**Métodos Estadísticos para IA**

**Resumen ejecutivo**

Este documento es una versión ampliada y mejorada de las diapositivas proporcionadas sobre **Estadística Descriptiva** aplicable a Inteligencia Artificial (IA). Está pensado como una monografía/tesina para presentar y exponer en clase: desarrolla cada tema con explicaciones teóricas, ejemplos reales y ejercicios prácticos (incluye sugerencias de código en Python para EDA). También contiene recomendaciones para su presentación en diapositivas y posibles extensiones hacia análisis inferencial y modelos de IA.

**Objetivos**

1. Explicar en detalle las técnicas de estadística descriptiva útiles en IA.
2. Enseñar a construir y a interpretar tablas de frecuencias, medidas de posición y dispersión.
3. Presentar los principales gráficos de la estadística descriptiva y su interpretación en problemas reales de IA.
4. Proveer ejemplos aplicados, fragmentos de código y tareas prácticas que sirvan como base para trabajos y exposiciones.

**Índice**

1. Introducción: por qué la estadística descriptiva importa en IA
2. Tipos de variables y su tratamiento
3. Construcción de tablas de frecuencias (con ejemplos)
4. Medidas de posición: definición, fórmulas y ejemplos reales
5. Medidas de dispersión: rango, IQR, varianza, desviación estándar y coeficiente de variación
6. Gráficos descriptivos: histogramas, ECDF, Q–Q plots, diagramas de caja
7. Interpretación práctica: EDA para proyectos de IA
8. Ejemplos aplicados (casos reales)
9. Recomendaciones para presentar la monografía y la exposición
10. Glosario y referencias

**1. Introducción: por qué la estadística descriptiva importa en IA**

Antes de entrenar un modelo de IA es imprescindible conocer la distribución, la dispersión, la presencia de valores atípicos y la relación entre variables. La estadística descriptiva (ED—Exploratory Data Analysis) permite:

* Resumir grandes volúmenes de datos de manera comprensible.
* Detectar errores, valores faltantes y outliers que condicionan modelos.
* Decidir transformaciones (por ejemplo: log, Box–Cox) para cumplir supuestos de modelos lineales o mejorar convergencia en entrenamiento.
* Seleccionar y escalar features apropiadamente.

**Ejemplo real:** Antes de entrenar un modelo de regresión para predecir el tiempo de entrega de pedidos, la EDA puede revelar que la variable "distancia" tiene una cola larga (pocos pedidos muy lejanos). Un transformado logarítmico reducirá la influencia de esos casos extremos y mejorará el ajuste.

**2. Tipos de variables y su tratamiento**

**Variable estadística:** característica observada en cada individuo o instancia.

**Tipos y ejemplos aplicados**

* **Variables cualitativas/categóricas** (no numéricas)
  + *Nominales*: categorías sin orden (p.ej. tipo de producto: "A", "B", "C").
  + *Ordinales*: categorías con orden (p.ej. nivel de satisfacción: "bajo", "medio", "alto").
* **Variables cuantitativas** (numéricas)
  + *Discretas*: toman valores enteros finitos (p.ej. número de compras por mes).
  + *Continuas*: pueden tomar cualquier valor dentro de un intervalo (p.ej. tiempo de respuesta en segundos, peso, altura, temperatura).

**Consejo práctico:** el tipo de variable determina las medidas y gráficas adecuadas (frecuencias y barras para categóricas; histogramas, boxplots y medidas de tendencia para numéricas).

**3. Construcción de tablas de frecuencias**

Una tabla de frecuencias resume cuántas observaciones caen en cada categoría o intervalo.

**Elementos de una tabla de frecuencias (para variables continuas)**

* **Intervalos de clase** (bins)
* **Frecuencia absoluta (f\_i)**: número de observaciones en el intervalo
* **Frecuencia absoluta acumulada (F\_i)**
* **Frecuencia relativa (h\_i = f\_i / N)**
* **Frecuencia relativa acumulada (H\_i = F\_i / N)**

**Cómo elegir intervalos**

* Calcular el rango: (R = x\_{max} - x\_{min}).
* Regla práctica para el número de intervalos: (k \approx \lceil \sqrt{n} \rceil) (entre 5 y 20 normalmente).
* Ancho de clase: (w = R / k).

**Ejemplo aplicado:** análisis de tiempos de sesión en una app (n = 1000). Calcular k ≈ 32, pero se puede optar por 10–20 bins para visualización.

**Ejemplo rápido en Python (con pandas)**

import pandas as pd

import numpy as np

# datos simulados

np.random.seed(0)

data = np.random.gamma(shape=2, scale=3, size=200)

s = pd.Series(data, name='tiempo\_sesion')

# tabla de frecuencias con 10 bins

freq\_table = pd.cut(s, bins=10).value\_counts().sort\_index()

rel = freq\_table / len(s)

pd.DataFrame({'fi': freq\_table, 'hi': rel})

**4. Medidas de posición**

**4.1 Media aritmética (muestra)**

Definición: (\bar{x}=\frac{1}{n}\sum\_{i=1}^n x\_i).

**Propiedades:** sensible a valores extremos.

**Ejemplo:** ingreso promedio mensual en una pequeña muestra: [20k, 22k, 21k, 1M] → la media estaría muy sesgada por el millonario; la mediana refleja mejor el "típico".

**4.2 Mediana**

Valor que deja el 50% de observaciones por debajo. Robusta frente a outliers.

**Ejemplo:** mediana de salarios evita la distorsión por salarios muy altos.

**4.3 Moda**

Valor con mayor frecuencia. Útil para variables categóricas (p.ej. el modelo de auto más vendido) o para detectar picos en distribuciones discretas.

**4.4 Cuartiles y percentiles**

* **Q1 (25%)**, **Q2 (50%) = mediana**, **Q3 (75%)**.
* **Percentil 90 (P90)**: el 90% de observaciones son ≤ P90.

**Ejemplo práctico:** en modelos de recomendación, usar P75 o P90 del tiempo de interacción para definir umbrales de usuario frecuente.

1. **Medidas de dispersión**

**5.1 Rango muestral**

(R = x\_{(n)} - x\_{(1)}). Fácil, pero sensible a outliers.

**5.2 Rango intercuartílico (IQR)**

(IQR = Q3 - Q1). Más robusto.

**Uso:** definir límites para detectar outliers: normalmente se consideran atípicos valores fuera de ([Q1 - 1.5\cdot IQR, Q3 + 1.5\cdot IQR])).

**5.3 Varianza muestral y desviación estándar**

* Varianza muestral: (s^2 = \frac{1}{n-1}\sum\_{i=1}^n (x\_i - \bar{x})^2).
* Desviación estándar: (s = \sqrt{s^2}).

**Interpretación:** mide la dispersión alrededor de la media. En ML, la desviación estándar sirve para estandarizar features (z–score).

**5.4 Coeficiente de variación (CV)**

(CV = s / \bar{x}) (sin unidades). Permite comparar variabilidad relativa entre variables con distinta escala.

**Ejemplo aplicado:** comparar la variabilidad relativa de latencia de red (ms) frente al número de requests (conteo). CV permite comparar.

1. **Gráficos descriptivos**

**6.1 Histogramas**

* Visualizan la estimación de la función de densidad.
* Pueden normalizarse (área total = 1) para compararse con distribuciones teóricas.

**Ejemplo:** histograma de edades de usuarios para detectar segmentos dominantes.

**Python (matplotlib/seaborn)**

import matplotlib.pyplot as plt

import seaborn as sns

sns.histplot(data=s, bins=20, stat='density')

**6.2 Distribución acumulada empírica (ECDF)**

* Grafica la fracción de datos ≤ x.
* Muy útil para comparar distribuciones empíricas y teóricas.

**Ejemplo:** comparar ECDF de tiempos de latencia entre servidores A y B.

**6.3 Q–Q plots**

* Comparan cuantiles muestrales con los cuantiles teóricos (p.ej. normales).
* Si los puntos quedan cerca de la recta, la normalidad es plausible.

**Importancia en IA:** algunos métodos (t-tests, ciertos intervalos de confianza, regresión lineal clásicamente) asumen normalidad de errores.

**6.4 Diagramas de caja (boxplots)**

* Muestran mediana, Q1, Q3, IQR y outliers (puntos fuera de LI/LS).
* Útiles para comparar distribuciones entre grupos.

**Ejemplo aplicado:** comparar consumo de combustible de autos por número de cilindros (mtcars); identificar outliers.

1. **Interpretación práctica: EDA en proyectos de IA**

**Pasos sugeridos**

1. **Revisión rápida:** df.info(), df.describe(), conteo de nulos.
2. **Distribuciones por variable:** histogramas / boxplots.
3. **Relaciones bivariadas:** scatterplots, tablas de contingencia para categóricas.
4. **Transformaciones:** log, raíz o escalado según necesidad.
5. **Detección/gestión de outliers:** decidir si imputar, truncar o dejar.
6. **Escalado de features:** StandardScaler (z–score), MinMaxScaler.

**Ejemplo aplicado (pipeline):** clasificación de churn — EDA revela que la variable "tiempo\_con\_el\_servicio" está muy sesgada; aplicar log(1 + x) mejora la performance del modelo tree-based y sobre todo de modelos lineales.

**8. Ejemplos aplicados — Casos de la vida real**

**Caso A: Predicción de demanda en e‑commerce**

* **Problema:** predecir demanda diaria para optimizar inventario.
* **ED:** histogramas por producto, ECDF de demanda, IQR para detectar días atípicos (promociones), CV para comparar volatilidad entre SKUs.
* **Decisión:** para SKUs con CV alto, usar modelos que capten heterocedasticidad (p.ej. modelos de series temporales con volatilidad).

**Caso B: Detección de fraude en transacciones**

* **ED:** resumir montos (media vs mediana), identificar outliers por IQR y límites, boxplots por región.
* **Decisión:** crear features binarias para grandes transacciones (por encima del P99) y escalar montos con log.

**Caso C: Salud — análisis de tiempos de recuperación**

* **ED:** ECDF para comparar tratamientos, Q–Q plots para comprobar normalidad, uso de mediana y IQR cuando distribución sesgada.

**9. Recomendaciones para la presentación/exposición**

* **Estructura clara:** introducir objetivos, metodología, resultados y conclusiones.
* **Visuales limpias:** 1 figura por diapositiva; evitar saturar texto.
* **Interpretación siempre:** mostrar la gráfica y añadir 1–2 bullet points sobre la interpretación y decisiones a tomar.
* **Código en apéndice:** incluir fragmentos de Python en un anexo o repositorio GitHub.
* **Ejercicio propuesto:** EDA de un dataset (p.ej. Blackmore.csv mencionado en diapositivas) con informe de 2 páginas.

**10. Glosario y fórmulas clave**

* Media: (\bar{x}=\frac{1}{n}\sum x\_i)
* Varianza muestral: (s^2=\frac{1}{n-1}\sum (x\_i-\bar{x})^2)
* IQR: (Q3 - Q1)
* CV: (s/\bar{x})