UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

"LOS VALORES FALTANTES EN LAS ENCUESTAS SOCIALES: COMPARACIÓN DE LOS ENFOQUES BAYESIANO OBJETIVO Y FRECUENTISTA."

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Académica en Estadística

SOFÍA DE LOS ÁNGELES BARTELS GÓMEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica

2021

Dedicatoria

A Dios, a mis padres que nos enseñaron que la educación es la herencia más importante y a mi esposo que me animó cuando más lo necesitaba.

Agradecimientos

A la Dra. Eiliana Montero Rojas, como directora de esta tesis por su paciencia, guía y apoyo en todo proceso de investigación; además por impulsarme para que presentara mis resultados preliminares en conferencias internacionales.

Al Dr. Ricardo Alvarado Barrantes, que como lector de tesis no solo contribuyó con sus valiosas observaciones, sino que se involucró en mi aprendizaje con el curso de las simulaciones que implementó.

A la Dra. Vanessa Smith Castro, por permitir que los datos de su proyecto de investigación se utilizaran en esta tesis, y que con valiosos aportes enriqueció el estudio.

A todas las personas que de una u otra manera contribuyeron en la culminación de este trabajo.

Esta Tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Estadística

GILBERT BRENES CAMACHO (FIRMA)

Firmado digitalmente por GILBERT BRENES CAMACHO (FIRMA)
Fecha: 2021.05.21 15:02:07 -06'00'

Dr. Gilbert Brenes Camacho
Representante del Decano
Sistema de Estudios de Posgrado

EILIANA MONTERO ROJAS (FIRMA)

Firmado digitalmente por EILIANA MONTERO ROJAS (FIRMA)
Fecha: 2021.05.21 13:59:57 -06'00'

Dra. Eiliana Montero Rojas
Profesora Guía

RICARDO ALVARADO BARRANTES (FIRMA)

Firmado digitalmente por RICARDO ALVARADO BARRANTES (FIRMA)
Fecha: 2021.05.24 16:59:49 -06'00'

Dr. Ricardo Alvarado Barrantes
Lector

VANESSA SMITH CASTRO (FIRMA)

Firmado digitalmente por VANESSA SMITH CASTRO (FIRMA)
Fecha: 2021.05.24 17:30:39 -06'00'

Dra. Vanessa Smith Castro
Lectora

JOHNNY ANGEL MADRIGAL PANA (FIRMA)

Firmado digitalmente por JOHNNY ANGEL MADRIGAL PANA (FIRMA)
Fecha: 2021.05.24 21:14:14 -06'00'

M.Sc. Johnny Madrigal Pana
Director Programa
de Posgrado en Estadística

SOFIA DE LOS ANGELES BARTELS GOMEZ (FIRMA)

Firmado digitalmente por SOFIA DE LOS ANGELES BARTELS GOMEZ (FIRMA)
Fecha: 2021.05.25 09:15:58 -06'00'

Sofia de los Ángeles Bartels Gómez
Sustentante

Tabla de Contenidos

Resumen

La aparición de datos faltantes en las Investigaciones Sociales es frecuente, este problema ha tomado relevancia en los últimos años por las implicaciones que tienen sobre la validez de las técnicas estadísticas empleadas, los resultados y las conclusiones que se reportan en los estudios. Además, su abordaje implica una serie de cuestionamientos que deben de considerarse para determinar si las observaciones han sido perdidas al azar o si su falta se debe a causas definibles, llevando esto a debatir si estos datos pueden o no ser tratados como ignorables, así como la conveniencia, facilidad y apropiado de utilizar algún método de imputación para su recuperación. Por lo anterior se plantea esta investigación, cuyo objetivo es comparar el enfoque Bayesiano Objetivo y Frecuentista para la imputación de valores faltantes en un contexto típico de encuestas de investigación social.

El estudio de simulación permitió evidenciar que, fijar por defecto en 10 la cantidad de iteraciones del mecanismo de imputación múltiple permite minimizar el tiempo de procesamiento estadísticos y no implica que se vea incrementado el sesgo. Además, se detecta la relevancia de la magnitud de la asociación entre la variable con el dato faltante y las independientes utilizadas para su predicción, en el nivel de precisión resultante del proceso de imputación múltiple. Entre mayor sea esta, menor es el error promedio introducido en los datos, independientemente del enfoque de estimación utilizado (frecuentista o bayesiano).

Finalmente, a partir de lo encontrado en esta investigación, se puede concluye que existen casos en los que la imputación múltiple no es recomienda, específicamente en aquellos donde la asociación entre la variable con datos faltantes y las predictoras sea moderada (menor a 0.55), esto porque se observan grandes diferencias entre resultados que arrojan los datos completos originales, y los que se generarían a partir de una imputación. Por lo que se recomienda incluir en el diseño de investigación un plan de teoría de los valores faltantes y valorar la implementación alternativa de incentivos a los participantes para minimizar la cantidad de datos faltantes.

Índice de Cuadros

Índice de Figuras

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Sofía de los Ángeles Bartels Gómez, con cédula de identidad 113810509, en mi condición de autor del TFG titulado Los valores faltantes en las encuestas sociales: Comparación de los enfoques Bayesiano Objetivo y Frecuentista.

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. SI [X] NO * [ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kérwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director(a) de Tesis o Tutor(a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:
Nombre Completo: Sofía de los Ángeles Bartels Gómez.
Número de Carné: A70910. Número de cédula: 113810509.
Correo Electrónico: sofia.bartels@ucr.ac.cr.
Fecha: 13/05/2021. Número de teléfono: 88971140.
Nombre del Director (a) de Tesis o Tutor (a): Eiliana Montero Rojas.

SOFIA DE LOS ANGELES BARTELS GOMEZ (FIRMA)
Firmado digitalmente por SOFIA DE LOS ANGELES BARTELS GOMEZ (FIRMA)
Fecha: 2021.05.25 09:17:45 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kérwá.

Capítulo I Problema y Objetivos

Este apartado describe el problema de investigación, así como la importancia de su abordaje. Además, detalla los objetivos planteados para contestar la pregunta de investigación y muestra un esquema de las variables dentro de la temática.

1.1. Problema

¿Cuál enfoque para la imputación de datos, Bayesiano objetivo o frecuentista, recupera de forma más precisa la información faltante en encuestas sociales?

1.2. Justificación

En un escenario perfecto de análisis de datos se contaría con variables en donde, para cada uno de los sujetos examinados, existe un valor válido y confiable capaz de brindar información de su comportamiento. Sin embargo, esto no siempre es posible, Howell, D. C. (2007) indica que el tratamiento de los valores faltantes ha sido un problema en la estadística, el cual ha pasado a primer plano en los últimos años, esto porque es un tópico regularmente analizado tanto en los censos como en las encuestas.

La falta de datos implica una serie de inconvenientes. En relación con esto Allison (2001) menciona que el problema radica en que los métodos estadísticos suponen que cuenta con la información completa sobre todas las variables incluidas en el análisis, lo cual tiene implicaciones en los resultados. Esto es reconocido por Acock (2005) quien enfatiza cómo a menudo sucede que el punto más débil en los estudios está dado por la calidad de los datos, y esta a su vez influye sobre el problema de investigación.

Por lo anterior, el abordaje de los datos faltantes debe ir desde la compresión de la existencia del problema al cómo establecer si las observaciones han sido perdidas al azar o si su falta se debe a causas definibles, llevando esto a cuestionarse si estos datos pueden o no ser tratados como ignorables, por lo que, lo ideal es entonces conocer los patrones de los datos faltantes, así como una justificación de por qué el patrón importa o no.

Existen diferentes métodos para lograr imputar los valores faltantes, entre ellos máxima verosimilitud y la imputación múltiple. Ambas opciones tienen propiedades estadísticas favorables, pero según Allison (2001) es esencial tener en cuenta las ventajas y desventajas presentes en estos métodos, y dependen de ciertas suposiciones fácilmente vulnerables. Ambos métodos pueden llevarse a cabo desde los paradigmas bayesiano y frecuentista.

Lo anterior coincide con lo que expone Acock (2005), quien señala que ambos métodos son aceptables para tratar los datos faltantes. No obstante, hace la observación de que estos enfoques están bien establecidos para el análisis de un solo nivel. Sin embargo, para el caso de datos multinivel, existe menos orientación disponible a partir de estudios anteriores sobre los datos ausentes (Larsen, 2011). Larsen realizó recientemente un estudio comparativo de los enfoques de Imputación múltiple (MI) y Máxima Verisimilitud (ML) en situaciones donde no había personas anidadas dentro de grupos, encontró que ambos enfoques eran relativamente similares en la recuperación del nivel 1, pero, para el segundo nivel de estimación, observó una mejor estimación de las variables que tenían valores faltantes al utilizar ML comparado con el enfoque de MI, esto debido a que el procedimiento de MI utilizado en su estudio no tuvo en cuenta los efectos aleatorios.

Esta investigación se entra en un caso típico de investigación social, para el desarrollo de la aplicación práctica del problema de investigación se trabajará con datos del Proyecto de investigación 723-B3-307 “Nuevas formas de medir viejas ideologías: el caso de los sexismos y sus implicaciones en el ámbito académico” de la investigadora principal Dra. Vanessa Smith Castro.

En esta aplicación se quiere modelar la relación entre el rendimiento en matemática con las variables del sexismo hostil y benevolente, la autoeficacia matemática, las habilidades generales de razonamiento y las variables sociodemográficas de sexo y edad, bajo un contexto de datos faltantes en la variable dependiente, y cuya escogencia de recuperación será recomendada a partir de simulaciones bajo escenarios frecuentistas y

bayesianos. La selección del tema sustantivo como caso aplicado se debe a la relevancia de estos datos para la toma de decisiones en las políticas a favor de la equidad.

La ONU Mujeres (2010) menciona que para logar economías más fuertes, alcanzar los objetivos de desarrollo, la sostenibilidad en convenidos internacionalmente y mejorar la calidad de vida de las mujeres, las familias y las comunidades, es fundamental empoderar a las mujeres para una participación plena en la vida económica en todos sus sectores.

Para el caso de Costa Rica se observan desigualdades en la participación de las mujeres en sectores económicos importantes, según el estudio de Bonilla & Uribe (2005) donde señalada que, en la Ciencia y la Tecnología, de acuerdo con datos del Censo 2000, se muestra una clara división por género según las ocupaciones: en las Ingenierías y afines la población está mayormente representada por hombres; las autoras mencionan, por ejemplo, que en Ingeniería Eléctrica el censo reporta 701 hombres y 29 mujeres; es decir, una razón por sexo de 24. En los ingenieros mecánicos hay 450 hombres censados contra 19 mujeres, cuya razón es también de 24. El censo reporta además la existencia de 855 ingenieros topógrafos hombres y solo 39 mujeres. Estos ejemplos exponen como las áreas de las Ingenierías han sido culturalmente masculinizadas, rasgo prevaleciente aun en la actualidad, lo anterior es señalado con regularidad; Clark Blickenstaff (2005) por ejemplo, coincide al respecto indicando cómo las mujeres están insuficientemente representadas en la ciencia, tecnología, ingeniería y matemáticas (STEM).

Todas las carreras mencionadas anteriormente cuentan con un componente en común dentro de los planes de estudio: las matemáticas. Bussey y Bandura (1999), citados en Brown y Leaper (2010), hacen una reflexión sobre la importancia de la representación de las mujeres en áreas de ciencia, tecnología, ingeniería y matemáticas, por cuanto estas áreas están asociadas con las ocupaciones dentro de los sectores económicos de los países en donde se obtienen los más altos salarios, aunado a un gran prestigio social.

La Universidad de Costa Rica no está alejada de esta realidad, por ejemplo, según Marín, Barrantes & Chavarría (2008) actualmente se gradúan mucho menos mujeres que

hombres en el pregrado y posgrado de la Escuela de Ciencias de la Computación e Informática (ECCI).

1.3. Población involucrada o afectada.

- Se trabajó con datos de 487 estudiantes matriculados en undécimo año en el año 2015 de 10 secundarias públicas seleccionadas al azar del conjunto de secundarias académicas diurnas de la GAM de Costa Rica.

1.4. Objetivo general:

- Comparar el enfoque Bayesiano Objetivo y Frecuentista para la imputación de valores faltantes en un contexto típico de encuestas de investigación social.

1.4.1. Objetivos específicos:

- Generar un marco de referencia teórico que sustente la investigación en términos de las bondades y limitaciones de los métodos Bayesiano Objetivo y Frecuentista para la imputación de datos faltantes.

- Generar dos escenarios perdidos completamente al azar o no perdidos al azar (MCAR, y MNAR) simulados para la presencia de valores faltantes, a partir de los datos recolectados en el Proyecto de investigación 723-B3-307 “Nuevas formas de medir viejas ideologías: el caso de los sexismos y sus implicaciones en el ámbito académico”.

- Comparar por medio de modelos de regresión la precisión de los enfoques Bayesiano Objetivo y Frecuentista en imputación múltiple para recuperar la información en los dos escenarios de valores faltantes.

- Proponer recomendaciones para la imputación de valores faltantes dependiendo del patrón que se presente en los datos.

1.5. Diagrama gráfico de la interrelación entre los factores independientes

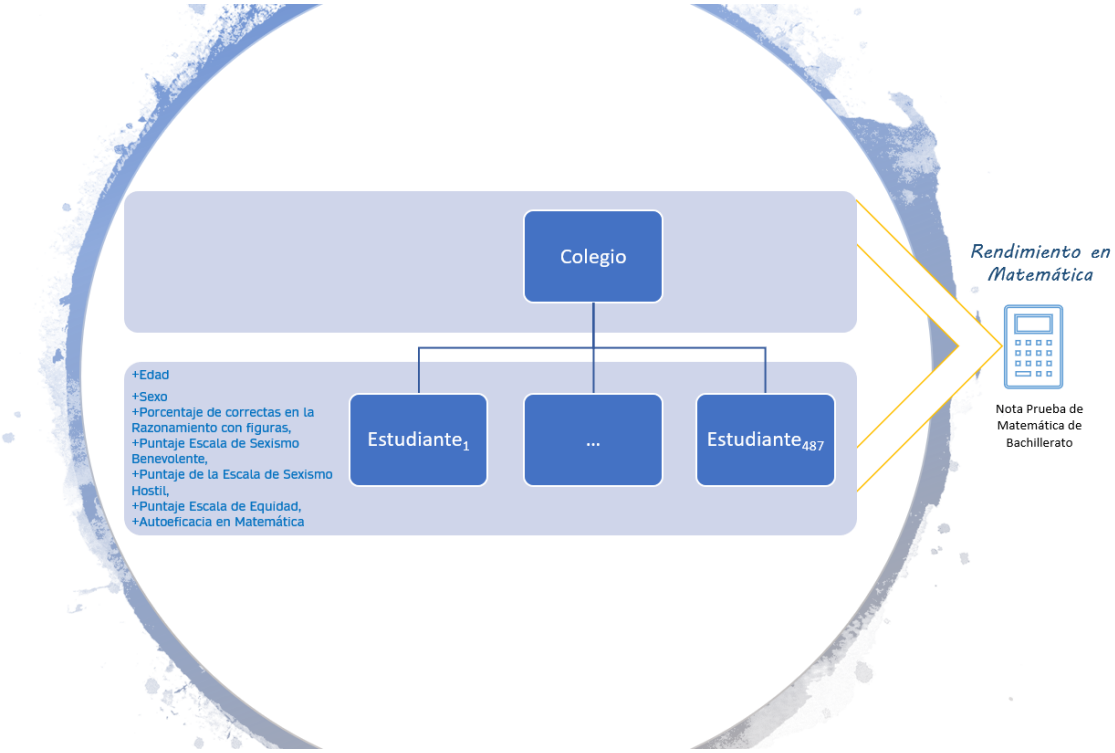


Figura 1. Diagrama de interrelación entre los factores independientes y dependientes

Capítulo II: Estado de la Cuestión

Este apartado aborda el marco teórico pertinente al análisis de datos faltantes, su relación con los paradigmas y así como los métodos utilizados para su recuperación.

2.1 Los datos faltantes en la investigación

La pérdida de información es un problema común enfrentado por los investigadores, las razones por la cuales no es un tema inusual se deben a diversos aspectos, Castro & Ávila (2006) mencionan que la proporción de ausencia de un ítem en un registro puede variar dependiendo del estudio, de la dificultad de llenar el cuestionario o de la medición de una variable. Sin embargo, dependerá del investigador considerar el registro como pérdida parcial o como pérdida total, esta decisión dependerá del número de ítems faltantes por individuo.

Además, la ausencia de respuestas constituye una de las mayores limitaciones de cualquier estudio, según lo comentan Carracedo-Martínez & Figueiras (2006), pues, desde el punto de vista conceptual, la falta de respuestas engloba dos aspectos relevantes: por un lado, la no participación de un sujeto en el estudio al no contestar el cuestionario, y, por el otro, los valores ausentes de personas que responden al cuestionario de forma incompleta al no contestar una o varias variables.

Así también, Kang (2013) advierte cuales son los cuatro principales problemas implicados en el faltante de datos: 1) la ausencia de datos reduce la potencia estadística, 2) los datos faltantes pueden causar sesgo en la estimación de parámetros, 3) se puede reducir la representatividad de las muestras, y 4) pueden complicar el análisis del estudio.

Por lo anterior, una solución por considerar, cuando se presentan datos faltantes, es buscar métodos para tratar de completar la información con valores plausibles. A dicho procedimiento se le denomina imputación. No obstante, es una tarea que debe ser

tomada con precaución. En este sentido, para Castro & Ávila (2006) los investigadores deben de determinar cuál es el porcentaje máximo de pérdida de ítems considerado como tratable mediante imputación o cuando considerar que los datos faltantes se deben a una mala recolección de información y, por tanto, la base de datos construida es muy defectuosa y no debe ser usada.

En la misma línea Laaksonen (2000), citado en (Galarza Guerrero, 2013), recalca la importancia de identificar la tasa de no respuesta con el propósito de determinar qué proceso se debe seguir, así, por ejemplo, si la tasa supera un tercio del total del tamaño muestral es considerada como elevada, aunque ha encontrado que otros investigadores son más conservadores y hablan de pérdidas máximas entre 1 y 20 por ciento. Sin embargo, recalca como al final dependerá mucho de la precisión del estudio, el área y objetivos de la investigación para considerar la selección de la técnica de imputación a utilizar, estas técnicas se abordarán más adelante.

2.1.1 Problemas de los datos faltantes

El no tener registros de información completos desencadena una serie de problemáticas difíciles de ser ignoradas, en este aspecto Der & Everitt (2013) señalan como, en cualquier estudio de investigación, el objetivo de los análisis es hacer inferencias válidas con respecto a una población de interés. Razón por la cual los datos faltantes amenazan este objetivo, en especial si se han perdido de una manera capaz de provocar que la muestra con información completa sea diferente de la población definida en la investigación, es decir, si los datos no presentes crean una muestra sesgada.

En esta misma dirección Cheung (2013) menciona que uno de los problemas más importantes de la pérdida de datos es la reducción de la potencia y precisión. Señala además la importancia de darse cuenta de cómo la cantidad de datos que faltan en un estudio está en función de los recursos invertidos y como estos recursos no son simplemente recursos financieros. Por ello señala la existencia de una relación inversa

entre la cantidad de recursos invertidos y el número de valores faltantes, es así como cuantos más recursos se introduzcan, es menor número de valores faltantes presentes.

Considerando lo anterior es importante reflexionar y decidir la forma en la cual se van a tratar los valores faltantes en la investigación. Esta decisión debe ser tomada con cautela, por cuanto se pueden generar problemas en la validez de los resultados. En este tema Acock (2005) hace hincapié en como la inadecuada elección de la estrategia para recuperar los valores faltantes puede producir estimaciones sesgadas, distorsiones estadísticas y conclusiones inválidas.

2.1.2 Tipos de datos faltantes

Unos de los puntos claves al tratar los valores faltantes es detectar correctamente el mecanismo de generación de valores ausentes al realizar la imputación. Si no se tiene en cuenta o si se detecta erróneamente, las imputaciones obtenidas pueden estar sesgadas y, por tanto, las conclusiones extraídas del estudio pueden ser incorrectas (Abellán de Andrés, 2015).

Por esta razón, y según lo menciona Cheung (2013), al seleccionar la táctica adecuada para tratar los valores faltantes se debe conocer el patrón de datos faltantes, este puede ser monótono o no monótono. Un modelo monótono se presenta, por lo general, debido a la deserción del estudio, de tal manera que algunos sujetos fueron medidos en las variables de exposición de los resultados a corto plazo, pero no para los resultados a largo plazo. Este patrón reduce el tamaño efectivo de la muestra, pero los valores no introducen un sesgo de selección.

Una forma de reconocer los patrones monótonos es verificar que la pérdida de información fue en cadena a partir de una variable determinada. En caso contrario se está en un patrón no monótono (Ver Cuadro 1).

Patrón de Valores faltantes

Monótono					No monótono				
A	B	C	D	…	A	B	C	D	…
*	*	*	*		*	*	-	*	
*	*	*	-		-	*	*	-	
*	*	-	-			…			
*	-	-	-		*	-	-	*	

Nota: A, B, C y demás, son variables. * es observada; - es no observada

Cuadro 1. Posibles ejemplos de datos faltantes, según tipo de patrón (Cheung, 2013)

Por otro lado, Acock (2005) menciona la existencia de varias clasificaciones de los valores faltantes. Para cada una de ellas hay una estrategia óptima al trabajar con valores que faltan. Estos son: por falta de definición de subpoblaciones, faltantes completamente al azar, faltantes al azar, y los valores faltantes que no pueden ignorarse.

En el caso de los valores faltantes por la definición de subpoblaciones, estos se deben a que algunos de los participantes del estudio se excluyeron del análisis desde el planteamiento de la investigación, esto por cuanto para variables determinadas no están en la subpoblación bajo investigación. La mayoría de las encuestas tienen definidos los códigos de los valores faltantes para distinguir a los participantes por ser tratados como desaparecidas por definición, de aquellos para quienes es apropiado imputar los valores, por lo general estos son definidos con el código NA (No Aplica).

El autor destaca que la definición de la subpoblación para el estudio, la eliminación de las personas que no encajan en este dominio y la imputación de los valores que faltan se debe hacer con mucho cuidado.

La segunda y tercera clasificación tiene relación con datos faltantes al azar. Es explicado por Cheung (2013) como completamente al azar (MCAR) si los mecanismos que conducen a los datos que faltan son independientes tanto de la de los datos (que faltan) como de los no observados, y faltantes al azar (MAR) si el valor perdido sólo depende de los datos observados.

Según Acock (2005) tener datos completamente al azar es raro, ya que está bien establecido que los hombres, los individuos de grupos minoritarios, las personas con altos ingresos, los de poca educación, y las personas deprimidas o ansiosas son menos propensas a responder a cada elemento de un cuestionario. Sin embargo, es razonable en algunos estudios tener datos faltantes completamente al azar desde el diseño, para explicar esto con claridad el autor señala el ejemplo de una entrevista de 100 ítems a niños entre edades de 5 y 9, donde esta cantidad de preguntas instituiría un problema de fatiga grave, razón por la cual el investigador puede seleccionar al azar a 20 elementos para cada niño y sólo se tienen respuestas para dichos elementos. La única limitación es la incertidumbre, introducida esta por el proceso de imputación, y esta incertidumbre reduce la potencia estadística en comparación con tener datos completos.

El supuesto MAR, cabe recalcar, es válido si se puede suponer como condicionalmente al azar al patrón de valores que faltan, teniendo en cuenta los valores observados en las variables del mecanismo. La selección de las variables del mecanismo es una tarea que se debe hacer meticulosamente.

Complementando lo anterior, Carracedo-Martínez & Figueiras (2006) mencionan que los valores no ignorarles no corresponden a valores faltantes al azar, y por ello implican un buen conocimiento del porque se da la perdida de información, debe ir acompañado de análisis de sensibilidad. Además de que despiertan la desconfianza en las conclusiones. (Ver cuadro 2).

Acrónimo	Denominación	Características	Ejemplo
DFEA	Datos faltantes estrictamente aleatorios(missing completely at random)	La probabilidad de que una respuesta a una variable Y sea dato faltante es independiente tanto de los valores de la variable X como de Y	Sea Y el peso y X la edad. Existe el mismo porcentaje de datos faltantes a cualquier edad
DFA	Datos faltantes aleatorios(missing at random)	La probabilidad de que una respuesta a una variable Y sea dato faltante es dependiente de los datos de la variable X pero independiente de los valores de Y	Sea Y el peso y X la sexo, uno de los dos sexos (por ejemplo, el femenino) tiene un porcentaje de datos faltantes mayor para la variable peso
DFNA	Datos faltantes no aleatorios(not missing at random)	La probabilidad de que una respuesta a una variable Y sea dato faltante es dependiente de los datos de la variable y, es decir existe sesgo	Sea Y el peso, los sujetos con mayor valor de peso tienen un porcentaje de datos faltantes más elevado en la variable peso

Cuadro 2. Mecanismos de datos faltantes según sus características Carracedo-Martínez, E., & Figueiras, A. (2006)

Todo lo anterior implica la pregunta de cómo identificar el patrón presente en los datos que faltan, Montenegro-Montenegro, Oh & Chesnut (2015) explican cómo se deben determinar los tipos de los patrones o mecanismos de datos faltantes, estos están vinculados con la relación existente entre los datos faltantes y los datos presentes en las variables, se tiene entonces:

- Faltantes completamente al azar:

$$\Pr(R = 1 \mid X, Y) = \Pr(R = 1)$$

En este caso se puede ver cómo el patrón establece que la probabilidad de encontrar valores faltantes en Y no está relacionada con alguna variable X presente en los datos. Este patrón por lo general es considerado como un supuesto en la mayoría de las técnicas.

- Faltantes al azar:

$$\Pr(R = 1 \mid X, Y) = \Pr(R = 1 \mid X)$$

- Faltantes no al azar (con patrón), sucede cuando no se cumple el supuesto de MNAR, la razón para la existencia de datos faltantes subyace en la variable en sí misma, lo cual significa que este mecanismo ocurre cuando los datos faltantes, en una determinada variable, ocurren debido a los niveles de los sujetos en esa variable (citado en Little et al., 2014)

2.1.3 Formas de estimación de los datos faltantes

Una vez establecido el tipo de dato faltante presente, Galván (2007), empleando lo expuesto por Little y Rubin, clasifica los métodos de imputación de la siguiente forma: Análisis de datos completos (Listwise), análisis de datos disponibles (pairwise), imputación por medias no condicionadas, imputación por medias condicionadas mediante métodos de regresión, máxima verosimilitud (EM) e imputación múltiple (IM).

Para Acock (2005), con referencia al primer tipo, el análisis de datos completos (Por lista o por eliminación) es la solución más común para los valores que faltan, es tan común que es el valor por defecto en los paquetes estadísticos estándar. Para utilizar esta forma de estimación se debe cumplir el supuesto de MCAR, la eliminación por lista tiene la desventaja de ser sensible a tamaños de la muestra pequeños, esto por cuanto se inflará el error y reducirá el nivel de significación. Por lo tanto, en los casos en los cuales se cumplan los supuestos de MCAR y se utilice este método significará aumentar el riesgo de un error tipo II. Este problema se hace más leve conforme se aumenta la muestra.

Cabe destacar que en los casos en donde no se cumpla el supuesto de MCAR, este método puede producir estimaciones sesgadas. Acock (2005) hace mención de que, en general, habrá presencia de sesgo, esto por cuanto los casos completos pueden no representar a la población. Además, es importante señalar como, en el caso multivariado, el sesgo

causado por la eliminación por lista puede exagerar, tener efectos significativos o subestimar los demás.

Para Acock (2005), en el caso de Pairwise, esta técnica utiliza toda la información disponible, esto en el sentido de que todos los participantes que respondieron a un par de variables se utilizan para estimar la covarianza entre las variables, independientemente de que respondan las otras variables o no. Este método tiene la dificultad de calcular los grados de libertad debido a que diferentes partes del modelo tienen diferentes muestras. Si se toma como tamaño el de la muestra usando la correlación presente en la mayoría de las observaciones serían un error y se exagera el poder estadístico; y al hacerlo tomando la muestra menor se reduciría el poder.

Para el método de EM, el autor menciona como los valores que faltan se imputan por valores por máxima verosimilitud. Este enfoque se basa en las relaciones observadas entre todas las variables e inyecta un grado de error aleatorio para reflejar la incertidumbre de la imputación. La imputación individual usada en EM es un avance importante respecto a los métodos tradicionales, pero, tiene un defecto inherente. Debido a la omisión de las posibles diferencias entre múltiples imputaciones, la imputación individual tiende a subestimar los errores estándar y por lo tanto sobrestimar el nivel de precisión.

Finalmente, la imputación múltiple permite la puesta en común de las estimaciones de los parámetros mejorados, esto debido a que múltiples imputaciones producen una solución algo diferente para cada imputación. Si estas soluciones fueron muy similares, esto sería evidencia para apoyar la imputación. Si estas soluciones diferían notablemente, entonces es importante incorporar esta incertidumbre en los errores estándar.

2.1.3.1 Imputación Múltiple

El método de imputación múltiple es descrito por Cano Berlanga (2016) como una herramienta estadística que permite el análisis de información truncada basándose en la simulación de Montecarlo y cadenas de Markov. Ésta ha tenido avances a partir de su creación, entre ellas determinar que la distribución t de Student no es consistente, y como alternativa consideran el uso de una mezcla de Normales. Además, Luo, Lawson, He, Elm & Tilley (2016) señalan que el algoritmo de Gibss permite extraer los valores faltantes de $p(Y_{mis}; Y_{obs})$ específicamente en las versiones actuales del parámetro.

La imputación múltiple se puede desarrollar bajo el paradigma bayesiano o frecuentista. Además, Galván (2007) menciona como, según Rubin (1987), la técnica de imputación múltiple tiene la ventaja de brindar buenos resultados aún en presencia de porcentajes de valores perdidos entre 30 y 50.

Para Guerra & Gallestey (2010) la imputación múltiple, según la literatura moderna, es una opción ventajosa para el trabajo en datos faltantes. Estas ventajas las asocian primero con su eficiencia, su versatilidad en relación con las inferencias estadísticas post-imputación, y después, en el caso bayesiano, con el hecho de que incorpora la incertidumbre en el proceso de imputación.

2.1.4. Estadística Bayesiana y Frecuentista en el análisis de valores faltantes

El enfoque frecuentista se basa en la frecuencia de ocurrencia de los eventos. Con relación a este tema Carranza & Kuzniak (2009) mencionan que se ingresa en el tópico de la ley de los grandes números, donde para un cierto tipo de experiencias susceptibles de reproducirse bajo las mismas condiciones (al menos mentalmente), la frecuencia de aparición de un evento dado se estabiliza progresivamente cuando el número de realizaciones crece considerablemente. En este caso, la probabilidad es una característica de la serie infinita, la proporción de apariciones del evento observado. Por el contrario, el enfoque bayesiano se centra en que la probabilidad representa una medida de certeza

sobre una proposición dada, teniendo en cuenta la información disponible. La probabilidad no es una característica del objeto sino la medida de la credibilidad establecida por un sujeto para una proposición dada. Asociada a esta noción se encuentra el teorema de Bayes (Bayes, 1763): las probabilidades atribuidas inicialmente a una proposición evolucionan en función de la llegada de nueva información sobre la proposición en cuestión, la tasa de cambio está dada por la fórmula de Bayes.

Tal y como lo indica Lee (1997), citado en Austin, Naylor y Tu (2001), el paradigma bayesiano permite combinar las creencias anteriores sobre las cuales subyacen los parámetros, con los datos observados para obtener distribuciones de probabilidad de los parámetros. La perspectiva bayesiana considera tanto los datos como los parámetros subyacentes usados para generar los datos como variables aleatorias.

En relación con las debilidades atribuidas al enfoque frecuentista se encuentra la mencionada por Bologna (2012), para quien una de las limitaciones que suele atribuirse es la del valor p en la prueba de hipótesis, pues depende del tamaño muestral. Una diferencia de medias de un valor absoluto dado puede ser significativa si proviene de una muestra de 300 casos, pero no serlo si la muestra solo tiene 30. En este mismo sentido Ayçaguer, Villegas & Fernández (2000) mencionan cómo, en este enfoque, las muestras demasiado pequeñas pueden no llegar a probar nada, y de igual manera las muestras demasiado grandes, esto por cuanto, si el tamaño de muestra es muy grande se sabe de antemano lo que la prueba de hipótesis frecuentista va a producir.

Aunque el enfoque Bayesiano no presenta el problema anterior, Ayçaguer, Villegas & Fernández (2000) si se menciona que su debilidad radica en que tiene más problemas para establecer consenso científico, específicamente indican:

“El bayesianismo se enreda con los precedentes; en cambio, en la aproximación frecuentista, el parroquialismo de utilizar solamente los datos facilita la consecución de consenso. El bayesianismo es vulnerable a intereses porque los intereses se pueden introducir en los análisis a través de los prioris; en cambio, el análisis en la aproximación

frecuentista se reduce a la parroquia de los datos porque la estadística no es únicamente útil para hacer inferencias científicas; la estadística tiene otra gran función: la de estandarizar, regularizar y ayudar a establecer políticas y toma de decisiones (p 487)"

Existe un problema en la susceptibilidad presente en las estimaciones bayesianas a las prioris establecidas, los autores mencionan que "son los pretenciosos los más peligrosos, porque la ignorancia no lo es tanto. Lo que es más terrible es un poquito de conocimiento, ya que le da fuerzas a una persona para llegar a pretender y expresar su pretensión a través de los prioris" (Ayçaguer, Villegas & Fernández, 2000, p 487)

Existe entonces un señalamiento por considerar como lo es la definición de prioris (información previa antes de tomar en cuenta los datos), ya que tener prioris poco informativas o planas por miedo de ser cuestionado sobre las conclusiones a las que se llega genera la pérdida la gran ventaja del enfoque, siendo esta precisamente la de incorporar el existente conocimiento previo.

Little (2006) realiza una comparación entre los paradigmas, donde concluye que el enfoque Frecuentista tiene la ventaja de evitar la necesidad de definir una distribución previa, y hace una clara separación de la función de información anterior en la formulación del modelo y la función de los datos en la estimación de parámetros. No obstante, señala como desventaja que el paradigma es incompleto, ambiguo e incoherente en algunos casos. Incompleto ya que tiene problemas al enfrentarse a muestras pequeñas, se enfoca en las propiedades asintóticas, aunque indica que este problema se ha tratado de corregir por medio de los métodos semiparamétricos; establece que es ambigua debido a que el conjunto de referencia para la determinación de las propiedades de muestreo repetido a menudo es ambiguo, y la teoría frecuentista sufre de diversidad de las estadísticas auxiliares, por último, menciona que es incoherente en el sentido de que viola el principio de verosimilitud, modelos y conjuntos de datos que conducen a la misma función de verosimilitud deben generar las mismas inferencias estadísticas.

En el caso de la estadística bayesiana se plantea el abordaje de muchas de las deficiencias de los planteamientos frecuentistas. Aborda el tema prescriptivo cuando se realiza una determinada elección del modelo y la distribución a priori. Sin embargo, la distribución a priori puede ser difícil de calcular, y se necesitan controles para garantizar que la distribución posterior es correcta, por lo que el paradigma bayesiano presenta la desventaja del reto de establecer el modelo completo en problemas complejos, establecer un modelo mal implica conclusiones erróneas. Además, se dice que esta es coherente en tanto se incorporan las creencias al incluir nueva información, pese a esto, se debe tener cuidado, ya que esto también puede ser una desventaja en el sentido de la confianza de los argumentos respaldados en el marco teórico, y si se tienen prioris no informativas se hace la crítica de que es caer en el paradigma frecuentista (Ver Cuadro 3)

Criterio	Paradigma	
	Frecuentista	Bayesiano
Inferencia bajo modelo asumido	Fuerte	Débil
Formulación del modelo / Evaluación	Débil	Fuerte

Cuadro 3. Fortalezas y debilidades de los Paradigmas Estadísticos, según criterio.

En el caso de la imputación de datos se expone que lo ideal es hacer una combinación de los métodos. Rubin (1996) desarrolló la imputación múltiple, la cual imputa más de un sorteo de la distribución predictiva de los valores, esta imputación combina las reglas para los datos que faltan. (MICR)

MICRs tiene la ventaja de estar basado en principios bayesianos, mientras que las críticas se centran en temas frecuentista sobre la estimación insesgada de la varianza de muestreo. Rubin (1996) señala que en situaciones en las cuales se utiliza la imputación múltiple, en conjuntos de datos de uso público, el modelo de imputación debe tener en

cuenta el hecho de que el usuario puede adoptar un modelo o análisis diferente. Rubin (1996) argumentó que el modelo de imputación debe ser relativamente "débil", en el sentido de incluir en vez de excluir covariables, argumentando que es mejor sacrificar un poco de eficiencia para evitar la imposición de un modelo de imputación fuerte en el usuario de la base de datos.

Viada, Bouza, Ballesteros, Fors, Robaina & Uranga (2014), citando también a Rubin (1983), exponen otros enfoques para el estudio de los datos faltantes, los cuales los clasificó como enfoque basado en la aleatorización, frecuentista, y enfoque basado en el modelo de superpoblaciones, y el enfoque Bayesiano. Entre las ventajas y desventajas entre los diferentes enfoques se destaca que el enfoque aleatorio es más simple de computar y es más robusto, sin embargo, el enfoque Bayesiano obtiene más precisión del intervalo de probabilidad para la variable a imputar.

Finalmente, la gran ventaja del tratamiento de valores ausentes, desde el punto de vista bayesiano, es que, al obtener una muestra de la distribución posterior de los valores ausentes, se puede, o bien hacer una imputación simple (media, mediana) o, si se prefiere, una imputación múltiple (Abellán de Andrés, 2015).

Capítulo III: Abordaje Metodológico

3.1 Fuentes de Información

Los datos provienen del Proyecto de investigación 723-B3-307 “Nuevas formas de medir viejas ideologías: el caso de los sexismos y sus implicaciones en el ámbito académico” de la investigadora principal la Dra. Vanessa Smith Castro.

3.2 Definición de variable(s) de estudio

Las variables son:

- La Prueba Nacional de Bachillerato en Matemática

  Esta prueba mide la competencia matemática conceptualizada como la habilidad para formular, emplear e interpretar las Matemáticas en una variedad de contextos. Incluye los temas de geometría, álgebra, estadística y probabilidades (Mena, 2015).

- Subescalas de Sexismo Benevolente y Sexismo Hostil del Inventario de Sexismo Ambivalente de Glick & Fiske (1996)

  “El inventario está conformado por 22 ítems que miden dos construcciones relacionadas con Sexismo Hostil y Sexismo Benevolente, respectivamente 11 para cada uno de ellos, en el caso de los ítems que corresponden al sexismo benevolente se mide la visión de las mujeres como criaturas delicadas, confinadas a roles limitados.” [Montero et al., 2017, p10]

- Autoeficacia Matemática

  Subescala de Confianza Personal para las Matemáticas compuesta por 11 ítems que son respondidos mediante una escala de Likert de 5 puntos (Montero Rojas, Moreira Mora, Zamora Araya & Smith Castro, 2017).

- Habilidades generales de razonamiento:

  Se utiliza la Prueba de Razonamiento con Figuras (PRF). Esta prueba fue desarrollada por el Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica (UCR), basada en el concepto de "Inteligencia Fluida" (Gf), definido junto con el concepto de "Inteligencia Cristalizada" (Gc) en la Teoría de la Inversión (Investment Theory) de Cattell (1963). (Montero Rojas, Moreira Mora, Zamora Araya & Smith Castro, 2017).

- Sexo del estudiante

  Variable cualitativa con nivel de medición nominal, con categorías de hombre y mujer.

- Edad del estudiante

  Variable de años cumplidos, con nivel de medición de razón.

- Colegio del estudiante

3.3 Evidencias de calidad de la medición para la(s) variable(s) del estudio

Las escalas aplicadas a la población de estudio muestran Alpha de Cronbach aceptables para fines de investigación. Por ejemplo, para la escala de Sexismo Benevolente 0.74 (Montero et al., 2017)

3.2.1. Variables relacionadas con rendimiento académico en matemáticas

Este apartado detalla como las variables seleccionadas se relacionan con el rendimiento en matemáticas.

3.2.1.1 El sexismo hostil y benevolente y su relación con el rendimiento académico en matemáticas

Según Garaigordobil Landazabal & Aliri Lazcano (2011) el sexismo es una forma de actitud discriminatoria dirigida a las personas por su pertenencia a un determinado sexo biológico, la cual está relacionada directamente con violencia hacia las mujeres, al asumir que deben existir diferencia en conductas y características con respecto de los hombres. el sexismo es clasificado en dos tipos el benevolente y el hostil.

El Sexismo Hostil, es un concepto cercano a la concepción tradicional que se tienen del sexismo, Glick y Fiske (1996) lo definen como un conjunto de actitudes y conductas discriminatorias hacia las mujeres basadas en la en la creencia de una supuesta inferioridad de las mujeres como grupo. Y por otro lado definen el Sexismo Benevolente como un conjunto de actitudes relacionadas entre sí hacia las mujeres, estereotipándolas en roles restringidos, aunque son subjetivamente positivas en tono de sentimientos, estás actitudes muestran a la mujer como sujetos agradables, pero, incompetentes para realizar tareas importantes. Por lo anterior es importante señalar que el Sexismo Hostil al ser evidentemente claro produce un fuerte rechazo en la sociedad, mientras que el sexismo benevolente debilita la resistencia de las mujeres ante el patriarcado.

En el caso del sexismo benevolente Fernández, Castro & Lorenzo (2004) indican que son actitudes que enfatizan en los roles tradicionales de la mujer acentuando su debilidad y necesidad de protección de los hombres. Garaigordobil Landazabal & Aliri Lazcano (2011) mencionan que este tipo de sexismo tiene la particularidad de que este tiene apariencias más discretas y ligeras de expresión, las cuales resultan menos

llamativas a aquellas observadas en el sexismo hostil, aunque de igual modo están caracterizadas por un tratamiento desigual y perjudicial hacia las mujeres.

Lo que es ampliado por Kilianski & Rudman (1998), citando a Glick & Fiske (1996) mencionan que el sexismo benevolente radica en tres fuentes: paternalismo protector, la diferenciación de género y la intimidad heterosexual.

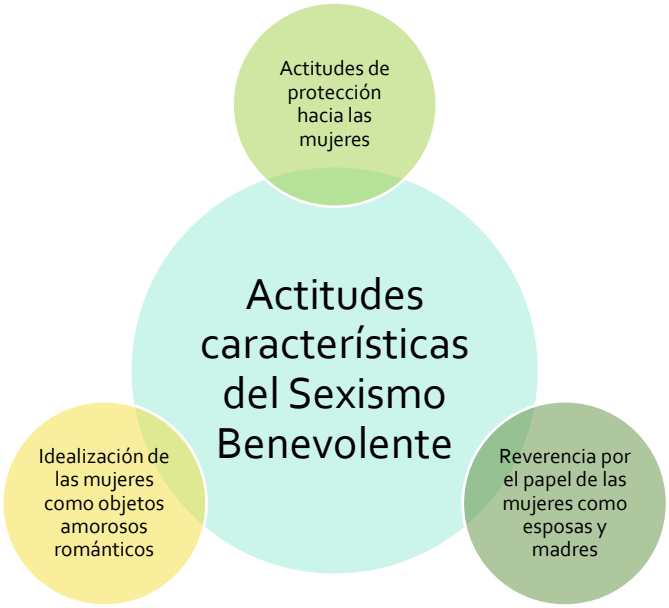


Figura 2. Actitudes características del Sexismo Benevolente Kilianski & Rudman (1998), citando a Glick & Fiske (1996)

El primer término hace referencia a actitudes relacionadas con necesidad de proteger y ayudar a las mujeres al ser individuos con las competencias necesarias, el segundo con acciones dirigidas a marcar diferencias entre hombres y mujeres, pero favoreciendo a las mujeres. Y la tercera con relación a los sentimientos.

En el caso de la diferenciación de género indica que el sexismo benevolente proviene de condiciones de género por aspectos biológicos (fuerza y tamaño). Aspectos relacionados con divisiones de roles basados en género los cuales permiten clasificar

a los individuos por medio de ideologías sociales permitiendo así la diferenciación de estatus.

La intimidad heterosexual es considerada la fuente más intensa de la ambivalencia de los hombres y las mujeres, reside en la forma en la cual llevan sus relaciones románticas, ya que, se visualiza a las mujeres como dos extremos: bondadosas o malvadas, conceptualizándolas como manipuladoras. (Glick & Fiske, 1996)

Estos constructos: sexismo Hostil y el sexismo Benevolente, están siendo estudiados para determinar su relación con el rendimiento académico en matemática. Hyde & Kling (2001), por ejemplo, mencionan la relación de estas variables con el logro en matemática, en donde, la revisión de literatura señala a los hombres como mejores a las mujeres, y que las diferencias radican en el tipo de habilidad evaluada.

Además, se refiere a la existencia de amenazas adicionales en el rendimiento académico por causa del sexismo benevolente, porque se ha identificado en las aulas, según estudios de diferentes investigadores, como se minimizan las capacidades de las mujeres, por ejemplo, dándole ejercicios más sencillos o atribuyendo el fracaso a falta de capacidad.

En esa misma línea de investigación, Brown y Leaper (2010) concluyen de su estudio que la percepción de las adolescentes sobre sexismo en matemáticas, y las ciencias en general, se asocia con una menor competencia percibida y valoración de las matemáticas y la ciencia, aun cuando se controla por sus calificaciones. Además, indica que la diferencia es mayor en niñas latinas, esto bajo la creencia de que las niñas latinas pueden ser ligeramente más susceptibles al sexismo académico en comparación con las niñas americanas o europeas. Este estudio utiliza un cuestionario aplicado a mujeres de entre 13 y 18 años de secundaria y campamentos de verano, y en los estudios se utilizan análisis de variancia.

Oswald & Harvey (2000) en una investigación anterior, ya señalaban, por medio de un análisis de varianza realizado con un diseño experimental a setenta y dos estudiantes de pregrado; que existe entre la amenaza del estereotipo y el medio ambiente hostil un efecto en el rendimiento matemático.

La importancia de identificar como el sexismo está dentro en los ambientes de las niñas llama a reflexionar su relación no solo con los profesores, si no en los hogares, especialmente con la madre, primer modelo. Con relación a este punto, la investigación de Montañés, de Lemus, Bohner, Megías, Moya &García-Retamero (2012) estudia la correlación entre la trasmisión de creencias sexistas benevolentes de madres a hijas y su afectación con el rendimiento académico. Su estudio recolecta información de 192 madres (con edades entre 31 y 57 años) e hijas (con edades entre 11 y 18 años) reclutados de 10 escuelas secundarias españolas diferentes, tanto de zonas rurales como ciudades.

Para el análisis de los datos utilizan análisis de rutas, de los cuales se concluye que el sexismo hostil en las madres correlaciona positivamente con el sexismo benevolente en sus hijas y negativamente con el número de materias aprobadas por los adolescentes. Destacan la posibilidad de que las madres menos sexistas aprecien el rendimiento académico de sus hijas como una vía para alcanzar una mayor independencia en la edad adulta.

3.2.1.2. Otras variables asociadas con el rendimiento académico en matemáticas

Existen variables que históricamente se han asociado al rendimiento en matemática, propias de los estudiantes, por ejemplo, la edad, la cual tiene un efecto negativo sobre el rendimiento en matemática, esto es apoyado por Backhoff Escudero, Sánchez Moguel, Peón Zapata & Andrade Muñoz (2010) quienes indican en su estudio que los estudiantes con una edad diferente a la normativa (edad correspondiente al grado cursado) tienen menor rendimiento en matemáticas que aquellos con edad normativa.

Además, el sexo donde estudios como el de Cervini (2002) señalan que los hombres tienen en promedio un rendimiento en matemática mejor que las mujeres; también, la habilidad de razonamiento que según Cerda, Ortega, Pérez, Flores & Melipillán (2011) concluyen en su estudio, los estudiantes con puntajes en la prueba de inteligencia lógica altos (test con ítems de tipo figurativo, incluyendo formas geométricas abstractas como puntos, líneas rectas o curvas, polígono y otros) tienen un mayor promedio en matemáticas.

Otra variable que se ha asociado al rendimiento en matemática es la autoeficiencia, con respecto al concepto de la autoeficiencia, Bandura (1986), citado por Contreras, Espinosa, Esguerra, Haikal, Polanía & Rodríguez (2005) define esta variable como la capacidad percibida que tienen las personas para hacer frente a situaciones específicas, en la cual se incluye el reconocimiento acerca de las propias capacidades para alcanzar los resultados y enfrentar desafíos.

Contreras, Espinosa, Esguerra, Haikal, Polanía & Rodríguez (2005) concluyen en su estudio que la autoeficacia resulta ser el mejor predictor para pronosticar el rendimiento académico, en especial cuando se analizan las áreas artísticas, sociales y matemáticas.

Pérez, Edgardo, & Cupani, Marcos, & Ayllón, Silvia (2005) concuerdan con lo anterior, ellos utilizan un modelo de regresión con el cual concluyen que las escalas de autoeficacia utilizadas en su modelo contribuyen positivamente a incrementar la explicación de Rendimiento Académico, especialmente en el área de matemática.

Sin embargo, es importante señalar que se deben incluir en los análisis, variables que contengan el contexto en el cual están inmersos los estudiantes, en este sentido Moreira Mora (2009) menciona la importancia de considerar el factor institucional como componente en el análisis del rendimiento en matemática, esto por cuanto ella encuentra diferencias en rendimiento asociadas a variables de la institución.

3.3. Modelos y técnicas estadísticas para el análisis

La técnica estadística utilizada para dar una interpretación al análisis sustantivo de la relación entre las variables el rendimiento académico y las dependientes es el análisis multinivel. Con este tipo de análisis se pretende abarcar el problema desde una perspectiva integral, porque se considera tanto al individuo como el contexto en el cual está inmerso. Llanos & Salas (2007) exponen que los modelos multinivel tienen la ventaja de eliminar las barreras producidas por modelos de un único nivel cuando las estructuras son jerárquicas, por ejemplo, la correlación entre los individuos en las estimaciones de los mínimos cuadrados ordinarios y las significaciones espurias de las covariables. Además, permite realizar una interpretación correcta de los resultados al no caer en falacia ecológica y atomística.

La especificación del modelo sería la siguiente. Dada una muestra de n estudiantes, se tiene información para cada alumno i perteneciente al colegio j sobre la Prueba Nacional de Bachillerato en Matemática como variable a explicar (yi j), y una serie de k variables explicativas, xij=(x1,ij , x2,ij ,…, xk,ij)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij} + \cdots + \beta_k x_{ij} + u_j + e_{ij}$$

En el caso del análisis de valores perdidos se utilizará los enfoques bayesiano y frecuentista en la técnica de Imputación múltiple. La imputación múltiple (MI) se puede definir como un proceso en el cual se sustituyen los datos faltantes con predicciones basadas en los datos observados. Al respecto Rubin (2004) menciona que este método posee las ventajas de que las imputaciones son elegidas aleatoriamente en el intento de representar la distribución de datos, lo cual produce un incremento en la eficiencia de la estimación.

Merkle (2011) indica que, como alternativa a los tradicionales métodos MI, los investigadores pueden manejar los datos faltantes y la estimación del modelo con pasos simultáneos, por medio de la imputación de forma secuencial con los datos que faltan y el muestreo de un modelo bayesiano-completo de datos a través de la cadena de Markov Monte Carlo (MCMC).

Las cadenas de Markov de Monte Carlos son, según Jackman (2000), uno de los desarrollos más importantes de la estadística en los últimos diez años, pues permite aplicar técnicas que en el pasado tenían muchas dificultades para su estimación.

Para evaluar la capacidad de recuperar los datos faltantes en los dos paradigmas con los métodos seleccionados se utilizará la simulación estadística, la cual consiste en obtener matrices de datos incompletas con diferentes técnicas de análisis para estimar un modelo posterior.

H. Masiel & G. Gugnoli, citado en Bú (1993), definen la simulación estadística como: “Técnica numérica para realizar experimentos en una computadora digital. Estos experimentos involucran cierto tipo de modelos matemáticos y lógicos que describen el comportamiento de sistemas de negocio, sociales, biológicos, físicos o químicos (p. 11)”

Asimismo, explica que una de ventajas de hacer una simulación es estudiar el efecto de los cambios en los sistemas al hacer cambios en el mismo, esto permite proponer cuál de

los escenarios admite un mejor resultado. Pero, esta técnica tiene el inconveniente de que requiere equipo computacional y tiempo.

Piera (2004) también indica: “El conocimiento que se obtiene en el desarrollo de un modelo de simulación es de gran interés para poder sugerir posibles mejoras de su rendimiento (p.14)”. Además, menciona como se pude llegar a encontrar respuesta a preguntas sobre cambios en las técnicas sin correr el riesgo de hacerlo en la realidad. Por lo que la simulación permite determinar cuáles herramientas estadísticas se pueden utilizar, bajo condiciones similares a partir de mejoras encontradas.

En el área de análisis simulación bayesiana uno de los usos más frecuentes es la imputación de datos faltantes, debido a los parámetros del modelo se puede calcular y se almacenan durante las iteraciones de un algoritmo MCMC. (Jackman, 2000)

Para llevar a cabo el proceso de simulación de la manera correcta se debe tener en consideración los pasos descritos por Bú (1993), iniciando por la definición del sistema que toma en cuenta la descripción de las variables involucradas y finalizando por la documentación, la cual permite replicar la simulación en caso de ser necesario y que evidencia el resultado de esta. (Ver Figura 3).

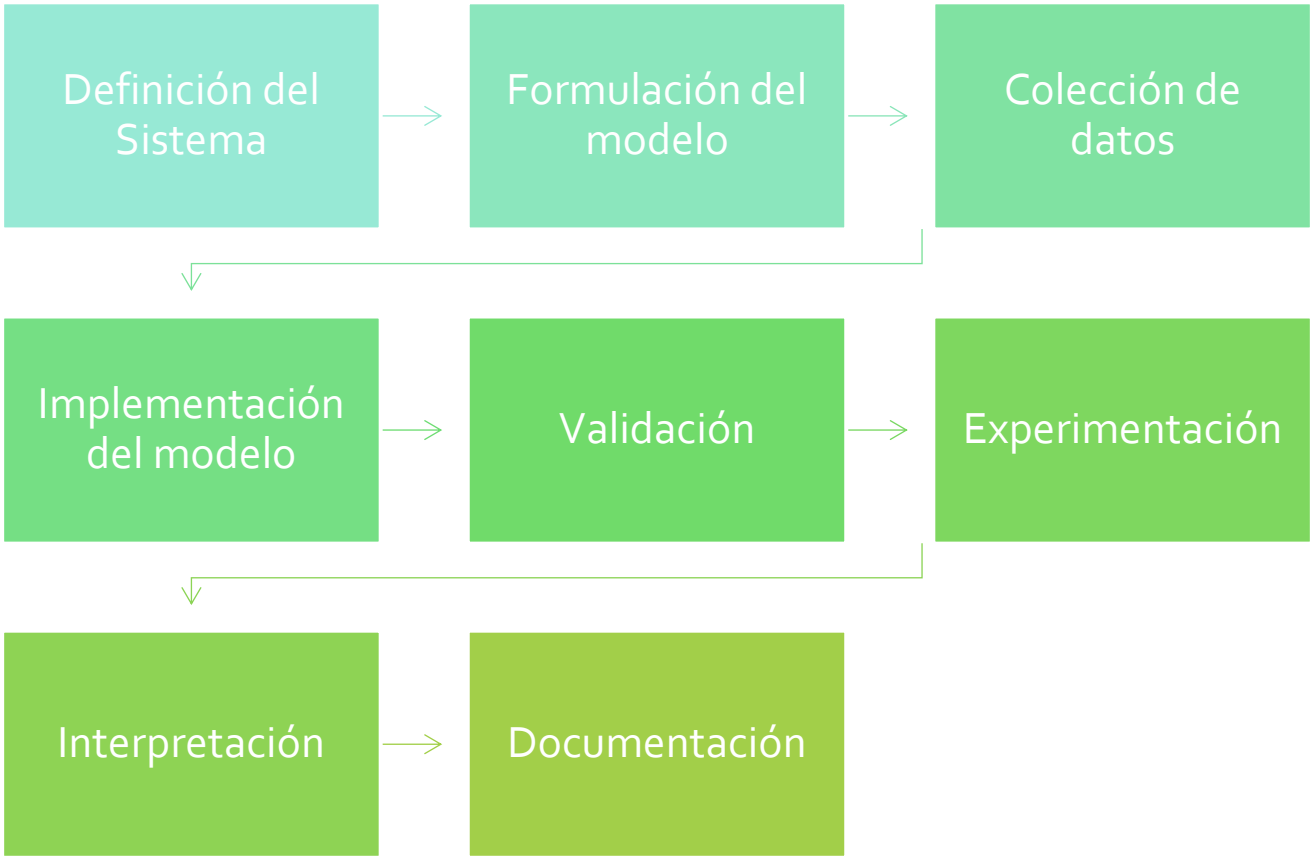


Figura 3. Pasos para el desarrollo del proceso de simulación Bú, R. C. (1993

3.4. Escenarios

Para los escenarios planteados se consideran tres aspectos: el patrón de pérdida de los datos (MCAR o MNAR), el porcentaje de valores perdidos en la variable en análisis (20% o 40 %) y la correlación entre la variable dependiente (variable con valores perdidos) y las variables predictoras (moderada: entre 0.20 menor a 0.55 o alta: mayor o igual 0.55), esto implica tener 8 escenarios diferentes, los cuales se evalúan tanto en el paradigma bayesiano como en el frecuentista, esto basado en la revisión de literatura del capítulo anterior, con lo cual se pude corroborar si la eficiencia del mecanismo no se ve afectado en porcentaje de faltantes altos, así como comprender aspectos relacionados con la violación del supuesto MCAR y la fuerza en la relación de las variables utilizadas en la recuperación en el mecanismo de imputación. Además, se considera el criterio experto en aspectos relacionados con la perdida en el patrón no al azar. (ver Figura 4).

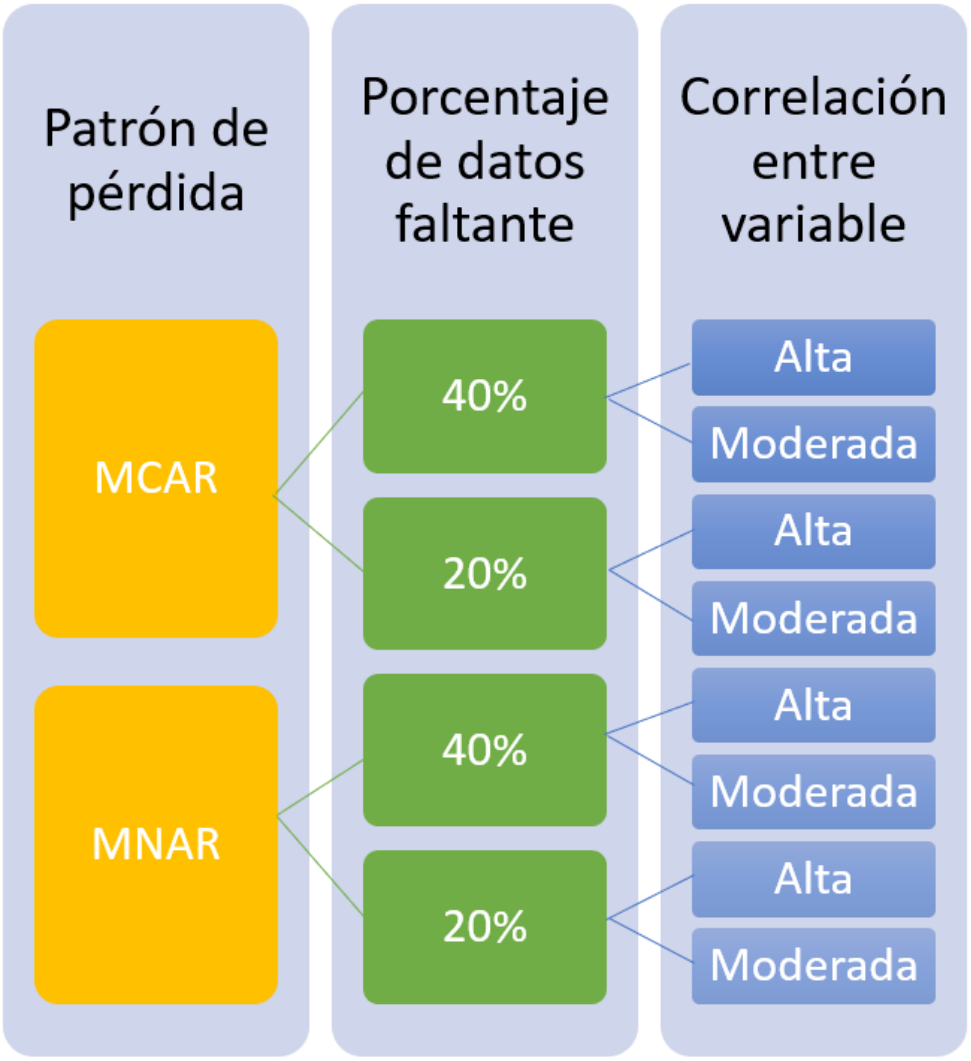


Figura 4. Variables consideradas en la construcción de los escenarios a simular

3.5. Software estadístico y herramientas informáticas

- STATA

Para la estimación de los modelos multinivel se utilizó el paquete estadístico STATA, este tiene la ventaja de ser muy familiar para los usuarios de análisis de datos, y, en el caso específico de la estimación de Modelos Multinivel, permitir el tratamiento de los datos sin necesidad de recurrir a softwares adicionales, estos modelos pueden utilizarse a partir de la versión 10. (Martínez-Garrido & Murillo, 2014).

- *R*

El lenguaje R se utiliza para la generación del código de las simulaciones, este lenguaje presenta la ventaja de permitir modificar por medio de comandos la cantidad de iteraciones y la tolerancia a la convergencia. Además, es software libre para la comunidad científica, permite usar, construir rutinas y modificar las existes. Odeh, Featherstone & Bergtold (2010).

Se utiliza en la imputación bayesiana los paquetes coda, mass y MCMCpack.

- Clúster Kabre

Para ejecutar las simulaciones de manera más ágil se utiliza el clúster Kabre, este es un supercomputador del Centro Nacional de Alta Tecnología. En general los clústeres son grupos de computadoras unidas en red, que por sus características permiten ejecutar rutinas de código en menor tiempo dada su capacidad de procesamiento.

Capítulo IV: Resultados

Este apartado detalla los principales hallazgos obtenidos a partir de la aplicación de la metodología implementada.

4.1. Estadísticos descriptivos del caso real

Los datos provienen del Proyecto de investigación 723-B3-307, durante la recolección de la información de los estudiantes al 21% no se le registra la nota que obtuvieron en la Prueba Nacional de Matemática de Bachillerato.

Al verificar cuales variables están relacionadas con la falta del dato basado en juicio experto, se observa que cuando la edad aumenta y también cuando aumenta el puntaje de razonamiento la pérdida es mayor (ver Figura 5)

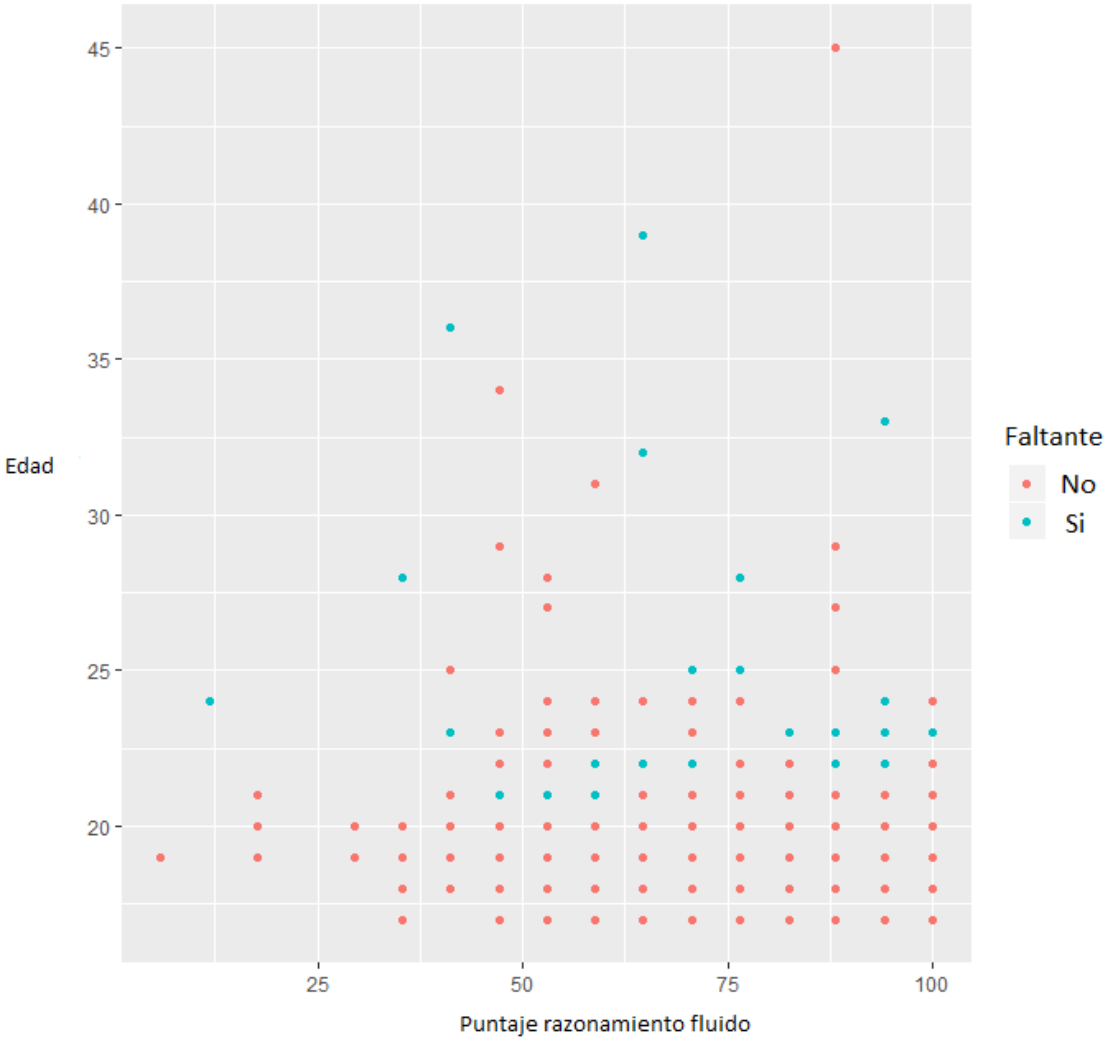


Figura 5. Edad y la prueba de razonamiento con figuras, con la falta de información en la nota de Matemática de Bachillerato, proyecto de investigación 723-B3-307, año 2015.

Además, se observa una asociación entre la variable sexo con tener faltante el dato de la nota de Matemática de Bachillerato, esto al examinar las probabilidades condicionales (Ver Cuadro 4)

- P(F/H)= 106/903 =0.12
- P(F/M)=85/903= 0.09

Sexo	Faltante en nota Matemática Bachillerato		Total
	No	Sí	
Hombre	347	106	453
Mujer	365	85	450
Total	712	191	903

Nota: Valor p = 0.09, eta de 0.55

Cuadro 4. Relación entre sexo del estudiante y tener faltante la nota de Matemática de Bachillerato, Proyecto de investigación 723-B3-307, año 2015.

Considerando lo anterior se construye un modelo de regresión logística para la predicción de si se tiene el faltante del dato de la nota de Matemática de Bachillerato, en el cual se evidencia que todas las variables son relevantes en el modelo, con mayor fuerza el puntaje de razonamiento con figuras y la edad. De lo anterior se deduce que el escenario real de pérdida es aquel donde esta no fue al azar y que existen variables asociadas.

Variable	Coeficientes no estandarizados		Coeficientes estandarizados	t	Valor p
	B	Error típ.	Beta		
(Constante)	91,005	5,829		15,612	<0,001
Puntaje en PRF	0,284	0,030	0,332	9,551	<0,001
Sexo	1,767	0,947	0,064	1,865	0,063
Edad	-1,724	0,244	-0,242	-7,063	<0,001

Nota: R^2 de 0,209, se cumplen los supuestos del modelo

Cuadro 5. Resultados del ajuste de la regresión logística para la predicción de faltantes en nota en Matemática de Bachillerato

De todos los registros presentes en el archivo de datos de 906 estudiantes del proyecto, solo se tiene información completa en todas las variables para 487 estudiantes. Dado que se quiere conocer la efectividad en la recuperación de la nota de matemática se utiliza este archivo para el análisis, y se verifican cuáles de las variables del archivo completo están asociadas desde la teoría con la nota de bachillerato en matemática. El análisis deja en evidencia que la edad, el porcentaje de respuestas correctas y la autoeficacia en matemáticas tienen correlaciones significativas con la nota en matemática y dentro de los parámetros establecidos en los escenarios como correlaciones moderadas, por lo cual se utilizarán dentro del mecanismo de imputación múltiple en un modelo de regresión, las otras variables quedan excluidas debido a que, si bien tienen relevancia teórica no cumplen con los parámetros establecidos en los escenarios. Si embargo, se consideran en el caso aplicado. (ver Cuadro 6)

Variable	Nota en Matemática de Bachillerato	Edad	Porcentaje de correctas en la Prueba de Razonamiento con figuras	Autoeficacia en Matemática	Puntaje en la escala de Sexismo Benevolente	Puntaje en la escala de Sexismo Hostil
Nota en Matemática de Bachillerato	1	-.228**	.403**	.367**	-.138**	-0,039
Edad	-.228**	1	-.091*	0,028	.113*	0,067
Porcentaje de respuestas correctas en la Prueba de Razonamiento con figuras	.403**	-.091*	1	.249**	-0,059	.101*
Autoeficacia en Matemática	.367**	0,028	.249**	1	0,041	.097*
Puntaje en la escala de Sexismo Benevolente	-.138**	.113*	-0,059	0,041	1	.412**
Puntaje en la escala de Sexismo Hostil	-0,039	0,067	.101*	.097*	.412**	1

Nota: **. La correlación es significativa en el nivel 0,01 (bilateral).
*. La correlación es significativa en el nivel 0,05 (bilateral).

Cuadro 6. Coeficientes de correlación de Pearson entre las variables que teóricamente están relacionadas con la nota de Matemática de Bachillerato.

4.2. Selección del mecanismo a utilizar en la imputación múltiple

El proceso de imputación múltiple requiere una definición del mecanismo de simulación a ser utilizado para producir las diferentes estimaciones de los valores faltantes, esto incluye tanto la cantidad de datos con la que se cuenta dentro de cada una de las simulaciones (matriz de datos), como la también la cantidad de simulaciones utilizadas.

Donde denota S_1 es la primera simulación y S_n la última, y para cada una de ellas se tiene la estimación del e_i, que denota el sesgo.

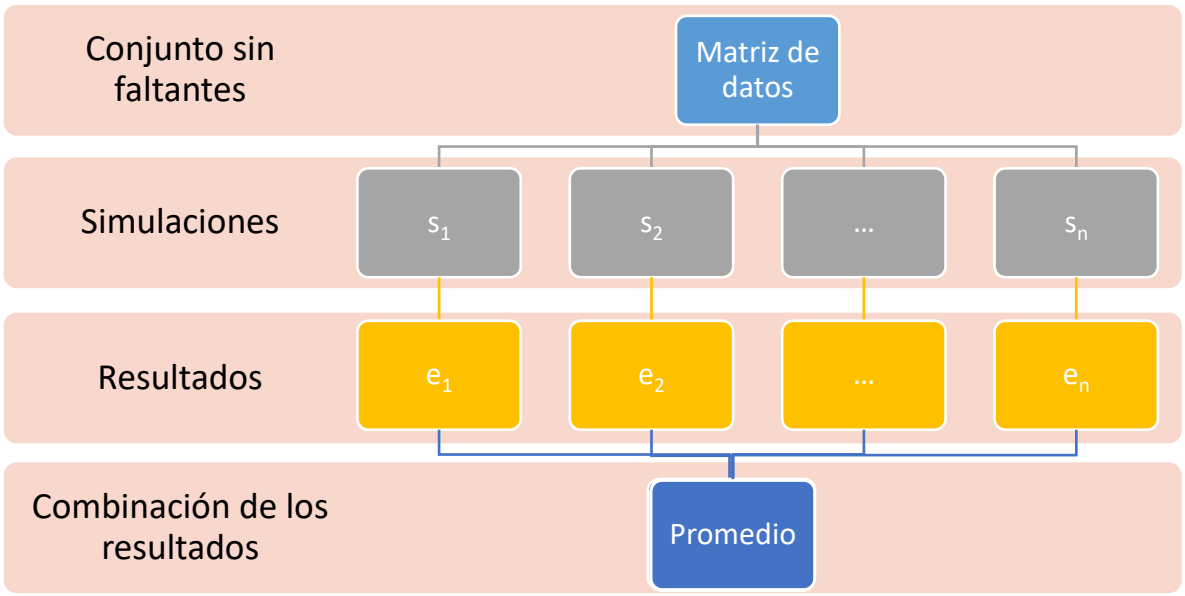


Figura 6. Revisión del mecanismo de imputación múltiple para el proceso de simulación en los escenarios

Para una primera etapa considerando el conjunto de datos reales, se analiza la cantidad de datos de la matriz completa usada para generar cada una de las simulaciones, y, posteriormente, el número de simulaciones posibles para combinar los resultados y obtener una única estimación de cada valor faltante, en distintos los escenarios de pérdida establecidos: 1- faltantes no al azar (MNAR) con un porcentaje de pérdida moderado, 2- faltante no al azar (MNAR) con un porcentaje de alto, 3- faltantes al azar (MAR) con un porcentaje de pérdida moderado, 4- faltantes al azar (MAR) con un porcentaje de pérdida moderado.

Es importante recordar que, si existen un gran número de valores faltantes en el archivo de datos, la matriz de datos completos utilizada para generar cada una de las simulaciones será limitada. Esto quiere decir, por ejemplo, que si se tiene un conjunto de datos original con 100 registros en los cuales para una variable se ha perdido el 40% de los datos, y se quiere recuperar la información por imputación múltiple, solo se tendrían 60 datos para seleccionar los m conjuntos de simulaciones. Entonces, si se decide, por ejemplo, trabajar con un mecanismo en el cual se toman solo un 50% de esos datos completos como la muestra s_i de la simulación, se tendrían finalmente solo 30 datos para generar por medio de una regresión, u otra técnica, el valor de los datos que faltan, y el número puede ser menor si la cantidad de datos completos fuera más pequeña. Es por ello que se consideran para el análisis solo aquellos mecanismos en los cuales haya como mínimo una selección del conjunto de datos en donde se contemple entre 50% y 90% de los casos disponibles, estos son seleccionados por medio de una muestra simple al azar sin remplazo.

En esta evaluación los resultados reflejan que, independientemente del porcentaje de datos utilizados en cada una de las m simulaciones, el resultado del error absoluto promedio es igual, por lo tanto, basado en esto, se define como mecanismo el seleccionar una muestra aleatoria del 60% de los datos completos, lo cual permite tener un mayor porcentaje de casos de los que se tendrían con 50%, esto minimiza los problemas de inferencia presentes en muestras pequeñas y, además, se minimizan los tiempos computacionales. (ver Cuadro 7)

Porcentaje	Estadístico	Escenario			
		% moderado de pérdida-MNAR	% alto de pérdida-MNAR	% moderado de pérdida-MCAR	% alto de pérdida-MCAR
50%	Promedio	9,43	9,50	9,51	9,54
	Varianza	0,36	0,15	0,41	0,15
60%	Promedio	9,41	9,51	9,51	9,53
	Varianza	0,35	0,14	0,41	0,15
70%	Promedio	9,43	9,54	9,54	9,55
	Varianza	0,37	0,13	0,41	0,15
80%	Promedio	9,45	9,53	9,50	9,54
	Varianza	0,39	0,14	0,39	0,15
90%	Promedio	9,44	9,50	9,50	9,57
	Varianza	0,40	0,14	0,40	0,14

Cuadro 7. Estadístico de error absoluto promedio y la varianza de las imputaciones por escenario, según el porcentaje de datos seleccionado de la matriz completa.

En el caso de la cantidad de simulaciones que se debe de utilizar para realizar la combinación de la información, con el fin de construir un único valor estimado, Rubín (1987), citado por Medina y Galván (2007), señala como el método de imputación múltiple es capaz de generar resultados robustos con pocas iteraciones, pero recomienda entre 5 y 10, cuando las tasas son inusualmente altas (tasas mayores a 20%).

Además, Badler, Alsina, Puigsubirá y Vitelleschi (2002) realizan una prueba para verificar la eficiencia en la estimación al usar diferentes cantidades de simulaciones (entre 3 a 20), en el método de imputación múltiple con el paquete SAS. Basados en los resultados obtenidos llegan a la conclusión de que no se justificaría trabajar con un número de conjuntos imputados mayores a 10, esto por cuanto el incremento relativo de la variancia es bajo. Lo anterior concuerda con la cantidad utilizada por defecto SPSS, donde el número de simulaciones por defecto es de 10.

Al realizar las pruebas con los datos reales en los escenarios establecidos, se concluye que, en general, el error promedio absoluto es muy similar entre las diferentes cantidades de simulaciones. Por lo que, considerando el promedio y la variancia en el estadístico de error, y las recomendaciones teóricas, se selecciona como mecanismo de imputación 10 simulaciones (Ver Cuadro 8).

Número de simulaciones	Escenario							
	% moderado de pérdida- MNAR		% moderado de pérdida- MCAR		% alto de pérdida- MNAR		% alto de pérdida- MCAR	
	Promedio	Varianza	Promedio	Varianza	Promedio	Varianza	Promedio	Varianza
5	9,44	0,38	9,52	0,37	9,50	0,14	9,52	0,37
10	9,47	0,39	9,53	0,41	9,52	0,15	9,53	0,41
15	9,47	0,39	9,54	0,38	9,52	0,15	9,54	0,38
20	9,45	0,38	9,52	0,40	9,53	0,14	9,52	0,40
25	9,45	0,38	9,52	0,39	9,53	0,15	9,52	0,39
30	9,47	0,37	9,53	0,36	9,51	0,14	9,53	0,36
35	9,45	0,39	9,52	0,38	9,52	0,14	9,52	0,38
40	9,42	0,44	9,53	0,40	9,52	0,13	9,53	0,40
45	9,42	0,39	9,48	0,42	9,51	0,14	9,48	0,42
50	9,43	0,41	9,48	0,42	9,51	0,14	9,48	0,42

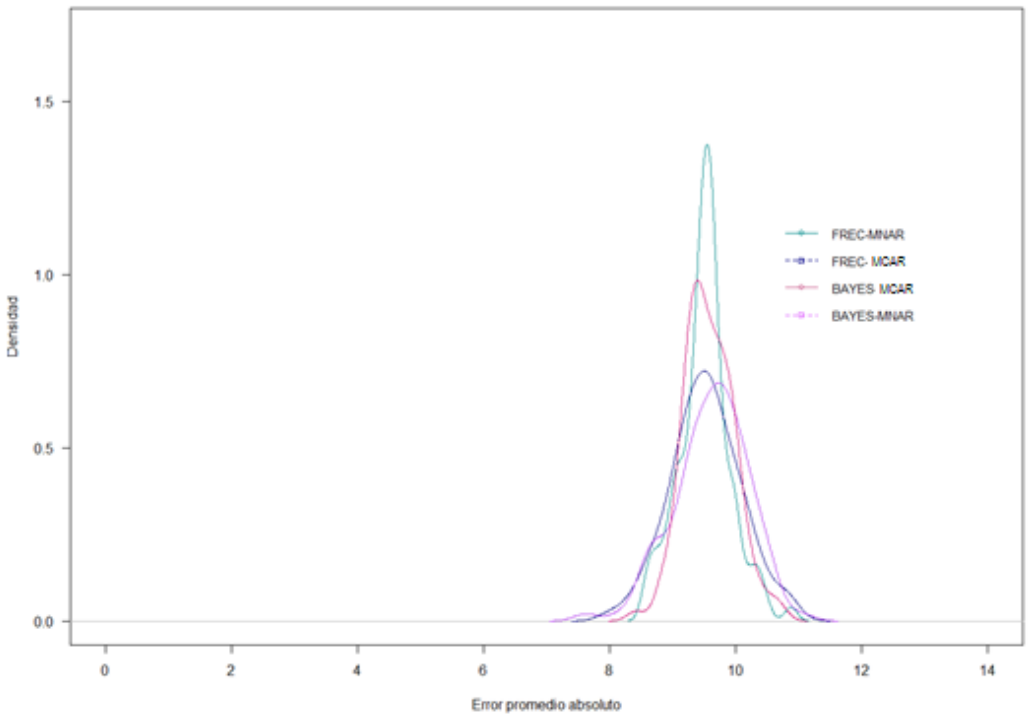
Cuadro 8. Estadísticos de error absoluto promedio y la varianza de las imputaciones por escenario, según el número de simulaciones utilizada en la imputación múltiple.

4.3. Eficiencia de la Imputación múltiple en los escenarios

A continuación, se detalla la eficiencia de los mecanismos de recuperación de datos bajo los paradigmas frecuentistas y bayesianos, en los escenarios establecidos.

4.3.1. Comparación de los escenarios bajo una correlación moderada (usando datos reales)

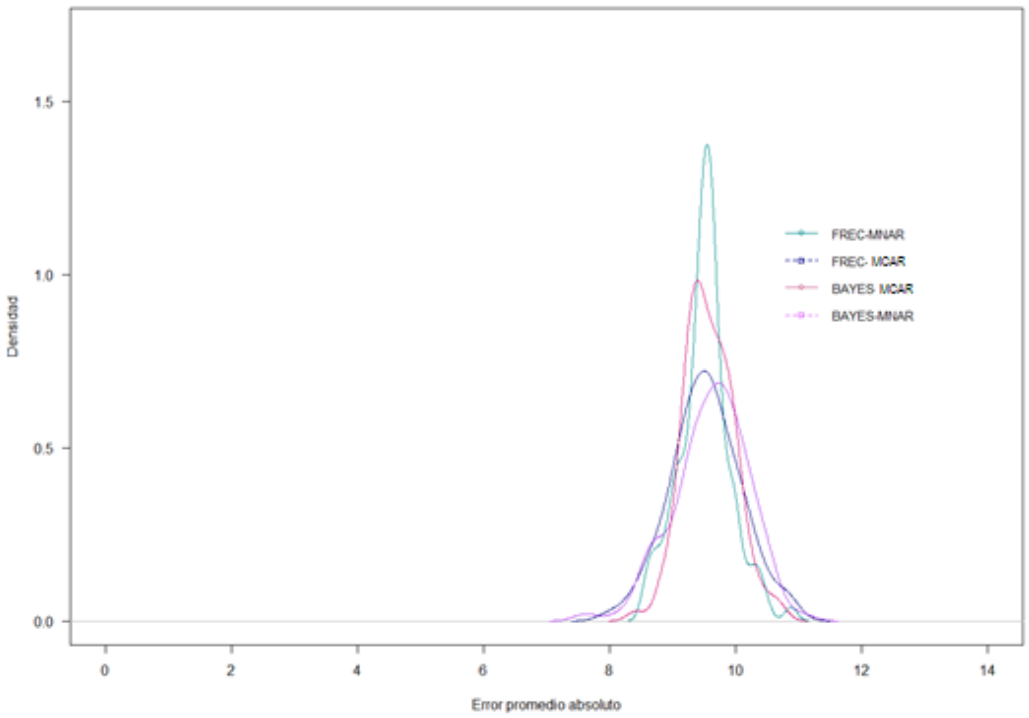
Al analizar el comportamiento del sesgo (error absoluto promedio) en el caso en que se tiene un escenario con una correlación moderada entre las variables dependientes e independiente (utilizando los datos reales), simulando una pérdida alta de datos (40 %) es posible apreciar un patrón de pérdida al alzar, bajo el paradigma frecuentista se observa una distribución centrada en un promedio un poco más a la izquierda (valores bajos), diferente al apreciado en el caso bayesiano. Al contrario, cuando se tienen patrones de faltantes que no fueron perdidos al azar las distribuciones son similares, aunque más consistentes en el caso frecuentista (ver Figura 7).



Nota: definiciones Frec frecuentista, MCAR patrón de perdida completamente al azar, MNAR patrón de perdida no al azar

Figura 7. Efectividad en la recuperación de los datos faltantes en el escenario de pérdida del 40%, con correlación entre las variables moderada, por paradigma y patrón de pérdida.

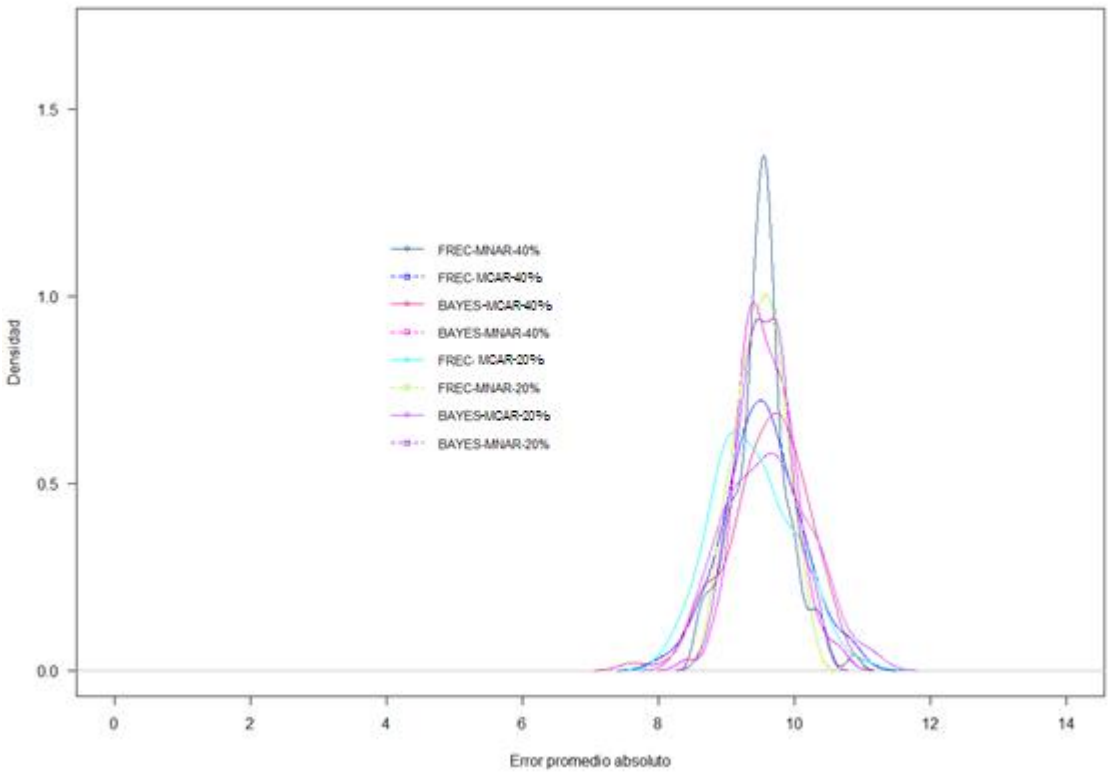
En el caso en que se tiene un escenario con una correlación moderada entre las variables dependientes e independiente (utilizando el escenario real), simulando una pérdida del 20 % se observa que, igual al caso anterior, para un patrón de pérdida “al alzar”, bajo el paradigma frecuentista, se observa una distribución centrada en un promedio un poco más a la izquierda (valores bajos), que el mostrado el caso bayesiano. Por otro lado, cuando se tienen patrones de faltantes que no fueron perdidos al azar no se observan diferencias (ver Figura 8).



Nota: definiciones Frec frecuentista, MCAR patrón de perdida completamente al azar, MNAR patrón de perdida no al azar

Figura 8. Efectividad en la recuperación de los datos faltantes en el escenario de pérdida del 20%, con correlación entre las variables moderada, por paradigma y patrón de pérdida

Cuando se comparan todos los escenarios, manteniendo constante correlación moderada entre las variables dependientes e independiente (utilizando el escenario real), se destaca una mejor recuperación en el caso frecuentista cuando hay presencia de valores perdidos del 20%, esto cuando se observa el comportamiento de los estadísticos. Sin embargo, para todos los casos la media de del error absoluto promedio es alta, está entre 7 a 12 puntos con respecto a la nota real (ver Figura 9 y Cuadro 9).



Nota: definiciones Frec frecuentista, MCAR patrón de perdida completamente al azar, MNAR patrón de perdida no al azar

Figura 9. Efectividad en la recuperación de los datos faltantes en los escenarios de correlación entre variables moderada entre variables con distintos porcentajes de pérdidas por paradigma

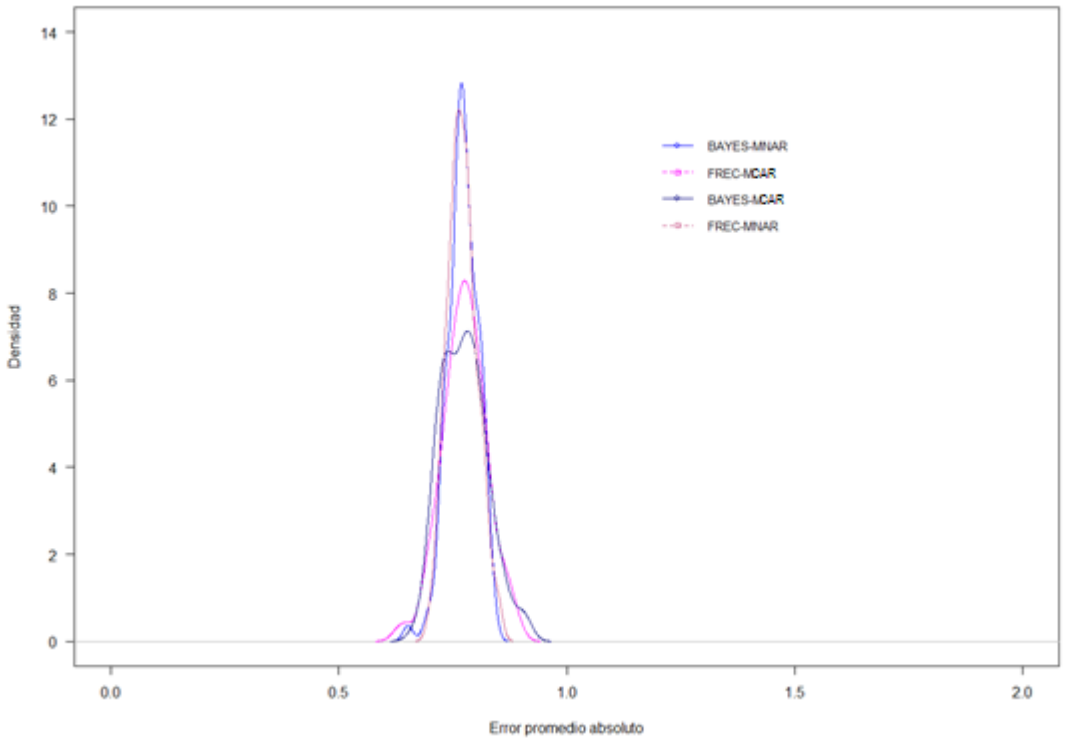
Escenario	Mínimo	Mediana	Promedio	Máximo	Varianza
BAYES-MCAR-20%	8,38	9,57	9,61	11,14	0,38
FREC-MCAR-20%	8,10	9,35	9,39	11,03	0,13
BAYES-MNAR-20%	8,64	9,56	9,55	10,40	0,14
FREC-MNAR-20%	8,64	9,50	9,49	10,25	0,35
BAYES-MCAR-40%	7,62	9,67	9,61	11,05	0,35
FREC-MCAR-40%	7,99	9,50	9,52	10,90	0,31
BAYES-MNAR-40%	8,41	9,53	9,58	10,70	0,16
FREC-MNAR-40%	8,57	9,51	9,51	10,88	0,17

Cuadro 9. Estadístico de error absoluto promedio por escenario bajo el caso de una correlación entre variables moderada

4.3.2. Comparación de los escenarios bajo una correlación alta (usando datos simulados)

Para comparar el efecto que tiene sobre el sesgo variar el un conjunto de datos, por uno donde la correlación entre los predictores y la variable que tiene faltantes sea alta, se simula este escenario con correlaciones entre Nota en Matemática de Bachillerato y los predictores mayores a |0.55|: Autoeficacia en Matemática (0.62), porcentaje de respuestas correctas en la Prueba de Razonamiento con figura (0.69) y edad (-0.58).

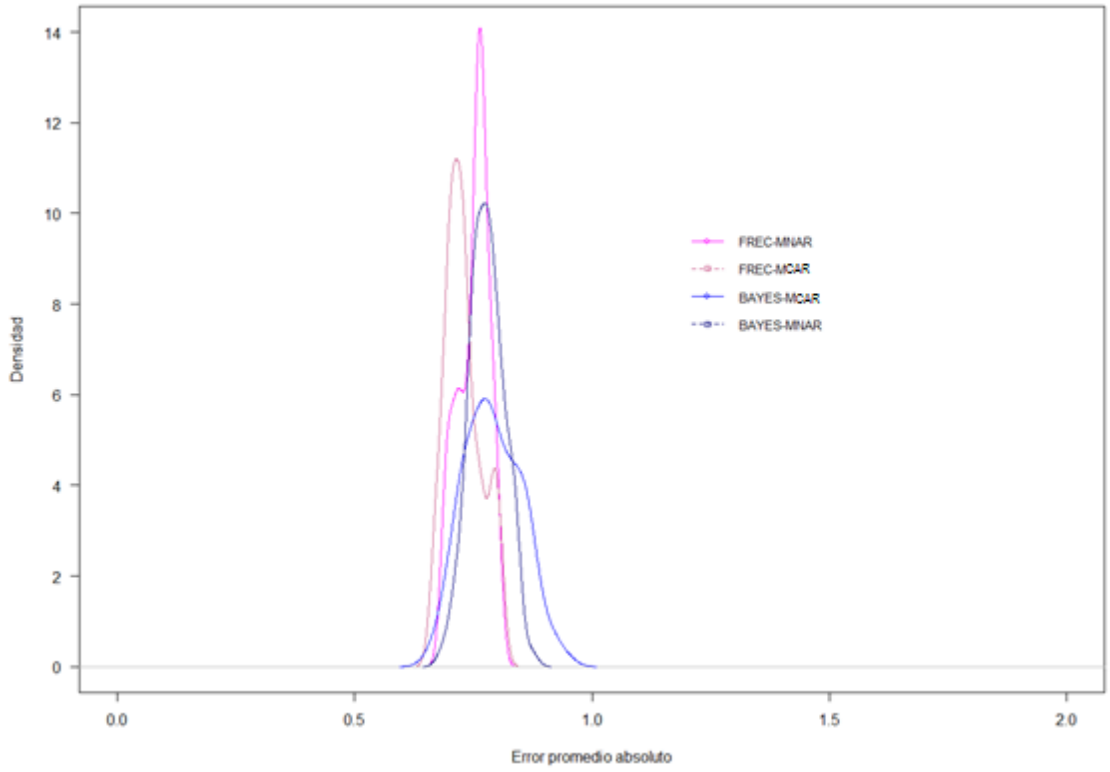
En estos escenarios, cuando se mantiene constante una pérdida del 40% de los datos, se observa como el promedio es muy similar, y que las distribuciones de los errores promedios entre simulaciones varían en la curtosis, ya que la concentración de los datos es mayor cuando se tienen datos faltantes que no fueron perdidos al azar independientemente del paradigma, que cuando se tienen faltantes bajo un caso al azar (ver Figura 10).



Nota: definiciones Frec frecuentista, MCAR patrón de perdida completamente al azar, MNAR patrón de perdida no al azar

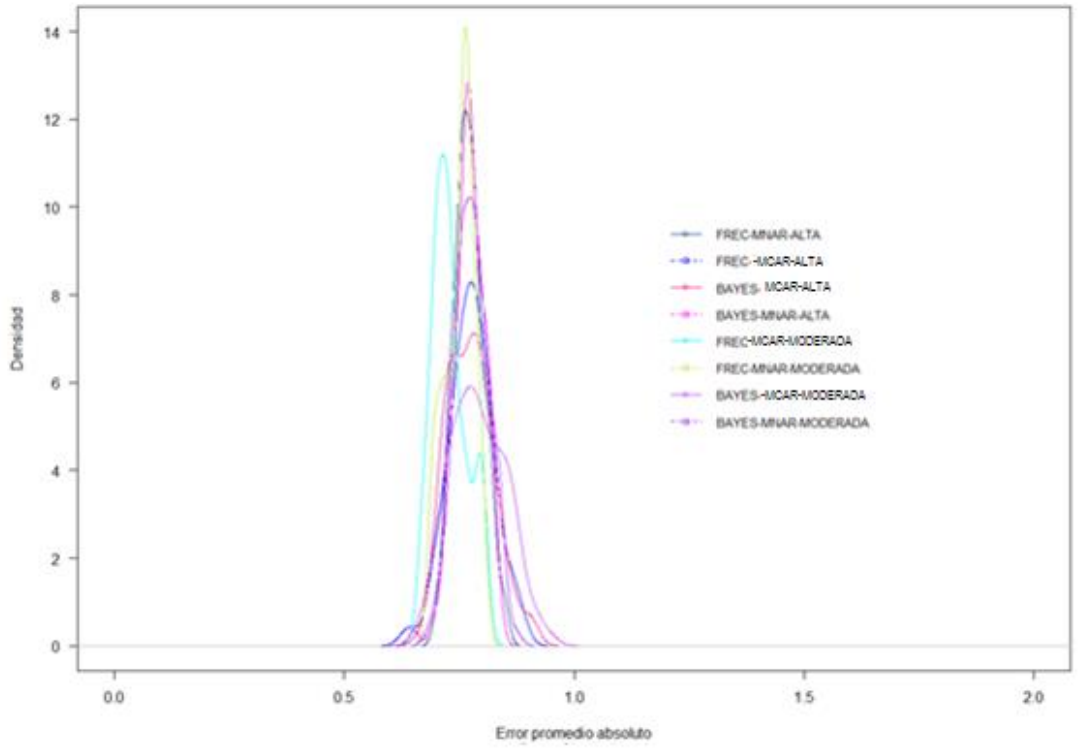
Figura 10. Efectividad en la recuperación de los datos faltantes en el escenario de pérdida del 40% con correlación entre las variables alta, por paradigma y patrón de pérdida

Al contrario del caso anterior, cuando se está en un escenario donde se mantiene constante un porcentaje de faltantes del 20% de los datos, se observan más diferencias entre las distribuciones del error promedio entre simulación, así, por ejemplo, en el caso de faltantes en un escenario al azar, el paradigma frecuentista los recupera con un error promedio menor que en el caso bayesiano; mientras que, bajo un escenario con faltantes que no fueron al azar, las distribuciones son más similares (ver Figura 11).



Nota: definiciones Frec frecuentista, MCAR patrón de perdida completamente al azar, MNAR patrón de perdida no al azar

Figura 11. Efectividad en la recuperación de los datos faltantes en el escenario de pérdida del 20% con correlación entre las variables alta, por paradigma y patrón de pérdida



Nota: definiciones Frec frecuentista, MCAR patrón de pérdida al azar, MNAR patrón de pérdida no al azar

Figura 12. Comparación de la efectividad en la recuperación de los datos faltantes en los escenarios de correlación entre variables alta entre variables con distintos porcentajes de pérdidas por paradigma

Escenario	Mínimo	Mediana	Promedio	Máximo	Varianza
BAYES-MCAR-20%	0,660	0,787	0,790	0,945	0,004
FREC-MCAR-20%	0,671	0,722	0,729	0,803	0,001
BAYES-MNAR-20%	0,684	0,778	0,778	0,873	0,001
FREC-MNAR-20%	0,690	0,761	0,750	0,800	0,001
BAYES-MCAR-40%	0,667	0,774	0,774	0,912	0,002
FREC-MCAR-40%	0,632	0,777	0,775	0,888	0,002
BAYES-MNAR-40%	0,651	0,771	0,772	0,837	0,001
FREC-MNAR-40%	0,699	0,768	0,771	0,850	0,001

Cuadro 10. Estadísticos para error absoluto promedio por escenario bajo el caso de una correlación entre variables alta

Cuando se comparan todos los escenarios manteniendo constante correlación alta entre las variables dependientes e independiente (utilizando los datos simulados), se destaca que existe una mejor recuperación en el caso frecuentista cuando hay presencia de valores perdidos del 20% (pérdida moderada), esto cuando se observa el comportamiento de los estadísticos, y es prácticamente igual el error promedio absoluto en todos los casos donde la pérdida fue alta (40%). Además, la recuperación en estos escenarios es mucho mejor, ya que mientras en los casos donde la correlación entre las variables predictoras de los faltantes y la variable con faltantes era moderada se tenía un error promedio de 7 a 12, en estos (correlación alta) es menor a un punto (ver Figura 12 y Cuadro 10).

4.4. La imputación de los datos en el caso real y sus implicaciones en el modelo de regresión

El objetivo general de esta sección es identificar la sensibilidad que tienen los resultados de un modelo de regresión multinivel a partir de las diferencias encontradas al utilizar datos completos y datos que fueron imputados.

Para ello se trabaja con la muestra del del Proyecto de investigación 723-B3-307, misma que se utiliza para el análisis de la recuperación de valores faltantes, la muestra está compuesta por 487 estudiantes que en 2015 cursaban el undécimo año en 10 colegios públicos diurnos del país (Ver cuadro 11).

Colegio	Absoluto	%
A	86	17,7
B	77	15,8
C	66	13,6
D	60	12,3
E	55	11,3
F	44	9,0
G	42	8,6
H	21	4,3
I	19	3,9
J	17	3,5
Total	487	100,0

Cuadro 11. Distribución de los estudiantes de la muestra por colegio, proyecto de investigación 723-B3-307, año 2015.

Para estos estudiantes se simula una pérdida del 21% de los datos en la variable de Nota en la prueba de Bachillerato de matemática, esto considerando el patrón de pérdida que presentaban los datos faltantes no son al azar. Posteriormente se imputan utilizando el mismo mecanismo recomendado para este contexto, bajo un paradigma frecuentista. Como resultado los datos reflejan una diferencia absoluta promedio de 9,2 puntos entre la nota verdadera del estudiante y la obtenida de la imputación múltiple (Ver Figura 15), promedio similar al obtenido en la elección del método de imputación. Además, al comparar el promedio de la variable con y sin imputación se observa que los promedio son iguales (Ver Figura 13).

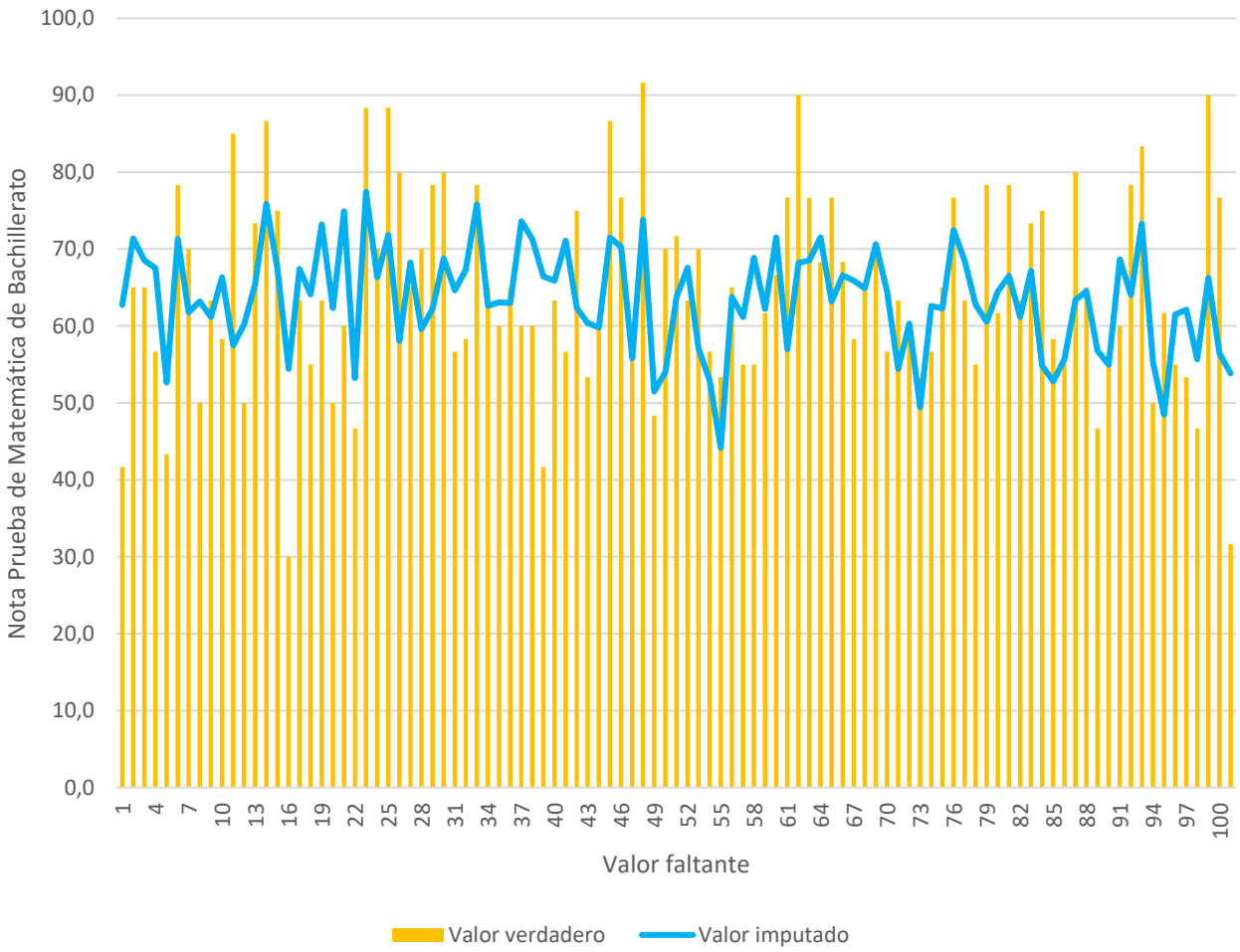
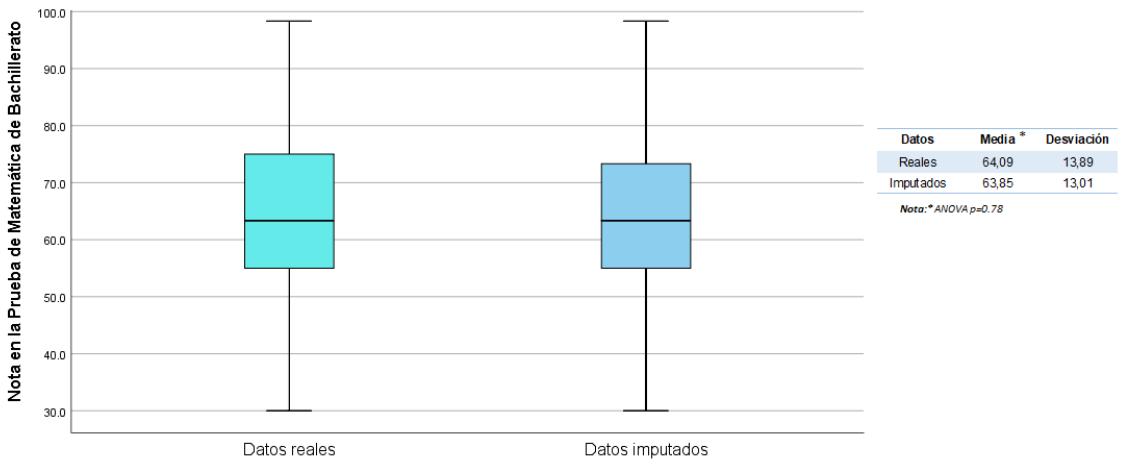


Figura 13. Valores verdaderos e imputados para los estudiantes en la Nota de Bachillerato en Matemática



Datos	Media *	Desviación
Reales	64,09	13,89
Imputados	63,85	13,01

Nota: ANOVA p=0,78*

Figura 14. Distribución de la Nota de Bachillerato en Matemática de los datos reales y de los datos con imputación

Colegio[1]	Nota Prueba de Matemática de Bachillerato [2]	Nota Prueba de Matemática de Bachillerato (imputado)[2]	Edad [2]	Porcentaje de correctas en la Razonamiento	Puntaje Escala de Sexismo Benevolente[2]	Puntaje de la Escala de Sexismo Hostil[2]	Puntaje Escala de Equidad	Autoeficacia en Matemática
G	64,21	65,41	16,45	72,13	2,63	2,65	4,46	3,64
J	64,61	64,50	16,82	67,82	3,15	2,67	4,03	4,05
E	60,91	61,59	16,36	66,42	3,30	3,10	4,31	3,79
H	70,64	68,85	16,43	63,31	3,03	2,81	4,36	3,95
D	56,56	56,73	17,03	65,59	3,21	3,08	4,48	3,94
I	49,65	50,94	17,42	61,30	3,43	3,09	4,69	3,85
A	69,34	67,74	16,78	65,60	2,90	2,73	4,50	3,87
C	67,25	67,21	16,77	68,36	2,80	2,89	4,26	3,98
I	69,28	68,15	16,43	67,51	3,21	2,91	4,72	4,09
J	62,29	62,15	16,56	63,25	3,06	2,79	4,01	3,71
Total	64,09	63,85	16,68	66,24	3,03	2,87	4,37	3,87

Nota:[1/] El nombre del colegio se anonimiza para mantener la confidencialidad requerida.
[2/] diferencias de promedios entre colegios.
Fuente: Proyecto de investigación 723-B3-307

Cuadro 12. Promedios de las variables en análisis por colegio

Una vez realizado el proceso de imputación se procedió a analizar el comportamiento de las variables entre los colegios. Los puntos en común presentes en estos colegios son: están ubicados en el Gran Área Metropolitana, son públicos y diurnos. Los resultados reflejan la presencia de diferencias entre los promedios de las variables por colegio. Por ejemplo, el colegio I obtuvo un promedio menor en la Nota de la Prueba de Bachillerato en Matemática que la mayoría de los colegios (con excepción del D), mismo comportamiento se observa con la variable imputada. Además, se puede observar cómo, aunque se imputaron valores en la variable simulada de la Nota de la Prueba de Bachillerato en Matemática, el promedio de los datos es muy similar al de la variable con los casos completos (ver Cuadro 12).

A partir lo anterior se construyen modelos de regresión multinivel que permitan considerar la estructura jerárquica de los datos en la predicción del rendimiento en matemáticas. Se tienen 2 modelos principales uno que tienen como variable dependiente los datos originales (completos) para todos los estudiantes y otro con variable dependiente la Nota de la Prueba de Bachillerato en Matemática con los datos imputados. Ambos modelos evidencian la importancia de incluir el componente jerárquico, ya que el colegio explica más de un 15% de la variabilidad total, sin embargo, es menor en el caso de la variable dependiente imputada (18% y 15% respectivamente). Este aspecto llama la atención, ya que señala desigualdades en los aprendizajes de secundaria de colegio a colegio.

Para comparar los modelos se revisa si las mismas variables del nivel del individuo son mantienen la misma magnitud en los coeficientes. Sin embargo, entre los modelos existen diferencias, ya que, por un lado, el modelo con datos imputados solo considera como relevantes la edad, Puntaje en la prueba de razonamiento con figuras y autoeficacia en matemática, mientras que, por el otro, el modelo con datos completos indica que la variable sexo no debe ser ignorada, y la magnitud del coeficiente sexo disminuye en casi la mitad entre el modelo con los datos completos y el que tiene datos imputados. (ver Cuadro 13)

Variable Estadístico	Modelo sin imputación		Modelo con imputación	
	Coef	*p*	*Coef*	*p*
Sexo	2,12	0,067	1,20	0,254
Edad	-2,93	0,000	-3,16	0,000
Puntaje PRF	0,20	0,000	0,21	0,000
Escala Sexismo Benevolente	-1,25	0,116	-1,30	0,074
Escala Sexismo Hostil	-0,97	0,223	-0,16	0,823
Autoeficacia en matemática	4,58	0,000	4,75	0,000
Intercepto	87,03	0,000	87,90	0,000
Varianza de los efectos aleatorios				
Colegio	21,99	12,325	14,84	8,581
Residual	120,19	7,837	100,09	6,526
Número d observaciones	487		487	
Número de grupos	10		10	
Log-Likelihood	-1862,76		-1818,04	

Cuadro 13. Estimación del modelo de regresión multinivel para la variable dependiente Nota en la Prueba de bachillerato en matemática, según si se utilizaron los datos completos o los imputados.

Variable	Modelo datos completos		Modelo datos imputados	
	Límite inferior	Límite superior	Límite inferior	Límite superior
Sexo	0,21	4,02	-0,53	2,94
Edad	-4,01	-1,85	-4,15	-2,17
Puntaje Razonamiento con figuras	0,15	0,25	0,17	0,25
Escala Sexismo Benevolente	-2,56	0,06	-2,49	0,1
Escala Sexismo Hostil	-2,28	0,33	-1,36	1,03
Autoeficacia en matemática	3,53	5,61	3,8	5,7
Intercepto	64,02	106,03	70,61	105,18

Cuadro 14. Intervalos de Confianza (90%) para los coeficientes de las variables del modelo, según si los datos utilizados están completos o se imputaron.

Otro criterio utilizado para verificar las similitudes entre los modelos es la de comparar los intervalos de confianza. De esta comparación se concluye que, aunque los intervalos se cruzan, en el caso de la variable sexo el intervalo para el caso de los datos completos no contiene el cero, por su parte el de los datos imputados sí (ver Cuadro 14). Además, si se compara el ajuste en ambos modelos, utilizando los criterios AIC y BIC, el ajuste es menor en el modelo de datos reales, y, para el caso del modelo con datos sin imputar, la proporción de variancia explicada para ambos niveles (R_1^2 , R_2^2) es menor que la dada por el modelo con datos imputados (ver Cuadro 15).

Modelo	AIC	BIC	R^{2*}	
Modelo sin imputación	3743,522	3781,216	Nivel 1	Nivel 2
			0.30	0.35
Modelo con imputación	3654,074	3691,769	0.35	0.41

Nota* R^2 Snijders y Bosker

Cuadro 15. Comparación de los estadísticos de ajuste entre el modelo con datos completos y datos imputados.

Capítulo V: Conclusiones y Recomendaciones

5.1 Conclusiones

1. Los resultados obtenidos señalan que es adecuado fijar en 10 la cantidad de iteraciones del mecanismo de imputación múltiple, tal y como se establece por defecto en algunos paquetes estadísticos, y como lo mencionaba Rubín (1987), citado por Medina y Galván (2007), debido a que se evidencia que el aumentar la cantidad de iteraciones no aporta en la reducción del sesgo e incrementa el tiempo de procesamiento estadísticos.

2. Basándose en los análisis de simulación se evidencia la relevancia de la magnitud de la asociación entre la variable con el dato faltante y las independientes utilizadas para su predicción, en el nivel de precisión resultante del proceso de imputación múltiple. Entre mayor sea esta, menor es el error promedio introducido en los datos, independientemente del enfoque de estimación utilizado (frecuentista o bayesiano).

3. Galván (2007) menciona que según Rubin (1987), la técnica de imputación múltiple tiene la ventaja de brindar buenos resultados aún en presencia de porcentajes de valores perdidos entre 30% y 50%. Sin embargo, aunque se concluye que tanto en porcentajes altos de valores faltantes (40%) como en el caso de porcentaje moderado (20%), el comportamiento del sesgo es similar, esto no implica que la eficiencia de la recuperación sea apropiada.

4. El análisis permite evidenciar, tanto en el caso de un porcentaje de alto de valores faltantes (40%) como en el caso de porcentaje moderado (20%), la inexistencia de diferencias sustantivas relevantes entre la utilización un enfoque frecuentista y uno bayesiano objetivo, pues, en ambos casos, el error absoluto promedio en la recuperación de los valores faltantes es similar (alrededor de 9 y 0.7 cada caso), lo que coincide con Austin, Naylor y Tu (2001) pues para él la ventaja del enfoque bayesiano radica en incorporar prioris informativas.

5. Bajo el enfoque frecuentista, considerando se tienen correlaciones altas (mayores o iguales a 0.55) entre las variables con valores faltantes y las predictoras, y un porcentaje de faltantes del 20%, se observa un sesgo promedio menor si el patrón de pérdida se produce al azar (MAR=0.72 /MNAR=0.75).

6. En el escenario de simulación que utilizó las condiciones reales de los datos observados se encuentra que el indicador del sesgo es alto (en promedio entre 7 y 12 puntos en la escala de 0 a 100). Estas condiciones son: 20% de datos faltantes en la variable de interés, una correlación moderada (0.20 a menor de 0.55) entre la variable con faltantes y los predictoras, y un patrón de datos faltantes no al azar.

7. Al estimar un modelo de regresión multinivel y comparar los resultados incluyendo o no los datos imputados con el enfoque frecuentista, se concluye que existen diferencias relevantes en las magnitudes de los coeficientes fijos del modelo. Por tanto, no se recomienda la imputación en este contexto. Esto posiblemente es debido a que los faltantes no siguen un patrón MCAR, y la variable sexo afecta la probabilidad de perdida de forma diferente para hombres y mujeres.

8. A partir del desarrollo de este estudio se puede concluir que la simulación es una herramienta que facilita la visualización de realidades más allá de la experimentada en un caso específico, razón por la cual brinda información adicional a los tomadores de decisiones, permitiendo así adelantarse a situaciones futuras. Sin embargo, a priori deben plantearse claramente los escenarios para lograr resultados óptimos, el objetivo final de un estudio de simulación debe ser alimentar la toma de decisiones para el uso de modelos estadísticos en contextos aplicados. En este caso particular se partió de condiciones que se presentan típicamente en la investigación social.

5.2. Recomendaciones

1. Para establecer los escenarios a utilizar en las simulaciones es necesario conocer el contexto de los datos reales, así como la teoría en la cual están inmersos, por lo tanto, se recomienda consultar especialistas que conozcan la problemática sustantiva del estudio aplicado que dio origen a ese conjunto de datos.

2. Para llevar a cabo el proceso imputación de múltiple se recomienda analizar las relaciones entre la variable que se desea imputar y las posibles variables predictoras, puesto que el nivel de asociación que presenten es crucial para la precisión de los resultados, pues como se evidencia en el estudio la imputación tiene una efectividad similar en presencia de porcentajes moderados y altos de faltantes (20% y 40%).

3. No se recomienda imputar en casos donde la asociación entre la variable con datos faltantes y las predictoras sea moderada (menor a 0.55), esto porque se observan grandes diferencias en la estimación resultante de las simulaciones. En el caso de la investigación aplicada que se analizó la diferencia promedio entre el valor real y el imputado fue aproximadamente de 9 puntos en la escala 0 a 100. En el contexto de investigación educativa se considera un error demasiado alto.

4. Se recomienda que las personas investigadoras incorporen en el diseño del estudio hipótesis específicas previas en cuanto al comportamiento de los datos faltantes, lo que podría llamarse la “teoría de los valores faltantes”. Estas son las hipótesis que sustentan el mecanismo por el cual se podrían producir valores faltantes. Asimismo, como parte de esa teoría, valoren hasta qué punto tiene sentido pensar que se cumplen los supuestos para realizar imputación múltiple. Esto a la luz del hecho de que en la investigación social es difícil cumplir con las condiciones de partida para implementar de manera adecuada la imputación múltiple y obtener datos imputados que resguarden la validez de los resultados.

5. Por lo anterior, se recomienda, además, valorar mecanismos alternativos que minimicen la posibilidad de generar valores faltantes en el proceso de investigación, por ejemplo, ofrecer incentivos a las personas para que participen.

6. Si se está en presencia de un patrón de valores faltantes que no es al azar (MNAR), pero se cuenta con variables predictoras altamente correlacionadas con la variable de interés (mayores o iguales a 0.55), se podría considerar el uso del método de imputación múltiple para recuperar la información faltante.

7. Si se presentan las condiciones apropiadas para realizar imputación múltiple y dado que no se encuentran diferencias bajo el enfoque frecuentista y bayesiano objetivo, se recomienda utilizar el frecuentista, por su mayor simplicidad.

8. Se sugiere continuar esta línea de investigación en torno a la imputación de valores faltantes con el método de imputación múltiple, en contextos de investigación social, incluyendo la contrastación de resultados, a partir del cambio de un enfoque bayesiano objetivo a uno subjetivo.

Referencias

Abellán de Andrés, C. (2015). *Tratamiento bayesiano de valores ausentes en datos espacio-temporales.* (Tesis, Universitat de Valencia). Recuperado de http://roderic.uv.es/bitstream/handle/10550/48578/Tesis.pdf?sequence=1

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*(4), 1012-1028.

Allison, P. D. (2001). *Missing data* (Vol. 136). Sage publications.

Austin, P. C., Naylor, C. D., & Tu, J. V. (2001). A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *Journal of evaluation in clinical practice*, *7*(1), 35-45.

Ayçaguer, L. S., Villegas, A. M., & Fernández, E. V. (2000). Debate sobre métodos frecuentistas vs bayesianos. *Gaceta Sanitaria*, *14*(6), 482-494.

Badler C., Alsina S., Puigsubirá S., Vitelleschi M. Imputación Múltiple con SAS para estimaciones a partir de Bases de Datos con Información Faltante. Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Séptima Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística. 2002.

Backhoff Escudero, E., Sánchez Moguel, A., Peón Zapata, M., & Andrade Muñoz, E. (2010). Comprensión lectora y habilidades matemáticas de estudiantes de educación básica en México: 2000-2005. *Revista electrónica de investigación educativa*, *12*(1), 1-15. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S160740412010000100004&lng=es&tlng=es.

Bologna, E. (2012). Tendencias en el análisis estadístico: Límites de la inferencia frecuencial y posibilidades del enfoque bayesiano. *Revista Evaluar*, *11*.https://revistas.unc.edu.ar/index.php/revaluar/article/view/2842

Bonilla, N. G., & Uribe, L. H. (2005). Mujer y ciencia en la Universidad de Costa Rica. *Voces Nuevas*, *2*(2), 79.

Brown, C. S., & Leaper, C. (2010). Latina and European American girls' experiences with academic sexism and their self-concepts in mathematics and science during adolescence. *Sex Roles*,*63*(11-12), 860-870.

Bú, R. C. (1993). *Simulación: un enfoque práctico*. Editorial Limusa. (México)

Cano Berlanga, S. (2016). *La imputación múltiple y su aplicación a series temporales financieras*. Doctoral. Universitat Rovira i Virgili. Departamentd'Economia.

Carracedo-Martínez, E., & Figueiras, A. (2006). Statistical processing of non-response in transversal epidemiological studies. *Salud Pública de México*, *48*(4), 341-347.

Carranza, P., &Kuzniak, A. (2009). Enfoque bayesiano "oculto" y enfoque frecuentista "ambiguo" en los manuales franceses de Première S y ES. *Teoría y aplicaciones del Análisis Estadístico Implicativo: Primera aproximación en lengua hispana* (pp.447-460). Universitat Jaume I de Castellón.

Castro, L. M. U., & Ávila, D. M. M. (2006). Una introducción a la imputación de valores faltantes. *Terra Nueva Etapa*, *22*(31), 127-151.

Cerda, G., Ortega, R., Pérez, C., Flores, C. y Melipillán, R. (2011). Inteligencia lógica y rendimiento académico en matemáticas: un estudio con estudiantes de Educación Básica y Secundaria de Chile. *Anales de Psicología*, 27(2),389-398. Disponible en: https://www.redalyc.org/articulo.oa?id=167/16720051015.

Cervini, R. (2002). Desigualdades socioculturales en el aprendizaje de la matemática y lengua de la educación secundaria en Argentina: Un modelo de tres niveles. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa*, 8(2), 135-158. Disponible en: https://www.redalyc.org/articulo.oa?id=916/91680201.

Cheung, Y. (2013). *Statistical analysis of human growth and development*. Boca Raton (FL): CRC Press; 2014.

Contreras, Francoise, y Espinosa, Juan Carlos, y Esguerra, Gustavo, y Haikal, Andrea, y Polanía, Alejandra, y Rodríguez, Adriana, y "Autoeficacia, ansiedad y rendimiento académico en

adolescentes." *Diversitas: Perspectivas en Psicología*, vol. 1, no. 2, 2005, pp.183-194. Redalyc, https://www.redalyc.org/articulo.oa?id=67910207

Clark Blickenstaff, J. (2005). Women and science careers: leaky pipeline or gender filter?.*Gender and education*, *17*(4), 369-386.

Der, G. & Everitt, B. (2013). *Applied medical statistics using SAS*. Boca Raton, FL: CRC Press.

Fernández, M., Castro, Y., & Lorenzo, M. (2004). EvolutionofHostileSexism and BenevolentSexism in a SpanishSample. *Social IndicatorsResearch, 66*(3), 197-211. Retrieved August 15, 2019, from http://www.jstor.org.ezproxy.sibdi.ucr.ac.cr:2048/stable/27522069

Galarza Guerrero, L. (2013). Comparación mediante simulación de los métodos EM e imputación múltiple para datos faltantes (Licenciatura). Universidad Nacional Mayor de San Marcos

Galván, M. (2007).*Imputación de datos: teoría y práctica* (Vol. 54). UnitedNationsPublications.

Garaigordobil, Maite, &Aliri, Jone (2011). Sexismo hostil y benevolente: relaciones con el autoconcepto, el racismo y la sensibilidad intercultural. *Revista de Psicodidáctica*, 16(2),331-350.

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, *70*(3), 491.

Guerra, J. B., &Gallestey, J. B. (2010). Imputación múltiple en variables categóricas usando data aumentación y árboles de clasificación. *InvestigaciónOperacional, 31*(2), 133-139.

Howell, D. C. (2007). The treatment of missing data. *The Sage handbook of social science methodology*, 208-224.

Hyde, J. S., & Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly, 25*(4), 364-378.

Jackman, S. (2000). Estimation and inference are missing data problems: Unifying social science statistics via Bayesian simulation. *Political Analysis*, *8*(4), 307-332.

Kang, H. (2013). The prevention and handling of the missing data.*Korean Journal of Anesthesiology*,*64*(5), 402–406. http://doi.org/10.4097/kjae.2013.64.5.402

Kilianski, S. E., & Rudman, L. A. (1998). Wanting it both ways: Do women approve of benevolent sexism? *Sex Roles, 39*(5-6), 333-352.

Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. *StructuralEquationModeling, 18*(4), 649-662.

Llanos, A. A., & Salas, M. M. (2007). La conveniencia del análisis multinivel para la investigación en salud: una aplicación para Costa Rica. *Población y Salud en Mesoamérica*, *4*(2), 6.

Marín, G., Barrantes, E. G., & Chavarría, S. (2008). Differences in Perception of Computer Sciences and Informatics due to Gender and Experience. *CLEI Electronic Journal*, *11*(2).

Martínez-Garrido, C., & Murillo, F. J. (2014). Programas para la realización de Modelos Multinivel. Un análisis comparativo entre MLwiN, HLM, SPSS y Stata. *REMA*. *RevistaElectrónica de MetodologíaAplicada, 9*(2), 1-24.

Merkle, E. (2011). A Comparison of Imputation Methods for Bayesian Factor Analysis Models. *Journal of Educational and Behavioral Statistics,36*(2), 257-276. Retrieved from http://www.jstor.org/stable/29789480

Mena, P. (2015). Desarrollo en la pruebanacional de bachillerato de Matemática: una necesidad [Development of the National High-schools Exit Test in Mathematics]. *Cuadernos de Investigación y Formación en EducaciónMatemática, 10*(13)*,* 53-66.

Montañés, P., de Lemus, S., Bohner, G., Megías, J. L., Moya, M., &Garcia-Retamero, R. (2012). Intergenerational transmission of benevolent sexism from mothers to daughters and its relation to daughters' academic performance and goals.*Sex Roles*,*66*(7-8), 468-478.

Montenegro-Montenegro, E., Oh, Y., &Chesnut, S. (2015). No le tema a los datos faltantes: enfoques modernos para el manejo de datos faltantes. *Actualidades en Psicología*, *29*(119), 29-42.

Montero Rojas, E., Moreira Mora, T. E., Zamora Araya, J. A. & Smith Castro, V. (2017). *Una nueva mirada teórica y metodológica a diferencias de género en pruebas de matemática: razonamiento, actitudes psicosociales y modelos multinivel.* [Manuscrito sin publicar]. Consejo Nacional de Rectores.

Moreira Mora, T. E. (2009). Factores endógenos y exógenos asociados al rendimiento en matemática: un análisis multinivel. *Revista Educación*, *33*(2), 61-80. Disponible en: https://www.redalyc.org/articulo.oa?id=440/44012058005

Muñiz, J., Elosua, P., &Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25*(2), 151-157.

Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician*, *60*(3), 213-223.

Luo, S., Lawson, A. B., He, B., Elm, J. J., & Tilley, B. C. (2016). Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. *Statistical Methods in Medical Research*, *25*(2), 821–837. http://doi.org/10.1177/0962280212469358

Odeh, O., Featherstone, A., &Bergtold, J. (2010). Reliability Of Statistical Software. *American Journal of Agricultural Economics, 92*(5), 1472-1489. Retrieved from http://www.jstor.org/stable/40931100

Oswald, D.L. & Harvey, R.D. *Hostile environments, stereotype threat, and math performance among undergraduate women.*Curr Psychol (2000) 19: 338. https://doi.org/10.1007/s12144-000-1025-5

Pérez, Edgardo, & Cupani, Marcos, & Ayllón, Silvia (2005). Predictores de rendimiento académico en la escuela media: habilidades, autoeficacia y rasgos de personalidad. *Avaliação Psicológica*, 4(1),1-11. ISSN: 1677-0471. Disponible en: https://www.redalyc.org/articulo.oa?id=3350/335027178002

Piera, M. À. (2004). *Modelado y simulación. Aplicación a procesos logísticos de fabricación y servicios* (Vol. 118). UniversitatPolitècnica de Catalunya. Iniciativa Digital Politècnica.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys (Vol. 81*). John Wiley &Sons, 2-80.

Anexos

Código del proceso de imputación

Código	Comentarios
	Imputación Bayesiana
	Escenarios con correlación entre variables simulada alta
base=read.csv(file ="/home/sbartels/tesis/base.csv", header=TRUE, sep=";")	
base	
attach(base)	
library(coda)	Proporciona funciones para resumir y trazar la salida de Monte Carlo (MCMC)
library(MASS)	Funciones y conjuntos de datos para admitir "Estadísticas aplicadas modernas con S"
library (MCMCpack)	Contiene funciones para realizar inferencias bayesianas inferencia utilizando simulación
mod2=lm(base$MATEBach~base$Autoeficaciamate+base$PORCPRF +base$Edad)	calcula los coeficientes del modelo a partir de la "base completa"
selecciondecasoseliminarMRA=function(por,variable)	Función que selecciona los elementos a eliminar de base completa escenario McAR
{	
n=length(variable)	
p=floor(n*por)	
CasosEliminar=sample(1:n,p,replace=F)	
CasosEliminar2=sort(CasosEliminar)	
return(CasosEliminar2)	
}	
selecciondecasoseliminarMNRA=function(variable,beta)	Función que selecciona los elementos a eliminar de base completa escenario MNAR
{	
variable=MATEBach	
porcent=0	
n=length(variable)	
variablefaltante=variable	
posicion=c()	
betas=c(-5.54,0.31,beta)	
x=cbind(1,Sexo,Edad)	
pl=x %*% betas	
pl	

```
  pi=1/(1+exp(-pl))

  pi

  CasosEliminar=rbinom(variable,1,pi)

  CasosEliminar2=which(CasosEliminar==1)

  porcent=mean(CasosEliminar)

  return(list(porcent,CasosEliminar2))

}
```

```
remplazofaltantesMRA=function(variable,CasosEliminar2,base2,numv)
```

Función que remplaza los valores perdidos MAR

```
{

  library(coda)

  library(MASS)

  n=length(variable)

  variablefaltante=variable

  j=1

  mod2 = MCMCregress ( base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$Edad, data = base2 )

  summary(mod2)

  for(i in 1:n){

    if(i == CasosEliminar2[j])

    {

variablefaltante[i]=summary(mod2)$statistics[1]+summary(mod2)$statistics[2]*Autoeficaciamate[i]+summary(mod2)$statistics[3]*PORCPRF[i]+summary(mod2)$statistics[4]*Edad[i]

      if(j<length(CasosEliminar2))

      {j=j+1}

    }

  }

  casosimputados=variablefaltante[CasosEliminar2]

  return(casosimputados)

}
```

```
remplazofaltantesMNRA=function(variable,CasosEliminar2,base2,numv)
```

Función que remplaza los valores perdidos MNAR

```
{

  library(coda)

  library(MASS)

  library ( MCMCpack )
```

```
  n=length(variable)

  variablefaltante=variable

  j=1
  mod2 = MCMCregress (
base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$
Edad, data = base2 )
  for(i in 1:n){

    if(i == CasosEliminar2[j])

    {

variablefaltante[i]=summary(mod2)$statistics[1]+summary(
mod2)$statistics[2]*Autoeficaciamate[i]+summary(mod2)$st
atistics[3]*PORCPRF[i]+summary(mod2)$statistics[4]*Edad[
i]
      if(j<length(CasosEliminar2))
      {j=j+1}
    }
  }

  casosimputados=variablefaltante[CasosEliminar2]

  return(casosimputados)
}
ImputacionMNRA=function(variableCamb,base,simulaciones,n
muestras,pos,b)                                          Funcion para la imputacion multiple MNRA
{


  vectcomparacion=c()                                    vector para guardar los valores verdaderos

  matrizdemedias=matrix(nrow=simulaciones,ncol=1)

  vectvarianza=c()                                       vector para guardar los valores verdaderos

  matrizvarianza=matrix(nrow=simulaciones,ncol=1)

  porcentajePerdida=c()                                  vector para guardar los valores verdaderos


  matrizporcentaje=matrix(nrow=simulaciones,ncol=1)


  for(h in 1:simulaciones)
  {


eliminadosC=selecciondecasoseliminarMNRA(variableCamb,b)  Llamado de la función que elimina los casos

      eliminados=eliminadosC[[2]]
      porcentajePerdida=eliminadosC[[1]]
      completos=variableCamb[eliminados]
      k=length(eliminados)


      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)
```

```
matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)                                  ciclo para estimar los valores perdidos
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)                                selecciono una muestra de completos
        base2=base[muestra,]                                  formo la base de completos
x=remplazofaltantesMNRA(variableCamb,eliminados,base2,po      remplazo los valores
s)
        matrizdevaloresperdidos[i,]=x                         guardo los valores en los campos eliminados
      }
      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      comparacion = abs(medias-completos)                     compara el valor imputado con el valor verdadero
vectcomparacion=sum(comparacion)/length(eliminados)
      var=(comparacion-vectcomparacion)*(comparacion-
vectcomparacion)
      vectvarianza= sum(var)/(length(eliminados)-1)
      matrizdemedias[h,]=vectcomparacion
      matrizvarianza[h,]=vectvarianza
      matrizporcentaje[h,]=porcentajePerdida
  }
return(list(matrizdemedias,matrizvarianza,matrizporcenta
je))
}
summary(MATEBach)                                             variable verdadera nota de bachillerato de
                                                              matematica
hist(MATEBach)
cor(MATEBach,Autoeficaciamate)
cor(MATEBach,PORCPRF)
cor(MATEBach,Edad)
mod=lm(MATEBach~Autoeficaciamate+PORCPRF+Edad)
summary(mod)
MATEBachSimu=91.10+4.90*Autoeficaciamate+0.21*PORCPRF-
3.60*Edad+rnorm(487,0,2)
```

MATEBachSimu=200+17*Autoeficaciamate+0.72*PORCPRF-19.5*Edad+rnorm(487,0,2)	
b=(98-30)/141	
a=98-b*105	
z=a+b*MATEBachSimu	
MATEBachSimu2=z	
summary(z)	
hist(z)	
a=cor(z,Autoeficaciamate)	
b=cor(z,PORCPRF)	
c=cor(z,Edad)	
correlaciones=c(a,b,c)	
base2=cbind(base,MATEBachSimu2)	
simulaciones=100	iteraciones de la simulación
nmuestras=10	iteraciones dentro de la simulación
pos=35	
betas=cbind(0.245,0.30)	
medias=matrix(nrow=simulaciones,ncol=2)	
varianza=matrix(nrow=simulaciones,ncol=2)	
porcentaje=matrix(nrow=simulaciones,ncol=2)	
porc=c()	
write.csv(correlaciones, file="/home/sbartels/tesis/correlacionesB.csv")	guardo los resultados correlaciones
write.csv(base2, file="/home/sbartels/tesis/base2B.csv")	
for(i in 1:length(betas)){	
impuN2=ImputacionMNRA(MATEBachSimu2,base2,simulaciones,nmuestras,pos,betas[i])	Llamado de la función de imputación caso MNRA
medias[,i]= t(impuN2[[1]])	
varianza[,i]= t(impuN2[[2]])	
porcentaje[,i]= t(impuN2[[3]])	
porc[i]= mean(impuN2[[3]])	
}	
write.csv(porcentaje, file="/home/sbartels/tesis/BAYESPNMARcorr.csv")	guardo los resultados del porcentaje de perdida
write.csv(varianza, file="/home/sbartels/tesis/BAYESVNMARcorr.csv")	guardo los resultados de la varianza del sesgo
write.csv(medias, file="/home/sbartels/tesis/BAYESMNMARcorr.csv")	guardo los resultados del media del sesgo
ImputacionMRA=function(variableCamb,PorcPerdidos,base,simulaciones,nmuestras,pos)	Función para la imputación múltiple MAR
{	
vectcomparacion=PorcPerdidos	vector para guardar los valores verdaderos
matrizdesimula=matrix(nrow=simulaciones,ncol=length(PorcPerdidos))	
vectvarianza=PorcPerdidos	vector para guardar los valores verdaderos

```
matrizvarianza=matrix(nrow=simulaciones,ncol=length(PorcPerdidos))
  for(h in 1:simulaciones)
  {
    for (j in 1:length(PorcPerdidos))                    ciclo para calcular el error
    {

eliminados=selecciondecasoseliminarMRA(PorcPerdidos[j],variableCamb)   eliminados de la variable verdadra de la nota MB
      completos=variableCamb[eliminados]
      k=length(eliminados)

      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)


matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)
      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)                             ciclo para estimar los valores perdidos
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)                           selecciono una muestra de completos
        base2=base[muestra,]                             formo la base de completos

x=remplazofaltantesMRA(variableCamb,eliminados,base2,pos)   remplazo los valores
        matrizdevaloresperdidos[i,]=x                    guardo los valores en los campos eliminados
      }
      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias


      comparacion = abs(medias-completos)                compara el valor imputado con el valor verdadero


vectcomparacion[j]=sum(comparacion)/length(eliminados)

      var=(comparacion-vectcomparacion[j])*(comparacion-vectcomparacion[j])

      vectvarianza[j]= sum(var)/(length(eliminados)-1)
    }
```

```
    matrizdesimula[h,]=vectcomparacion
    matrizvarianza[h,]=vectvarianza
  }
  return(list(matrizdesimula,matrizvarianza))
}
```

```
MATEBach
```

variable verdadera nota de bachillerato de matematica

```
simulaciones=100
PorcPerdidos=c(porc[1],porc[2])
```

porcentaje de valores perdidos

```
nmuestras=10
pos=35
mediasMAR=matrix(nrow=simulaciones,ncol=2)
varianzaMAR=matrix(nrow=simulaciones,ncol=2)
impu=ImputacionMRA(MATEBachSimu2,PorcPerdidos,base2,simulaciones,nmuestras,pos)
```

Llamado a la función de imputación MAR

```
impu
mediasMAR=impu[[1]]
varianzaMAR=impu[[2]]
write.csv(varianzaMAR, file="/home/sbartels/tesis/BAYESVMARcorr.csv")
```

guardo los resultados de la varianza del sesgo

```
write.csv(mediasMAR, file="/home/sbartels/tesis/BAYESMMARcorr.csv")
```

guardo los resultados del media del sesgo

Escenarios con correlación entre variables moderada

```
base=read.csv(file ="/home/sbartels/tesis/base.csv", header=TRUE, sep=";")
base
attach(base)
library(coda)
library(MASS)
library (MCMCpack)
mod2=lm(base$MATEBach~base$Autoeficaciamate+base$PORCPRF+base$Edad)
selecciondecasoseliminarMRA=function(por,variable)
```

Función que selecciona los elementos a eliminar de base completa MRA

```
{
  n=length(variable)
  p=floor(n*por)
  CasosEliminar=sample(1:n,p,replace=F)
  CasosEliminar2=sort(CasosEliminar)
  return(CasosEliminar2)
}
selecciondecasoseliminarMNRA=function(variable,beta)
```

Función que selecciona los elementos a eliminar de base completa MNAR

```
{
  variable=MATEBach
  porcent=0
```

```
  n=length(variable)
  variablefaltante=variable
  posicion=c()
  betas=c(-5.54,0.31,beta)
  x=cbind(1,Sexo,Edad)
  pl=x %*% betas
  pl
  pi=1/(1+exp(-pl))
  pi
  CasosEliminar=rbinom(variable,1,pi)
  CasosEliminar2=which(CasosEliminar==1)
  porcent=mean(CasosEliminar)
  return(list(porcent,CasosEliminar2))
}
```

```
remplazofaltantesMRA=function(variable,CasosEliminar2,base2,numv)
{
  library(coda)
  library(MASS)
  n=length(variable)
  variablefaltante=variable
  j=1
  mod2 = MCMCregress ( base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$Edad, data = base2 )
  summary(mod2)

  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {

variablefaltante[i]=summary(mod2)$statistics[1]+summary(mod2)$statistics[2]*Autoeficaciamate[i]+summary(mod2)$statistics[3]*PORCPRF[i]+summary(mod2)$statistics[4]*Edad[i]

      if(j<length(CasosEliminar2))
      {j=j+1}


    }
  }

  casosimputados=variablefaltante[CasosEliminar2]

  return(casosimputados)
}
```

Función que reemplaza los faltantes MAR

```
remplazofaltantesMNRA=function(variable,CasosEliminar2,base2,numv)
```

Función que reemplaza los faltantes MNAR

```
{
  library(coda)
  library(MASS)
  library ( MCMCpack )
  n=length(variable)
  variablefaltante=variable
  j=1


  mod2 = MCMCregress (
base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$
Edad, data = base2 )
  summary(mod2)


  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {


variablefaltante[i]=summary(mod2)$statistics[1]+summary(
mod2)$statistics[2]*Autoeficaciamate[i]+summary(mod2)$st
atistics[3]*PORCPRF[i]+summary(mod2)$statistics[4]*Edad[
i]
      if(j<length(CasosEliminar2))
      {j=j+1}


    }
  }
  casosimputados=variablefaltante[CasosEliminar2]
  return(casosimputados)
}
ImputacionMNRA=function(variableCamb,base,simulaciones,n
muestras,pos,b)                                          Función para la imputación múltiple MNAR
{
  vectcomparacion=c()                                    vector para guardar los valores verdaderos
  matrizdemedias=matrix(nrow=simulaciones,ncol=1)
  vectvarianza=c()                                       vector para guardar los valores verdaderos
  matrizvarianza=matrix(nrow=simulaciones,ncol=1)
  porcentajePerdida=c()                                  vector para guardar los valores verdaderos
  matrizporcentaje=matrix(nrow=simulaciones,ncol=1)
  for(h in 1:simulaciones)
  {
eliminadosC=selecciondecasoseliminarMNRA(variableCamb,b) Llamado a la función que elimina casos
```

```
      eliminados=eliminadosC[[2]]
      porcentajePerdida=eliminadosC[[1]]
      completos=variableCamb[eliminados]
      k=length(eliminados)

      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)


matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)
      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)
        base2=base[muestra,]

x=remplazofaltantesMNRA(variableCamb,eliminados,base2,po
s)                                                        Llamado a la función de que remplaza faltantes
        matrizdevaloresperdidos[i,]=x
      }
      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)

      comparacion = abs(medias-completos)                 compara el valor imputado con el valor verdadero


vectcomparacion=sum(comparacion)/length(eliminados)
      var=(comparacion-vectcomparacion)*(comparacion-
vectcomparacion)
      vectvarianza= sum(var)/(length(eliminados)-1)
      matrizdemedias[h,]=vectcomparacion
      matrizvarianza[h,]=vectvarianza
      matrizporcentaje[h,]=porcentajePerdida
  }

return(list(matrizdemedias,matrizvarianza,matrizporcenta
je))
}
MATEBach
simulaciones=100
nmuestras=10
```

```
pos=6
betas=cbind(0.245,0.30)
medias=matrix(nrow=simulaciones,ncol=2)
varianza=matrix(nrow=simulaciones,ncol=2)
porcentaje=matrix(nrow=simulaciones,ncol=2)
porc=c()
for(i in 1:length(betas)){

impuN2=ImputacionMNRA(MATEBach,base,simulaciones,nmuestras,pos,betas[i])
  medias[,i]= t(impuN2[[1]])
  varianza[,i]= t(impuN2[[2]])
  porcentaje[,i]= t(impuN2[[3]])
  porc[i]= mean(impuN2[[3]])
}
porc
porcentaje
varianza
medias
write.csv(porcentaje,
file="/home/sbartels/tesis/BAYESPNMAR.csv")
write.csv(varianza,
file="/home/sbartels/tesis/BAYESVNMAR.csv")
write.csv(medias,
file="/home/sbartels/tesis/BAYESMNMAR.csv")
ImputacionMRA=function(variableCamb,PorcPerdidos,base,simulaciones,nmuestras,pos)
{
  vectcomparacion=PorcPerdidos

matrizdesimula=matrix(nrow=simulaciones,ncol=length(PorcPerdidos))
  vectvarianza=PorcPerdidos

matrizvarianza=matrix(nrow=simulaciones,ncol=length(PorcPerdidos))
  for(h in 1:simulaciones)
  {
    for (j in 1:length(PorcPerdidos))
    {

eliminados=selecciondecasoseliminarMRA(PorcPerdidos[j],variableCamb)
      completos=variableCamb[eliminados]
      k=length(eliminados)

      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)
```

Llamado a la función de imputación MNAR

guardo los resultados de la varianza del sesgo

guardo los resultados del media del sesgo

Funcion para la imputacion multiple MRA

```
matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)
      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)
        base2=base[muestra,]

x=remplazofaltantesMRA(variableCamb,eliminados,base2,pos
)
        matrizdevaloresperdidos[i,]=x
      }
      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias
      comparacion = abs(medias-completos)                  compara el valor imputado con el valor verdadero

vectcomparacion[j]=sum(comparacion)/length(eliminados)

      var=(comparacion-vectcomparacion[j])*(comparacion-
vectcomparacion[j])

      vectvarianza[j]= sum(var)/(length(eliminados)-1)
    }
    matrizdesimula[h,]=vectcomparacion
    matrizvarianza[h,]=vectvarianza
  }
  return(list(matrizdesimula,matrizvarianza))
}
MATEBach
simulaciones=100
PorcPerdidos=c(porc[1],porc[2])
nmuestras=10
pos=6
mediasMAR=matrix(nrow=simulaciones,ncol=2)
varianzaMAR=matrix(nrow=simulaciones,ncol=2)
impu=ImputacionMRA(MATEBach,PorcPerdidos,base,simulacion
es,nmuestras,pos)
```

```
impu

mediasMAR=impu[[1]]

varianzaMAR=impu[[2]]

write.csv(varianzaMAR,
file="/home/sbartels/tesis/BAYESVMAR.csv")

write.csv(mediasMAR,
file="/home/sbartels/tesis/BAYESMMAR.csv")
```

guardo los resultados de la varianza del sesgo

guardo los resultados del media del sesgo

Imputación Frecuentista

Escenarios con correlación entre variables simulada alta

```
base=read.csv(file ="/home/sbartels/tesis/base.csv",
header=TRUE, sep=";")

base

attach(base)

mod2=lm(base$MATEBach~base$Autoeficaciamate+base$PORCPRF
+base$Edad)

selecciondecasoseliminarMRA=function(por,variable)

{
  n=length(variable)
  p=floor(n*por)
  CasosEliminar=sample(1:n,p,replace=F)
  CasosEliminar2=sort(CasosEliminar)
  return(CasosEliminar2)
}
```

Función que selecciona los elementos a eliminar de base completa escenario MAR

```
selecciondecasoseliminarMNRA=function(variable,beta)

{
  variable=MATEBach
  porcent=0
  n=length(variable)
  variablefaltante=variable
  posicion=c()
  betas=c(-5.54,0.31,beta)
  x=cbind(1,Sexo,Edad)
  pl=x %*% betas
  pl
  pi=1/(1+exp(-pl))
  pi
  CasosEliminar=rbinom(variable,1,pi)
  CasosEliminar2=which(CasosEliminar==1)
  porcent=mean(CasosEliminar)
  return(list(porcent,CasosEliminar2))
}
```

Función que selecciona los elementos a eliminar de base completa escenario MNAR

```
remplazofaltantesMRA=function(variable,CasosEliminar2,ba
se2,numv)
```

Función que reemplaza los faltantes MAR

```
{
  n=length(variable)
  variablefaltante=variable
  j=1

mod2=lm(base2[,numv]~base2$Autoeficaciamate+base2$PORCPR
F+base2$Edad)
  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {

variablefaltante[i]=mod2$coef[1]+mod2$coef[2]*Autoeficac
iamate[i]+mod2$coef[3]*PORCPRF[i]+mod2$coef[4]*Edad[i]

      if(j<length(CasosEliminar2))
      {j=j+1}
    }
  }
  casosimputados=variablefaltante[CasosEliminar2]
  return(casosimputados)
}
remplazofaltantesMNRA=function(variable,CasosEliminar2,b
ase2,numv)
{
  n=length(variable)
  variablefaltante=variable
  j=1

mod2=lm(base2[,numv]~base2$Autoeficaciamate+base2$PORCPR
F+base2$Edad)

  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {

variablefaltante[i]=mod2$coef[1]+mod2$coef[2]*Autoeficac
iamate[i]+mod2$coef[3]*PORCPRF[i]+mod2$coef[4]*Edad[i]

      if(j<length(CasosEliminar2))
      {j=j+1}
    }
  }
  casosimputados=variablefaltante[CasosEliminar2]
  return(casosimputados)
}
```

Función que remplaza los valores perdidos MNAR

```
ImputacionMNRA=function(variableCamb,base,simulaciones,nmuestras,pos,b)
{


  vectcomparacion=c()                                   vector para guardar los valores verdaderos

  matrizdemedias=matrix(nrow=simulaciones,ncol=1)

  vectvarianza=c()                                      vector para guardar los valores verdaderos

  matrizvarianza=matrix(nrow=simulaciones,ncol=1)

  porcentajePerdida=c()                                 vector para guardar los valores verdaderos

  matrizporcentaje=matrix(nrow=simulaciones,ncol=1)

  for(h in 1:simulaciones)
  {


eliminadosC=selecciondecasoseliminarMNRA(variableCamb,b)

      eliminados=eliminadosC[[2]]
      porcentajePerdida=eliminadosC[[1]]
      completos=variableCamb[eliminados]
      k=length(eliminados)

      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)


matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)

      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)
        base2=base[muestra,]

x=remplazofaltantesMNRA(variableCamb,eliminados,base2,pos)
        matrizdevaloresperdidos[i,]=x
      }
      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias
```

Función para la imputación múltiple MNRA

```
      comparacion = abs(medias-completos)


vectcomparacion=sum(comparacion)/length(eliminados)

      var=(comparacion-vectcomparacion)*(comparacion-
vectcomparacion)

      vectvarianza= sum(var)/(length(eliminados)-1)


      matrizdemedias[h,]=vectcomparacion

      matrizvarianza[h,]=vectvarianza

      matrizporcentaje[h,]=porcentajePerdida

  }


return(list(matrizdemedias,matrizvarianza,matrizporcenta
je))

}

summary(MATEBach)

hist(MATEBach)

cor(MATEBach,Autoeficaciamate)

cor(MATEBach,PORCPRF)

cor(MATEBach,Edad)

mod=lm(MATEBach~Autoeficaciamate+PORCPRF+Edad)

summary(mod)

MATEBachSimu=91.10+4.90*Autoeficaciamate+0.21*PORCPRF-
3.60*Edad+rnorm(487,0,2)

MATEBachSimu=200+17*Autoeficaciamate+0.72*PORCPRF-
19.5*Edad+rnorm(487,0,2)

b=(98-30)/141

a=98-b*105

z=a+b*MATEBachSimu

MATEBachSimu2=z

summary(z)

hist(z)

a=cor(z,Autoeficaciamate)

b=cor(z,PORCPRF)

c=cor(z,Edad)

correlaciones=c(a,b,c)

base2=cbind(base,MATEBachSimu2)

names(base2)

simulaciones=100

nmuestras=10

pos=35

betas=cbind(0.245,0.30)

medias=matrix(nrow=simulaciones,ncol=2)

varianza=matrix(nrow=simulaciones,ncol=2)
```

compara el valor imputado con el valor verdadero

Variable simulada con correlacionada alta

```
porcentaje=matrix(nrow=simulaciones,ncol=2)
porc=c()
write.csv(correlaciones,
file="/home/sbartels/tesis/correlacionesF.csv")
write.csv(base2, file="/home/sbartels/tesis/base2.csv")


for(i in 1:length(betas)){

impuN2=ImputacionMNRA(MATEBachSimu2,base2,simulaciones,n
muestras,pos,betas[i])
  medias[,i]= t(impuN2[[1]])
  varianza[,i]= t(impuN2[[2]])
  porcentaje[,i]= t(impuN2[[3]])
  porc[i]= mean(impuN2[[3]])
}
write.csv(porcentaje,
file="/home/sbartels/tesis/PNMARFCORR.csv")
write.csv(varianza,
file="/home/sbartels/tesis/VNMARFCORR.csv")
write.csv(medias,
file="/home/sbartels/tesis/MNMARFCORR.csv")
ImputacionMRA=function(variableCamb,PorcPerdidos,base,si
mulaciones,nmuestras,pos)
{
  vectcomparacion=PorcPerdidos

matrizdesimula=matrix(nrow=simulaciones,ncol=length(Porc
Perdidos))
  vectvarianza=PorcPerdidos

matrizvarianza=matrix(nrow=simulaciones,ncol=length(Porc
Perdidos))
  for(h in 1:simulaciones)
  {
    for (j in 1:length(PorcPerdidos))
    {

eliminados=selecciondecasoseliminarMRA(PorcPerdidos[j],v
ariableCamb)
      completos=variableCamb[eliminados]
      k=length(eliminados)

      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)


matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)

      dim(matrizdevaloresperdidos)
      pos=pos
```

guardo los resultados del porcentaje de perdida

guardo los resultados de la varianza del sesgo

guardo los resultados del media del sesgo

Función para la imputación múltiple MAR

```
      attach(base)


      for (i in 1:nmuestras)
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)                    selecciona una muestra de completos
        base2=base[muestra,]                      forma la base de completos

x=remplazofaltantesMRA(variableCamb,eliminados,base2,pos    reemplaza los valores
)
        matrizdevaloresperdidos[i,]=x             guardo los valores en los campos eliminados
      }


      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias


      comparacion = abs(medias-completos)         compara el valor imputado con el valor verdadero


vectcomparacion[j]=sum(comparacion)/length(eliminados)

      var=(comparacion-vectcomparacion[j])*(comparacion-
vectcomparacion[j])

      vectvarianza[j]= sum(var)/(length(eliminados)-1)


    }
    matrizdesimula[h,]=vectcomparacion
    matrizvarianza[h,]=vectvarianza
  }


  return(list(matrizdesimula,matrizvarianza))
}
MATEBach                                          variable verdadera nota de bachillerato de
                                                  matematica
simulaciones=100
PorcPerdidos=c(porc[1],porc[2])                   porcentaje de valores perdidos
nmuestras=10
pos=35
mediasMAR=matrix(nrow=simulaciones,ncol=2)
varianzaMAR=matrix(nrow=simulaciones,ncol=2)
```

```
impu=ImputacionMRA(MATEBachSimu2,PorcPerdidos,base2,simulaciones,nmuestras,pos)
impu
mediasMAR=impu[[1]]
varianzaMAR=impu[[2]]

write.csv(varianzaMAR,
file="/home/sbartels/tesis/VMARFrecuentistaCORR.csv")
```

guardo los resultados de la varianza del sesgo

```
write.csv(mediasMAR,
file="/home/sbartels/tesis/MMARFrecuentistaCORR.csv")
```

guardo los resultados del media del sesgo

Escenarios con correlación entre variables moderada

```
base=read.csv(file ="/home/sbartels/tesis/base.csv",
header=TRUE, sep=";")

base
attach(base)
mod2=lm(base$MATEBach~base$Autoeficaciamate+base$PORCPRF
+base$Edad)
```

Función que selecciona los elementos a eliminar de base completa MRA

```
selecciondecasoseliminarMRA=function(por,variable)
{
  n=length(variable)
  p=floor(n*por)
  CasosEliminar=sample(1:n,p,replace=F)
  CasosEliminar2=sort(CasosEliminar)
  return(CasosEliminar2)
}
```

Función que selecciona los elementos a eliminar de base completa MNAR

```
selecciondecasoseliminarMNRA=function(variable,beta)
{
  variable=MATEBach
  porcent=0
  n=length(variable)
  variablefaltante=variable
  posicion=c()
  betas=c(-5.54,0.31,beta)
  x=cbind(1,Sexo,Edad)
  pl=x %*% betas
  pl
  pi=1/(1+exp(-pl))
  pi
  CasosEliminar=rbinom(variable,1,pi)
  CasosEliminar2=which(CasosEliminar==1)
  porcent=mean(CasosEliminar)
  return(list(porcent,CasosEliminar2))
}
```

```
remplazofaltantesMRA=function(variable,CasosEliminar2,base2,numv)
{
  n=length(variable)
  variablefaltante=variable
  j=1


mod2=lm(base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$Edad)


  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {

variablefaltante[i]=mod2$coef[1]+mod2$coef[2]*Autoeficaciamate[i]+mod2$coef[3]*PORCPRF[i]+mod2$coef[4]*Edad[i]

      if(j<length(CasosEliminar2))
      {j=j+1}


    }
  }
  casosimputados=variablefaltante[CasosEliminar2]
  return(casosimputados)
}
```

Función que reemplaza los faltantes MAR

```
remplazofaltantesMNRA=function(variable,CasosEliminar2,base2,numv)
{
  n=length(variable)
  variablefaltante=variable
  j=1


mod2=lm(base2[,numv]~base2$Autoeficaciamate+base2$PORCPRF+base2$Edad)


  for(i in 1:n){
    if(i == CasosEliminar2[j])
    {

variablefaltante[i]=mod2$coef[1]+mod2$coef[2]*Autoeficaciamate[i]+mod2$coef[3]*PORCPRF[i]+mod2$coef[4]*Edad[i]

      if(j<length(CasosEliminar2))
      {j=j+1}
```

Función que reemplaza los faltantes MNAR

```
    }
  }
  casosimputados=variablefaltante[CasosEliminar2]
  return(casosimputados)
}



ImputacionMNRA=function(variableCamb,base,simulaciones,nmuestras,pos,b)
{

  vectcomparacion=c()
  matrizdemedias=matrix(nrow=simulaciones,ncol=1)
  vectvarianza=c()
  matrizvarianza=matrix(nrow=simulaciones,ncol=1)
  porcentajePerdida=c()
  matrizporcentaje=matrix(nrow=simulaciones,ncol=1)
  for(h in 1:simulaciones)
  {


eliminadosC=selecciondecasoseliminarMNRA(variableCamb,b)
      eliminados=eliminadosC[[2]]
      porcentajePerdida=eliminadosC[[1]]
      completos=variableCamb[eliminados]
      k=length(eliminados)
      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)


matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)
      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)
```

Función para la imputación múltiple MNAR

vector para guardar los valores verdaderos

vector para guardar los valores verdaderos

vector para guardar los valores verdaderos

Llamado a la función que elimina casos

selecciona una muestra de completos

```
        base2=base[muestra,]                                    forma la base de completos

x=remplazofaltantesMNRA(variableCamb,eliminados,base2,pos)      reemplaza los valores
        matrizdevaloresperdidos[i,]=x                           guardo los valores en los campos eliminados
      }

      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias
      comparacion = abs(medias-completos)                       compara el valor imputado con el valor verdadero

vectcomparacion=sum(comparacion)/length(eliminados)
      var=(comparacion-vectcomparacion)*(comparacion-
vectcomparacion)
      vectvarianza= sum(var)/(length(eliminados)-1)
      matrizdemedias[h,]=vectcomparacion
      matrizvarianza[h,]=vectvarianza
      matrizporcentaje[h,]=porcentajePerdida
  }

return(list(matrizdemedias,matrizvarianza,matrizporcentaje))
}
MATEBach                                                        variable verdadera nota de bachillerato de
                                                                matematica
simulaciones=100
nmuestras=10
pos=6
betas=cbind(0.245,0.30)
medias=matrix(nrow=simulaciones,ncol=2)
varianza=matrix(nrow=simulaciones,ncol=2)
porcentaje=matrix(nrow=simulaciones,ncol=2)
porc=c()
for(i in 1:length(betas)){

impuN2=ImputacionMNRA(MATEBach,base,simulaciones,nmuestras,pos,betas[i])
  medias[,i]= t(impuN2[[1]])
  varianza[,i]= t(impuN2[[2]])
  porcentaje[,i]= t(impuN2[[3]])
  porc[i]= mean(impuN2[[3]])
}
```

```
porc
porcentaje
varianza
medias

write.csv(porcentaje,
file="/home/sbartels/tesis/FRECUENTISTAPNMAR.csv")          guardo los resultados del porcentaje de perdida


write.csv(varianza,
file="/home/sbartels/tesis/FRECUENTISTAVNMAR.csv")          guardo los resultados de la varianza del sesgo


write.csv(medias,
file="/home/sbartels/tesis/FRECUENTISTAMNMAR.csv")          guardo los resultados del media del sesgo

ImputacionMRA=function(variableCamb,PorcPerdidos,base,si
mulaciones,nmuestras,pos)
{
  vectcomparacion=PorcPerdidos


matrizdesimula=matrix(nrow=simulaciones,ncol=length(Porc
Perdidos))
  vectvarianza=PorcPerdidos


matrizvarianza=matrix(nrow=simulaciones,ncol=length(Porc
Perdidos))
  for(h in 1:simulaciones)
  {
    for (j in 1:length(PorcPerdidos))                       ciclo para calcular el error
    {


eliminados=seleccionDecasosEliminarMRA(PorcPerdidos[j],v
ariableCamb)
      completos=variableCamb[eliminados]
      k=length(eliminados)


      matrizdecomparacion=matrix(nrow=nmuestras,ncol=k)



matrizdevaloresperdidos=matrix(nrow=nmuestras,ncol=k)

      dim(matrizdevaloresperdidos)
      pos=pos
      attach(base)
      for (i in 1:nmuestras)                                ciclo para estimar los valores perdidos
      {
        base1=base[-eliminados,]
        t=nrow(base1)
        nm=floor(0.6*t)
        muestra=sample(1:t,nm)                              selecciona una muestra de completos
        base2=base[muestra,]                                forma la base de completos
```

```
x=remplazofaltantesMRA(variableCamb,eliminados,base2,pos
)                                                          reemplaza los valores
        matrizdevaloresperdidos[i,]=x                      guardo los valores en los campos eliminados
      }


      matrizdevaloresperdidos
      dim(matrizdevaloresperdidos)
      medias=apply(matrizdevaloresperdidos,2,mean)
      medias
      comparacion = abs(medias-completos)                  compara el valor imputado con el valor verdadero


vectcomparacion[j]=sum(comparacion)/length(eliminados)

      var=(comparacion-vectcomparacion[j])*(comparacion-
vectcomparacion[j])

      vectvarianza[j]= sum(var)/(length(eliminados)-1)
    }
    matrizdesimula[h,]=vectcomparacion
    matrizvarianza[h,]=vectvarianza
  }
  return(list(matrizdesimula,matrizvarianza))
}
MATEBach
simulaciones=100
PorcPerdidos=c(porc[1],porc[2])
nmuestras=10
pos=6
mediasMAR=matrix(nrow=simulaciones,ncol=2)
varianzaMAR=matrix(nrow=simulaciones,ncol=2)
impu=ImputacionMRA(MATEBach,PorcPerdidos,base,simulacion
es,nmuestras,pos)
impu
mediasMAR=impu[[1]]
varianzaMAR=impu[[2]]

write.csv(varianzaMAR,
file="/home/sbartels/tesis/FRECUENTISTAVMAR.csv")          guardo los resultados de la varianza del sesgo


write.csv(mediasMAR,
file="/home/sbartels/tesis/FRECUENTISTAMMAR.csv")          guardo los resultados del media del sesgo
```

Código para los análisis gráficos comparativos

```
B1=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos reales/bayes.csv", header=TRUE, sep=",")

colnames(B1)=c("id","Bayes1_Moderado","Baye1_Alto")

B2=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos reales/bayesn.csv", header=TRUE, sep=",")

colnames(B2)=c("id","Bayes2_Moderado","Bayes2_Alto")

F1=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos reales/frecu.csv", header=TRUE, sep=",")

colnames(F1)=c("id","Frecuentista1_Moderado","Frecuentista1_Alto")

F2=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos reales/frecun.csv", header=TRUE, sep=",")

colnames(F2)=c("id","Frecuentista1_Moderado","Frecuentista1_Alto")

errores=cbind(B1[-1],B2[-1],F1[-1],F2[-1])

colnames(errores)=c("B1_M","B2_M","F1_M","F2_M","B1_A","B2_A","F1_A","F2_A")

summary(errores)

MNMARM=apply(errores,2,mean)

VNMARM=apply(errores,2,var)

attach(errores)

print(plot(density(F2_A),
freq=F,las=1,col="cyan4",main="Comparación del error según paradigma y patrón perdida" ,sub= "Datos reales-Perdida 40%",xlab = "Error promedio absoluto"),lwd=6)

lines(density(F1_A),col="navy")

lines(density(B1_A),col="darkorchid2")

lines(density(B2_A),col="deeppink3")

legend(10.7, 1.2, c("FREC-MNAR","FREC-MAR","BAYES-MAR","BAYES-MNAR"), cex=0.8, col=c("cyan4", "navy", "darkorchid2","deeppink3"),

pch=21:22, lty=1:2, box.lty=0)

print(plot(density(F2_M),
freq=F,las=1,col="cyan4",main="Comparación del error según paradigma, porcentaje de perdida y patrón" ,sub= "Datos reales-Perdida 20%",xlab = "Error promedio absoluto"),lwd=6)

lines(density(F1_M),col="blue4")

lines(density(B1_M),col="hotpink3")

lines(density(B2_M),col="mediumorchid1")

legend(10.1, 0.9,c("FREC-MNAR","FREC-MAR","BAYES-MAR","BAYES-MNAR"), cex=0.8, col=c("cyan4", "blue4", "hotpink3","mediumorchid1"),

pch=21:22, lty=1:2, box.lty=0)
```

Gráfico para comparar resultado escenario correlaciones moderada y perdida 40%, por paradigma

Gráfico para comparar resultado escenariocorrelaciones moderada y perdida 20%, por paradigma

```
B1=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos correlacionalta/bayes.csv", header=TRUE, sep=",")

colnames(B1)=c("id","Bayes1_Moderado","Baye1_Alto")

B2=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos correlacionalta/bayesn.csv", header=TRUE, sep=",")

colnames(B2)=c("id","Bayes2_Moderado","Bayes2_Alto")

F1=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos correlacion alta/frecu.csv", header=TRUE, sep=",")

colnames(F1)=c("id","Frecuentista1_Moderado","Frecuentista1_Alto")

F2=read.csv(file ="C:/Users/sofia.bartels/Desktop/Datos correlacion alta/frecun.csv", header=TRUE, sep=",")

colnames(F2)=c("id","Frecuentista1_Moderado","Frecuentista1_Alto")

errores=cbind(B1[-1],B2[-1],F1[-1],F2[-1])

colnames(errores)=c("B1_M","B2_M","F1_M","F2_M","B1_A","B2_A","F1_A","F2_A")

summary(errores)

MNMARM=apply(errores,2,mean)

VNMARM=apply(errores,2,var)

attach(errores)

print(plot(density(B2_A), freq=F,las=1,col="deeppink3",main="Comparación del error según paradigma y patrón perdida" ,sub= "Datos simulados-Perdida 40%",xlab = "Error promedio absoluto"),lwd=6)

lines(density(F1_A),col="navy")

lines(density(B1_A),col="darkorchid2")

lines(density(F2_A),col="cyan4")

legend(10.7, 1.2, c("FREC-MNAR","FREC-MAR","BAYES-MAR","BAYES-MNAR"), cex=0.8, col=c("cyan4", "navy", "darkorchid2","deeppink3"),

pch=21:22, lty=1:2, box.lty=0)

print(plot(density(F2_M), freq=F,las=1,col="cyan4",main="Comparación del error según paradigma, porcentaje de perdida y patrón" ,sub= "Datos simulados-Perdida 20%",xlab = "Error promedio absoluto"),lwd=6,)

lines(density(F1_M),col="blue4")

lines(density(B1_M),col="hotpink3")

lines(density(B2_M),col="mediumorchid1")
```

Gráfico para comparar resultado escenario correlaciones alta y perdida 40%, por paradigma

Gráfico para comparar resultado escenario correlaciones alta y perdida 20%, por paradigma

```
legend(10.1, 0.9,c("FREC-MNAR","FREC-MAR","BAYES-
MAR","BAYES-MNAR"), cex=0.8, col=c("cyan4", "blue4",
"hotpink3","mediumorchid1"),

pch=21:22, lty=1:2, box.lty=0)
```

Código del modelo de regresión multinivel

```
xtmixedMATEBach Sexo Edad PORCPRF bene_totalhostil_totalAutoeficaciamate || Colegio: ,reml var
level(90)
mltrsq
estimates store m1
xtmixedMATEBachImputada Sexo Edad PORCPRF bene_totalhostil_totalAutoeficaciamate || Colegio:
,remlvar
mltrsq
estimates store m2
estimates store m1 m2
estimates stats m 1m2
xtmixedZMATEBachZSexoZEdad ZPORCPRF Zbene_totalZhostil_totalZAutoeficaciamate|| Colegio: , reml
var level(90)
xtmixedZMATEBachImputadaZSexoZEdad ZPORCPRF Zbene_totalZhostil_totalZAutoeficaciamate|| Colegio: ,
reml var level(90)
```