Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

¿Qué son los datos faltantes?

- En muchas bases de datos, no todas las variables están observadas para todas las unidades.
- A estas ausencias se les denomina **datos faltantes** (*missing data*).
- Su manejo es crucial porque:
  - Puede introducir sesgo en las estimaciones.
  - Reduce el tamaño efectivo de muestra.
  - Afecta inferencias y predicciones.

Posibles causas de datos faltantes

- **Errores de medición o captura:** fallas en sensores, registros incompletos en encuestas.
- **Pérdida de información:** páginas extraviadas, errores en la digitalización o transmisión de datos.
- **No respuesta en encuestas:** participantes omiten preguntas sensibles (ingresos, salud, religión).
- **Diseño del estudio:** variables no medidas en todas las unidades por restricciones de costo o logística.
- **Abandono en estudios longitudinales:** sujetos que dejan de participar en encuestas de seguimiento.

Efectos sobre los métodos estadísticos

- **Reducción del tamaño efectivo de muestra:** pérdida de poder estadístico.
- **Sesgo en las estimaciones:** si los datos no son MCAR (Missing Completly At Random), los parámetros pueden estar distorsionados.
- **Cambios en la varianza:** la imputación inadecuada puede subestimar o sobreestimar la incertidumbre.
- **Validez de la inferencia:** tests de hipótesis y construcciones de intervalos pueden ser inválidos.
- **Limitaciones en modelos predictivos:** la calidad de predicción se ve afectada si los patrones de faltantes no se consideran.

Ejemplo ilustrativo: Sesgo por datos faltantes

Encuestas de ingresos

- En una encuesta socioeconómica, las personas con ingresos más altos suelen omitir la pregunta de salario.
- Si eliminamos estos casos, la muestra queda sesgada hacia ingresos más bajos.
- Resultado: la media estimada de ingresos es significativamente menor a la real.

Ensayos clínicos

- Pacientes con efectos secundarios severos tienden a abandonar el estudio.
- El análisis de solo los que completan el tratamiento subestima la tasa de efectos adversos.

Ejemplo ilustrativo: Pérdida de eficiencia

Estudio longitudinal

- En un estudio de salud a 10 años, algunos participantes faltan en ciertas visitas.
- Aunque los datos sean MCAR, la eliminación de casos reduce el tamaño muestral efectivo.
- Consecuencia: intervalos de confianza más amplios y menor poder para detectar efectos.

Análisis multivariado

- En un análisis de componentes principales, si algunas variables tienen valores faltantes, se pierden observaciones completas.
- Esto reduce la precisión en la estimación de las correlaciones y de las componentes.

Mecanismos de datos faltantes

- **MCAR: Missing Completely At Random**
  - El patrón de faltantes es completamente independiente de los datos.
  - Ejemplo: en una encuesta, algunas páginas se extravían por error de impresión, sin relación con las respuestas.
- **MAR: Missing At Random**
  - La probabilidad de ausencia depende de los valores observados, pero no de los valores faltantes.
  - Ejemplo: en una encuesta de ingresos, los datos de salario faltan más en personas jóvenes, pero dentro de cada grupo de edad, la ausencia es aleatoria.
- **MNAR: Missing Not At Random**
  - La probabilidad de ausencia depende directamente de los valores faltantes.
  - Ejemplo: en la misma encuesta de ingresos, las personas con salarios muy altos tienden a no reportarlos; aquí el dato faltante está relacionado con el valor real de la variable omitida.

Estrategias de manejo

- **Eliminación**: quitar casos con valores faltantes.
- **Imputación simple**: media, mediana, moda, *hot-deck*.
- **Imputación múltiple**: varios conjuntos imputados bajo un modelo probabilístico.
- **Máxima verosimilitud**: ajuste directo considerando los faltantes.
- ***Métodos de ML**: *kNN*, árboles, redes neuronales.

Mecanismo MCAR

Sea $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ y $\mathbf{R}$ la matriz indicadora de faltantes.

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \theta) = P(\mathbf{R})$$

- La probabilidad de que falte un dato no depende de los valores de $\mathbf{Y}$.
- Bajo MCAR, la muestra observada es representativa de la población.
- La eliminación de casos completos produce estimadores no sesgados (aunque menos eficientes).

MCAR $\Rightarrow$ factoriza e ignorable para θ

Sea $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, $\mathbf{R}$ el patrón de faltantes y θ parámetros del modelo de datos, ψ del mecanismo de faltantes.

Definición (MCAR).

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \theta, \psi) = P(\mathbf{R} \mid \psi).$$

Proposición (factorización). Bajo MCAR,

$$p(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = p(\mathbf{Y} \mid \theta)\, p(\mathbf{R} \mid \psi).$$

Corolario. La *verosimilitud de datos observados* para θ es

$$L_{\text{obs}}(\theta) \propto \int p(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}} \mid \theta)\, d\mathbf{Y}_{\text{mis}} = p(\mathbf{Y}_{\text{obs}} \mid \theta).$$

Así, la inferencia sobre θ puede basarse en $p(\mathbf{Y}_{\text{obs}} \mid \theta)$, ignorando $p(\mathbf{R} \mid \psi)$.

MCAR $\Rightarrow$ la submuestra observada es representativa

Supongamos $Y_1, \ldots, Y_n \overset{iid}{\sim} F_\theta$, e indicadores $R_i \in \{0, 1\}$ con MCAR: $R_i \perp Y_i$.

Proposición. La distribución condicional de los observados coincide con la poblacional:

$$\mathcal{L}(Y_i \mid R_i = 1) = \mathcal{L}(Y_i) = F_\theta.$$

Implicación. Cualquier estadístico que dependa sólo de la ley marginal de Y mantiene su interpretación si se calcula con la submuestra observada (p. ej., media poblacional).

Listwise deletion bajo MCAR: insesgado pero menos eficiente

Sea $\mu = \mathbb{E}[Y]$, $\sigma^2 = \mathrm{Var}(Y)$ y $n_{\mathrm{obs}} = \sum_{i=1}^{n} R_i$. El estimador de casos completos es $\bar{Y}_{\mathrm{obs}} = \frac{1}{n_{\mathrm{obs}}} \sum_{i=1}^{n} R_i Y_i$.

Insesgo (condicional). Dado el patrón R con $n_{\mathrm{obs}} \geq 1$,

$$\mathbb{E}(\bar{Y}_{\mathrm{obs}} \mid R) = \frac{1}{n_{\mathrm{obs}}} \sum_{i:R_i=1} \mathbb{E}(Y_i \mid R_i = 1) = \frac{1}{n_{\mathrm{obs}}} \sum_{i:R_i=1} \mu = \mu,$$

pues MCAR $\Rightarrow \mathbb{E}(Y_i \mid R_i = 1) = \mathbb{E}(Y_i) = \mu$. Por lo tanto, $\mathbb{E}(\bar{Y}_{\mathrm{obs}}) = \mu$.
Varianza (condicional). Dado R,

$$\mathrm{Var}(\bar{Y}_{\mathrm{obs}} \mid R) = \frac{\sigma^2}{n_{\mathrm{obs}}}.$$

Como $n_{\mathrm{obs}} \leq n$ típicamente, la varianza aumenta: pérdida de eficiencia por menor tamaño muestral efectivo.

Resumen formal de los tres puntos

1. **Independencia de R y factorizar:** bajo MCAR, $p(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = p(\mathbf{Y} \mid \theta)p(\mathbf{R} \mid \psi)$.
2. **Representatividad:** $Y \mid (R = 1) \overset{d}{=} Y \Rightarrow$ la submuestra observada preserva la ley poblacional.
3. **Listwise deletion:** $\mathbb{E}(\bar{Y}_{\text{obs}}) = \mu$ pero $\mathrm{Var}(\bar{Y}_{\text{obs}}) = \sigma^2/n_{\text{obs}}$, lo que implica *insesgo con pérdida de eficiencia*.

Mecanismo MAR: Missing At Random

Sea $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ y $\mathbf{R}$ el patrón de faltantes.

Definición (MAR).

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \theta, \psi) = P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \psi),$$

es decir, la probabilidad de ausencia depende de los *valores observados* pero no de los valores faltantes.

Intuición. Condicionando en lo que sí vemos ($\mathbf{Y}_{\text{obs}}$), el mecanismo de faltantes deja de depender de los datos que no vemos ($\mathbf{Y}_{\text{mis}}$).

Factorización e ignorabilidad de MAR para inferir θ

Modelo de selección:

$$p(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = p(\mathbf{Y} \mid \theta)\, p(\mathbf{R} \mid \mathbf{Y}, \psi).$$

Bajo **MAR**:

$$p(\mathbf{R} \mid \mathbf{Y}, \psi) = p(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \psi).$$

Verosimilitud de datos observados para θ:

$$L(\theta; \mathbf{Y}_{\text{obs}}, \mathbf{R}) = \int p(\mathbf{Y} \mid \theta)\, p(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \psi)\, d\mathbf{Y}_{\text{mis}} = p(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \psi)\; p(\mathbf{Y}_{\text{obs}} \mid \theta).$$

Conclusión. Para *frecuentista*, la parte que depende de ψ es un factor multiplicativo que no afecta la optimización en θ:

$$L(\theta; \mathbf{Y}_{\text{obs}}, \mathbf{R}) \propto p(\mathbf{Y}_{\text{obs}} \mid \theta).$$

Para *Bayesiano*, si además el prior factoriza $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$ (*distinctness of parameters*), el mecanismo es *ignorable* para inferir θ.

Representatividad: condicional vs. marginal

Bajo **MAR**, la submuestra observada es representativa *condicionando en* $\mathbf{Y}_{obs}$:

$$\mathcal{L}(\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}, \mathbf{R}) = \mathcal{L}(\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}).$$

Pero en general **no** es representativa *marginalmente*. Implicaciones:

- Estimar una **media marginal** con casos completos puede ser sesgado.
- Estimar un **modelo condicional** (p.ej., $\mathbb{E}[Y \mid X]$) puede ser consistente si la ausencia depende sólo de X (observado) y el modelo está bien especificado.

Estrategias recomendadas bajo MAR

- **Máxima verosimilitud** (EM): integra $\mathbf{Y}_{\text{mis}}$ y usa toda la información de $\mathbf{Y}_{\text{obs}}$.
- **Imputación múltiple (MI)**: genera m datasets completos bajo un modelo compatible con MAR.
- **Modelos completos condicionados en observados**: p.ej. regresiones, GLMs y modelos multivariados donde la ausencia depende de covariables observadas.

Notas prácticas: (i) MAR es una suposición no testable directamente; analizar patrones ayuda a justificarla. (ii) Incluir en el modelo de imputación todas las variables relacionadas con la ausencia y con el resultado reduce sesgo.

Ejemplos ilustrativos (MAR)

Encuesta de ingresos

Faltan salarios con mayor probabilidad en personas jóvenes y con menor escolaridad (ambas observadas). *MAR:* R depende de X (edad, escolaridad), no del salario faltante condicionalmente a X.

Estudio clínico

Pacientes con IMC alto (observado) omiten medidas de glucosa. Condicionando en IMC, la falta de glucosa no depende del valor real de glucosa.

Series temporales

En sensores ambientales, se pierde lectura cuando la batería baja (voltaje observable). Faltantes dependen del voltaje observado, no del valor perdido condicionalmente al voltaje.

Mecanismo MNAR: Missing Not At Random

Sea $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ y $\mathbf{R}$ el patrón de faltantes.

Definición (MNAR).

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \theta, \psi) \text{ depende explícitamente de } \mathbf{Y}_{\text{mis}}.$$

- La probabilidad de ausencia depende de los *valores no observados*.
- El mecanismo de faltantes **no es ignorables** para inferir θ.
- No existe factor multiplicativo que elimine $p(\mathbf{R} \mid \cdot)$ de la verosimilitud.

Consecuencias de MNAR

- **No representatividad:** la submuestra observada distorsiona sistemáticamente la distribución poblacional.
- **Sesgo inevitable:** eliminar casos o asumir MAR introduce sesgo en estimadores.
- **Necesidad de modelado explícito:** se debe construir un modelo conjunto para los datos y el proceso de ausencia.

Verosimilitud completa:

$$L(\theta, \psi; \mathbf{Y}_{\text{obs}}, \mathbf{R}) = \int p(\mathbf{Y} \mid \theta)\, p(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi)\, d\mathbf{Y}_{\text{mis}}.$$

La parte de $p(\mathbf{R} \mid \cdot)$ **no puede ignorarse**.

Modelos bajo MNAR

Dos enfoques principales:

- **Modelos de selección (Heckman, Diggle-Kenward):**

$$p(\mathbf{Y}, \mathbf{R}) = p(\mathbf{Y} \mid \theta)\, p(\mathbf{R} \mid \mathbf{Y}, \psi).$$

  Modelan conjuntamente los resultados y el mecanismo de no-respuesta.

- **Modelos de patrones de mezcla (pattern-mixture):**

$$p(\mathbf{Y}, \mathbf{R}) = p(\mathbf{R})\, p(\mathbf{Y} \mid \mathbf{R}),$$

  separando la distribución de los datos por patrones de faltantes.

Ambos requieren suposiciones fuertes, difíciles de validar empíricamente.

Ejemplos ilustrativos (MNAR)

Encuesta de ingresos

Los individuos con ingresos muy altos tienden a no responder la pregunta de salario. El hecho de faltar depende del propio valor de ingreso no observado.

Estudio clínico

Pacientes con glucosa extremadamente elevada son más propensos a abandonar el estudio. La ausencia depende directamente del nivel real de glucosa (faltante).

Límites de detección

En estudios ambientales, concentraciones por debajo de cierto umbral no se registran. La probabilidad de ausencia depende del valor real de la variable (bajo el límite).

¿Qué hacer con datos MNAR?

- **Reconocer la dificultad:** distinguir entre MAR y MNAR es prácticamente imposible con los datos observados.
- **Modelar explícitamente el mecanismo:**
  - *Modelos de selección* (p.ej., Heckman, Diggle–Kenward).
  - *Modelos de patrones de mezcla* (pattern-mixture models).
  - *Modelos de sensibilidad* que exploran escenarios plausibles.
- **Usar información externa:** estudios auxiliares, expertos o datos adicionales pueden ayudar a especificar el mecanismo de no respuesta.
- **Análisis de sensibilidad:** evaluar cómo cambian las conclusiones bajo distintas suposiciones sobre el mecanismo MNAR.
- **En práctica aplicada:** si no es posible justificar un modelo MNAR, se suele asumir MAR pero **declarando la limitación** y explorando escenarios alternativos.

Conclusiones sobre mecanismos de datos faltantes

- **MCAR**: faltantes completamente al azar.
  - Submuestra representativa.
  - Métodos simples (casos completos) son insesgados, pero menos eficientes.
- **MAR**: faltantes al azar condicionalmente a los observados.
  - El mecanismo puede ignorarse en la inferencia si se modela correctamente.
  - Métodos recomendados: máxima verosimilitud, imputación múltiple.
- **MNAR**: faltantes no al azar.
  - No ignorables; se requiere modelado explícito o análisis de sensibilidad.
  - Conclusiones dependen fuertemente de los supuestos.

Mensaje clave: comprender el mecanismo es tan importante como elegir el método estadístico.

Bibliografía recomendada

Rubin, D. B. (1976).

Inference and missing data.

Biometrika, 63(3), 581–592.

Little, R. J. A., & Rubin, D. B. (2002).

Statistical Analysis with Missing Data (2nd ed.). Wiley.

Schafer, J. L. (1997).

Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.

van Buuren, S. (2018).

Flexible Imputation of Missing Data (2nd ed.). Chapman & Hall/CRC.