


Imputación de Datos

- La imputación de datos es clave para manejar datos faltantes en análisis.
- Evita sesgos y mejora la calidad de los resultados.

 por conectiva oficial



Viabilidad de la imputación

% Faltantes	Acción recomendada
< 5%	Imputar sin problema
5%-20%	Imputar con precaución
20%-40%	Imputar solo si es importante y hay buena justificación, de lo contrario eliminar la variable
> 40%	Mejor eliminar o tratar como categoría aparte

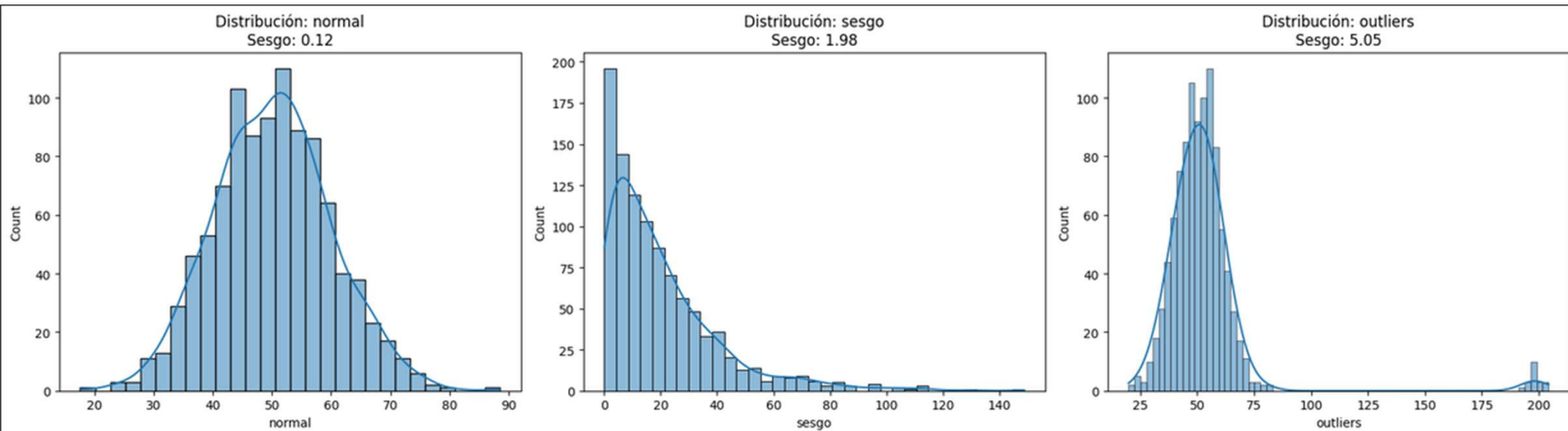
Métodos de imputación según el contexto

Imputación simple

- Para datos numéricos: Usar la media, mediana o moda.
- Para datos categóricos: Usar la categoría más frecuente o crear una nueva categoría llamada "Desconocido".

Imputación avanzada

- Regresiones: Usar variables relacionadas para predecir los valores faltantes.
- Modelos basados en vecinos cercanos (KNN): Ideal si los datos tienen patrones complejos.
- Técnicas basadas en aprendizaje automático: Por ejemplo, árboles de decisión, Random Forest entre otros.



- **Media:** si los datos siguen una distribución normal o se acerca a ella. Es sensible a valores extremos, un solo dato puede distorsionar el valor
- **Mediana:** Se usa cuando los datos están sesgados o tienen outliers. Si embargo no tiene en cuenta todo el conjunto de datos como lo tiene la media
- **Moda:** Es útil para datos categóricos o datos numéricos donde un valor específico se repite con mucha frecuencia. Si no hay un valor que predomine o si todos tiene la misma frecuencia, la moda es inútil

Como analizar la distribución de los datos

- Gráficos: histogramas, box plot, diagramas de densidad
- Pruebas estadísticas: coeficiente de asimetría, test de normalidad

```
# Test de normalidad
from scipy.stats import shapiro
stat, p = shapiro(df['columna'])
print(f"Estadístico W = {stat:.4f}")
print(f"p-valor = {p:.4f}")
```

Nota: si es > 0.05 Los datos parecen seguir una distribución normal



Como analizar la distribución de distribución de los datos

Asimetría

```
from scipy.stats import skew
```

```
asimetria = df['Nota'].skew()
```

```
print(f"Asimetría (skewness): {asimetria:.4f}")
```

Skewness	Interpretación	Forma de la distribución
≈ 0	Distribución aproximadamente simétrica	Como la normal
> 0 (positiva)	Sesgo a la derecha (cola más larga hacia la derecha)	Valores extremos altos
< 0 (negativa)	Sesgo a la izquierda (cola más larga hacia la izquierda)	Valores extremos bajos
> 1 o < -1	Asimetría alta o severa	Muy inclinada
0.5 a 1 o -0.5 a -1	Asimetría moderada	
-0.5 a 0.5	Asimetría ligera o despreciable	Muy cerca a simétrica



Naturaleza de los datos

- **MCAR (Completamente Aleatorios)**: Los datos faltantes no dependen de ninguna variable.
- **MAR (Aleatorios Condicionales)** : Los datos faltantes dependen de otras variables conocidas (por ejemplo, ingresos afectan si falta el sexo).
- **NMAR (No Aleatorios)**: Los datos faltantes dependen de la propia variable faltante (por ejemplo, alguien no indica su sexo porque es sensible al tema).



Pasos para imputar datos

1. Analizar los datos faltantes

- Identifica el porcentaje de valores faltantes.
- Verifica si los datos faltantes están distribuidos de forma aleatoria o tienen un patrón MCAR, MAR, NMAR.

2. Entender la variable afectada:

- ¿Es categórica o numérica?
- ¿Cuál es su distribución (usando histogramas o proporciones)?

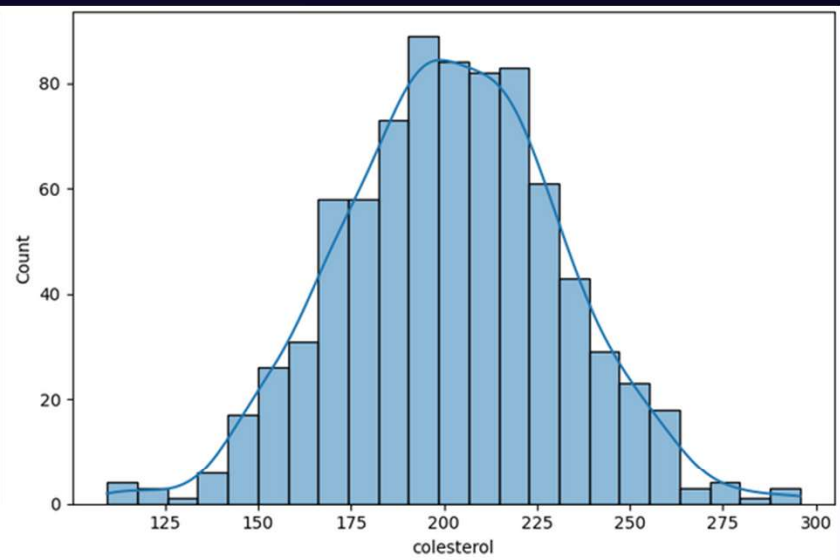
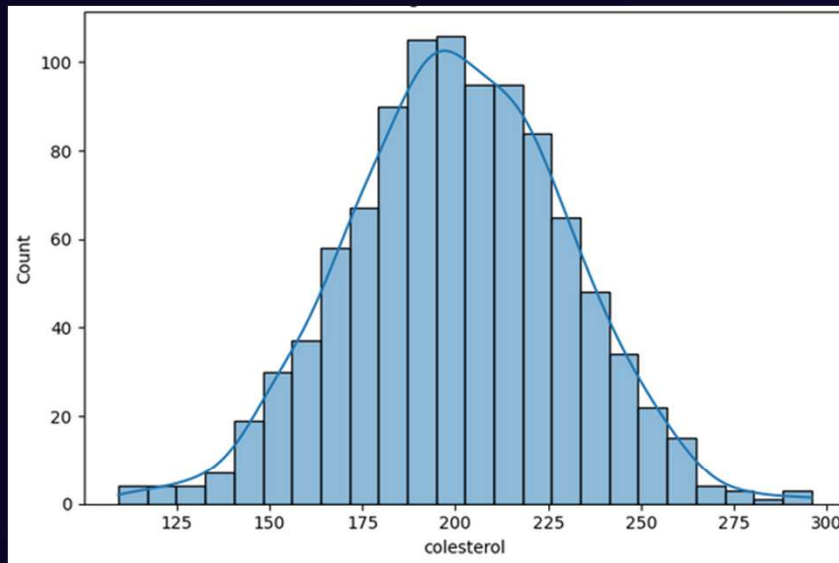
3. Seleccionar un método:

- Si los datos parecen normales → probar con la media.
- Si hay outliers o sesgo → probar con la **mediana**.
- Si los datos son categóricos → probar con la **moda** o imputación condicional (relacionada con otras variables)

4. Validar los resultados:


- Compara cómo cambian las distribuciones antes y después de la imputación para asegurarse que no estás distorsionando el análisis.

Pasos para imputar datos



Análisis de Datos

- Resumen de análisis

 por conectiva oficial



ANOVA

¿El método de estudio afecta la nota?

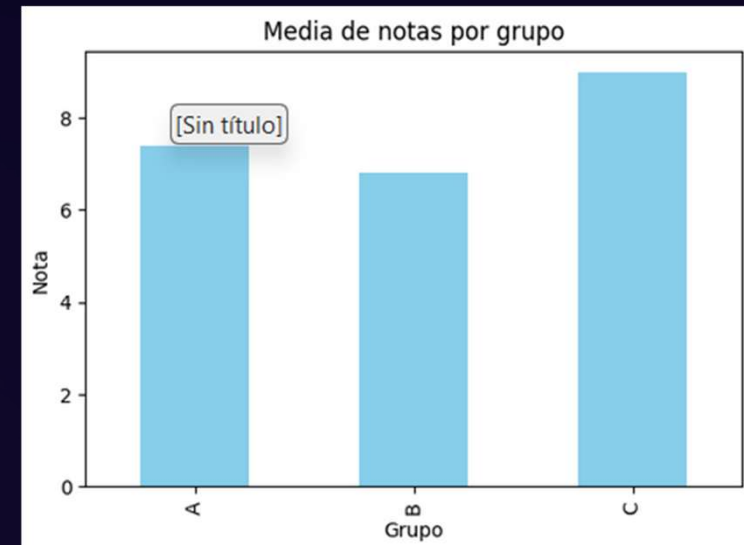
- Si las **medias** de los grupos son **similares**, entonces **NO** hay efecto → se acepta la hipótesis nula (H_0).
- Si las **medias** son **diferentes**, entonces **SÍ** hay efecto → se acepta la hipótesis alternativa (H_1).

¿Qué compara?

- Compara la **variabilidad entre grupos** (diferencias de medias) vs. la **variabilidad dentro de los grupos** (dispersión individual).
- Si las medias son **muy diferentes** entre grupos en comparación con su dispersión → H_1 .
- Si las medias son **similares** (o la variación dentro del grupo es muy grande) → H_0 .

Nota

- Una **diferencia entre medias grande** sugiere que los grupos **se comportan de forma distinta** (por ejemplo, un método de enseñanza funciona mejor).
- Una **diferencia pequeña** sugiere que los grupos son **similares**.



TUKEY

- Diferencia de medias = grupo 2 – grupo 1
- Si meandiff es positivo, entonces group2 tiene mayor media que group1.
- Si meandiff es negativo, entonces group2 tiene menor media que group1

group1	group2	meandiff	p-adj	lower	upper	reject
A	B	-0,6	0.5676	-2.1403	0.9403	False
A	C	1,6	0.0417	0.0597	3.1403	True
B	C	2,2	0.0065	0.6597	3.7403	True

- El grupo C tiene en promedio 1.6 puntos más que el grupo A.
- El grupo C tiene en promedio 2.2 puntos más que el grupo B.
- El grupo C tiene notas más altas que A y B

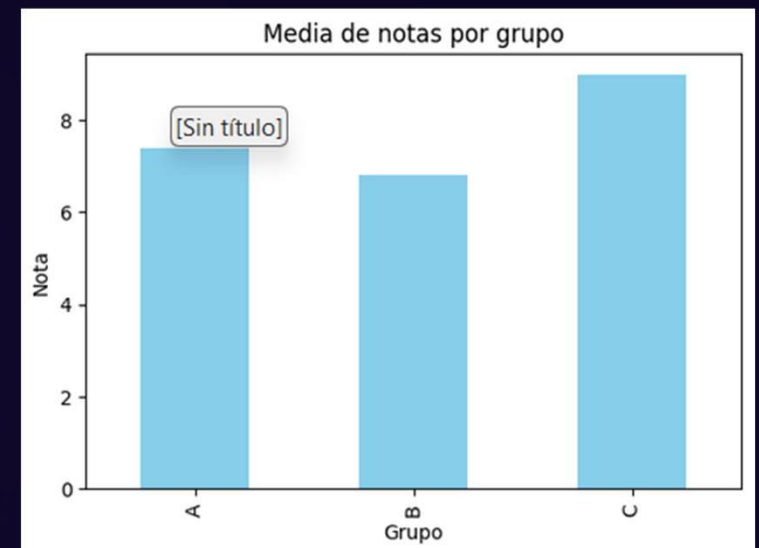


Tabla comparativa

Método	Tipo de variables	Qué mide	Cuándo usarlo
Cramér's V	Categóricas	Fuerza de la asociación entre 2 variables categóricas	Para medir la asociación entre 2 variables categóricas
Phik	Categóricas	Fuerza de la asociación corregida entre 2 variables categóricas	Para medir la asociación entre 2 variables categóricas con corrección ante distribuciones desbalanceadas
Chi ²	Categóricas	Independencia entre 2 variables categóricas	Para verificar si dos variables categóricas son independientes (no tienen asociación)
ANOVA	Categórica (independiente) y numérica (dependiente)	Diferencia entre medias de más de 2 grupos categóricos	Para comparar si las medias de más de 2 grupos (categóricos) son significativamente diferentes entre sí
Regresión	Numéricas (continuas)	Relación cuantitativa entre variables	Para evaluar si una o más variables numéricas afectan de manera significativa a una variable dependiente numérica