**PRIMER EJERCICIO OBLIGATORIO PARA SEMINARIO IMPUTACIÓN DE RESPUESTAS: MECANISMOS**

**Brian Norman Peña Calero**

**Responde este ejercicio en este Word y entrega también anexado el script de R.**

**Abrir el archivo Ejercicio1.dat desde R**

El objetivo de este ejercicio es familiarizarte con cómo inspeccionar los valores perdidos de un archivo antes de empezar a trabajar con él con técnicas inferenciales (regresiones, ANOVAS, SEM, etc.). Utilizaremos las herramientas de R que hemos visto en clase. Además, también se pretende con este ejercicio que sepas falsar que la pérdida de tus datos se ocasiona por mecanismos MCAR.

En este archivo encontrarás 779 estudiantes seleccionados al azar de carreras técnicas en el que recogemos 8 variables. ID es el número de identificador del sujeto. REND es la nota media en la carrera (primera convocatoria a la que se presentaron al examen). HESTUDIO contiene el percentil de horas de estudio, MOT el percentil en una escala de motivación por los contenidos aprendidos en su grado, SEXO (0 = mujer y 1 = varón), SELEC = la nota de selectividad, STUAULA son los estudiantes/aula de su grado y GASTO es el gasto en miles de euros por año (matrícula, piso, materiales, etc.).

1.1) Con la función md.pattern del paquete mice, contesta a la siguientes preguntas.

¿Cuántos casos tenemos sin valores perdidos (casos completos)?

Se tiene un total de 172 casos con valores completos en todas las variables.

1.2) ¿Cuál es el patrón más común de valores perdidos que tenemos (sin contar el de casos completos)? ¿Cuántos casos tiene ese patrón y qué variable/es tienen en ese patrón valores perdidos? Si quitásemos o eliminásemos del archivo a esa/as variable/es que implican a ese patrón, ¿cuántos casos completos tendríamos?

Excluyendo los casos completos, el patrón más frecuente de valores perdidos corresponde a la **variable MOT con 103 casos**. Este patrón se caracteriza por tener únicamente valores perdidos en MOT y valores completos en el resto de las variables: SEXO, STUAULA, SELEC, REND, GASTO y HESTUDIO.

Si elimináramos dicha variable, estos 103 casos podrían considerarse completos (ya que no tendrían ningún otro valor faltante). Así, al sumar estos 103 casos a los 172 casos completos iniciales, obtendríamos un total de **275 casos completos**.

1.3) ¿Consideras que se puede asumir que la pérdida de datos **para la variable REND** en este archivo es MCAR? Justifica tu respuesta (*nota*: conviene que generes primero una variable binaria con códigos 1 si tenemos valor válido en REND y 0 si tenemos valor perdido en REND. Te puedes ayudar del script visto en clase)

Para evaluar si la pérdida de datos en la variable REND puede asumirse como MCAR (Missing Completely At Random), se creó una variable indicadora binaria (REND\_obs) que toma el valor 1 cuando el dato de REND está presente y 0 cuando está ausente. A partir de esta variable, se realizaron comparaciones entre los grupos con y sin datos en REND utilizando pruebas t de Student para variables numéricas y una prueba chi-cuadrado para la variable categórica SEXO:

* **GASTO**: los estudiantes con datos perdidos en REND presentan un gasto significativamente menor (M = 1.78) que aquellos con datos presentes (M = 2.09; t(35.71) = -3.05, p = .004, d = -0.56).
* **MOT (Motivación)**: se observa una diferencia marcada en motivación, siendo más baja en el grupo con datos faltantes (M = 34.21) frente al grupo observado (M = 51.48; t(38.85) = -3.43, p = .001, d = -0.61).
* **HESTUDIO (Horas de estudio)**: los estudiantes con datos válidos en REND reportan significativamente más horas de estudio (M = 51.29) que aquellos con datos perdidos (M = 30.76; *t*(33.58) = -4.45, *p* < .001, d = -0.78).
* **SELEC (Nota de selectividad)**: se detecta una diferencia significativa, con menor nota media en el grupo con datos perdidos (M = 7.81) que en el grupo observado (M = 8.22; t(73.11) = -3.13, p = .002, d = -0.41).
* **STUAULA**: no se identifica una diferencia estadísticamente significativa en el número de estudiantes por aula, aunque la tendencia es contraria (mayor cantidad en el grupo con datos perdidos; t(62.69) = 1.69, p = .096, d = 0.28).
* **SEXO**: la prueba chi-cuadrado no revela asociación alguna entre el sexo y la presencia de datos perdidos en REND (χ²(1) = 0.013, p = .910).

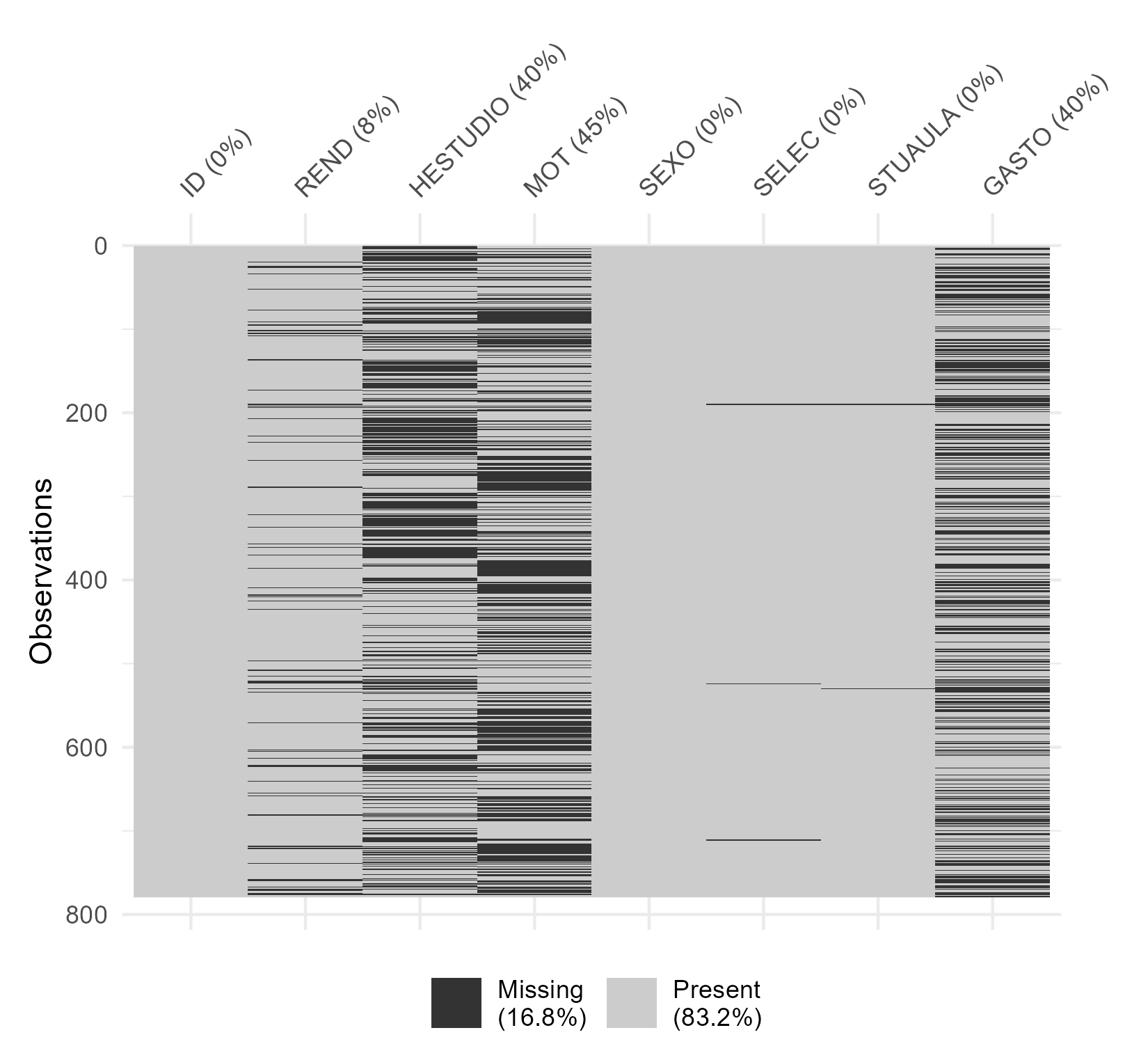
A partir de los análisis realizados, se concluye que la pérdida de datos en la variable REND no **puede considerarse MCAR** (Missing Completely At Random). Las asociaciones significativas con variables observadas como GASTO, MOT, HESTUDIO y SELEC, junto con tamaños del efecto moderados, sugieren que la probabilidad de que un dato falte depende de información disponible.

1.4) ¿Cuál es la media de horas de estudio (medida en percentiles) para los casos con un valor válido en REND? ¿Y para los casos con valores perdidos en REND? ¿Hay diferencias significativas en las horas de estudio entre los dos grupos?

Los resultados muestran que los estudiantes con datos válidos en REND tienen una media de **51.29** en HESTUDIO y aquellos con datos perdidos presentan una media de **30.76**. Esta diferencia es estadísticamente significativa (*t*(33.58) = -4.45, *p* < .001), con un tamaño del efecto grande (d = -0.78), lo que indica que la cantidad de horas de estudio está fuertemente asociada a la presencia de datos en REND. Dicho de otra forma, los estudiantes que reportan menos horas de estudio tienen mayor probabilidad de tener datos perdidos en REND, lo cual refuerza la evidencia de que la pérdida no es completamente aleatoria.

1.5) Para visualizar el patrón de valores perdidos univariado utiliza la función vis\_miss()del paquete naniar. Pega el grafico y comenta cuál es la variable del archivo que tiene más valores perdidos (informa del porcentaje)

La variable con mayor proporción de valores perdidos es MOT (Motivación), con 352 casos faltantes, lo que representa un 45.2% del total de observaciones.



**Mecanismos de valores perdidos. Abrir el archivo Ejercicio2.dat desde R**

Este ejercicio tiene como objetivo poner en práctica tu comprensión sobre los mecanismos que generan los valores perdidos: MCAR, MAR y MNAR.

Tenemos dos variables, IQ = Cociente intelectual y REND = Rendimiento. Supongamos que la media poblacional de Rendimiento es μ = 4,89, su desviación típica poblacional es σ = 2,34, la correlación entre ambas variables (poblacional) es ρ = 0,50.

Vamos a ver la recuperación de estos parámetros bajo MCAR, MAR y MNAR haciendo nuevamente cinco réplicas (o, si lo deseas, puedes hacer muchas más utilizando la estructura for) con cada uno de estos mecanismos. El objetivo es generar pérdida de datos en la variable rendimiento.

Condiciones: en todas las condiciones queremos un 50% de valores perdidos en Rendimiento.

Para MCAR generamos valores perdidos en Rendimiento a partir de una variable aleatoria uniforme (0,1) de forma que si el valor es < 0,50 tenemos un valor perdido en rendimiento y, en caso contrario, un valor válido.

Para MAR generamos valores perdidos en rendimiento de la siguiente manera. Para el primer cuartil de CI la probabilidad de valor perdido en Rendimiento = 0,80, para el segundo cuartil de CI la probabilidad de valor perdido en Rendimiento = 0,60, tercer cuartil de CI una probabilidad de 0,40 y para el cuarto cuartil de CI una probabilidad de 0,20.

Para MNAR generamos valores perdidos en rendimiento de la siguiente manera. Para el primer cuartil de rendimiento la probabilidad de valor perdido será de 0,80, probabilidad de 0,60 para el segundo cuartil de rendimiento, probabilidad de 0,40 para el tercer cuartil y probabilidad de 0,20 para el cuarto cuartil.

2.1) Justifica brevemente que esta forma de generar valores perdidos es, efectivamente, MCAR, MAR y MNAR.

* **MCAR (Missing Completely At Random):** En este caso, los valores perdidos en REND se generan a partir de una variable aleatoria uniforme independiente de cualquier otra variable del conjunto de datos. Es decir, la probabilidad de que un valor esté ausente no depende ni de REND ni de IQ, ni de ninguna característica observable o no observable. Por tanto, se cumple la definición de MCAR, ya que la pérdida es completamente aleatoria.
* **MAR (Missing At Random):** En esta condición, la probabilidad de pérdida en REND depende del nivel de CI (IQ), pero no del propio valor de REND. Se asignan probabilidades diferenciadas de valores perdidos en función del cuartil de CI: 0.80, 0.60, 0.40 y 0.20. Esto se alinea con definición de MAR, ya que los valores perdidos en REND dependen de otra variable observada (IQ), pero no del valor que tendría REND si no estuviera ausente.
* **MNAR (Missing Not At Random):** En este caso, la probabilidad de que un valor esté perdido en REND depende del propio valor de REND, es decir, de la variable que presenta la pérdida. Se asignan probabilidades decrecientes de pérdida según el cuartil de REND: 0.80, 0.60, 0.40 y 0.20. Esto corresponde a una condición MNAR, ya que el mecanismo de pérdida no puede explicarse completamente con variables observadas, sino que depende de la variable que presenta los datos faltantes.

2.2) Rellena la siguiente tabla (o si has utilizado más réplicas, pon directamente las medias obtenidas para el conjunto de réplicas):

Se realizó la introducción de los mecanismos con 1000 réplicas y se resumieron por la media:

|  |  |  |  |
| --- | --- | --- | --- |
| **Mecanismo** | **Media** | **Desviación Típica** | **Correlación Rend-CI** |
| MAR | 5.387 | 2.281 | 0.456 |
| MCAR | 4.889 | 2.340 | 0.494 |
| MNAR | 5.859 | 2.122 | 0.457 |
| Poblacional | 4.89 | 2.34 | 0.50 |

2.3) Comenta brevemente los resultados encontrados.

Al analizar los resultados obtenidos bajo los tres mecanismos de pérdida (MCAR, MAR y MNAR), se observan diferencias importantes en la recuperación de los parámetros poblacionales:

* **MCAR (Missing Completely At Random)**: Los valores estimados son muy cercanos a los parámetros poblacionales conocidos (media ≈ 4.89, desviación ≈ 2.34, correlación ≈ 0.50). Esto es esperable, ya que bajo MCAR la pérdida es completamente aleatoria y no introduce sesgos en las estimaciones si se trabaja con casos completos.
* **MAR (Missing At Random)**: Se observa una sobreestimación de la media del rendimiento (5.39 vs. 4.89) y una ligera subestimación de la correlación (0.456 vs. 0.50). Esto ocurre porque los valores perdidos en REND están asociados a IQ, lo que afecta la representatividad de los casos observados.
* **MNAR (Missing Not At Random)**: Es el caso más problemático. La media de rendimiento es claramente sobreestimada (5.86), la desviación típica subestimada (2.12 vs. 2.34), y la correlación también se ve alterada. Esto se debe a que los valores perdidos dependen directamente de la propia variable REND, lo cual genera estimaciones sesgadas.

Anexo Script

##%######################################################%##  
# #  
#### Ejercicio 1 ###  
#### Brian Norman Peña Calero ###  
#### Seminario de Valores Perdidos ###  
# #  
##%######################################################%##  
  
# Carga de Paquetes -------------------------------------------------------  
  
library(tidyverse)  
library(mice)  
  
# Importar Datos Ejercicio 1 ----------------------------------------------  
  
carreras <- read\_delim("Ejercicio1.dat")  
carreras  
  
  
# Pregunta 1.1, 1.2 -------------------------------------------------------  
  
md.pattern(carreras)  
  
  
# Pregunta 1.3 y 1.4 ------------------------------------------------------  
  
## 1. Crear variable binaria  
carreras$REND\_obs <- ifelse(is.na(carreras$REND), 0, 1)  
  
## 2. Comparaciones de medias entre grupos (t-test)  
  
carreras |>   
 pivot\_longer(  
 cols = c(SELEC:GASTO, HESTUDIO:MOT),  
 names\_to = "Var\_cont",  
 values\_to = "valores"  
 ) |>   
 group\_nest(Var\_cont) |>   
 mutate(  
 test = map(  
 data,   
 ~ statsExpressions::two\_sample\_test(  
 data = .x,  
 x = REND\_obs,  
 y = valores,  
 effsize.type = "d"  
 )  
 )  
 ) |>   
 unnest(test) |>   
 select(  
 Var\_cont, mean.parameter1:p.value,  
 cohen\_size = estimate  
 ) |>   
 insight::format\_table() |>   
 insight::export\_table()  
  
## 3. Comparación en variable categórica (chi-cuadrado)  
tabla\_sexo <- table(carreras$SEXO, carreras$REND\_obs)  
chisq.test(tabla\_sexo)  
  
  
# Pregunta 1.5 ------------------------------------------------------------  
  
carreras |>   
 select(-REND\_obs) |>   
 naniar::vis\_miss() +  
 theme(  
 plot.margin = margin(l = 10, r = 30)  
 )  
  
## Exportar gráfico  
# ggsave(  
# "univariado\_missing\_ej1.png",  
# dpi = 300,  
# width = 5.5,  
# height = 5,  
# bg = "white"  
# )  
  
carreras |>   
 select(-REND\_obs) |>   
 naniar::miss\_var\_summary()  
  
  
# Importar Datos Ejercicio 2 ----------------------------------------------  
  
rendimiento <- read\_delim("Ejercicio2.dat")  
  
  
# Pregunta 2 -------------------------------------------------------------  
  
## Función para generar valores perdidos en REND  
generar\_missing <- function(data, mecanismo = c("MCAR", "MAR", "MNAR")) {  
 data <- data  
 mecanismo <- match.arg(mecanismo)  
   
 if (mecanismo == "MCAR") {  
 prob <- runif(nrow(data))  
 data$REND[prob < 0.5] <- NA  
 }  
   
 if (mecanismo == "MAR") {  
 q\_iq <- ntile(data$IQ, 4)  
 probs <- c(0.8, 0.6, 0.4, 0.2)[q\_iq]  
 data$REND[runif(nrow(data)) < probs] <- NA  
 }  
   
 if (mecanismo == "MNAR") {  
 q\_rend <- ntile(data$REND, 4)  
 probs <- c(0.8, 0.6, 0.4, 0.2)[q\_rend]  
 data$REND[runif(nrow(data)) < probs] <- NA  
 }  
   
 return(data)  
}  
  
## Función para extraer estadísticas de casos completos  
resumen\_estadistico <- function(data) {  
 data <- drop\_na(data, REND)  
 tibble(  
 mean = mean(data$REND),  
 sd = sd(data$REND),  
 cor = cor(data$REND, data$IQ)  
 )  
}  
  
## Función para hacer 1000 réplicas por mecanismo  
replicar <- function(mecanismo, n\_rep = 1000, datos\_originales) {  
 replicate(n\_rep, {  
 datos\_miss <- generar\_missing(datos\_originales, mecanismo)  
 resumen\_estadistico(datos\_miss)  
 }, simplify = FALSE) |> bind\_rows()  
}  
  
## Ejecutar para cada mecanismo con la data  
resultados\_mcar <- replicar("MCAR", datos\_originales = rendimiento)  
resultados\_mar <- replicar("MAR", datos\_originales = rendimiento)  
resultados\_mnar <- replicar("MNAR", datos\_originales = rendimiento)  
  
## Unir resultados en tabla larga  
tabla\_replicas <- bind\_rows(  
 MCAR = resultados\_mcar,  
 MAR = resultados\_mar,  
 MNAR = resultados\_mnar,  
 .id = "Mecanismo"  
) %>%  
 mutate(Replica = rep(paste0("R", 1:1000), 3)) %>%  
 relocate(Mecanismo, Replica)  
  
tabla\_replicas |>   
 group\_by(Mecanismo) |>   
 summarise(  
 across(  
 where(is.numeric),  
 mean  
 )  
 )