i, y_i son puntos muestrales individuales indexados con i.
- $\bar{x}$ denota la media muestral definida por $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ (análogamente para $\bar{y}$).

Coeficiente de correlación de Spearman

Mide la relación monotónica (no necesariamente lineal) entre dos variables. También va de -1 a 1 y es útil cuando los datos no necesariamente siguen una relación lineal.

La fórmula del coeficiente de correlación de Spearman (r_s) es:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

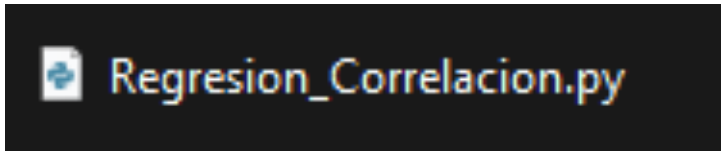
donde:

- d_i es la diferencia entre los rangos de cada par de datos.
- n es el número de pares de datos.

https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Spearman

Correlación y regression: Cálculo

¿Cómo cálculo la correlación y la regresión?



Métodos de Muestreo

¿Qué son los métodos de muestreo?

Los métodos de muestreo son **técnicas** utilizadas **para seleccionar una muestra representativa** de una población más grande. La muestra se estudia para **inferir o hacer conclusiones sobre la población completa**. Los métodos de muestreo son fundamentales en la investigación estadística y en diversas áreas como las ciencias sociales, la medicina, la economía, etc.

Es más económico y rápido estudiar una muestra que toda la población.

Un muestreo bien hecho puede proporcionar resultados precisos.

Facilita el estudio de poblaciones muy grandes que serían difíciles de estudiar en su totalidad.

Una muestra adecuada puede ofrecer datos útiles y precisos para el análisis.

Métodos de Muestreo: Tipos

¿Tipos de métodos de muestreo?

Muestreo Simple

Cada miembro de la población tiene la misma probabilidad de ser seleccionado. Es como si se seleccionara una muestra al azar de un "sombrero" donde cada elemento de la población está presente.

Ventajas

- Fácil de entender y aplicar.
- Resulta en una muestra representativa si se hace correctamente

Desventajas

- Puede ser difícil de implementar con grandes poblaciones.
- No garantiza la representatividad si la población tiene subgrupos significativos.

Ejemplo

Supongamos que queremos estudiar los hábitos de lectura de los estudiantes en una escuela con 1,000 estudiantes. Si seleccionamos 100 estudiantes al azar, estamos realizando un muestreo aleatorio simple.

Métodos de Muestreo: Tipos

¿Tipos de métodos de muestreo?

Muestreo Estratificado

La población se divide en subgrupos (estratos) que comparten características similares, y se toma una muestra aleatoria de cada estrato proporcionalmente a su tamaño en la población.

Ventajas

- Más preciso y representativo que el muestreo aleatorio simple, especialmente cuando hay diferencias importantes entre los estratos.
- Permite comparaciones entre diferentes subgrupos.

Desventajas

- Requiere conocer las características de la población para poder dividirla en estratos.
- Puede ser más complicado y costoso de implementar.

Ejemplo

Si queremos estudiar la satisfacción laboral en una empresa con 500 empleados, podríamos dividir a los empleados en estratos basados en el departamento (por ejemplo, marketing, ventas, recursos humanos) y luego seleccionar aleatoriamente un número proporcional de empleados de cada departamento.

Métodos de Muestreo: Tipos

¿Tipos de métodos de muestreo?

Muestreo por Conglomerados

La población se divide en grupos (conglomerados) y se seleccionan aleatoriamente uno o más conglomerados completos. Luego, se estudian todos los miembros de los conglomerados seleccionados o se toma una muestra de ellos.

Ventajas

- Más económico y fácil de implementar cuando la población está dispersa geográficamente.
- Útil para estudios a gran escala.

Desventajas

- Menos preciso que el muestreo estratificado si los conglomerados no son homogéneos.
- Puede introducir un sesgo si los conglomerados seleccionados no son representativos de la población.

Ejemplo

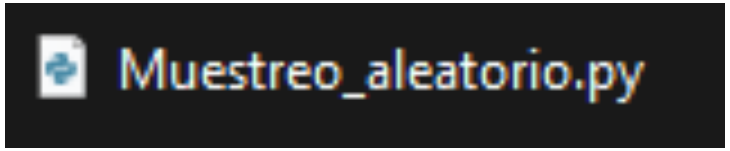
Si queremos estudiar el rendimiento académico de los estudiantes de secundaria en una ciudad, podríamos dividir las escuelas en conglomerados y seleccionar aleatoriamente varias escuelas (conglomerados). Luego, podríamos estudiar todos los estudiantes de las escuelas seleccionadas.

*Los conglomerados se refiere a grupos completos

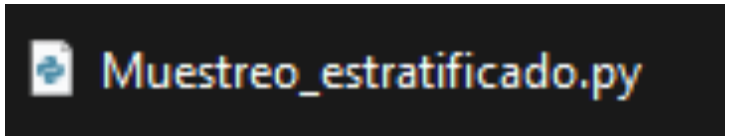
Métodos de Muestreo

¿Cómo se realizan los diferentes muestreos con python?

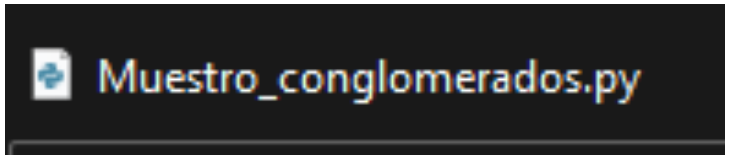
Muestreo Simple



Muestreo Estratificado



Muestreo por Conglomerados



Métodos de Muestreo: Tamaños y errores

Tamaño de la muestra y error

El **tamaño** de la muestra es un factor **crítico en cualquier método de muestreo** y tiene un impacto significativo en la precisión y la confiabilidad de los resultados

Efectos del Tamaño de la Muestra

Precisión: Un tamaño de muestra mayor tiende a proporcionar estimaciones más precisas de los parámetros de la población. Esto se debe a que una muestra más grande reduce la variabilidad de la estimación.

Representatividad: Una muestra mayor es más probable que capture la diversidad de la población, lo que la hace más representativa.

Coste y Tiempo: Recoger y analizar una muestra más grande puede ser más costoso y llevar más tiempo.

Métodos de Muestreo: Tamaños y errores

Relación entre el Tamaño de la Muestra y el Error de Muestreo

Error Estándar: El error estándar es una medida de la **precisión** de la estimación de la muestra. **Disminuye a medida que aumenta el tamaño de la muestra**. Matemáticamente, el error estándar de la media (σ_x̄) se calcula como:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

donde σ es la desviación estándar de la población y
n es el tamaño de la muestra.

Error_std.py

Aplicaciones de la estadística en IA

Normalización de datos

La normalización de datos transforma los valores de las variables en un rango común, generalmente entre 0 y 1 o -1 y 1.

¿Por qué es útil?

Diferentes características en un conjunto de datos pueden tener **diferentes unidades y escalas**. Por ejemplo, una característica podría representar ingresos en miles de dólares, mientras que otra podría representar la edad en años. Si estas características no se normalizan, **las diferencias en la escala pueden afectar negativamente el rendimiento de los algoritmos de aprendizaje automático.**

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Aplicaciones de la estadística en IA

Estandarización de datos

La estandarización de datos transforma los valores para que tengan una media de 0 y una desviación estándar de 1.

¿Por qué es útil?

Al igual que la normalización, la estandarización es crucial cuando las características tienen diferentes escalas. Esto es especialmente importante para algoritmos que asumen que los datos están centrados en torno a cero y tienen una variabilidad constante.

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

Donde:
μ es la media de la característica y
σ es la desviación estándar.

Aplicaciones de la estadística en IA

Gestión de datos “faltantes” – Missing Data

Los datos faltantes ocurren cuando no hay valores disponibles para algunas observaciones en el conjunto de datos.

¿Por qué es útil?

Los algoritmos de IA generalmente requieren conjuntos de datos completos. Los datos faltantes pueden sesgar los resultados y reducir la precisión del modelo.

¿Cómo se gestionan?

Hay varias técnicas estadísticas para manejar datos faltantes:

1.Eliminación: Se eliminan las filas o columnas con datos faltantes. Esta técnica es simple pero puede resultar en la pérdida de información valiosa.

2.Imputación: Se rellenan los valores faltantes con estimaciones. Algunas técnicas de imputación comunes son:

1. **Media/Mediana/Moda**: Rellenar con la media, mediana o moda de la columna.
2. **Regresión**: Utilizar un modelo de regresión para predecir los valores faltantes.
3. **Imputación por KNN (K-Nearest Neighbors)**: Rellenar con valores de observaciones similares.

Aplicaciones de la estadística en IA

Limpieza de Datos

La limpieza de datos es el proceso de identificar y corregir o eliminar datos corruptos, incorrectos, o irrelevantes.

¿Por qué es útil?

Los datos sucios pueden llevar a modelos de IA inexactos o sesgados. La limpieza de datos asegura que el modelo aprenda patrones correctos y generalice bien.

¿Cómo se realiza?

- **Detección y eliminación de valores atípicos (outliers)**: Los valores atípicos pueden distorsionar los resultados. Se pueden detectar usando métodos como el rango intercuartílico (IQR) o desviaciones estándar.
- **Corrección de errores tipográficos y duplicados**: Corregir errores de entrada de datos y eliminar registros duplicados.
- **Transformación de datos**: Asegurarse de que los datos estén en el formato correcto. Por ejemplo, convertir fechas a un formato uniforme.

Aplicaciones de la estadística en IA

Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar la capacidad de generalización de un modelo de IA. Consiste en dividir el conjunto de datos en múltiples subconjuntos y entrenar/validar el modelo en diferentes combinaciones de estos subconjuntos.

¿Por qué es útil?

Ayuda a evitar el sobreajuste (*overfitting*), asegurando que el modelo generalice bien a datos no vistos, en lugar de memorizar el conjunto de entrenamiento.

¿Cómo se realiza?

La forma más común es la validación cruzada **k-fold**:

k-fold: Se divide el conjunto de datos en k subconjuntos (folds).
El modelo se entrena en k-1 folds y se valida en el fold restante.
Este proceso se repite k veces, usando cada fold como conjunto de validación una vez.

Aplicaciones de la estadística en IA

División de Datos en Entrenamiento, Validación y Prueba

Es una práctica común dividir el conjunto de datos en tres subconjuntos distintos:

1. **Entrenamiento (Training Set)**: Se utiliza para entrenar el modelo.

2. **Validación (Validation Set)**: Se utiliza para ajustar los *hiperparámetros* y evaluar el modelo durante el entrenamiento.
3. **Prueba (Test Set)**: Se utiliza para evaluar el rendimiento final del modelo en datos no vistos.

¿Por qué es útil?

Esta división asegura que el modelo se entrene y ajuste correctamente y que se evalúe de manera imparcial en datos no vistos.

Aplicaciones de la estadística en IA

Métricas de Evaluación de los modelos

Las métricas de evaluación son herramientas estadísticas utilizadas para medir el rendimiento de los modelos de IA, especialmente en problemas de clasificación. Las métricas más comunes son la precisión (accuracy), el recall (sensibilidad o exhaustividad) y el F1-score.

Precisión / Accuracy

La precisión mide la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el total de predicciones.

$$\text{Precisión} = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall (Sensibilidad o Exhaustividad)

El recall mide la proporción de verdaderos positivos que se identifican correctamente. Es especialmente útil cuando es importante capturar todos los positivos.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-score

El F1-score es la media armónica de la precisión y el recall, proporcionando un balance entre ambas métricas.

$$\text{F1-score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$