Universitat Oberta de Catalunya

25 años

Missing data: Imputación múltiple en bases de datos pequeñas

Juan Hernández-Villena, Lic.

Trabajo final de Máster de Bioinformática y Bioestadística

Marco teórico

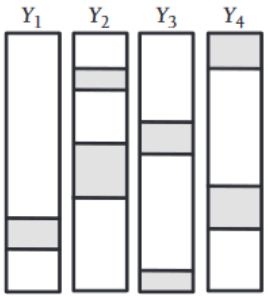
- ¿Qué son?
- Tipos de patrones
- Tipos de mecanismos
- Tipos de tratamientos

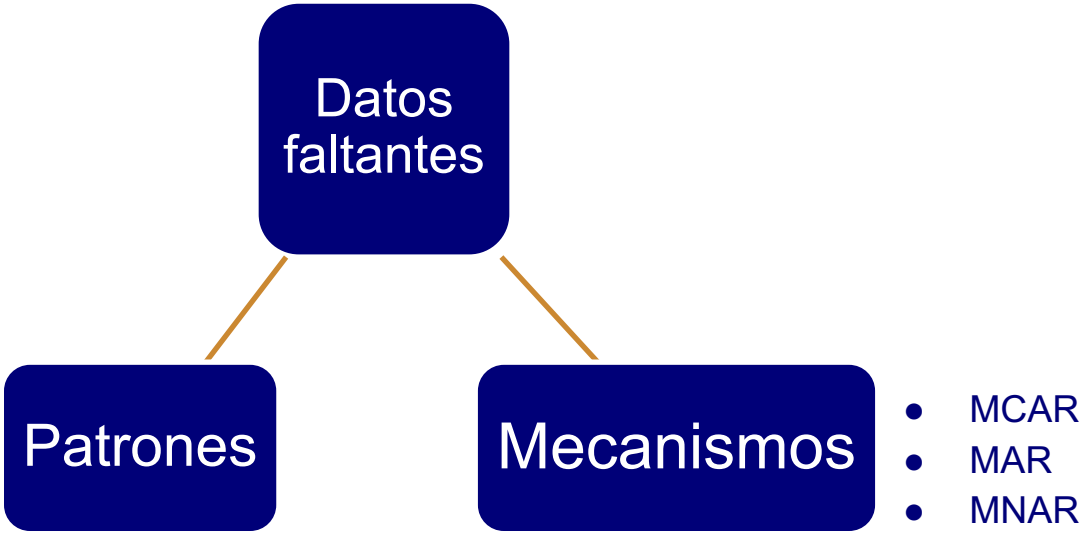
Datos faltantes

Un dato faltante o *missing*, es un valor no observado que, en caso de que hubiese sido observado, presentaría un valor significativo para el análisis (Little y Rubin, 2020)

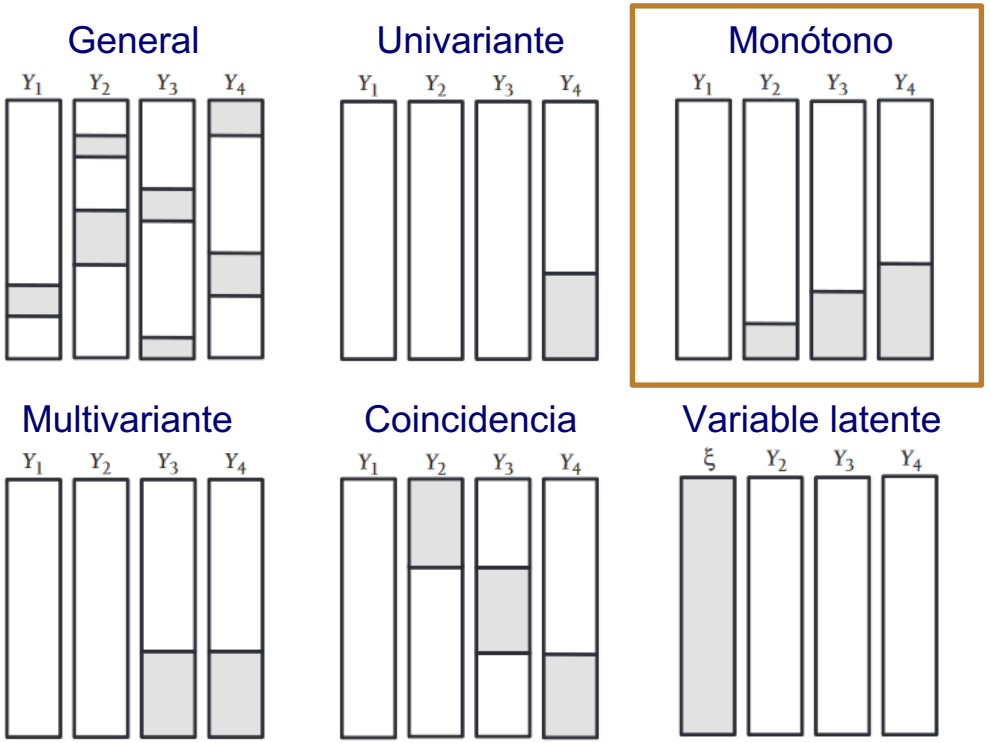
ID	Sexo	Salario	Peso_t1	Peso_t2	…	Peso_tn
1	M	100	90	NA	…	NA
2	NA	1000	90	NA	…	NA
3	F	NA	76	71	…	62
4	M	2000	100	NA	…	NA
…	…	…	…	…	…	…
n	M	50	NA	NA	…	NA

Teoría de datos faltantes





Patrones de datos faltantes



*Enders CK. Applied missing data analysis. Guilford press; 2010

Mecanismos de *missings:* Independiente

MCAR: *missing completely at random*

ID	Sexo	Salario	Peso_t1	Peso_t2
1	M	100	90	73
2	NA	1000	90	90
3	F	NA	76	71
4	M	2000	100	100
…	…	…	…	…
n	M	50	85	60

Mecanismos de *missings:* Dependiente

MAR: *Missing at random*

ID	Sexo	Salario	Peso_t1	Peso_t2
1	M	100	90	NA
2	F	1000	90	90
3	F	3000	76	71
4	M	2000	100	100
…	…	…	…	…
n	M	50	NA	NA

Mecanismos de *missings:* Dependiente

MNAR: *Missing not at random*

ID	Sexo	Salario	Peso_t1	Peso_t2
1	M	100	90	73
2	F	1000	90	NA
3	F	3000	76	71
4	M	2000	100	NA
…	…	…	…	…
n	M	50	85	60

Tratamientos de *missings*

Métodos de eliminación

- Método de análisis de datos completos (*Listwise deletion*)
- Método de análisis de casos disponibles (*Pairwise deletion*)

Métodos de imputación no estocásticos:

- Observación previa (LVCF)
- Imputación por media
- Imputación por regresión

Métodos de imputación estocásticos:

- Estimación por máxima verosimilitud
- Imputación múltiple

Tratamientos de *missings*

Métodos de eliminación

- Método de análisis de datos completos (*Listwise deletion*)
- Método de análisis de casos disponibles (*Pairwise deletion*)

Métodos de imputación no estocásticos:

- Observación previa (LVCF)
- Imputación por media
- Imputación por regresión

Métodos de imputación estocásticos:

- Estimación por máxima verosimilitud
- Imputación múltiple

Tratamientos de *missings*

Métodos de eliminación

- Método de análisis de datos completos (*Listwise deletion*)
- Método de análisis de casos disponibles (*Pairwise deletion*)

Métodos de imputación no estocásticos:

- Observación previa (LVCF)
- Imputación por media
- Imputación por regresión

Métodos de imputación estocásticos:

- Estimación por máxima verosimilitud
- Imputación múltiple

Imputación múltiple : *Predictive mean matching* (*PMM*)

Predictive mean matching (*PMM*)

- Emplea un procedimiento *hot deck* (uso de donantes)

Sexo	Salario	Peso_t1
M	100	90
M	2000	100
F	3000	76
M	50	NA

- Alternativa *Midastouch*

Flujo de trabajo

- Generación los escenarios
- Evaluación del patrón
- Evaluación del mecanismo
- Tratamiento (IM : PMM)

Base de datos* : Efavirenz vs Lopinavir/r (n = 116)

Cuadro 3.1: Variables seleccionadas para las bases de datos

Variable	Tipo	Descripción
Grupo	Categórica	Binaria (EFV / LVP/r)
Sexo	Categórica	Binaria (Masculino / Femenino)
Edad	Numérica	Por año
tpo_vih_meses	Numérica	Tiempo desde el contagio (mes)
factor_riesgo_total	Categórica	Categorías de factor de riesgo (4)
DC4A_	Numérica	Temporal (0,12,24,36,48 semanas)
CargaViral_	Numérica	Temporal (0,12,24,36,48 semanas)

*Echeverria y col., 2010. DOI: https://doi.org/10.1016/j.antiviral.2009.11.008

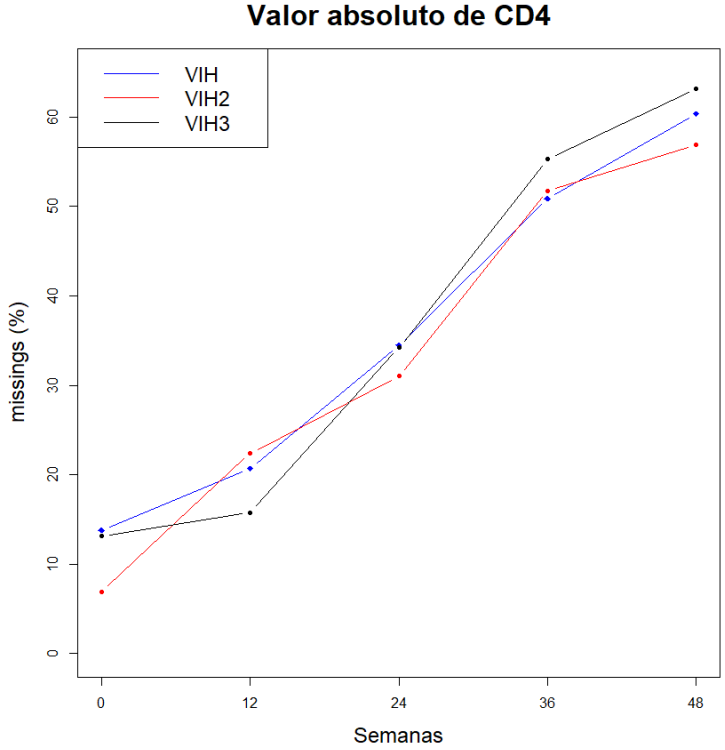
Generación de escenarios: Tamaño muestral

Cuadro 4.1: Características de las bases de datos (BD)

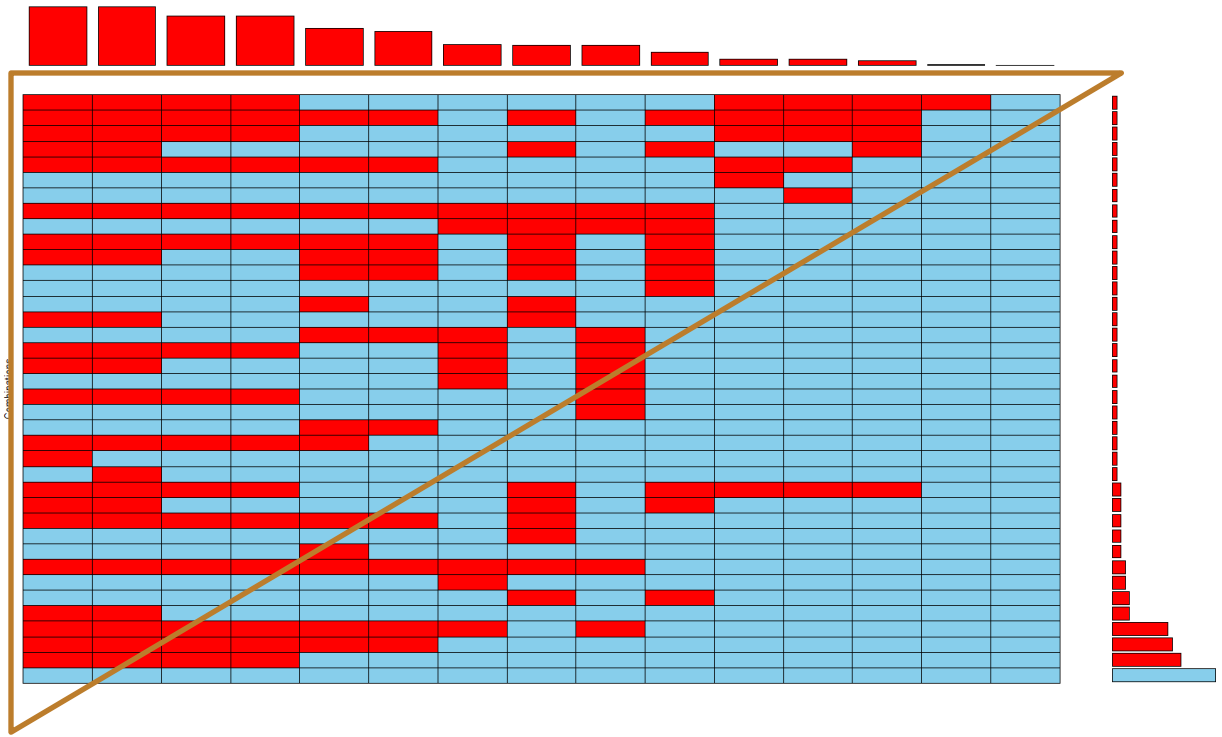
BD	N° Observaciones	N° missings	Missings(%)
VIH	116	452	26.0
VIH2	58	212	24.4
VIH3	38	147	25.8

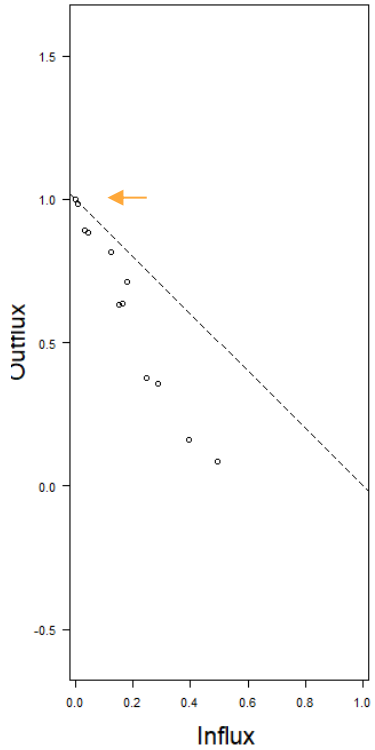
Cuadro 4.2: Porcentajes de *missings* de cada base de datos

Variable	VIH	VIH2	VIH3
CargaViral_48	60.34	58.62	63.16
CD4A_48	60.34	56.9	63.16
CargaViral_36	50.86	51.72	55.26
CD4A_36	50.86	51.72	55.26
CargaViral_24	37.93	36.21	36.84
CD4A_24	34.48	31.03	34.21



Evaluación del patrón





*Van Buuren S. Flexible imputation of missing data. CRC press; 2018.

Mecanismo de datos faltantes

- Datos sin distribución normal
- Prueba no paramétrica MCAR (Test de Jamshidian, Jalal y Jansen)

- VIH1 ($p > 0.05$)
- Conocimiento sobre el origen de los datos

- No es de tipo **MNAR**

*Jamshidian y col., 2014. DOI: 10.18637/jss.v056.i06

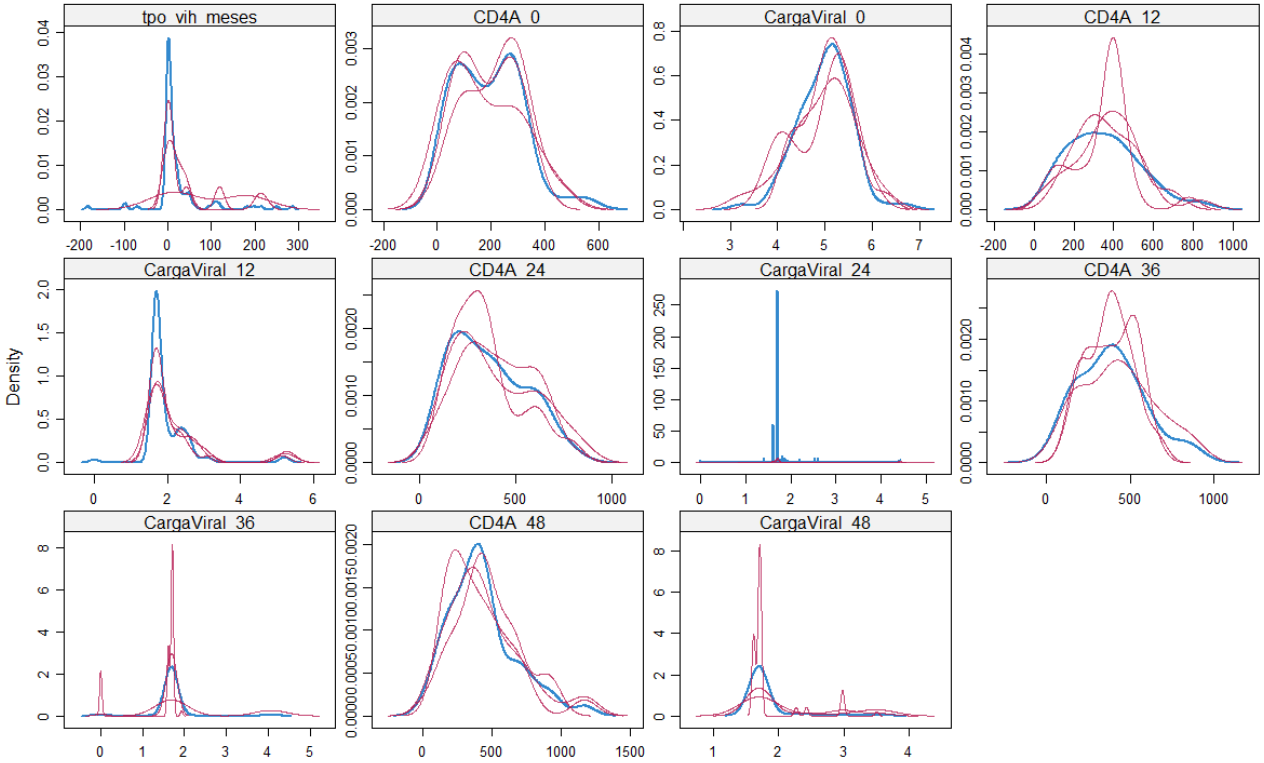
Imputación múltiple (m = 3 / iteraciones = 5 / método = *PMM*)

- Por defecto
  - Variable binaria -> método regresión logística (*logreg)*
  - Variable categórica -> método regresión logística multinomial *(polyreg)*
  - Variable numérica -> método *PMM*

- *PMM*
  - Todas las variables -> método *PMM*

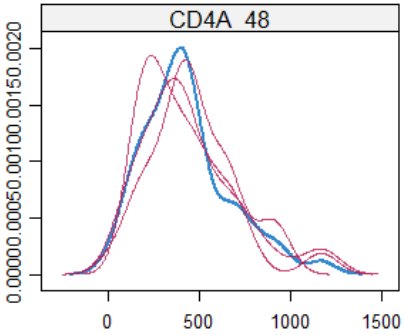
- *Midastouch*
  - Variable binaria -> método regresión logística (*logreg)*
  - Variable categórica -> método regresión logística multinomial *(polyreg)*
  - Variable numérica -> método *Midastouch*

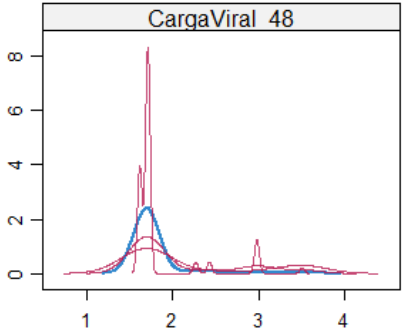
Imputación: *Densityplot*



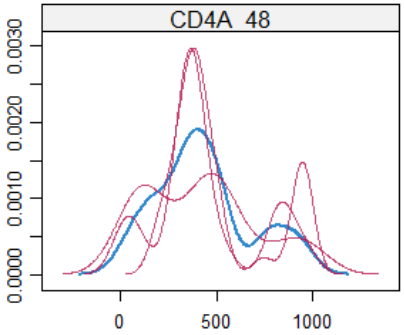
Imputación: *Densityplots*

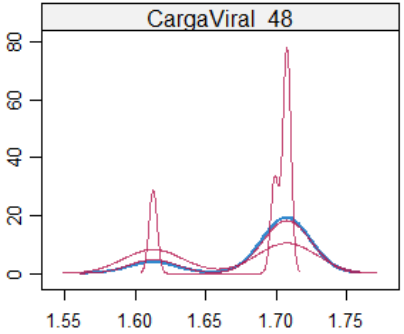
VIH1 (116)



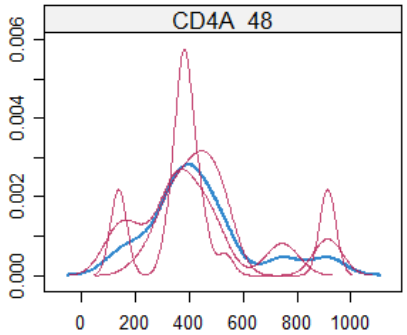


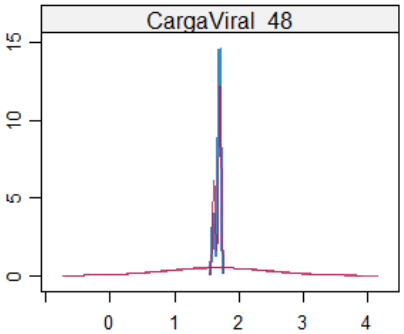
VIH2 (58)





VIH3 (38)





Comparación de resultados

		CD4A (semana 48)		Log_{10} Carga viral (semana 48)	
Base de datos	N	P valor	Z	P valor	Z
VIH_defecto	116	0.671	-0.428	0.476	0.701
VIH2_defecto	58	0.911	-0.117	0.895	-0.269
VIH3_defecto	38	**0**	3.461	**0.042**	-2.356
VIH_PMM	116	0.229	1.207	0.646	0.538
VIH2_PMM	58	0.923	0.101	0.079	-1.869
VIH3_PMM	38	**0.002**	3.078	1	0.413
VIH_Midas	116	0.879	0.155	0.262	-1.114
VIH2_Midas	58	0.354	0.936	1	0.066
VIH3_Midas	38	0.315	1.02	1	-0.594

Conclusiones

- Sobre la teoría de datos faltantes y su tratamiento…
- Sobre los resultados del trabajo…
- Sobre el aporte de este trabajo…
- Líneas a futuro

Agradecimientos

- Dra. Núria Pérez Álvarez
- Dr. Antoni Pérez Navarro
- Miembros del jurado

Referencias relevantes

- Little RJA, Rubin DB. Statistical analysis with missing data. Tercera edición ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2020.
- Van Buuren S. Flexible imputation of missing data. CRC press; 2018.
- Kleinke K. Multiple imputation by predictive mean matching when sample size is small. Methodology. 2018.