Missing data analysis in longitudinal data. How to analyze it?

Jorge J. Curto García

Universitat Oberta de Catalunya

M0.185 - TFM-Estadística y Bioinformática
Máster Universitario en Bioinformática y Bioestadística
UOC-UB

Nombre Profesor/a: **Núria Pérez**
PRA: Alexandre Sánchez Pla

Barcelona, 10 de Enero 2018

OBJETIVOS

Objetivo general

- Contextualizar el problema que generan los datos perdidos en el análisis de datos longitudinales y describir los métodos actuales disponibles para abordar dicho problema.

Objetivos específicos

- Caracterización de los estudios con datos longitudinales.
- Definición de datos faltantes
- Contextualización del problema de los datos perdidos en el análisis de los datos longitudinales.
- Identificación los métodos actuales de tratamiento de datos perdidos y determinar las bondades y limitaciones de estos métodos.
- Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos longitudinales en el ámbito de la biomedicina.

ENFOQUE

Sección teórica

- Recopilación bibliográfica acerca del análisis de datos longitudinales y la problemática de los datos perdidos
- Revisión exhaustiva de los distintos métodos actuales para el tratamiento de los datos faltantes
- Búsqueda de librerías disponibles en R

Sección practica

- Búsqueda online de una base de datos longitudinales
- Generación de un informe estadístico dinámico con el análisis descriptivo e inferencial de los datos originales, generación de datos perdidos y el análisis del tratamiento de los datos faltantes (R, Rstudio y Markdown)

DATOS LONGITUDINALES

- "En general" obtendremos datos longitudinales cuando la variable/variables de interés se recoge en mediciones sucesivas a lo largo del tiempo, porque el interés se halla en analizar el cambio en función del tiempo.
- Los estudios de datos longitudinales pueden ser observacionales o de intervención y tanto prospectivos como retrospectivos.
- 3 elementos clave [1]:
  - El seguimiento
  - Más de 2 medidas
  - Análisis que tenga en cuenta dichas medidas.

- **OBJETIVO:** *analizar los "patrones interindividuales de cambio intraindividual"*[2]

(1) Delgado M., Llorca J. Estudios longitudinales: concepto y particularidades. Rev. Esp. Salud Pública, mar.-abr. 2004, vol.78, no.2, p.141-148. ISSN 1135-5727. http://www.monografias.com/trabajos902/estudios-longitudinales/estudios-longitudinales.shtml#ixzz4bCQlxJyU

(2) Nesselroade, J. R. y Baltes, P. B. (Eds.) (1979). Longitudinal research in the study of behaviour and development. New York: Academic Press.

MÉTODOS DE ANÁLISIS DE DATOS LONGITUDINALES

- Englobados en el contexto de los modelos lineales generalizados
- Herramientas convencionales de regresión:
  - relacionar el **efecto** con las diferentes exposiciones
  - Correlación de las **medidas entre sujetos**.
  - Tratar covariables dependientes del tiempo
- Difiere en función de la naturaleza de la variable respuesta

DATOS PERDIDOS (datos faltantes/missing data)

- Foco de dificultad en el análisis de datos longitudinales
- En la investigación (clínica/social/biomédica, etc) es lo habitual (lo excepcional es su ausencia)
- Los motivos de aparición son variados (falta de consentimiento, problema técnico, no asistencia, etc)
- Tipos (Rubin 1976):
  - **MCAR** (missing completely at random): Datos perdidos completamente al azar
  - **MAR** (missing at random): Datos perdidos al azar
  - **MNAR** (missing not at random): Datos perdidos no ignorables o no debidos al azar

TRATAMIENTO DE DATOS PERDIDOS

- En el desarrollo de un estudio de investigación:
  - **DISEÑO:** Prever la posible presencia de datos faltantes en el cálculo del tamaño muestral
  - **RECOGIDA DE DATOS:** Adecuada monitorización para aumentar la calidad de los datos
  - **ANÁLISIS DE DATOS:** Identificar casos/variables afectadas por su presencia y los posibles patrones, para decidir el tipo de tratamiento que se lleva a cabo previo al análisis de los objetivos del estudio.

TRATAMIENTO DE DATOS PERDIDOS: MÉTODOS

- **ELIMINACIÓN:** eliminación de los casos/eliminación por pares
- **IMPUTACIÓN SIMPLE:** sustitución por la media/ regresión simple/ vecino más cercano/ regresión estocástica /Imputación por la observación previa (LOCF Last Observation Carried Forward)
- **MÉTODOS MODERNOS:** Algoritmo de expectación-maximización (EM)/ Algoritmo de imputación múltiple (MI) /Fully Conditional Specification (FCS)/ Full Information Maximum Likelihood (FIML)

TRATAMIENTO DE DATOS PERDIDOS: IMPUTACIÓN MÚLTIPLE

- 3 FASES
  - IMPUTACIÓN: Se generan *m* copias del conjunto de datos
    - Diversas estrategias, destacan Markov Chain Monte Carlo o el algoritmo de Equiparación de media predicitiva (PMM)

  - ANÁLISIS: En cada copia del conjunto de datos se lleva a cabo el análisis (con técnicas estadísticas)
  - COMBINACIÓN DE LOS RESULTASDOS: Se obtiene una estimación global como promedio de las *m* estimaciones.

TRATAMIENTO DE DATOS PERDIDOS: BONDADES Y LIMITACIONES

En general, los métodos de imputación no aleatoria (no estocásticos) reducen variabilidad y se ha de asumir el mecanismo MCAR o MAR.

- **Eliminación:** Perdida de eficacia por reducción del tamaño muestral, requiere asumir MCAR y la muestra sigue siendo insesgada.
- **Media:** reducción de varianza y produce sesgo.
- **Regresión:** reduce el sesgo pero también la variabilidad.
- **LOCF:** Se distorsiona el comportamiento a lo largo del tiempo afectando a diferencias entre grupos y las estimaciones de los parámetros.

los ***métodos estocásticos (imputación aleatoria)*** generan múltiples conjuntos de datos basados en los valores observados y reduciendo la posibilidad de que se produzca sesgo estadístico y a su vez maximizan la variabilidad

LIBRERIAS R

Existen multitud de librerías para el tratamiento de valores perdidos:

- **MissingDataGUI, VIM, VIMGUI,MICE, Amelia, HMISC, MI.**
- **VIM:** Herramientas muy útiles para la visualización gráfica de los datos faltantes y datos imputados
- **MICE:** Paquete muy completo con múltiple opciones desde no estocástica/no aleatoria (media) a modelos de imputación aleatoria para datos continuos (predictive mean matching PMM, normal), datos binarios (regresión logística), datos categóricos no ordenados (regresión logística politómica) y datos categóricos ordenados (odds proporcionales).

EJEMPLIFICACIÓN: PLAN DE ANÁLISIS

- Selección, presentación y análisis descriptivo de la base de datos seleccionada.
- Generación de nuevas bases de datos (dataframes en R) con datos faltantes generados de manera aleatoria, 3 escenarios: 10%, 20% y 30%.
- Tratamiento de los datos faltantes para cada una de las bases de datos generadas anteriormente.
- Reproducción de los análisis originales publicados por los autores (base de datos completa).
- Reproducción de los análisis originales publicados por los autores con las bases de datos generadas tras el tratamiento de los datos faltantes.
- Comparación de los resultados de los dos pasos anteriores.

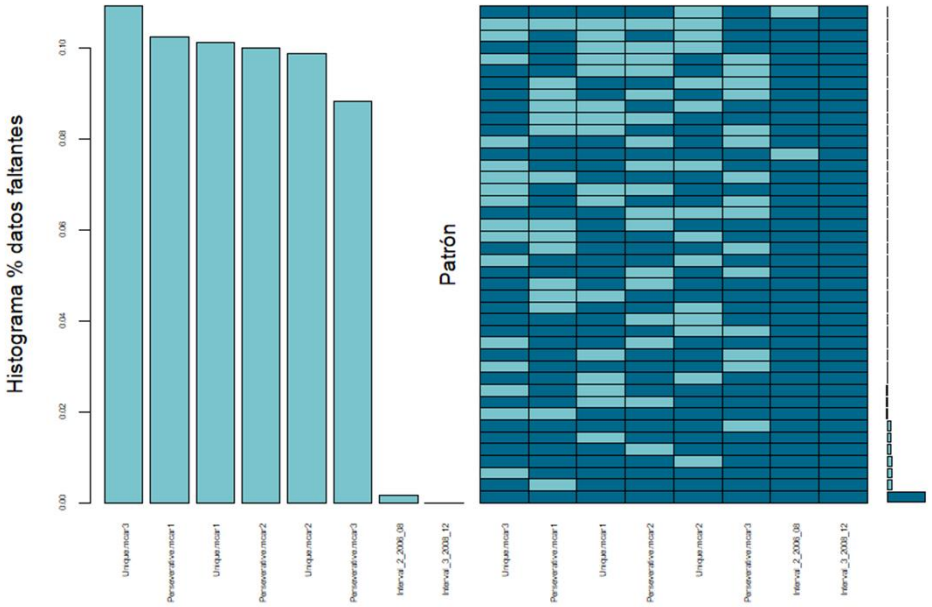
EJEMPLIFICACIÓN: ANÁLISIS ORIGINAL

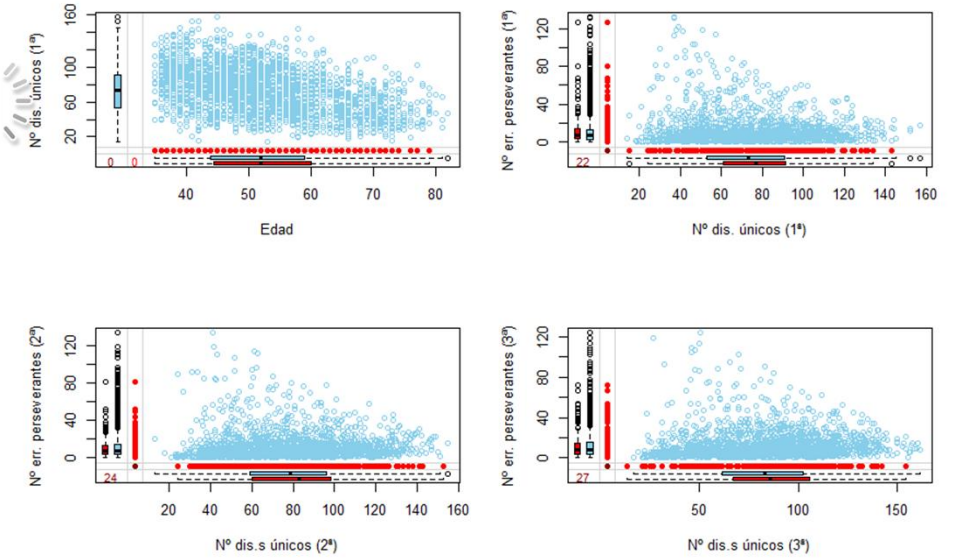
- **Base de datos: Estudio PREVEND** es de acceso público en "The Dryad Digital Repository", **Prueba RFFT** (prueba cognitiva que evalúa la función ejecutiva FE), 2 variables principales: Nº de diseños únicos y Nº de errores perseverantes (3 mediciones consecutivas).
- Los autores reportaron, en 2515 participantes, un aumento significativo ($p_{tendencia}<0.001$) en el número medio (DE) de diseños únicos en el RFFT incrementándose de 73 (26) en la primera medición, a 79 (27) en la segunda medición y a 83 (26) en la tercera. Además dicho aumento se asoció negativamente con la edad (disminuyó con 0,23 por incremento de un año) y no se halló relación con el nivel educativo.
- Resultados similares se obtuvieron para los errores perseverantes

Se reprodujeron los mismos resultados que los autores.

EJEMPLIFICACIÓN: GENERACIÓN Y ANÁLISIS DE DATOS FALTANTES (ESCENARIO 10%)

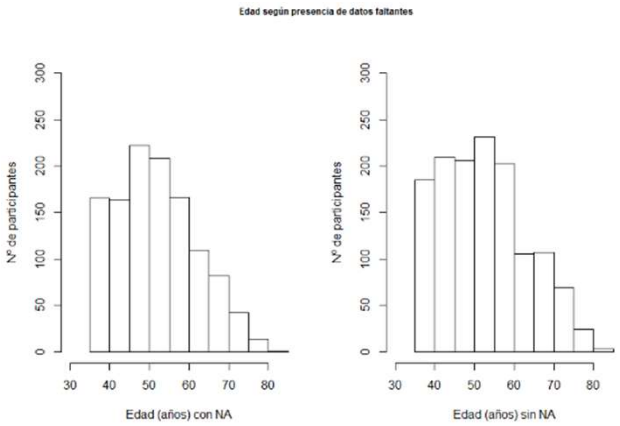
1) Se generaron aleatoriamente 3 escenarios: 10%, 20% y 30% de datos faltantes en las variables principales. No se hallan patrones (paquete VIM)

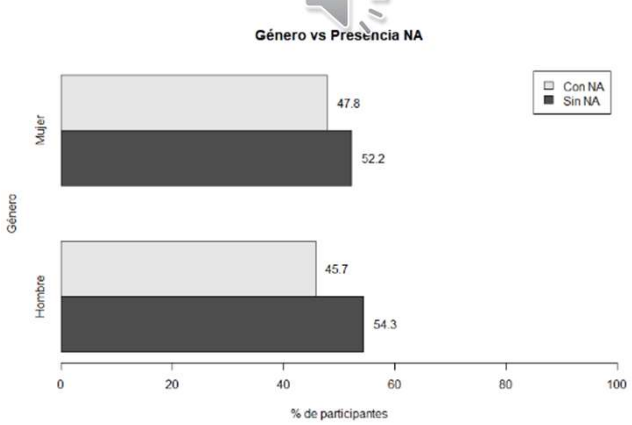


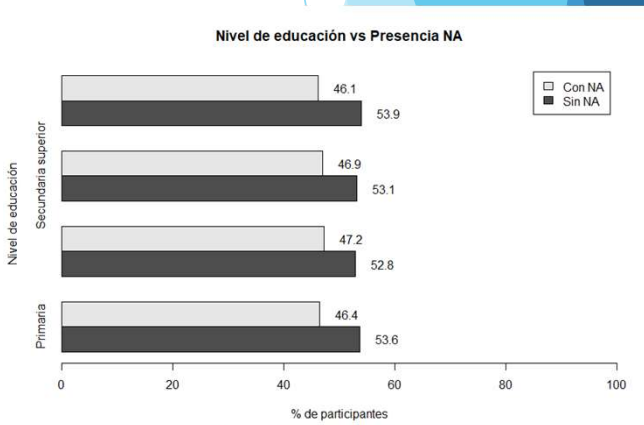


EJEMPLIFICACIÓN: GENERACIÓN Y ANÁLISIS DE DATOS FALTANTES (ESCENARIO 10%)

Se corroboró la ausencia de patrones de los datos faltantes y de relación entre la presencia y ausencia de datos faltantes con las características basales. Además se obtuvo un resultado no significativo al aplicar el test Little´s MCAR-test, que permitió asumir que los datos faltantes generados aleatoriamente en el escenario del 10% son MCAR (se obtuvieron resultados similares para el escenario del 20%, y 30% y también para la variable *nº de errores perseverantes*).



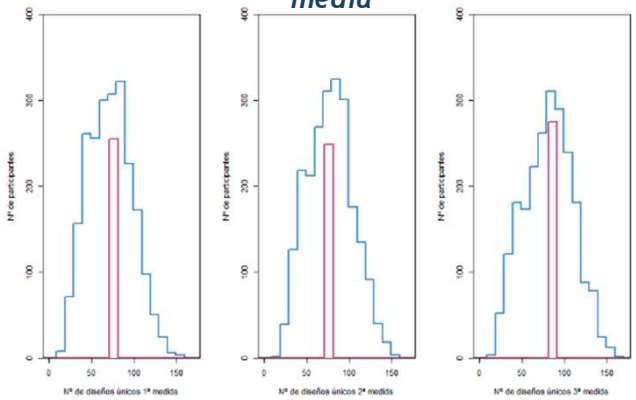




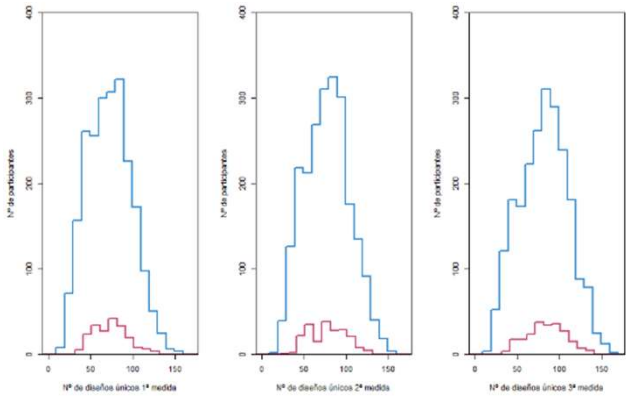
EJEMPLIFICACIÓN: TRATAMIENTO DE DATOS FALTANTES (ESCENARIO 10%)

- 4 métodos: Eliminación de casos, media *(con mice)*, regresión *(con mice)*, imputación múltiple (con mice, PMM, 5 iteraciones)

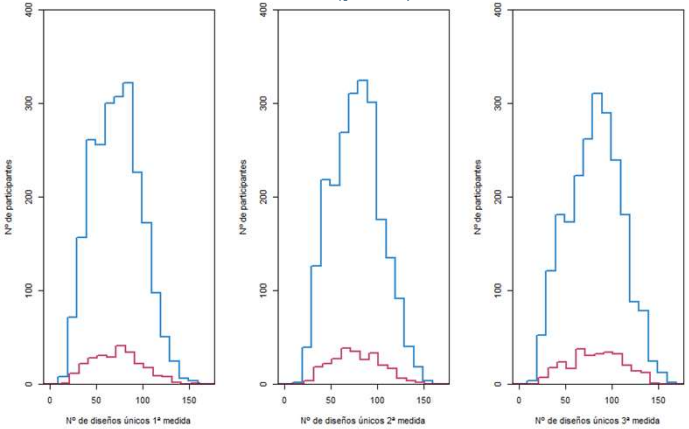
media

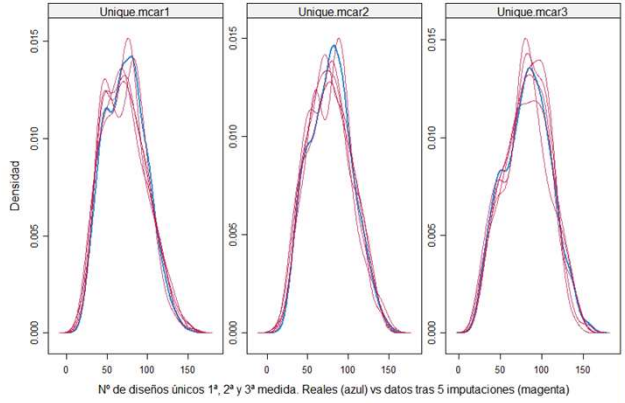


regresión



IM (pmm)





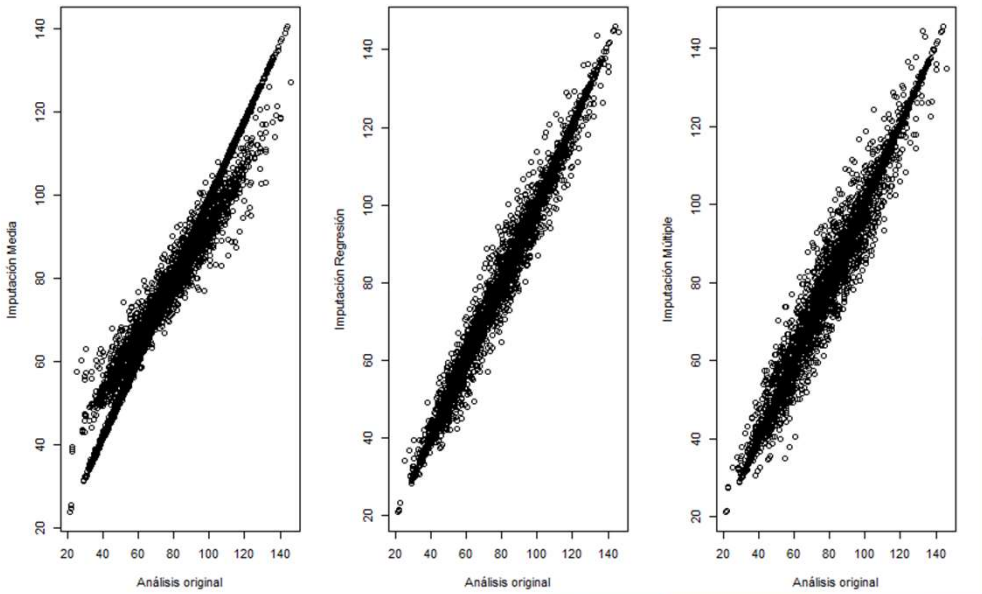
EJEMPLIFICACIÓN: ANÁLISIS DE RESULTADOS ORIGINALES VS IMPUTADOS (ESCENARIO 10%)

Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	Modelo 3	Modelo 3	Modelo 3	Modelo 3	Modelo 3
Variables	B (DE), IC95% (p-valor)	B (DE), IC95% (p-valor)	B (DE), IC95% (p-valor)	B (DE), IC95% (p-valor)	B (DE), IC95% (p-valor)
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.69 (0.062) , -0.81 to -0.57 (<0.001)	-0.61 (0.053), -0.72 to -0.51 (<0.001)	-0.67 (0.051), -0.76 to -0.57 (<0.001)	-0.66 (0.053), -0.76 to -0.56 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	1.14 (0.853) , -0.53 to 2.82 (0.18)	1.01 (0.68), -0.33 to 2.34 (0.139)	1.27 (0.719), -0.14 to 2.68 (0.077)	1.16 (0.723), -0.25 to 2.58 (0.108)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	6.66 (1.866), 3 to 10.32 (<0.001)	5.63 (1.469), 2.75 to 8.51 (<0.001)	6.57 (1.555), 3.52 to 9.62 (<0.001)	6.7 (1.562), 3.64 to 9.77 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	12.24 (1.896), 8.52 to 15.96 (<0.001)	10.94 (1.496), 8.01 to 13.87 (<0.001)	12.76 (1.583), 9.66 to 15.86 (<0.001)	12.84 (1.59), 9.72 to 15.95 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	24.56 (1.846), 20.94 to 28.18 (<0.001)	21.97 (1.462), 19.1 to 24.84 (<0.001)	25.04 (1.547), 22 to 28.07 (<0.001)	25.14 (1.555), 22.09 to 28.19 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	24.56 (1.207), 13.97 to 18.71(0.18)	15.2 (1.098), 13.04 to 17.35(<0.001)	16.64 (0.952), 14.77 to 18.51(<0.001)	16.9 (1.021), 14.9 to 18.9(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.22 (0.023), -0.27 to -0.18 (<0.001)	-0.2 (0.02), -0.24 to -0.16 (<0.001)	-0.23 (0.018), -0.26 to -0.19 (<0.001)	-0.23 (0.019), -0.27 to -0.19 (<0.001)
AIC	**65103.7**	**47020.5***	**65528.5**	**64291.2**	**65187.4**
Varianza residual	14.3	14.2*	15.1	13.2	14.1

EJEMPLIFICACIÓN: CONCLUSIONES

Teniendo en cuenta los 3 escenarios:

- IM (basada en el algoritmo PMM) de la función “mice”, presenta resultados robustos y presenta estimaciones cercanas a los resultados originales.
- Cabe destacar que con el método de regresión (método norm.predict,“mice”), la imputación es eficaz en general (datos faltantes cumplen MCAR).
- Con el método de sustitución por la media (método mean,“mice”), la imputación es peor que las anteriores y hay que tener en cuenta el problema de baja variabilidad que generará en los datos.
- Utilizar métodos de eliminación cuando el mecanismo de datos perdidos MCAR, conlleva pérdida de eficacia aunque la muestra permanezca insesgada.



MATERIAL PRESENTADO

- PRESENTACIÓN ORAL (FORMATO PPT)
- MEMORIA (FORMATO PDF)
- INFORME ESTADÍSTICO (FORMATO PDF)
- INFORME ESTADÍSTICO DINÁMICO REPRODUCIBLE (FORMATO MARKDOWN-RSTUDIO)