Manejo de Datos Faltantes

Fundamentos teóricos para el análisis económico

Prof. Nicolás Sidicaro
Clase 11 - Septiembre 2025

Objetivos de la clase

- Comprender la naturaleza y clasificación de datos faltantes
- Identificar los mecanismos que generan información ausente
- Evaluar el impacto de diferentes estrategias de imputación
- Reconocer casos específicos en análisis económico
- Desarrollar criterios para seleccionar métodos apropiados

La realidad de los datos faltantes en economía

Situaciones frecuentes

- **EPH**: Non-respuesta en preguntas sensibles (ingresos)
- **Censos**: Formularios incompletos, hogares ausentes
- **Datos administrativos**: Sistemas que no conversan entre sí
- **Encuestas empresariales**: Empresas que no responden
- **Series temporales**: Revisiones, cambios metodológicos

¿Por qué importa?

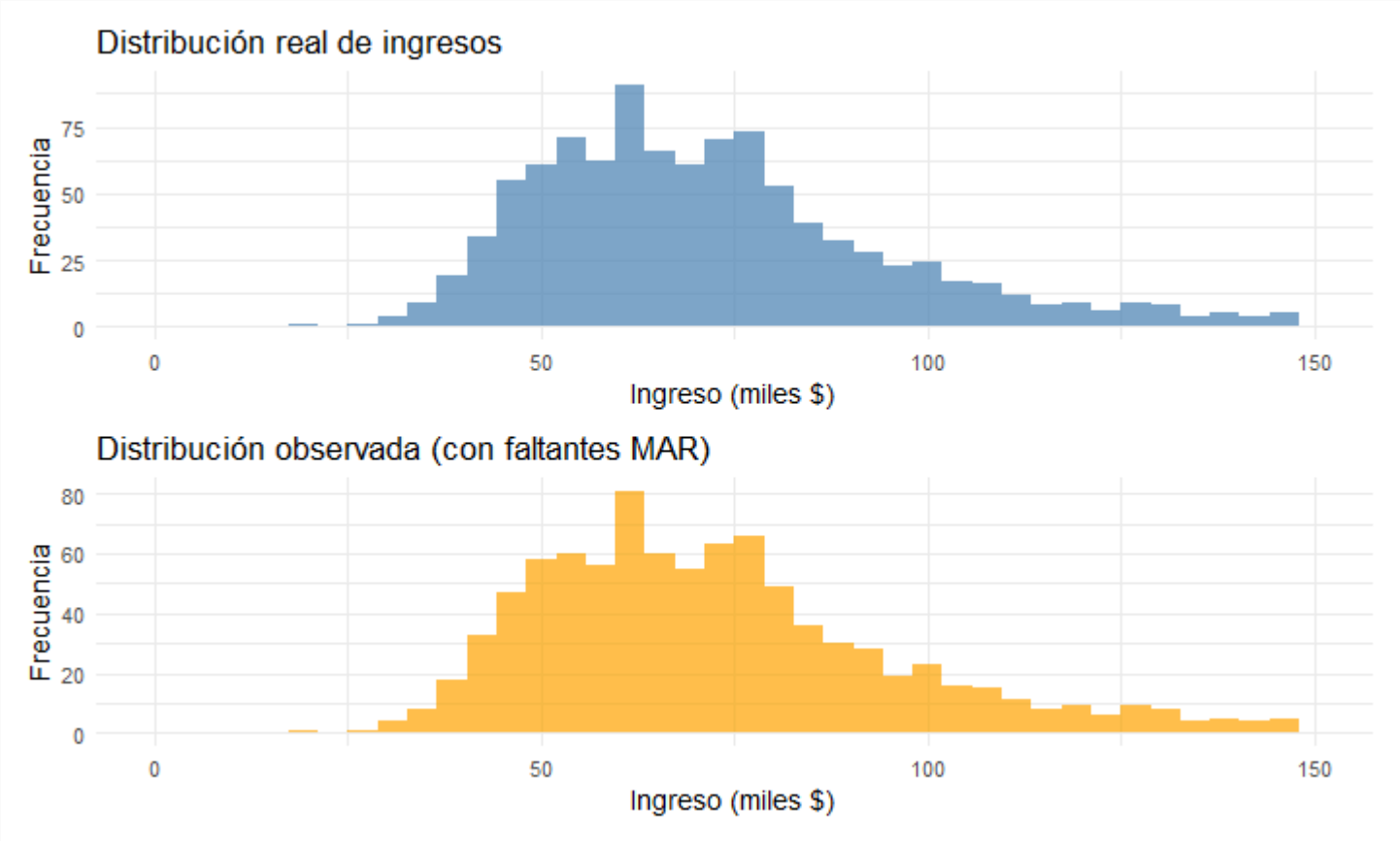
Problema estadístico

- Pérdida de información
- Sesgos en estimaciones
- Poder estadístico reducido

Problema económico

- Políticas basadas en información incompleta
- Decisiones de inversión erróneas
- Malentendidos sobre desigualdad

Ejemplo motivador: Encuesta de ingresos



Pregunta: ¿Qué conclusiones erróneas podríamos sacar si ignoramos los datos faltantes?

Clasificación de datos faltantes

Los tres mecanismos fundamentales

Clasificación de Rubin (1976)

Cada observación tiene una **probabilidad** de estar ausente. El proceso que gobierna estas probabilidades se llama **mecanismo de datos faltantes**.

MCAR: Missing Completely at Random

- Probabilidad igual para todos los casos
- Independiente de datos observados y no observados

MAR: Missing at Random

- Probabilidad depende solo de datos observados
- Independiente del valor faltante en sí mismo

MNAR: Missing Not at Random

- Probabilidad depende del valor faltante
- El más complejo de manejar

MCAR: Missing Completely at Random

Definición técnica

La probabilidad de que un dato esté ausente es **independiente** de todos los valores (observados y no observados).

$$P(\text{faltante}|X_{\text{obs}}, X_{\text{falta}}) = P(\text{faltante})$$

Ejemplos en economía

Casos típicos de MCAR:

- Fallo técnico del sistema de captura
- Encuestador enfermo en días aleatorios
- Error de digitalización aleatorio
- Pérdida de formularios por accidente
- Problemas de conexión en encuestas online

¿Por qué es poco común?

- Las personas **no** responden al azar
- Los sistemas **no** fallan al azar
- Las empresas **no** reportan al azar
- Casi siempre hay una razón sistemática

Implicancia estadística:

- Los casos completos son una muestra aleatoria válida
- Análisis sin sesgo (pero menos potencia)

MAR: Missing at Random

Definición técnica

La probabilidad de que falte un dato depende **solo** de la información observada, no del valor faltante en sí.

$$P(\text{faltante}|X_{\text{obs}}, X_{\text{falta}}) = P(\text{faltante}|X_{\text{obs}})$$

Ejemplos económicos realistas

Escenario	Variable faltante	Depende de (observado)	NO depende de	Mecanismo
Encuesta de ingresos	Ingresos	Nivel educativo	Monto del ingreso	Primaria → no responde ingresos
Censo empresarial	Ventas	Tamaño empresa	Monto de ventas	PyMES → no reportan ventas
ENGHO - Gastos	Gasto en salud	Edad del jefe	Monto del gasto	Jóvenes → no reportan gastos salud
Survey financiero	Inversiones	Sector económico	Monto inversión	Servicios → no reportan inversión

Clave conceptual: Podemos **predecir** la probabilidad de no respuesta usando datos observados, pero esta probabilidad no depende del valor que queremos medir.

MNAR: Missing Not at Random

Definición técnica

La probabilidad de que falte un dato depende del **propio valor faltante**, incluso después de controlar por todas las variables observadas.

$$P(\text{faltante}|X_{\text{obs}}, X_{\text{falta}}) \neq P(\text{faltante}|X_{\text{obs}})$$

Casos típicos:

- **Ingresos muy altos**: No quieren declarar por evasión
- **Empresas en crisis**: No responden encuestas
- **Desempleados de larga duración**: Evitan preguntas laborales
- **Actividades informales**: No declaran ventas
- **Depresión/salud mental**: Afecta participación en estudios

¿Por qué es problemático?

- No podemos predecir quién no responde
- Los valores faltantes están **sistemáticamente sesgados**
- Métodos estándar de imputación **no funcionan**
- Requiere información externa o modelos complejos

Detectar MNAR es difícil:

- Por definición, no observamos los valores faltantes
- Necesitamos conocimiento del dominio
- Análisis de sensibilidad es crucial

Comparando los tres mecanismos

Impacto de diferentes mecanismos en estadísticas observadas

Tipo	Media (SD)	% Faltante
Datos completos	73,212 (15,793)	0%
MCAR	73,197 (15,813)	13%
MAR	73,188 (15,660)	9%
MNAR	69,341 (15,287)	33%

Observación clave: Solo MCAR preserva las estadísticas originales. MAR y MNAR introducen sesgos sistemáticos.

Métodos para manejar datos faltantes

Estrategias principales

1. Eliminación (Deletion methods)

Complete Case Analysis (Listwise)

- Eliminar todas las observaciones con algún faltante
- Simple pero potencialmente sesgado
- Válido solo bajo MCAR

Pairwise deletion

- Usar todos los datos disponibles para cada análisis
- Inconsistente entre análisis
- Problemático para modelos complejos

¿Cuándo usar eliminación?

- Datos MCAR confirmado
- Porcentaje de faltantes muy bajo (< 5%)
- Muestra muy grande
- Análisis exploratorio inicial

Método	N observaciones	% Información perdida	Sesgo potencial	Complejidad
Datos originales	1,000	0%	Ninguno	-
Complete case	650	35%	Alto si no MCAR	Muy baja
Pairwise	Variable	10-35%	Moderado	Baja

Estrategias principales (cont.)

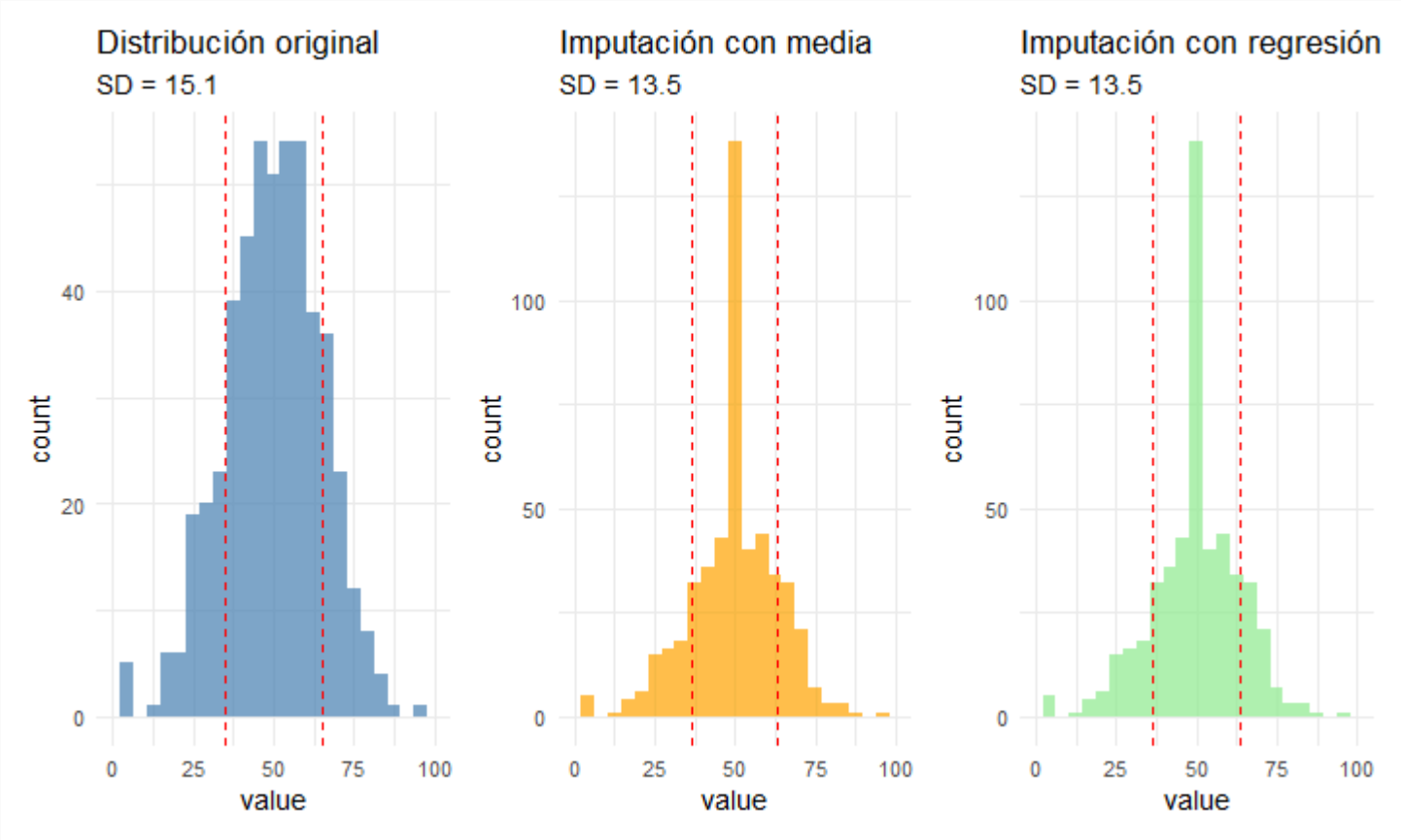
2. Imputación Simple (Single Imputation)

Reemplazar cada valor faltante con **un solo valor estimado**.

Método	Tipo de dato	Ventajas	Limitaciones	Cuándo usar
Media/Mediana/Moda	Numérico/Categórico	Muy simple, mantiene N	Subestima varianza	Exploración inicial
Last Observation Carried Forward	Series temporales	Simple para series	Asume estabilidad	Datos longitudinales estables
Regresión lineal	Numérico	Usa relaciones entre variables	Subestima varianza	Relaciones lineales claras
Hot-deck	Cualquiera	Mantiene distribución observada	Requiere variables similares	Datos categóricos complejos
K-nearest neighbors	Cualquiera	Flexible, no asume distribución	Computacionalmente intensivo	Relaciones no lineales

Problema fundamental: La imputación simple **no captura la incertidumbre** sobre los valores imputados.

El problema de la varianza subestimada



Problema: Los métodos de imputación simple **subestiman la variabilidad real** y pueden llevar a conclusiones erróneas sobre significancia estadística.

Estrategias principales (cont.)

3. Imputación Múltiple (Multiple Imputation)

Reconoce la **incertidumbre** inherente en los valores imputados.

Proceso de 3 pasos:

1. Imputación

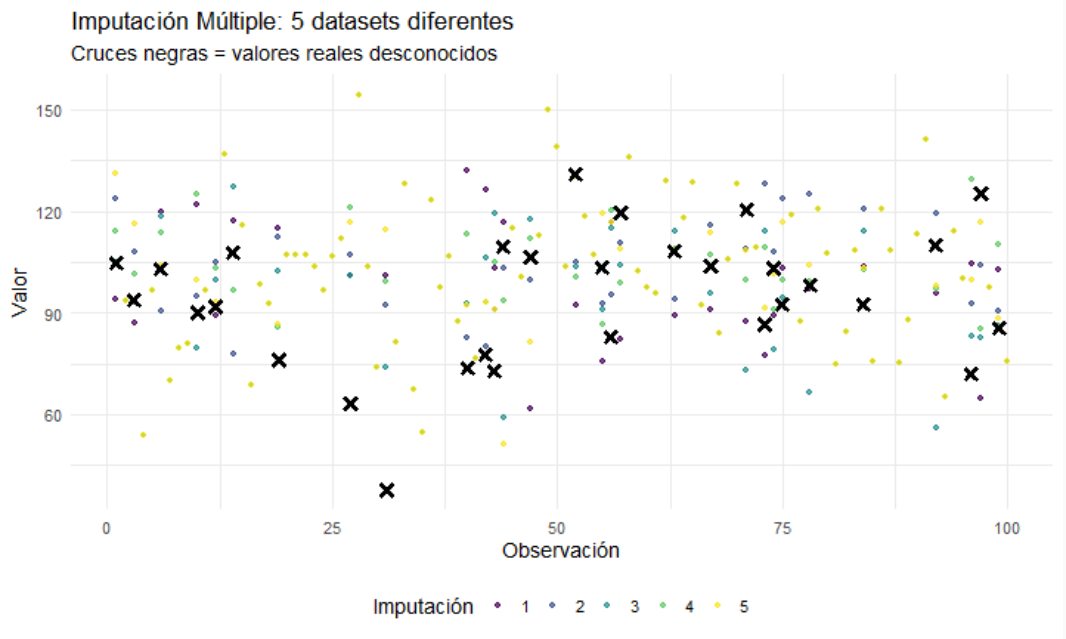
- Crear M datasets completos (típicamente M=5-10)
- Cada dataset tiene valores imputados diferentes
- Refleja incertidumbre sobre valores verdaderos

2. Análisis

- Correr el análisis deseado en cada dataset
- Obtener M conjuntos de resultados

3. Combinación (Pooling)

- Combinar resultados usando reglas de Rubin
- Incorpora variabilidad dentro y entre imputaciones



Ventajas de la Imputación Múltiple

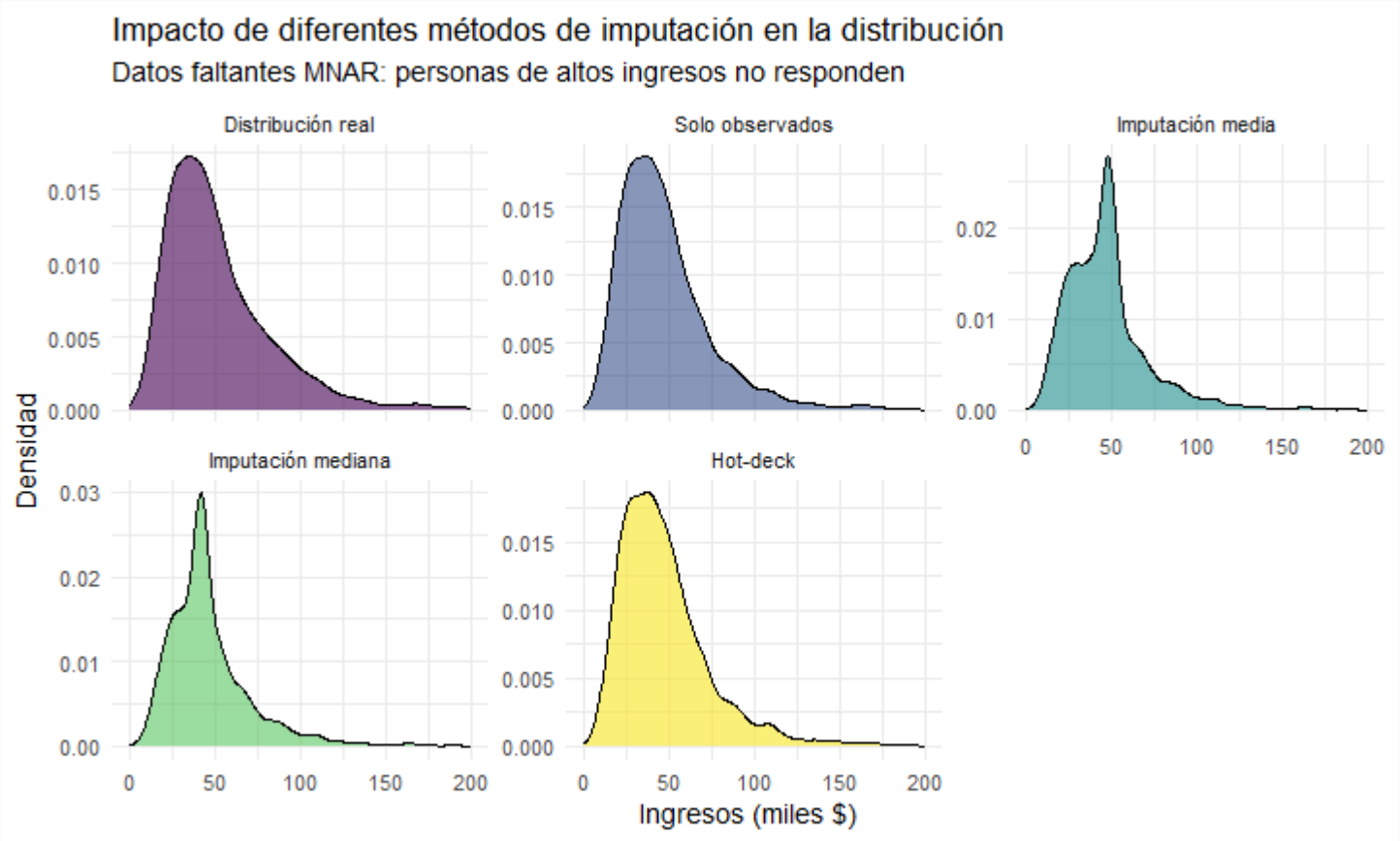
Comparación de métodos: Estimación del efecto de x1 sobre y

Método	Coeficiente x1	Error estándar	N efectivo	Sesgo
Valor real	0.500	NA	400	Ninguno
Complete case	0.619	0.057	271	Moderado
Mean imputation	0.617	0.061	400	Alto
Multiple imputation	**0.446**	**0.037**	**400**	**Mínimo**

Ventaja clave: MI proporciona estimaciones menos sesgadas y errores estándar apropiados que reflejan la incertidumbre de la imputación.

Impacto en distribuciones y análisis

Transformación de distribuciones según método



Observación crítica: Cada método produce una distribución diferente, afectando conclusiones sobre desigualdad, pobreza y políticas públicas.

Estadísticas descriptivas comparadas

Impacto en estadísticas descriptivas según método de imputación

metodo	N	Media	Mediana	Desv. Std	Coef. Variación	Q1	Q3	% > 80k	Diferencia Media	% Cambio Varianza
Distribución real	1000	52845	44009	34650	0.656	29794	66019	16.0	0	0.0
Solo observados	858	48493	41837	30524	0.629	28723	58872	11.2	-4352	-11.9
Imputación media	1000	48493	46849	28272	0.583	30862	54789	9.6	-4352	-18.4
Imputación mediana	1000	47548	41837	28367	0.597	30862	54789	9.6	-5297	-18.1
Hot-deck	1000	48310	41928	29730	0.615	28560	58446	10.8	-4535	-14.2

Implicaciones para política económica

- **Media subestimada**: Políticas de transferencias mal calibradas
- **Varianza reducida**: Subestimación de desigualdad
- **Percentiles sesgados**: Definición errónea de "clase media"
- **% > umbral**: Políticas focalizadas mal dirigidas

Casos específicos en análisis económico

1. Encuesta a hogares (Simil EPH, pero con gastos)

Problemas típicos de datos faltantes en EPH

Variable	% Faltante típico	Tipo probable	Razón principal	Sesgo esperado	Estrategia recomendada
Ingresos laborales	15-25%	MNAR	Evasión fiscal	Subestimación	MI con variables auxiliares
Ingresos no laborales	30-40%	MNAR	Actividades informales	Subestimación severa	Modelos complejos/externos
Horas trabajadas	5-10%	MAR	Dependiente de situación laboral	Leve	MI estándar
Gastos del hogar	20-35%	MNAR	Información sensible	Subestimación	MI + análisis sensibilidad
Nivel educativo	2-5%	MCAR	Error de captura	Mínimo	Eliminación o imputación simple

2. Series macroeconómicas

Desafíos específicos:

- **Revisiones estadísticas**: Datos preliminares vs definitivos
- **Cambios metodológicos**: Series no comparables
- **Frecuencias mixtas**: Datos anuales vs trimestrales vs mensuales
- **Estacionalidad**: Patrones que afectan imputación

Criterios para seleccionar estrategia

Factores técnicos:

- Tamaño de la muestra disponible
- Complejidad del análisis posterior
- Recursos computacionales
- Experiencia del equipo

Factores sustantivos:

- Conocimiento del dominio
- Consecuencias de decisiones erróneas
- Transparencia requerida
- Replicabilidad del estudio

Guía de selección de métodos según contexto

	Porcentaje y tipo de datos faltantes			
Escenario	**< 5% faltantes**	**5-20% MAR**	**5-20% MNAR**	**> 20% cualquier tipo**
Exploración inicial	Eliminación	Simple	Simple/Documentar	Recolectar más datos
Análisis descriptivo	Eliminación/Simple	MI	MI + sensibilidad	Externa
Modelos predictivos	MI	MI	MI + externa	Externa
Inferencia causal	MI	MI	Externa	Externa
Política pública	MI + sensibilidad	MI + sensibilidad	Externa + sensibilidad	Rediseño estudio
Investigación académica	MI	MI	Externa	Externa

Regla de oro: Siempre documentar y justificar la estrategia elegida. El método "perfecto" no existe, pero la transparencia sí.

Síntesis y recomendaciones

Puntos clave de la clase

Conceptos fundamentales

1. **Los datos faltantes no son aleatorios**: Casi siempre hay un patrón sistemático
2. **MCAR es muy raro**: La mayoría de casos reales son MAR o MNAR
3. **La imputación simple subestima varianza**: Puede llevar a conclusiones erróneas
4. **MNAR requiere información externa**: Los métodos estándar no funcionan

Recomendaciones prácticas

Antes de decidir método:

- Investigar el **proceso** que genera los faltantes
- Consultar literatura del dominio específico
- Analizar patrones de faltantes visualmente
- Considerar consecuencias de decisiones erróneas

Para análisis económico:

- **Exploración**: Simple imputation está bien
- **Descriptivos**: MI si >10% faltante
- **Inferencia**: MI siempre
- **MNAR**: Información externa + sensibilidad
- **Documentar**: Siempre justificar elección

Para la próxima clase

Implementación práctica

Veremos la **aplicación en R** de los conceptos discutidos hoy:

- Diagnóstico de patrones de faltantes con `VIM`
- Imputación múltiple con `mice`
- Análisis de sensibilidad
- Validación de resultados
- Casos específicos de EPH y datos económicos

Preparación recomendada

- Revisar material de R-bloggers y R4DS linkado
- Pensar en problemas de datos faltantes en sus áreas de interés
- Considerar estrategias para sus proyectos finales

La próxima clase veremos cómo implementar estas estrategias en la práctica

Referencias y lecturas adicionales

1. **Little, R.J.A. and Rubin, D.B.** (2019). *Statistical Analysis with Missing Data*, 3rd Edition. Wiley.

2. **Van Buuren, S.** (2018). *Flexible Imputation of Missing Data*, 2nd Edition. CRC Press.

3. **Schafer, J.L.** (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

4. **Rubin, D.B.** (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

5. **Carpenter, J. and Kenward, M.** (2013). *Multiple Imputation and its Application*. Wiley.

Recursos online:

- R-bloggers: Handling Missing Data in R
- R for Data Science: Missing Values
- mice package documentation