ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

MEDICIÓN DE LA EFECTIVIDAD DE TÉCNICAS DE IMPUTACIÓN PARA DATOS FALTANTES

Trabajo de Integración Curricular

Tipo: Proyecto de Investigación

Presentado para optar al grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTORES: JAMILTON DANIEL VINUEZA CHALCO

GALO ALEXANDER MASAQUIZA ARAGON

DIRECTOR: Ing. PABLO JAVIER FLORES MUÑOZ MSc.

Riobamba - Ecuador

2021

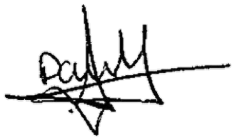
©2021, Jamilton Daniel Vinueza Chalco & Galo Alexander Masaquiza Aragón

Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento, siempre y cuando se reconozca el Derecho de Autor.

Nosotros, Vinueza Chalco Jamilton Daniel, Galo Alexander Masaquiza Aragón, declaramos que el presente trabajo de titulación es de nuestra autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

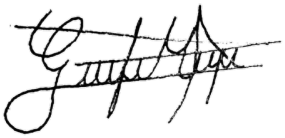
Como autores asumimos la responsabilidad legal y académica de los contenidos de este trabajo de titulación; El patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 6 de agosto de 2021



Jamilton Daniel Vinueza Chalco

140071320-0



Galo Alezander Masaquiza Aragón

160067765-0

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE ESTADÍSTICA INFOMÁTICA

El Tribunal del Trabajo de Integración Curricular certifica que: El Trabajo de Integración Curricular, Tipo: Proyecto de Investigación, **MEDICIÓN DE LA EFECTIVIDAD DE TÉCNICAS DE IMPUTACIÓN PARA DATOS FALTANTES**, realizado por los señores: **JAMILTON DANIEL VINUEZA CHALCO Y GALO ALEXANDER MASAQUIZA ARAGON**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Integración Curricular, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA**
Ing. Hector Salomón Mullo Guaminga MSc, **PRESIDENTE DEL TRIBUNAL**	HECTOR SALOMON MULLO GUAMINGA Firmado digitalmente por HECTOR SALOMON MULLO GUAMINGA Fecha: 2021.10.15 14:35:23 -05'00'	23 de agosto de 2021
Ing. Pablo Javier Flores Muñoz MSc. **DIRECTOR DEL TRABAJO DE TITULACIÓN**	PABLO JAVIER FLORES MUNOZ Firmado digitalmente por PABLO JAVIER FLORES MUNOZ Fecha: 2021.10.18 09:38:07 -05'00'	23 de agosto de 2021
Dr. Rubén Antonio Pazmiño Maji MSc. **MIEMBRO DEL TRIBUNAL**	Firmado electrónicamente por: **RUBEN ANTONIO PAZMINO MAJI**	23 de agosto de 2021

DEDICATORIA

A mis padres, Anita Chalco y Luis Vinueza, por ser mi gran inspiración, mis consejeros de vida y sobre todo ser mi ejemplo de constancia, dedicación y fortaleza diaria. A mis hermanas, Verónica, Valeria y Paula, quienes han sido un pilar fundamental en el transcurso de mi vida, por su apoyo y cariño en todo momento. A mi tía Dalila Chalco y Mami Rosita, por el cariño y dedicación que me ha demostrado, sobre todo en sus consejos, ayudándome a ser una persona íntegra y de principios. A mis tíos Diego y Adrián, por haber sido mis compañeros de vida en estos años, demostrando que todo se puede conseguir si existe dedicación y esfuerzo. A mi prima Kelly, por ser mi mejor amiga y acompañarme en todo el trayecto de mi vida universitaria. A mis primos, Ruby, Michelly, José Luis y Angie, por ser mis amigos y recordarles que siempre están en mi pensamiento. Y finalmente, a mis queridos sobrinos, Luis y Leyre, por enseñarme a amar y cambiar mi vida.

Jamilton

Dedico esta investigación con mucha fe, amor y respeto a Dios, ya que, con su cuidado, fortaleza y sabiduría ha sabido guiar cada paso que doy. Al Sr. Galo Masaquiza y a la Sra. Silvia Aragón, mis padres amados y ejemplo de vida, que han sabido inculcar en mí, los valores necesarios para ser una persona de bien. Mi hermana Janine Masaquiza, por brindarme su apoyo incondicional en todo momento. A mi novia Fernanda Cando e hija Briana Masaquiza, por apoyarme en cada decisión tomada y ser un pilar fundamental en mi desarrollo personal y profesional.

Galo

AGRADECIMIENTO

Un profundo agradecimiento a Dios, quién nos ha brindado salud y sabiduría para poder culminar con éxito la tan anhelada carrera, a nuestras familias por el esfuerzo y apoyo incondicional para poder cumplir esta gran meta propuesta.

A la Escuela Superior Politécnica de Chimborazo por habernos formado éticamente con los más altos estándares académicos, para poder servir a la sociedad de forma responsable. de manera análoga a los docentes quienes se han esforzado para brindarnos una educación.

De manera especial, brindar un sincero agradecimiento al ingeniero Pablo Flores, tutor de la presente investigación, por su tiempo y ayuda absoluta al impartir los conocimientos teóricos y metodológicos necesarios de manera excepcional.

Jamilton & Galo

TABLA DE CONTENIDOS

CAPÍTULO II

CAPÍTULO III

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El presente trabajo de titulación tuvo por objetivo medir la efectividad en términos de precisión y calidad de estimación que presentan distintas técnicas de imputación para datos faltantes, provenientes de una distribución normal. A partir del método de Montecarlo, se creó una matriz bivariada estructurada por datos observados y por datos perdidos, donde los valores faltantes fueron desarrollados a través de un modelo establecido. Se simularon 100.000 veces muestras representativas de tamaño 5, 10, 30 y 100 trabajando con diversos porcentajes de pérdida de información para los escenarios: Faltantes completamente aleatorios (MCAR), Faltantes aleatorios (MAR) y Faltantes no aleatorios (MNAR). Se aplicaron las técnicas de imputación por eliminación, media, mediana y regresión lineal, en la cual se diagnosticó el ajuste de los datos a través de una medida de precisión y se verificó si los datos imputados mantienen sus propiedades de estimación de insesgadez y mínima varianza, utilizando los estimadores de media y varianza. Mediante el uso del software RStudio se determinó qué regresión lineal es la más precisa en muestras a partir de 30, mientras la media y mediana en muestras pequeñas como 5 por obtener valores más cercanos a los datos reales. La insesgadez de la media, demuestra que la mejor técnica es la imputación por regresión lineal, debido a que su propiedad se mantiene en muestras a partir de 30. En la insesgadez de la varianza la técnica más viable es la eliminación en los escenarios MAR y MCAR para muestras de 30 y 100, mientras para MNAR en muestras de cualquier tamaño. Conforme a la mínima varianza de la media y varianza, la técnica que arrojó una varianza inferior en la mayoría de los contextos es la regresión lineal. Se recomienda ampliar el estudio utilizando técnicas de imputación múltiple y machine learning para diagnosticar mejores resultados.

Palabras clave: <ESTADÍSTICA>, <MÉTODO DE MONTECARLO>, <IMPUTACIÓN DE DATOS>, <PRECISIÓN DE AJUSTE>, <PROPIEDADES DEL ESTIMADOR>.

LEONARDO FABIO MEDINA NUSTE

Firmado digitalmente por LEONARDO FABIO MEDINA NUSTE
Fecha: 2021.10.20 09:13:26 -05'00'

UNIDAD DE PROCESOS TÉCNICOS
DBRA
DIRECCIÓN DE BIBLIOTECAS Y RECURSOS DEL APRENDIZAJE
ESPOCH

1813-DBRA-UTP-2021

SUMMARY/ABSTRACT

The objective of this research work was to measure the effectiveness in terms of precision and quality of estimation presented by different imputation techniques for missing data, coming from a normal distribution. From the Monte Carlo method, a bivariate matrix structured by observed data and by missing data was created, where the missing values were developed through an established model. Representative samples of size 5, 10, 30 and 100 were simulated 100,000 times working with different percentages of information loss for the scenarios: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The imputation techniques by elimination, mean, median and linear regression were applied, in which the adjustment of the data was diagnosed through a precision measure and it was verified if the imputed data maintain their estimation properties of unbiasedness and minimum variance., using the mean and variance estimators. Using the RStudio software, it was determined which linear regression is the most accurate in samples from 30, while the mean and median in small samples such as 5 to obtain values closer to the real data. The unbiasedness of the mean shows that the best technique is the imputation by linear regression, since its property is maintained in samples from 30 onwards. In the unbiasedness of the variance, the most viable technique in MAR and MCAR is elimination for samples of 30 and 100, while for MNAR in samples of any size. According to the minimum variance of the mean and variance, the technique that yielded a lower variance in most contexts is linear regression. It is recommended to extend the study using multiple imputation techniques and machine learning to diagnose better results.

Keywords: <STATISTICS>, <MONTE CARLO METHOD>, <DATA IMPUTATION>, <ADJUSTMENT PRECISION>, <ESTIMATOR PROPERTIES>.

INTRODUCCIÓN

Todo proceso investigativo requiere de un eficiente análisis de datos que permita desarrollar nuevos conocimientos y obtener conclusiones precisas para alcanzar los objetivos propuestos. Actualmente, todas las industrias del mercado existente, realizan estudios previos para construir, implementar e innovar nuevos mecanismos de mejora a través del análisis de datos. Partiendo de la recolección, transformación, limpieza y modelado, se estiman parámetros calculando los estadísticos previstos y se ajusta un modelo preestablecido para aceptar o rechazar las hipótesis de interés y tomar decisiones confiables. El proceso se detiene cuando en la matriz de información militan datos faltantes, también denominados "*missing data"* debido a diversos factores que aparecen al momento de la recolección y como consecuencia producen alteraciones significativas en los resultados con bajo nivel de precisión. Problema que un analista está en la obligación de solucionar (Batanero and Godino, 2001, p. 5).

Existen diversos tipos de bases de datos, con variables de tipo cuantitativo o cualitativo y de extensa o corta dimensión, que para obtenerlas como tal, deben pasar previamente por un proceso de relección de datos. Las fuentes de recolección de datos pueden ser Primarias o Secundarias. En la primera, la recolección se realiza interactuando de manera directa con el sujeto objeto de estudio, aplicando encuestas, entrevistas o simplemente a través de la observación, valorando parámetros como el objeto de estudio, las posibilidades de acceso, el tamaño de la población o muestra, los recursos con los que se cuenta y el origen de los datos. La fuente de recolección secundaria es desarrollada por terceras personas, la información ya se encuentra organizada y analizada probablemente. El problema empieza cuando la recolección de datos sufre anomalías y no se completa eficientemente según lo esperado, dificultando su proceso de análisis posterior (de Tiratel, 2000, p. 66).

Dado que existen varios tipos de datos faltantes, es necesario identificar a qué tipo de pérdida se enfrentan. En 1976, Donald B. Rubin, agrupó los inconvenientes de datos faltantes encontrados en tres categorías al cual los llamó "Mecanismos de falta de respuesta", mencionando que cada punto de dato tiene una probabilidad de faltar (Rubin, 1976). Los mecanismos designados son los siguientes: Faltantes completamente aleatorios (MCAR), la probabilidad de obtener un dato perdido es la misma para todos casos o individuos; Faltantes aleatorios (MAR), la probabilidad de faltar es la misma como en MCAR, pero dentro de grupos definidos; Faltantes no aleatorios (MNAR), la probabilidad varía por varias razones, se desconoce. En muchos de los casos, los investigadores desconocen de estos mecanismos y proceden a ignorar estos faltantes (Van Buuren, 2018, pp. 23-30).

La magnitud de este problema es grave, infringe reglas estadísticas, contribuyendo a deducciones poco fiables de un tratamiento, por lo cual, es necesario buscar alternativas óptimas para arreglar

dicho percance. Una solución conocida y perfeccionada a lo largo de las décadas por varios autores es la Imputación, una técnica que devuelve valores efectivos si se lo aplica adecuadamente. Son varias las técnicas que desglosa la imputación, que dependen del tipo de variable, magnitud, entre otros. Métodos que trabajan desde el relleno con la media, hasta procesos que requieren de algoritmos avanzados de clasificación por *Machine Learning*. El investigador, por ende, está en la capacidad de elegir la mejor alternativa para el relleno de sus datos con el respectivo estudio previo (Dagnino, 2014, p. 34).

Un estudio importante lo realizó Abdul Lateef Bello en 1993, el cual ejecutó un proyecto de simulación con el objetivo de comparar algunos métodos de imputación midiendo la precisión de cada una, en la estimación de la media de la variable de interés con su respectiva varianza. Posteriormente existieron también otros distinguidos autores que manejaron una metodología similar que, a través de la recolección de una base de datos, escogieron aleatoriamente valores faltantes y emplearon técnicas de imputación simples y múltiples para la generación de resultados y poder compararlos (Bello, 1993, p. 14).

En el contexto investigativo, se puede aclarar que los expertos recurrieron a simulación, que se puede definir como una técnica de muestreo estadístico, cuyo proceso es controlado y utiliza un modelo para generar respuestas de tipo probabilístico. Al incurrir en el mundo de análisis de datos de medición se pueden encontrar principalmente dos técnicas de simulación, el primero se denomina el Método Montecarlo y el segundo es el llamado Método de remuestreo bootstrap, ambos son apropiados para estudiar las propiedades estadísticas de las mediciones (Bello, 1993, p. 15).

El surgimiento y actualizaciones de sistemas informáticos (Software) es indispensable en el manejo de datos, ellas permiten el control sistematizado y ordenado de pocos o abundantes datos, son capaces de inferir y resolver problemas del ámbito académico y laboral bajo situaciones de incertidumbre, logrando de este modo obtener resultados y tomas de decisiones más acertadas.

En el año 2018, se remonta una investigación bajo únicamente un mecanismo de respuesta, en ella se desarrolló una comparación de métodos de imputación al trabajar con respuestas censuradas en un diseño de experimento bivariado realizado por los autores Zúñiga Maldonado y Hernández Ripalda (Zúñiga Maldonado, Hernández Ripalda and Jiménez García, 2018, p. 21).

El presente estudio es primordial debido a que no existen estudios previos realizados con la misma metodología. El estudio mide la efectividad en términos de precisión y calidad de estimación que presentan distintas técnicas de imputación simple para datos faltantes. A partir de una base de datos normal multivariante simulada mediante los mecanismos de respuesta, se eliminan datos y se obtiene una base de datos con missing. Se aplican las técnicas de imputación simples seleccionadas y a través del nivel de precisión de ajuste se compara la base de datos con valores faltantes y la base de datos original. De la base de datos imputada se comprueba si cumplen con las propiedades de los estimadores: insesgadez y mínima varianza. Para todo el desarrollo del

presente capítulo se utiliza el software estadístico R-Studio donde se indican las funciones utilizadas, para la generación de datos perdidos, para cada técnica de imputación, nivel de precisión de ajuste y propiedades de los estimadores. Se presenta las funciones del R – Package de las técnicas de imputación de datos. Se recalca que éste estudio solo es posible en un escenario de simulación, mas no con datos de la vida real, debido a que nunca se sabrá cuál es el valor faltante y se imposibilita conocer el nivel de precisión de la imputación.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. Antecedentes

Recolectar una base de datos completa y manejable es la herramienta principal de todo investigador para poder realizar inferencias y, por ende, toma de decisiones. Sin embargo, la pérdida de datos ha afectado en todo ámbito posible a través de las décadas. Varios autores han estudiado la forma de llenar dichos vacíos, desarrollando y experimentando con diversos procesos llamados "métodos de imputación". La situación se complica debido a la existencia de varios tipos de datos, diferencias en magnitudes y numerosidad, por ello, es necesario establecer métodos que ayuden de forma eficiente de acuerdo a la situación individual. En los últimos tiempos, se han desarrollado con la ayuda del avance de la computación, nuevas formas de estudiar los datos faltantes, obteniéndose una variedad de técnicas basadas en diferentes enfoques según las características de la data (Castro and Ávila, 2006, p. 8). Los primeros aportes en imputación se realizaron en 1932 por Samuel Stanley Wilks (Wilks, 1932, p. 6), quien propuso el reemplazo de los datos faltantes por la media de los datos presentes de la variable. Este método, puede ser aplicado cuando existen pocos datos faltantes, ya que tienden a distorsionar la distribución de las variables. Con el avance tecnológico de los sistemas computacionales se iniciaron investigaciones sobre el método, particularmente en las décadas de los setenta y ochenta. Donald B. Rubin, hizo una distinción entre MAR (valores perdidos o faltantes de manera aleatoria) y MCAR (valores perdidos o faltantes de manera completamente aleatoria). En el caso de MAR, los datos perdidos dependen de los valores observados, pero no de las variables perdidas propiamente, es decir, la pérdida de información de la variable no depende de ella misma, mientras que en MCAR los datos perdidos no dependen de otros valores observados ni de otros datos perdidos (Rubin, 1976, pp. 23-25). Otro aporte al tratamiento de la "no respuesta" es el método *Listwise*, el cual fue propuesto por "Hemel". Es usado al tratar con grandes conjuntos de datos, eliminando la fila o columna en la posición del dato faltante, para obtener una base con menos datos, pero completa. No es recomendable para conjuntos de datos pequeños debido a que se pierde una cantidad directamente proporcional al número de datos perdidos (Hemel *et al.*, 1987, p. 18). Con el avance tecnológico se ha buscado crear o mejorar los métodos de relleno de datos, basados en diferentes modelos o algoritmos como el análisis de componentes principales, análisis factorial, redes neuronales, entre otras (Castro and Ávila, 2006, p. 4). De tal forma, se necesita llegar a diferentes criterios que deben ser abordados de acuerdo a la distribución, estructura, numerosidad y demás parámetros de interés del conjunto de datos, alcanzando una solución confiable para cada enfoque. Necesidad que se evidencia además en ámbitos como Minería de datos, Gestión documental y áreas donde se

trabajen día a día con extensas cantidades de datos, en que una decisión errónea relacionada al mal manejo de datos faltantes puede provocar tomas de decisiones poco acertadas, ineficientes que alteran el análisis y con ello la productividad de la investigación. Por lo tanto, es ineludible conocer profundamente las propiedades y características de cada técnica de imputación en base a la naturaleza de los datos, para aplicar y entregar una información eficaz (Amón Uribe, 2010, p. 21). Los datos faltantes no siempre se aprecian en su justa medida. Fernando Medina y Marco Galván advierten sobre la falta de consciencia por parte de los usuarios y aún de muchos investigadores, sobre las implicaciones estadísticas que conlleva trabajar con datos faltantes o aplicar procedimientos de imputación o sustitución de información deficientes: "La aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio"(Medina and Galván, 2007, p. 43). Los procedimientos utilizados con frecuencia restringen el poder explicativo de los modelos y a su vez, provocan graves sesgos que alteran las relaciones existentes entre las variables estudiadas, subestimando, además, las propiedades de la varianza (Medina and Galván, 2007, p. 44). Aitor Puerta Goicoechea, resume ciertos criterios a tomar en consideración para seleccionar la técnica adecuada para imputar. Primero menciona que debemos estudiar el tipo de variable a imputar, en el caso cualitativo las categorías y en el caso cuantitativo los intervalos. Los parámetros a estimar, distinguir adecuadamente los valores y el fin investigativo para de acuerdo a ello definir si se imputa desde lo más básico como con la media o mediante procesos más complejos como de árboles de clasificación. La tasa de no respuesta dado que, si el porcentaje de no respuesta es alta, no existe confiabilidad alguna y finalmente, recomienda el utilizar al máximo la información auxiliar disponible para deducir información y en casos posibles, hallar grupos homogéneos respecto a una variable auxiliar con alto grado de correlación para poder encontrar registros similares (Goicoechea, 2002, p. 4). Sharon L. Lohr, refiriéndose a la importancia de los procedimientos de imputación señala que ésta no radica solo en reducir el sesgo por las ausencias de respuestas, sino también para producir un conjunto de datos "limpios" sin datos faltantes. Son muchas las técnicas de imputación que han surgido, sobre todo desde la década de los setenta, que emplean enfoques univariantes y multivariantes. A pesar de estos avances, no se ha encontrado una metodología capaz de reproducir la data o que pueda resolver en forma totalmente satisfactoria el tratamiento de los datos faltantes, debido, generalmente, a problemas en cuanto a las alteraciones de la distribución de los datos, alteración en la relación de las variables, sesgo en las estimaciones, inflación de la varianza, entre otros, razón por la cual, aún se sigue investigando en busca de mejorar las técnicas existentes (Lohr, Velasco and Alfredotr, 2000, p. 25).

1.2. Planteamiento del problema

1.2.1. Enunciado del problema

Uno de los problemas más habituales al depurar bases de datos es la imputación de valores perdidos. Para determinar la técnica más efectiva al trabajar con una distribución normal, es necesario que exista una adecuada precisión de ajuste, que mide la dispersión entre los datos reales con los imputados, estableciendo que, a menor diferencia, existe una mejor precisión. El siguiente punto, es estudiar que no se pierdan las propiedades de los estimadores. Un estimador es insesgado cuando su sesgo es nulo, es decir, cuando su esperanza es igual al parámetro que se anhela estimar. En cuanto a la segunda propiedad se emplea el Teorema de la cota de Cramér-Rao, en el que si la varianza alcanza su cota o límite inferior se sigue manteniendo la propiedad de mínima varianza.

1.2.2. Formulación

El proyecto busca dar respuesta a la siguiente pregunta:
¿Qué técnicas son las más efectivas para la imputación de valores faltantes?

1.3. Justificación

El inicio de un análisis estadístico implica una revisión y depuración de las bases de datos a trabajar. El investigador se encuentra constantemente con falta de respuestas en el proceso de recolección de información por diversas causas, lo que afecta en la confiabilidad al procesar resultados. Es necesario comparar las diferentes técnicas de imputación de datos al medir su nivel de precisión y comprobar las propiedades de los estimadores para complementar dichos vacíos, y de esta forma investigar cómo afectan en el proceso de inferencia estadística.

1.4. Objetivos

1.4.1. Objetivo general

Medir la efectividad en términos de precisión y calidad de estimación que presentan distintas técnicas de imputación para datos faltantes, provenientes de una distribución normal.

1.4.2. Objetivos específicos

- Establecer con base a estudios previos las técnicas de imputación que se usaran en el presente trabajo.
- Determinar la precisión de ajuste que presentan las distintas técnicas de, acuerdo a la variable en estudio y el porcentaje de imputación.
- Estudiar las propiedades de estimación que presentan los nuevos datos imputados a través de algoritmos de simulación.
- Crear un R-Package que permita el fácil y comprensible usos de las distintas técnicas estudiadas.

1.5. Conceptos puntuales sobre Inferencia Estadística

- Población: Es un conjunto de sujetos homogéneos, de los que se extraen ciertas conclusiones a través del estudio de características observables. En el ámbito estadístico, población también se denomina aquella distribución que sigue la variable objeto de estudio, como ejemplo, se puede acotar que nos encontramos ante una población normal, cuando la variable analizada sigue una distribución normal (Faraldo Roca, Pedro and Pateiro López, Beatriz, 2012, p. 76).
- Parámetro: Es una cantidad numérica obtenida a partir de una población, que resume ciertas características. En la distribución normal pueden ser la media y la varianza (Faraldo Roca, Pedro and Pateiro López, Beatriz, 2012, p. 22).
- Estadístico: Es una cantidad numérica obtenida a partir de una muestra, que resumen ciertos aspectos. En la distribución normal pueden ser la media muestral y la varianza muestral (Faraldo Roca, Pedro and Pateiro López, Beatriz, 2012, p. 25).
- Estimador: Son estadísticos que, a través de análisis, aproximan un parámetro. Por ejemplo, si θ es el parámetro objeto de estudio, su estimador es denotado por $\hat{\theta}$. En la distribución normal, la media muestral actúa como estimador de la media poblacional ($\bar{X} = \hat{\mu}$) y la varianza muestral como estimador de la varianza poblacional ($s^2 = \hat{\sigma}^2$) (Faraldo Roca, Pedro and Pateiro López, Beatriz, 2012, p. 45).
- Método de muestreo: Es un proceso para elegir una muestra de la población total (también llamado universo) con el fin de poder analizar ciertos datos deseados (Faraldo Roca, Pedro and Pateiro López, Beatriz, 2012, p. 45).

1.6. Distribución Normal

Se la conoce también como la "Campana de Gauss" debido a que Carl Friedrich Gauss formuló una ecuación de la curva, en la que se destacan dos parámetros, su media μ y su desviación estándar σ, por ende, la densidad de ésta distribución es dada por:

$$\boldsymbol{f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty} \tag{1}$$

Que fija una curva similar a una campana (Pértegas Díaz and Pita Fernández, 2001, pp. 4-6).

1.6.1 Propiedades de la distribución normal:

- Posee una única moda, que coincide con su media y su mediana.
- La curva normal es asintótica al eje de abscisas. Por tanto, cualquier valor entre $-\infty$ y $+\infty$ es teóricamente posible, recalcando que el área total bajo la curva es, por ende, igual a 1.
- Presente simetría con respecto a su media μ, por tal razón, para dichas variables existe un 50% de probabilidad de observar un valor mayor que la media y otro 50% de observar un valor inferior.
- El recorrido entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación típica (σ). Cuanto mayor sea σ, más aplanada será la curva de la densidad.
- El área bajo la curva comprendido entre los valores situados aproximadamente a dos desviaciones estándar de la media es igual a 0.95.
- Los parámetros μ y σ dan la forma a la campana de Gauss, ya que, la posición de la campana es dada por la media y el grado de apuntamiento de la curva es dada por la desviación estándar.

Visto lo anterior, se aclara que existe una familia de distribuciones normales con una forma común, que difieren por los valores establecidos para la media y varianza. Como ejemplo, se tiene la distribución normal estándar que manifiesta una media 0 y varianza 1. De esta forma, la ecuación que define su densidad se puede obtener de la ecuación (1) o simplificado se tiene:

$$\boldsymbol{f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right), -\infty < z < \infty} \tag{2}$$

Se resalta que, a partir de cualquier variable X que siga una distribución $N(\mu, \sigma)$, se consigue otra característica Z con una distribución normal estándar, ejerciendo la siguiente transformación:

$$\boldsymbol{Z = \frac{X-\mu}{\sigma}} \tag{3}$$

1.7. Estimador

Un estimador es un estadístico que aproxima un parámetro. Un estadístico es cualquier función de las observaciones de una muestra aleatoria, es por lo tanto una variable aleatoria. Se denomina estimador de un parámetro θ a cualquier función de una muestra $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ que genera valores aproximados de θ, donde ese valor obtenido se determina como estimación (Casco Fernández, 2008, p. 8).

1.7.1. Propiedades de los estimadores

Las propiedades más relevantes y utilizadas son:

- Estimador insesgado (Insesgadez): Un estimador de un parámetro θ es insesgado si su valor esperado es θ, es decir, $\hat{\theta}$ es insesgado si $E[\hat{\theta}] = \theta$. Además, se recalca también que, a la diferencia $E[\hat{\theta}] - \theta$ se denomina sesgo del estimador (Casco Fernández, 2008, p. 45).

$$\boldsymbol{sesgo}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta} \qquad (4)$$

- Varianza de un estimador (Mínima varianza): Si existen varios estimadores de un parámetro, el mejor, que en otras palabras es el más eficiente, es aquel que presente menor varianza. Como ejemplo se observa que, si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores de θ, entonces:

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

Se tiene que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ (Borovkov, 1984, p. 9).

El teorema de Cramér-Rao determina que la varianza de un estimador insesgado $\hat{\theta}$ de un parámetro θ es, como mínimo:

$$\boldsymbol{VAR}(\widehat{\boldsymbol{\theta}}) \geq \frac{\mathbf{1}}{\boldsymbol{nE}\left[\left(\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\theta}}\ln \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\right)^{\mathbf{2}}\right]} \qquad (5)$$

Donde $f(x;\theta)$ representa la función de densidad de probabilidad de la muestra $X = (X_1, X_2, \ldots, X_n)^t$ en función del parámetro θ (Si un estimador alcanza esta cota mínima, entonces se dice que el estimador es de mínima varianza) (Nogales, 1998, p. 32).

1.7.2. Vector de medias

Es un vector de dimensión p, donde sus componentes son las medias de cada una de las p variables. Se obtiene a través del promedio de las medidas de cada elemento, que son vectores (Peña, 2002, p. 21).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad (6)$$

1.7.2.1. Propiedades

Según Carlos M. Cuadras (Cuadras, 1996, p. 176), las propiedades son:

- $M(x^t) = \left(M(x)\right)^t$
- Linealidad, si x es un vector estadístico p-dimensional, $b \in \mathbb{R}^m$ y $A \in Mat_{nxp}(IR)$, entonces $M(\underline{b} + AX) = \underline{b} + AM(\underline{x})$
- Aditividad, si $\underline{x}$ y $\underline{y}$ son vectores estadísticos p-dimensionales, entonces: $M(A\underline{X} + B\underline{Y}) = AM(\underline{X}) + BM(\underline{Y})$

1.7.3. La matriz de varianza

La varianza mide la variabilidad respecto a la media. La relación lineal entre dos variables se mide por la covarianza. La covarianza entre dos variables (x_j, x_k) se calcula:

$$S_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad (7)$$

Y mide su dependencia lineal. Esta información para una variable multivariante puede presentarse de esta forma:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})` \quad (8)$$

Que es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables (Cuadras, 1996, p. 23).

1.7.3.1. Propiedades

Según Carlos M. Cuadras (Cuadras, 1996), las propiedades son:

- Es semidefinida positiva, esto es: $\forall\ \underline{y} \in \mathbb{R}^p, \underline{y}^t S_{xy} \geq 0$
- $tr(S_x) \geq 0$
- Det$(S_x) \geq 0$
- Los autovectores de S_x son no negativos

1.8. Regresión Logística

(AGRESTI, 2002) La regresión logística estudia valores distribuidos binomialmente como,

$$Y_i \sim B(p_i, n_i), para\ i = 1, \dots, m,$$

Donde n_i que son los ensayos Bernoulli son conocidos y las probabilidades de éxito p_i son desconocidas. Los *logits* de las probabilidades binomiales desconocidas son modeladas como una función lineal de los X_i.

$$\boldsymbol{Logit(p) = \log\left(\frac{p}{1-p}\right)} \quad (9)$$

Como $Log_e x = Inx$ entonces,

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = \psi_0 + \sum_{j=1}^{k} \psi_j p_j$$

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = \psi_0 + \psi_1 p_1 + +\psi_2 p_2$$

La función logaritmo natural $y = Inx$ es la función inversa de la función exponencial $y = e^x$
Por tanto, la función inversa *Logit* es:

$$\boldsymbol{Logit^{-1}(p) = \frac{e^p}{1+e^p}} \quad (10)$$

1.9. Datos Faltantes

Los datos faltantes se presentan en cualquier ámbito de la investigación. Se denominan como datos no disponibles que durante un análisis resultan significativos en cuanto al reporte de resultados. Hay muchas razones específicas que permiten estos faltantes, cuyo malestar se genera porque introducen sesgos que invalidan los resultados. La pérdida se datos se han distribuido en tres escenarios distintos, los cuales son MCAR, MAR Y MNAR (Dagnino, 2014, p. 32).

1.9.1. Mecanismos de datos faltantes (Escenarios)

Como se lo mencionó, los problemas de datos faltantes son clasificados en tres escenarios (MCAR, MAR y MNAR), donde cada dato tiene una probabilidad de faltar. El proceso que representa estas probabilidades se denomina mecanismo de datos faltantes o de falta de respuesta (Van Buuren, 2018, p. 31).

1.9.1.1. MCAR

Los datos que faltan completamente al azar (MCAR) ocurren cuando la probabilidad de faltar es la misma para todos los casos, lo que implica que las razones de los valores faltantes no están relacionadas con los datos. Van Buuren brinda un ejemplo, la cual es tomar una muestra aleatoria

de una población, donde cada miembro tiene la misma posibilidad de ser incluido en la muestra (Van Buuren, 2018, p. 33).

Este escenario requiere que la probabilidad de pérdida de los datos de una variable no de dependa de otras variables medidas. De forma resumida, este mecanismo representa una situación para la cual la ausencia de datos es independiente de las variables de investigación, es exclusivamente al azar (Van Buuren, 2018, p. 33).

1.9.1.2. MAR

Los datos que faltan al azar (MAR) ocurren cuando la probabilidad de faltar es la misma solo dentro de los conjuntos establecidos por los valores observados. Un ejemplo es cuando tomamos una muestra de una población, donde la probabilidad de ser conocida depende de alguna propiedad conocida (Van Buuren, 2018, p. 37).

MAR refiere cuando la ausencia de valores se encuentra asociada a las variables independientes del estudio, pero no a la dependiente, un ejemplo claro es cuando se desarrolla un test de aptitud a unos alumnos y a los que superan una nota de corte establecida se les hace otro test más complejo mientras que a los demás no. De modo que, éstos tienen datos perdidos para la segunda variable y se debe a las observaciones de la primera (Van Buuren, 2018, p. 37).

1.9.1.3. MNAR

Existe una falta no aleatoria (MNAR) cuando ni MCAR ni MAR se mantienen. La probabilidad de faltar sucede por varias razones que se desconocen. Estos casos se dan cuando la pérdida de datos se debe a la variable dependiente, y posiblemente a alguna variable independiente. Un ejemplo es cuando en una entrevista le preguntas a un individuo por su renta mensual y éste no responde debido a que tiene una renta muy alta (Van Buuren, 2018, p. 39).

1.9.2. Patrones de perdida de datos

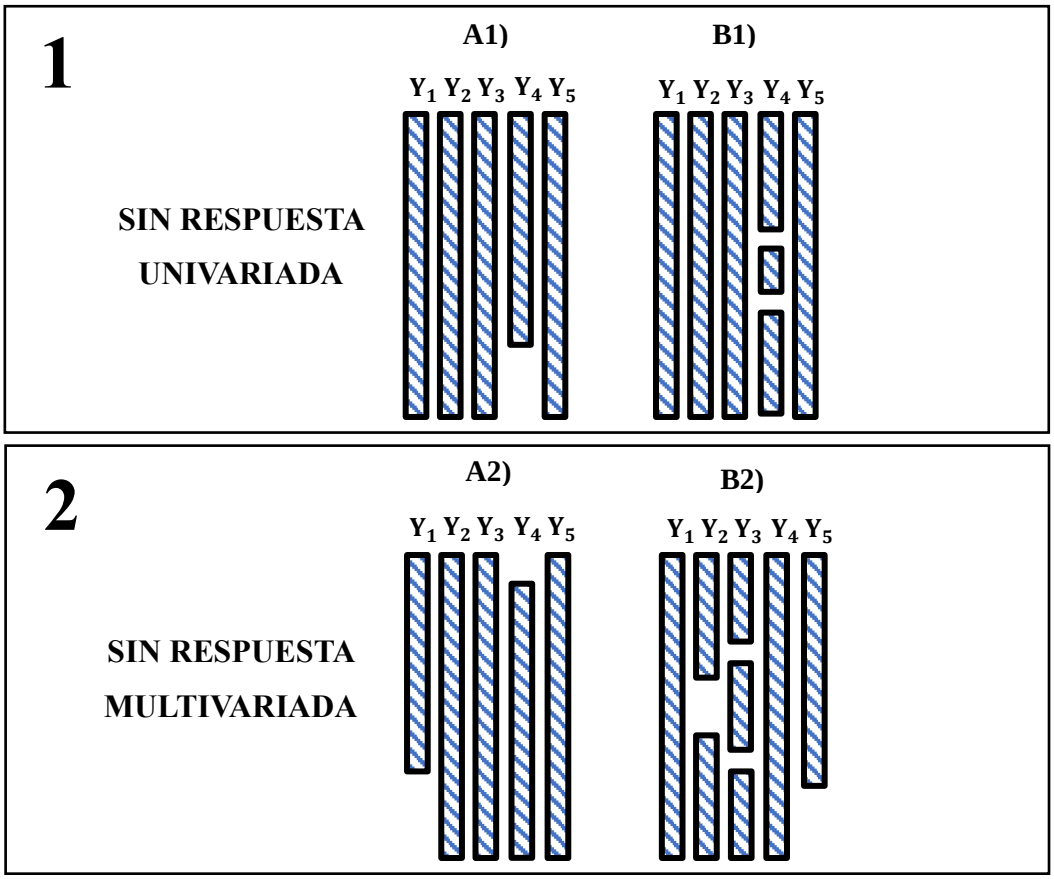


Figura 1-1: Patrones de datos faltantes

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

La pérdida de información puede seguir un patrón univariado o multivariado, siendo de diversa numerosidad de acuerdo al ámbito investigativo.

El cuadro 1 representa a los datos perdidos sin respuesta univariada, es decir, cuando únicamente existe pérdida de información en una de las variables. La figura A1) establece la falta de información ya sea al inicio o al final de la variable, pero de forma consecutiva. La figura B1) determina que la falta de datos puede existir en cualquier posición de la variable afectada.

El cuadro 2 representa a los datos perdidos sin respuesta multivariada, es decir, cuando la pérdida de información afecta a más de una variable. La figura A2) muestra la pérdida de datos ya sea al inicio o final en al menos dos variables. La figura B2) señala la existencia de datos faltantes en diversas posiciones en al menos dos variables.

1.10. Imputación

La imputación es un conjunto de técnicas que utiliza diversos medios o métodos para rellenar información faltante univariada o multivariada, de tipo cualitativa o cuantitativa. Es utilizada en todo campo para solucionar aquellos vacíos que interfieren en una investigación o toma de decisiones (Medina and Galván, 2007, p. 43).

1.10.1. Técnicas de imputación de datos faltantes

A lo largo del tiempo se han formulado diversas formas de clasificación, como las Técnicas fundamentadas plenamente en información del exterior, las cuales se sustentan en variables relacionadas con encuestas adjuntas a bases de datos ajenas o que poseen reglas previamente establecidas, entre las que se encuentran las tablas Look-up o los llamados métodos deductivos. Son muy conocidas además las Técnicas determinísticas, las cuales aparecen cuando al trabajar con múltiples unidades bajo idénticas condiciones, provocan las mismas respuestas. Dentro de este grupo se pueden nombrar a las basadas en las Medidas de tendencia central, como imputación con la media, mediana o moda, por redes neuronales o imputación fundamentadas en modelos de series de tiempo. Existen también las Técnicas aleatorias, en las que al repetir el mismo método en una unidad bajo idénticas condiciones, éstas provocan resultados diferentes porque son producto del azar. Entre las que se encuentra la imputación por regresión logística o la imputación por regresión aleatoria. Finalmente, también existe una alternativa que no se considera como tal técnica, sino un arreglo rápido para trabajar con una base de datos completa, entre ellas sobresale el Análisis de datos completos y el Análisis con datos disponibles.

De esta forma, a través de una revisión sistemática y aplicativa, se han revisado y seleccionado algunas de las técnicas principales, desglosando en cuatro grupos: Imputación por Eliminación, Imputación Simple, Imputación Múltiple e Imputación Machine Learning (Perez *et al.*, 2017, p. 32).

1.10.2. Imputación por eliminación

1.10.2.1. Análisis de datos completos (Listwise)

Este análisis reconoce los espacios o datos vacíos existentes en cualquier variable y elimina toda la fila sin importar la existencia de mucha más información, para de esta forma obtener una base de datos completa y manejable, es decir, selecciona únicamente las observaciones que disponen de información completa para todas las variables, sin embargo, dicho método puede sesgar la estimación de parámetros o acarrear inconsistencias a la investigación dado que se pierde información valiosa y reduce la dimensión de la misma, afectando gravemente al conjunto activo

de variables. Es un arreglo rápido que permite realizar estudios con una matriz de datos completa, pero no garantiza la confiabilidad de los mismos (García, 2011, p. 32).

1.10.2.2. Análisis con datos disponibles (Pairwise deletion)

Se considera como un Análisis de datos completos, pero por variable. El método elimina los espacios vacíos únicamente dentro de las variables involucradas, es decir, realiza el proceso independientemente solo en las variables con información incompleta. La ventaja es el poder utilizar toda la información disponible, pero se generan diferentes tamaños muestrales, dificultando el estudio (García, 2011, p. 6).

1.10.3. Imputación simple

La imputación simple consiste en reemplazar cada valor ausente por una estimación de dicho valor, es decir imputar cada dato ausente con un único valor. De este modo se consigue completar la matriz de datos y se trabaja con valores completos. Se realiza la imputación simple completando la base de datos y distribuyéndola posteriormente, en lugar de distribuir los datos incompletos y cada usuario verse obligado a realizar una imputación según su propósito de estudio (Van Buuren, 2018). Existen ciertas desventajas de imputar un único valor por cada dato ausente, la principal es que al asignar un único valor por cada dato faltante no se tiene presente la variabilidad asociada a la incertidumbre de los valores ausentes, generando sesgos significativos en los resultados del estudio (Van Buuren, 2018, p. 35).

1.10.3.1. Imputación de la media

La imputación derivada de la media radica en el cálculo de la media aritmética para cada una de las variables que presentan valores faltantes, y posteriormente utilizar ese valor para reemplazar todos los espacios faltantes que tiene la variable correspondiente (Van Buuren, 2018, p. 36).
Es una solución rápida para una matriz incompleta, pero esta técnica altera la distribución de ciertas maneras, por cuanto subestima la varianza, perturba las relaciones entre las variables, sesga casi cualquier estimación, sin embargo, este método es el más común (Van Buuren, 2018, p. 37).

1.10.3.2. Imputación de la mediana

La imputación por la mediana a través de los valores observados completos reemplaza el valor de la mediana en los espacios faltantes, es decir, reemplaza el valor que se encuentra en la mitad al estar ordenado de menor a mayor, generando de igual forma un serio problema de sesgo, al mismo tiempo que subestima la varianza (Parra, 1995, p. 20).

1.10.3.3. Imputación de regresión

La imputación a través de la regresión añade el conocimiento de otras variables con el objetivo de producir imputaciones más inteligentes. El método consiste en crear un modelo a partir de los datos observados, en donde, realiza predicciones para los espacios incompletos, calculando bajo el modelo ajustado establecido (Van Buuren, 2018, p. 39).

Se pueden omitir las observaciones con datos incompletos y con los completos ajustar la ecuación de regresión para predecir y sustituir dichos espacios faltantes. Sin embargo, al igual que otras técnicas tiene desventajas puesto que sobreestima la asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2. En resumen, la imputación de regresión maneja de forma artificial y a conveniencia las relaciones en los datos, es una receta para falsas relaciones positivas y espurias (Van Buuren, 2018, p. 39).

1.11. Nivel de precisión

$$\boldsymbol{P.A} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_{obs}} - \boldsymbol{y_{imp}})^{\mathbf{2}} \quad (11)$$

Se propone utilizar la formula anterior, ya que se pretende usar una medida para ver qué tan diferentes son los datos observados mediante los datos imputados, mientras mayor sea la diferencia más distintos son los datos, entonces la mejor técnica es la que presente una diferencia menor, en esta medida se refleja la dispersión con respecto a la media de los datos (Osuna, Ferreras and Núñez, 1991, p. 21).

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Tipo de investigación

La presente investigación es:

- Según el método de investigación es cuantitativa, debido a que se busca comparar técnicas de imputación de datos faltantes numéricos.
- Según el objetivo es aplicativa, ya que se enfoca en la búsqueda, aplicación y consolidación de conocimientos o teorías para dar respuesta al problema concreto.
- Según el nivel de profundización en el objeto de estudio es exploratoria – explicativa por cuanto se examina el problema y se trata de resolverla, de igual forma se trabaja con una variable explicativa.
- Según la manipulación de variables es no experimental, ya que los datos faltantes son simulados y adecuados a cada mecanismo de respuesta.
- Según el tipo de inferencia es deductiva debido que se estipulan premisas aceptadas como punto de partida de la cual se extraen conclusiones.

2.2. Análisis estadístico

2.2.1. Instrumentos de procesamiento y análisis de información

Para el desarrollo del presente trabajo de titulación, se utilizó el Entorno de Desarrollo Integrado RStudio (versión 1.2. 1335), para la simulación, procesamiento, análisis y obtención de resultados, mediante el software estadístico R (versión 3.6.3.). La codificación del trabajo es UTF-8, para la conservación del idioma español.

2.2.2. Simulación inicial de datos

A partir del método de Montecarlo se creó un conjunto de datos multivariantes (bivariantes) pertenecientes a una distribución normal con una matriz estructurada por $Y = (Y_{observadas}, Y_{missing})$, en la cual se simuló 100.000 veces, muestras representativas de tamaño $n = 5, n = 10, n = 30$ y $n = 100$ para los 3 mecanismos de falta de respuesta que son MCAR (Faltantes completamente aleatorios), MAR (Faltantes aleatorios) y MNAR (Faltantes no aleatorios).

En cada muestra simulada se aplicaron las cuatro técnicas de imputación antes mencionadas, las cuales son: por eliminación, media, mediana y regresión lineal.

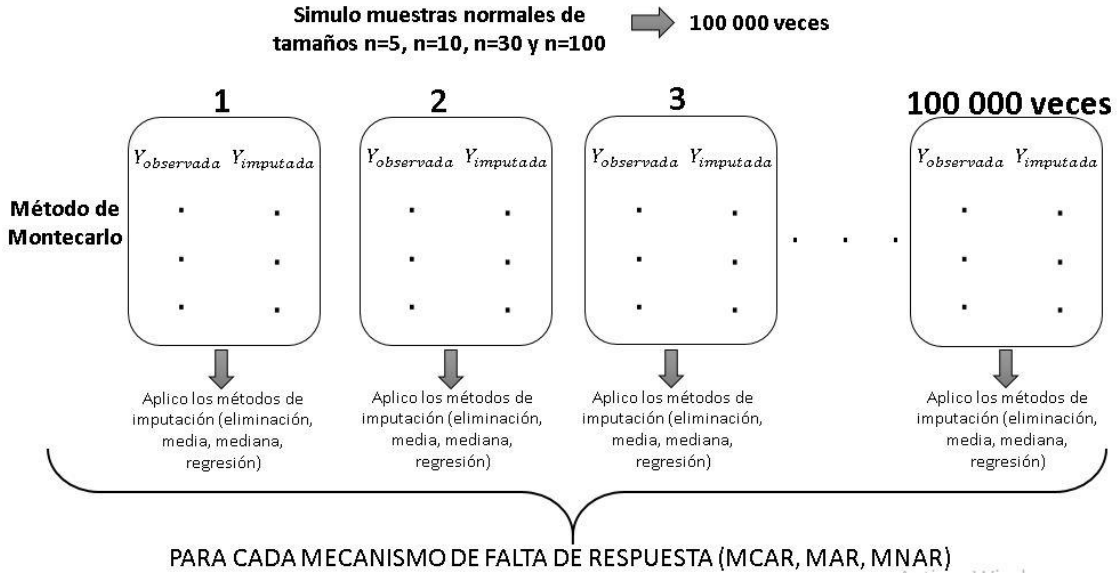


Figura 1-2: Estructura simulación e imputación de datos

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

A la nueva muestra imputada se le realiza dos estudios. Primero se diagnostica que tan bien se ajustan los datos imputados a los datos observados a través de una medida de precisión de ajuste, esto es para analizar qué tan diferentes son los datos observados de los datos imputados, sabiendo que mientras más pequeña es la diferencia más parecidos son. Se recalca que, mientras mayor sea el tamaño de simulaciones, mejora el nivel de precisión de la estimación en consecuencia, a la ley de los grandes números

El segundo análisis es el verificar si los datos imputados no pierden sus propiedades de estimación.

- La primera propiedad estudiada es la insesgadez, es decir, se comprueba si al realizar la imputación, la $\bar{x}$ y la s^2 siguen siendo buenos estimadores insesgados de μ y σ^2 respectivamente.
- Finalmente, la segunda propiedad estudiada es la de mínima varianza, la cual se comprueba a través de un límite propuesto por la cota de Cramér-Rao.

Se recalca que los dos análisis completos se realizan para cada Mecanismo de falta de respuesta. Se simuló un conjunto de datos multivariado para poder aplicar todas las técnicas de imputación, especialmente la imputación simple, ya que se necesita de una variable auxiliar que se correlacione con las demás variables, para generar iteraciones más confiables bajo un modelo. Del mismo modo se necesita de una variable auxiliar para cumplir con las propiedades del modelo

de datos perdidos y poder generarlos. En vista de que muchos métodos de estadística inferencial deben cumplir el supuesto de normalidad se decidió trabajar con un conjunto que siga esta distribución. Recordemos que este supuesto forma parte de muchos métodos estándares como el análisis factorial, componentes principales, análisis discriminante, entre otros.

2.2.3. Generación de datos faltantes

A través del libro titulado "Flexible Imputation of Missing Data" propuesto por Stef van Buuren, vamos a estudiar la distribución de probabilidad de la posición en donde se van a generar los valores perdidos (missing) en la base de datos. Mediante esta distribución de probabilidad podemos simular los 3 tipos de mecanismos de falta de respuesta los cuales son MCAR, MAR y MNAR.

La matriz R guarda las posiciones de los missings en la variable Y. La distribución de R depende de $Y = (Y_{observados}, Y_{missing})$. Si ψ contiene los parámetros de los datos perdidos, entonces la expresión general del modelo de datos perdidos es:

$$P(R|Y_{obs}, Y_{mis}, \psi)$$

Entonces en general decimos que R es una distribución de probabilidad que va a depender de los Y_{obs}, de los Y_{mis} y de ψ, donde éste último, es la parte totalmente aleatoria, la parte que no podemos modelar.

- Los datos son MCAR si:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|\ \psi)$$

La probabilidad de encontrar un missing depende solo del parámetro ψ.

- Los datos son MAR si:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|Y_{obs}, \psi)$$

La probabilidad de encontrar un missing depende de la información observada.

- Los datos son MNAR si:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi)$$

La probabilidad de encontrar un missing depende de la información observada y perdida.

Sabiendo lo anterior, se recalca que los datos $Y = (Y_1, Y_2)$, donde Y_1 son los datos observados y Y_2 los datos perdidos, se extraen de una distribución normal bivariada, entonces los missing se crean en Y_2 usando el modelo de datos perdidos siguiente:

$$\boldsymbol{P(R_2 = 0) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}} \psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}} \psi_2} \qquad (12)$$

Con diferentes ajustes de parámetros para la tripla formada por $\psi = (\psi_0, \psi_1, \psi_2)$.

Ahora fijamos los modelos individuales con esta nueva expresión.

- Para MCAR:

En el modelo podemos lograr que la probabilidad de $R_2 = 0$ dependa solo de ψ_0 realizando lo siguiente:

$$\begin{matrix} \psi_0 & \psi_1 & \psi_2 \\ \downarrow & \downarrow & \downarrow \end{matrix}$$
$$\psi_{MCAR} = (0.5, 0, 0)$$

entonces se obtiene el modelo:

$$\boldsymbol{MCAR: P(R_2 = 0) = 0.5}$$

Demostración:

$$P(R_2 = 0) = 0.5 + \frac{e^{Y_1}}{1 + e^{Y_1}} 0 + \frac{e^{Y_2}}{1 + e^{Y_2}} 0$$
$$P(R_2 = 0) = 0.5$$

Se concluye que los datos completamente aleatorios MCAR no dependen ni de los datos observados ni de los perdidos, entonces para calcular las posiciones donde se van a colocar los missing simplemente depende de una distribución de Bernoulli con probabilidad 0.5, se generan aleatoriamente números de 0 a 1 indistintamente, donde la probabilidad de tener 0 es del 50% y la probabilidad de tener 1 también es del 50%.

Nota: Para calcular otros porcentajes solo se modifica el ψ_0.

- Para MAR:

Un dato es missing aleatorio si la probabilidad de encontrar un valor perdido depende de los datos observados, por lo tanto, tenemos que:

$$\psi_{MAR} = (0, 1, 0)$$

Se obtiene el modelo:

$$\boldsymbol{MAR: logit(P(R_2 = 0)) = Y_1}$$

Donde $logit(p) = logit\left(\frac{p}{1-p}\right)$ *para cualquier* $0 < p < 1$ *es la función logit. En la práctica, es más conveniente trabajar con la función inversa logit:* $logit^{-1}(x) = \left(\frac{e^x}{1+e^x}\right)$

Demostración:

$$P(R_2 = 0) = 0 + \frac{e^{Y_1}}{1 + e^{Y_1}} * 1 + \frac{e^{Y_2}}{1 + e^{Y_2}} 0$$
$$P(R_2 = 0) = \frac{e^{Y_1}}{1 + e^{Y_1}} * 1$$

Se sabe que es una función exponencial, entonces para volver la variable a la original se calcula la función logit:

$$logit(P(R_2 = 0)) = logit\left(\frac{e^{Y_1}}{1 + e^{Y_1}} * 1\right)$$

Se utiliza logit porque la distribución normal que sigue una Y_1 es una distribución completamente asimétrica, quiere decir que, alrededor de la media el 50% de estos valores están a la derecha y el

otro 50% a la izquierda, entonces al sacar logit todos los valores de Y_1 de esta razón se convierten en valores entre 0 y 1 y envía el 50% de los datos a la derecha y el otro 50% a la izquierda.

$$logit\big(P(R_2 = 0)\big) = logit\left(\frac{e^{Y_1}}{1 + e^{Y_1}} * 1\right) \approx 0.5$$ ← p "porcentaje de missing"

Nota: Si quiero tener otro porcentaje de missing realizo el cambio de $\psi_1 = 1$, realizando:

$$0.5\,\psi_1 = p$$

$$\psi_1 = \frac{p}{0.5}$$

Por ejemplo, si quiero que $p = 0.2$ entonces:

$$\psi_1 = \frac{0.2}{0.5} = 0.4$$

Por lo tanto, la expresión general sería:

$$logit\left(\frac{e^{Y_1}}{1 + e^{Y_1}} * 0.4\right) \approx 0.2$$

El ejemplo concluye que existirá un 20% de 0 (datos perdidos) y un 80% de 1 (datos completos). De esta forma depende de p el porcentaje, donde éste debe ser ≤ 0.5 (50%) para que no exceda la probabilidad (1).

- Para MNAR:

$$\psi_{MNAR} = (0, 0, 1)$$

Se obtiene el modelo:

$$\boldsymbol{MNAR: logit(P(R_2 = 0)) = Y_2}$$

La demostración es la misma que en el caso anterior pero ahora trabajando con:

$$P(R_2 = 0) = \cancel{0} + \frac{e^{Y_1}}{1 + \cancel{e^{Y_1}}}\cancel{0} + \frac{e^{Y_2}}{1 + e^{Y_2}} * 1$$

$$P(R_2 = 0) = \frac{e^{Y_2}}{1 + e^{Y_2}} * 1$$

Nota: A la expresión de igual forma se le construye el modelo logit y se calcula. Se recalca que para cambiar el porcentaje de missing se debe cambiar ahora el $\psi_2 = 1$.

2.2.4. Porcentaje de valores perdidos

El porcentaje de valores perdidos adecuado o tratable para no considerarlo como una mala recolección de información es discutida por el investigador. El porcentaje de pérdidas depende de la magnitud del estudio, es decir, para las ciencias médicas, el factor más importante es la precisión, por lo que se debería considerar un porcentaje de pérdidas mínimo, en cambio para las ciencias sociales se puede trabajar con porcentajes de pérdidas más altas, está decisión depende

solo del investigador. Se ha decidido trabajar con porcentajes de pérdidas de datos como: 0.05, 0.1, 0.25 y 0.5.

2.2.5. Técnicas de imputación de datos

El presente trabajo de titulación tiene como objetivo principal medir la efectividad de distintas técnicas de imputación, con esto en mente se decidió trabajar con técnicas que los investigadores usan comúnmente sin conocer si son factibles de aplicarlos o no, también si se pueden aplicar a un conjunto de datos que provenga de una distribución normal.

Las técnicas de imputación que se aplicaron a los conjuntos de datos MCAR, MAR y MNAR fueron:

- Eliminación por listas, se encargó de eliminar los datos perdidos y trabajar sin ellos.
- La imputación de la media, esta técnica busca los datos perdidos y los reemplaza por la media aritmética.
- La imputación de la mediana, esta técnica busca los datos perdidos y los reemplaza por la mediana.
- Regresión lineal, en esta técnica se construyó el modelo lineal solo de los datos observados, y se predijo los datos faltantes mediante el modelo.

Se trabajaron con estas técnicas de imputación debido a que son las más utilizadas, al momento que un investigador se encuentra con una base de datos que presenta valores perdidos por lo general sabe usar una de estas técnicas y mediante este trabajo se compara para el mejor uso de cada una de ellas.

2.2.6. Nivel de precisión

El nivel de precisión se midió mediante la fórmula (11) de la sección del Marco teórico esta medida nos indica que tan diferentes son los datos observados a comparación de los datos imputados, se interpreta mediante la técnica que arroje una diferencia menor y la dispersión con respecto a la media de los datos.

2.2.7. Propiedades de estimadores

En este paso se verifica si los estimadores siguen manteniendo sus propiedades (insesgadez y mínima varianza).

Los estimadores son:

- $\overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}$ (13)

- $\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ (14)

2.2.7.1. Insesgadez

El sesgo del estimador es la diferencia entre su esperanza matemática y el parámetro poblacional, si el sesgo del estimador es nulo se lo denomina estimador insesgado.

Se trabajó con dos estimadores ($\bar{x}$ y s^2)

- $E(\overline{x}) - \mu = 0$
- $E(s^2) - \sigma^2 = 0$

2.2.7.2. Mínima Varianza

Otra de las propiedades de los estimadores es la mínima varianza, donde la Cota de Cramér-Rao expresa una cota inferior para la varianza de un estimador insesgado.

El teorema de la cota de Cramér-Rao dice:

Sea $X_1, X_2, \dots, X_m \sim v.a.i.i.d$ con función de probabilidad $f(x;\theta)$. Si $\hat{\theta}$ es un estimador insesgado de θ, entonces se cumple que:

$$\boldsymbol{VAR(\widehat{\theta}) \geq \frac{1}{n\,E\left(\frac{\partial}{\partial\theta} Ln\,(f(x;\theta))^2\right)} = \frac{1}{-n\,E\left(\frac{\partial^2}{\partial\theta^2} Lnf(x;\theta)\right)}} \quad (15)$$

Si hay dos estimadores insesgados, el mejor de ellos es el que presenta una varianza inferior.

- Cota de Cramer Rao para el estimador insesgado $\overline{x}$:

Se demuestra que la $\overline{x}$ alcanza la cota de Cramér-Rao:

La cota de la $\overline{x}$, $VAR\,(\overline{x}) = \frac{\sigma^2}{n}$, entonces:

$$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2): f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$VAR(\hat{\mu}) \geq \frac{1}{-nE(\frac{\partial^2}{\partial\mu^2} \ln(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}))}$$

$$VAR(\hat{\mu}) \geq \frac{1}{-nE(\frac{\partial^2}{\partial\mu^2} \ln(\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x-\mu)^2}{2\sigma^2}))}$$

$$VAR(\hat{\mu}) \geq \frac{1}{-nE(\frac{\partial}{\partial\mu}(+\frac{(x-\mu)}{\sigma^2}))}$$

$$VAR(\hat{\mu}) \geq \frac{1}{-nE(\frac{-1}{\sigma^2})}$$

$$VAR(\hat{\mu}) \geq \frac{1}{\frac{n}{\sigma^2}}$$

$$VAR(\hat{\mu}) \geq \frac{\sigma^2}{n}$$

No existe un estimador de $\hat{\mu}$ que tenga una varianza inferior que $\frac{\sigma^2}{n}$, entonces la varianza de la media alcanza la Cota de Cramer Rao.

- Cota de Cramer Rao para el estimador insesgado σ^2:

Sea $X_1, X_2, \dots, X_n$ $una\ va\ i.i.d \sim N(\mu, \sigma^2)$, la cota de Cramér-Rao para σ^2 es:

$$= E\left[\frac{\partial^2}{\partial(\sigma^2)^2} \ln(\frac{1}{\sqrt{2\pi}} . \frac{1}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right]$$

$$= E\left[\frac{\partial^2}{\partial(\sigma^2)^2} \ (\ln(\frac{1}{\sqrt{2\pi}}) + \ln(\frac{1}{(\sigma^2)^{\frac{1}{2}}}) - \frac{(x-\mu)^2}{2\sigma^2})\right]$$

$$= E\left[\frac{\partial}{\partial(\sigma^2)} \ (-(\sigma^2)^{\frac{1}{2}}(\frac{1}{2(\sigma^2)^{\frac{1}{2}}}) + \frac{(x-\mu)^2}{2(\sigma^2)^2})\right]$$

$$= E\left(\frac{\partial}{\partial\sigma^2} \ (-\frac{1}{2\sigma^2}) + \frac{(x-\mu)^2}{2\sigma^4})\right)$$

$$= E\left(\frac{1}{2\sigma^4} - \frac{(x-\mu)^2 2\sigma^2}{2(\sigma^4)^2}\right)$$

$$= E\left(\frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}\right)$$

$$= E\left(\frac{1}{2\sigma^4} - \frac{E(x-\mu)^2}{\sigma^6}\right)$$

$$= \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6}$$

$$= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4}$$

$$= -\frac{1-2}{2\sigma^4}$$

$$= -\frac{1}{2\sigma^4}$$

Por tanto:

$$VAR(\hat{\sigma}^2) \geq \frac{1}{-n(-\frac{1}{2\sigma^4})}$$

$$VAR(\hat{\sigma}^2) \geq \frac{1}{\frac{n}{2\sigma^4}}$$

$$VAR(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}$$

NOTA:

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ tiene varianza → $VAR(s^2) = \frac{2\sigma^4}{n-1}$ **Es insesgado pero no alcanza la cota de cramér-rao (minima varianza) porque** $\frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$
- $s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ tiene varianza → $VAR(s^{*2}) = \frac{2\sigma^4}{n}$ **Es insesgado y si alcanza la cota de cramér-rao (minima varianza) pero el valor de μ en la vida real no se lo conoce.**

2.2.8. Visualización de resultados

Los resultados se visualizan mediante gráficos del paquete ggplot2 del software estadístico R. Se presentaron casos donde se tuvo que obtener el logaritmo neperiano (ln) de los resultados, ya que estos presentaron una escala superior a la del gráfico.

Se realizaron gráficos para la presentación de la medida de precisión de ajuste y para conocer si cumplen con las propiedades de insesgadez y mínima varianza.

CAPÍTULO III

3. RESULTADOS Y DISCUSIÓN DE LOS RESULTADOS

3.1. Insesgadez

3.1.1. Insesgadez de la media

El estimador de la media aritmética cuando no presenta datos faltantes es un estimador insesgado de la media poblacional, es decir, su sesgo es nulo. Se comprueba si se mantiene esta propiedad a partir de la aplicación de los diferentes escenarios, el tamaño de la muestra y la proporción de datos faltantes.

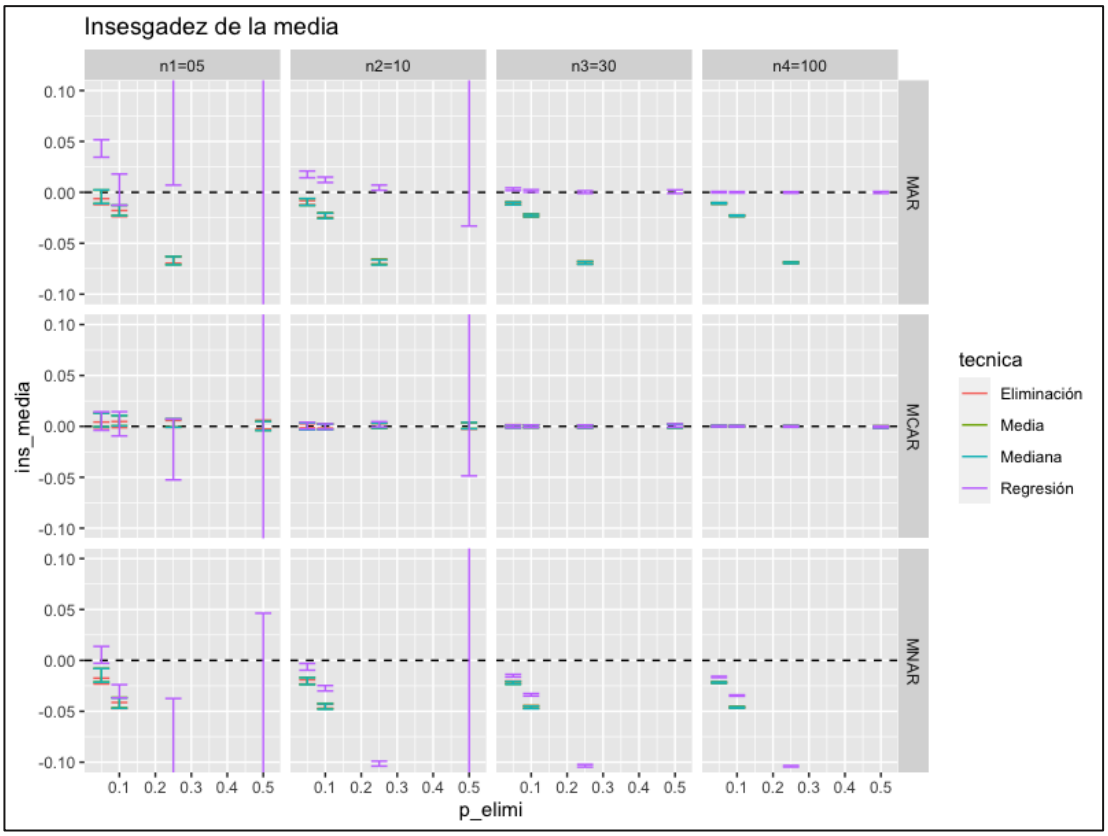


Figura 1-3: Intervalos de confianza de la media. Insesgadez de la media

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

En la Figura 1-3 se puede visualizar en cada gráfico una línea entrecortada en el eje $y = 0$, donde nos indica que, si el intervalo de confianza del estimador cae dentro de esta constante cumple con la propiedad de insesgadez.

Entonces:

- En el escenario MAR, la técnica más viable de imputación es la Regresión, ya que cumple con la propiedad de insesgadez en las muestra $n = 100$ con una pérdida de datos del 5%, 10%, 25% y 50%, en la muestra $n = 30$ con un 10%, 25% y 50%, en $n = 5$ y $n = 10$ lo hace para el 50%. Mientras que en las técnicas de imputación de la media y mediana mantienen esta propiedad con un 5% de pérdida de datos en la muestra $n = 5$.
- Para el escenario MCAR, las técnicas de imputación de eliminación, media, mediana y regresión mantienen la propiedad de insesgadez del estimador de la media para muestras de $n = 5, n = 10, n = 30$ y $n = 100$ con una pérdida de datos 5%, 10%, 25% y 50%.
- En el escenario MNAR el comportamiento es diferente, la única técnica y la más viable es la Regresión, debido a que cumple con la propiedad en la muestra $n = 5$ con un 10% y 50% de datos perdidos y en $n = 10$ con el 50%.

3.1.2. Insesgadez de la varianza

La varianza es un estimador insesgado de la varianza poblacional cuando no presenta ningún dato faltante y su distribución es normal, aplicando diferentes proporciones de datos faltantes, tamaños de muestra y escenarios de datos se comprueba si la insesgadez de la varianza se mantiene.

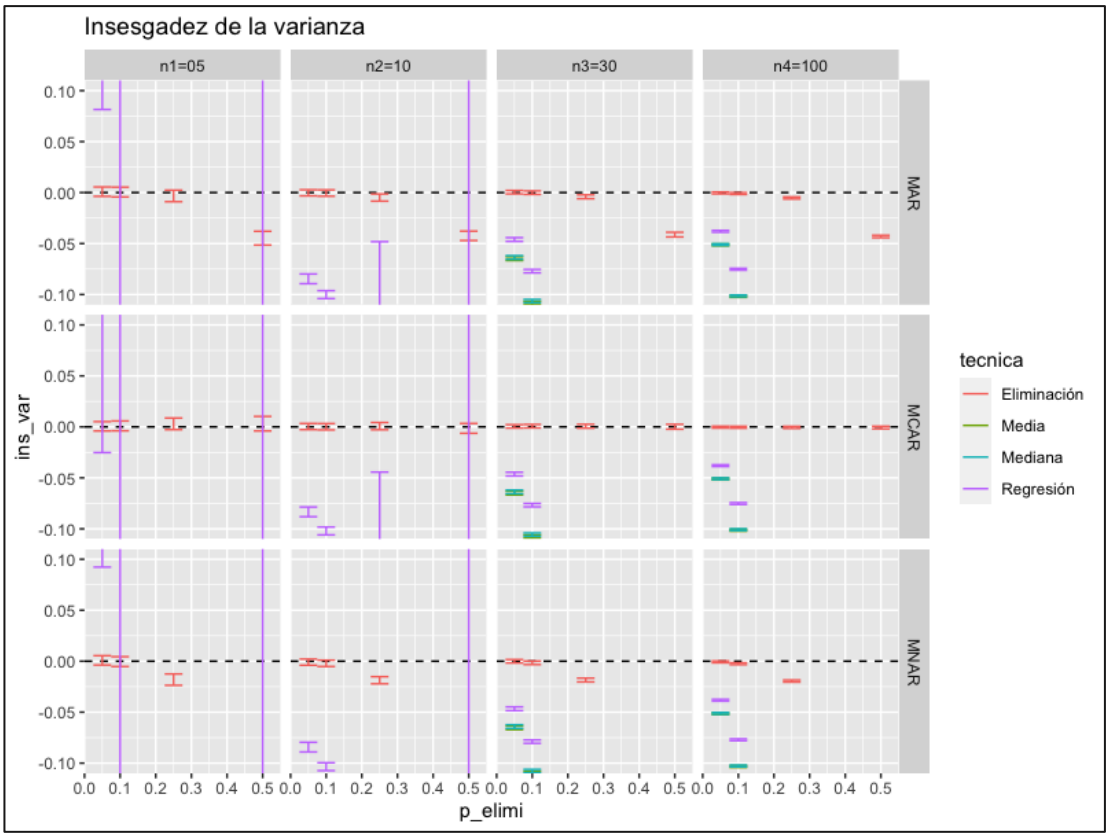


Figura 2-3: Intervalos de confianza de varianza. Insesgadez de la varianza

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

En la figura 2-3 se puede observar:

- Para el escenario MAR, la técnica de imputación de eliminación arrojó mejores resultados, ya que cumple con la propiedad en la muestra $n = 5$ con una proporción de pérdida de datos del 5%, 10%, y 25% y en las muestras $n = 10, n = 30\ y\ n = 100$ con un 5% y 10%, en la técnica de Regresión se mantiene la propiedad $n = 5$ para un 10% y 50% de datos faltantes y en $n = 10$ para un 50%, mientras que en las técnicas restantes no mantienen esta propiedad debido a una posible inflación en la varianza al momento de la simulación de los datos perdidos.
- En cuanto al escenario MCAR, el comportamiento es similar a MAR, pero en este caso técnica de eliminación mantiene la propiedad de insesgadez en las muestras n=5, $n = 10, n = 30\ y\ n = 100$ con un 5%, 10%, 25% y 50% de datos perdidos.
- En el caso de MNAR, la técnica de eliminación a comparación con las demás mantiene la propiedad en las muestras de n=5, $n = 10, n = 30\ y\ n = 100$ con un 5% y 10% de datos pérdidos y en la técnica de regresión se mantiene la propiedad $n = 5$ para un 10% y 50% de datos faltantes y en $n = 10$ para un 50%.

3.2. Mínima varianza

La mínima varianza es la siguiente propiedad a comparar con diferentes tamaños de muestra, proporción de datos perdidos y escenarios.

3.2.1. Variabilidad de la media

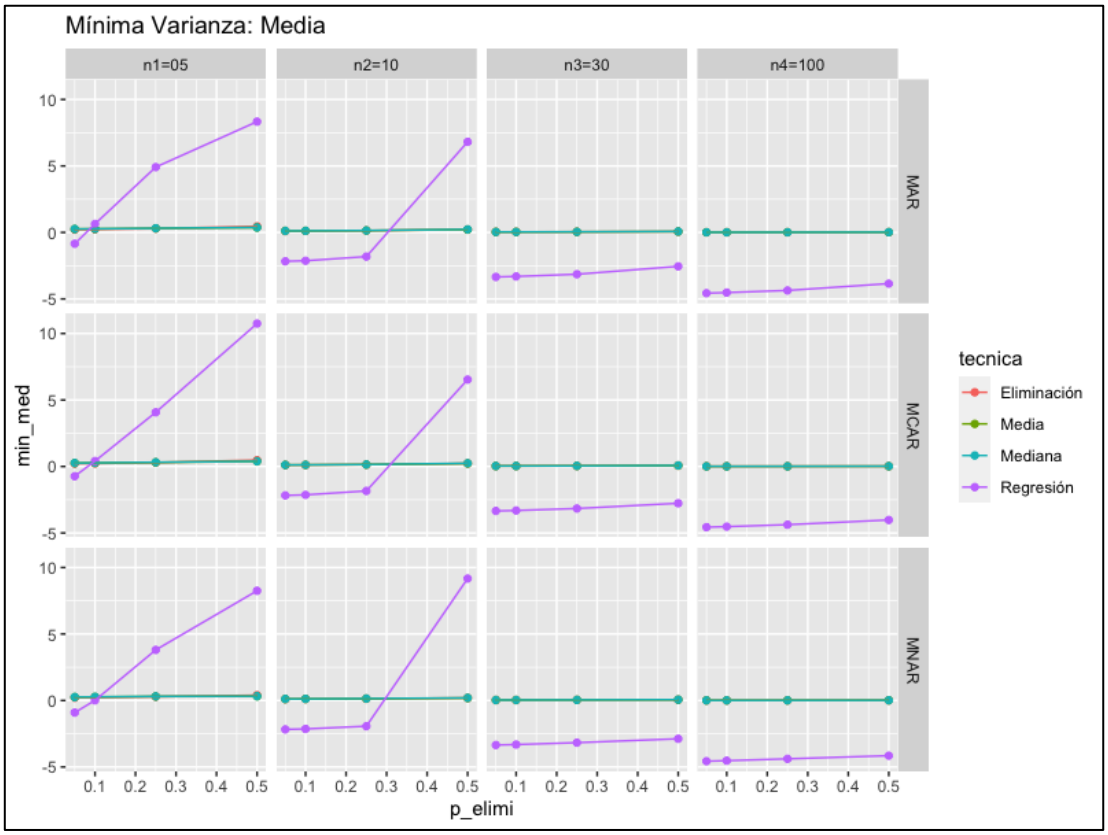


Figura 3-3: Mínima de la varianza. Variabilidad de la media

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

En la figura 3-3. se puede observar que:

- Para los tres escenarios de MAR, MCAR y MNAR, la técnica de regresión arroja una menor varianza en las muestras $n = 30$ y $n = 100$ con un 5%, 10%, 25% y 50% de datos perdidos, en la muestra $n = 10$ con un 5%, 10% y 25% y en la muestra de $n = 5$ con un 5%. Mientras que en las técnicas de eliminación, media y mediana arrojan una varianza inferior en $n = 5$ con un 10%, y 25% de datos faltantes. Hay que tener en cuenta que conforme aumenta el porcentaje de datos faltantes la variabilidad de la media va ascendiendo y en este caso se considera a la regresión una posible mejor técnica de imputación.

3.2.2. Variabilidad de la varianza

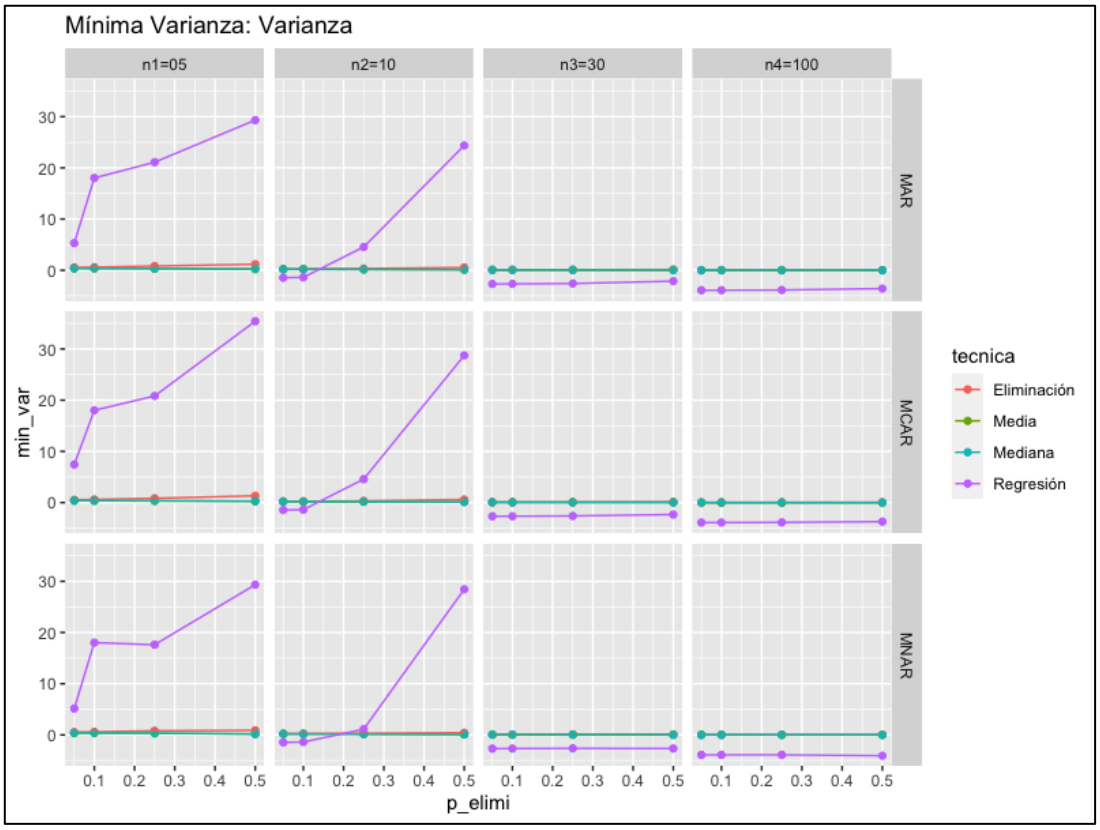


Figura 4-3: Mínima varianza. Variabilidad de la varianza

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

En la figura 3-4.:

- Se observa que en los escenarios MAR, MCAR y MNAR, la técnica de regresión arroja una varianza inferior en las muestras $n = 30\ y\ n = 100$ con un 5%, 10%, 25% y 50% de datos perdidos y en $n = 10$ con un 5% y 10%, considerando una mejor técnica a comparación de las demás, ya que las técnicas de eliminación, media y mediana arrojan una varianza mínima en la muestra $n = 5$ con un 5%, 10%, 25% y 50% de datos perdidos. Conforme aumenta el porcentaje de pérdida de datos aumenta la variabilidad de la varianza.

3.3. Medida de precisión de ajuste

La medida de precisión de ajuste nos indica la diferencia de los datos observados y los datos imputados, se considera la mejor técnica de imputación de datos que arroje una medida inferior.

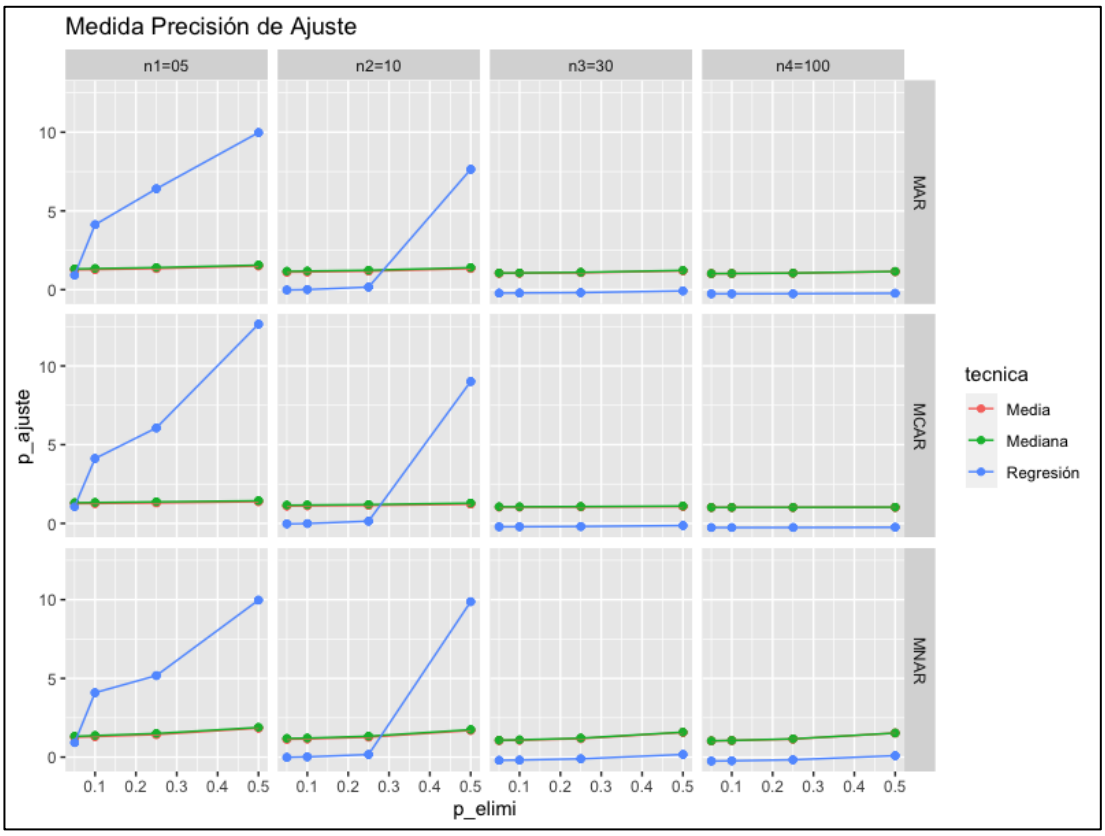


Figura 5-3: Medida de precisión de ajuste

Realizado por: Jamilton Vinueza y Galo Masaquiza, 2021.

Se puede observar en la figura 3-5. que:

- En los escenarios MAR, MCAR y MNAR, la técnica de regresión arroja una medida de precisión inferior en las muestras $n = 30$ y $n = 100$ con un 5%, 10%, 25% y 50% de datos perdidos, en $n = 10$ con un 5%, 10% y 25% y $n = 5$ con un 5%. Se considera la técnica más efectiva, ya que en las técnicas de imputación de la media y mediana se visualiza una medida superior, pero en la muestra $n = 5$ con un 10%, 25% y 50% de datos perdidos esta medida inferior a la técnica de Regresión, se puede observar que conforme aumenta el porcentaje de pérdida de datos aumenta la medida de precisión es menos efectiva.

CONCLUSIONES

Los estudios realizados por distinguidos autores como Abdul Lateef Bello, Fernando Medina, Marco Galván, entre otros., tienen un objetivo semejante al presente proyecto, al trabajar con diversos métodos de imputación en situaciones de incertidumbre, sin embargo, manejan otra metodología. Entre los investigadores coinciden algunas de las técnicas empleadas en sus estudios, pero concluyen que la selección de éstas es a criterio propio, por lo tanto, las técnicas de imputación seleccionadas son: Eliminación por listas, imputación de la media, imputación de la mediana y regresión lineal.

Mediante el análisis de la medida de precisión de ajuste, se determina que en los escenarios MAR, MCAR y MNAR la técnica de regresión lineal es la más idónea para trabajar con muestras a partir de $n = 10$. Para muestras como $n = 5$ la técnica de imputación más precisa es la media y la mediana, debido a que arrojó un valor más cercano a los datos reales.

Al estudiar las propiedades de los estimadores, se concluye que, respecto a la insesgadez de la media, la mejor técnica en los escenarios de MAR, MCAR y MNAR es la imputación por regresión lineal, debido a que sus propiedades se mantienen en muestras a partir de $n = 30$ para niveles de contaminación de 5%, 10%, 25% y 50% respectivamente, mientras que en las técnicas restantes no. En la insesgadez de la varianza la técnica más viable es la eliminación en los escenarios de MAR y MCAR para muestras de $n = 30\ y\ n = 100$ con el 5%, 10%, 25% y 50% de datos faltantes, para MNAR en muestras de $n = 5, n = 10, n = 30\ y\ n = 100$ con un 5% y 10%. Conforme a la mínima varianza de la media y varianza, la técnica que arrojó una varianza inferior en la mayoría de contextos como $n = 10\ y\ n = 30$ para datos con escenarios de MAR, MCAR y MNAR es también la técnica de regresión lineal.

Al final se desarrolló satisfactoriamente el paquete propuesto, con nombre “IDF” el cual replica n veces un conjunto de p variables de naturaleza cuantitativa que sigue una distribución normal, a cada conjunto simula datos faltantes dependiendo de la proporción que el usuario quiera trabajar. Una vez que genera los conjuntos con los datos faltantes, el paquete usa técnicas de imputación como la eliminación, la media, mediana y regresión lineal. Finalmente, el paquete obtiene una medida de precisión de ajuste que sirve para comparar cada técnica y obtener la mejor, también comprueba si cumplen con las propiedades de insesgadez y mínima varianza para los estimadores de la media y varianza.

RECOMENDACIONES

Se sugiere ampliar este estudio, trabajando con técnicas de imputación múltiple y de machine learning, para compararlas y diagnosticar mejores técnicas para su uso en diversos campos de investigación.

En vista de que se presentó un comportamiento diferente en la técnica de regresión lineal, se aconseja profundizar esta práctica con una base de datos con mayor número de variables.

Si un investigador desea imputar una matriz con muestras de $n \geq 30$ y escenarios MAR, MCAR y MNAR se recomienda utilizar la técnica de regresión lineal, manejando porcentajes de errores considerables. Para muestras pequeñas con un bajo porcentaje de datos perdidos se aconseja utilizar la técnica de la media o mediana, dado que en ambos casos las respectivas técnicas arrojan una precisión de ajuste más exacta y por ello, los datos imputados se acercan más a los datos reales.

Para comprobar la propiedad de insesgadez de la media se sugiere trabajar con la técnica de regresión lineal en escenarios de MAR y MCAR, debido a que su propiedad de insesgadez se sigue manteniendo al exponerlo en diversos parámetros. En el caso de la insesgadez de la varianza, la técnica recomendable es la eliminación para escenarios de MAR, MCAR y MNAR en muestras pequeñas.

GLOSARIO

Missing data: Término en inglés que se define como datos faltantes o valores que no están disponibles y que serían significativos para un posible análisis si fueran observados (Ware *et al.*, 2012).

Machine learning: Término en inglés que se define como aprendizaje automático, la cual es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante (El Naqa and Murphy, 2015, p. 32).

Mecanismo de falta de respuesta MCAR: Siglas de *Missing completely at random* que significa Pérdida completamente al azar. Refiere a que la probabilidad de faltar es la misma para todos los casos, por ende, las causas de los datos faltantes no están relacionadas con los datos (Van Buuren, 2018, p. 34).

Mecanismo de falta de respuesta MAR: Siglas de *Missing at random* que significa Pérdida al azar. Refiere a que la probabilidad de faltar es la misma solo dentro de los grupos definidos por los datos observados (Van Buuren, 2018, p. 65).

Mecanismo de falta de respuesta MNAR: Siglas de *Not missing at random* que significa Pérdida no al azar. Refiere a que la probabilidad de faltar varía por razones que desconocemos (Van Buuren, 2018, p. 23).

Simulación de Montecarlo: Es una técnica que combina conceptos estadísticos (muestreo aleatorio) con la capacidad que tienen los ordenadores para generar números pseudo-aleatorios y automatizar cálculos (Javier Faulín, 2005, p. 23).

Remuestreo bootstrap: Es una técnica de remuestreo que se fundamenta en la distribución Empírica y la simulación de Montecarlo, lo que permite obtener estimaciones de medidas de precisión, así como la realización de pruebas de hipótesis (Serrano Carvajal, 2018, p. 43).

BIBLIOGRAFÍA

AGRESTI, A. (2002) *Categorical Data Analysis*. New Jersey: JHON WILEY & SONS, INC. Available at: https://books.google.es/books?hl=es&lr=&id=hpEzw4T0sPUC&oi=fnd&pg=PR7&dq=Agresti,+Alan.+(2002).+Categorical+Data+Analysis.+New+York:+Wiley-Interscience.+ISBN+0-471-360937.&ots=nsCh9odpQ5&sig=ScpT4mqmQoWe4vJUvwUUYXB5t9o#v=onepage&q&f=false.

AMÓN URIBE, I. (2010) 'Guía metodológica para la selección de técnicas de depuración de datos', *Escuela de Ingeniería de Materiales* [Preprint].

BATANERO, C. AND GODINO, J. (2001) 'Análisis de datos y su didáctica', *Departamento de Didáctica de la Matemática de la Universidad de Granada* [Preprint]. Available at: https://d1wqtxts1xzle7.cloudfront.net/38741046/ANALISIS_DE_DATOS_Y_SU_DIDACTICA.pdf?1442013970=&response-content disposition=inline%3B+filename%3DANALISIS_DE_DATOS_Y_SU_DIDACTICA_Profeso.pdf&Expires=1626755913&Signature=SunUeCw-fXDiSHYRLNJficaK9ukwDU8mH6TERxO0YIo1rmVVBYEjkL5apOl380NiPIJGkmJlK6s23YBfHd21xLz1pR61JPzV06V4PIb0Py9q5bWUzFWadtIx0DsWvpVduym1jft5h3ReZL8KmXXmrKIkSMPRF71Weyp5wxW7hWkv1mnXKEUFInkyys2todf~j8IEk~tCWrIcuUSZhWrZP3hmwtV1n33OFL5UhZru64mtmf4D7X2YdfW6TMvkDOhtMN7KEvnfp4HWnETCUd8skkEe1Y~YJUDat9~R~9mzhu6x7v2LgQTuwr6GiiH4aniSFBOYuxxSBOjZcJJDQ6ykOw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

BELLO, A.L. (1993) 'Choosing among imputation techniques for incomplete multivariate data: a simulation study', *Communications in Statistics-Theory and Methods*, 22(3), pp. 853–877.

BOROVKOV, A.A. (1984) *Estadística matemática*.

CASCO FERNÁNDEZ, I. (2008) 'Estimación de parámetro', in. Madrid: Universidad Carlos III de Madrid, pp. 1–6. Available at: http://halweb.uc3m.es/esp/Personal/personas/icascos/esp/resumen_muestreo.pdf.

CASTRO, L.M.U. AND ÁVILA, D.M.M. (2006) 'Una introducción a la imputación de valores perdidos', *Terra. Nueva Etapa*, 22(31), pp. 127–151.

CUADRAS, C.M. (1996) *Nuevos métodos de análisis multivariante*. CMC Edicions.

DAGNINO, J. (2014) 'Datos faltantes (missing values)', *Rev Chil Anest*, 43, pp. 332–334.

EL NAQA, I. AND MURPHY, M.J. (2015) 'What Is Machine Learning?', in El Naqa, I., Li, R., and Murphy, M.J. (eds) *Machine Learning in Radiation Oncology: Theory and Applications*. Cham: Springer International Publishing, pp. 3–11. doi:10.1007/978-3-319-18305-3_1.

FARALDO ROCA, PEDRO AND PATEIRO LÓPEZ, BEATRIZ (2012) *Estadística y metodología de la investigación*. España: USC UNIVERSIDAD SANTIAGO DE COMPOSTELA. Available at: http://eio.usc.es/eipc1/base/basemaster/formularios-php-dpto/materiales/mat_g2021103120_estadisticatema4.pdf.

GARCIA, D.O. (2011) 'Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo'. Available at: http://eamo.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_616.pdf.

GOICOECHEA, A.P. (2002) 'Imputación basada en árboles de clasificación', *Eustat. Available in: http://www. eustat. es/documentos/datos/ct*, 4.

HEMEL, J.B. *ET AL.* (1987) 'Stepwise deletion: a technique for missing-data handling in multivariate analysis', *Analytica chimica acta*, 193, pp. 255–268.

JAVIER FAULÍN, Á.A.J. (2005) *Simulación de Monte Carlo con Excel, http://www.cyta.com.ar*. Técnica Administraitva, ISSN 1666-1680. Available at: http://www.cyta.com.ar/biblioteca/bddoc/bdlibros/monte_carlo/monte_carlo.htm?iframe=true&width=95%&height=95% (Accessed: 12 July 2021).

LOHR, S.L., VELASCO, P. AND ALFREDOTR, O. (2000) *Muestreodiseño y análisis*.

MEDINA, F. AND GALVÁN, M. (2007) *Imputación de datos: teoría y práctica*. Cepal. Available at: https://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590_es.pdf?sequence=1&isAllowed=y.

NOGALES, A.G. (1998) *Estadística matemática*. Universidad de Extremadura. Servicio de Publicaciones.

OSUNA, J.R., FERRERAS, M.L. AND NÚÑEZ, A. (1991) 'Inferencia estadística, niveles de precisión y diseño muestral', *Reis*, (54), pp. 139–162.

PARRA, J.M. (1995) 'Estadística descriptiva e inferencial I', *Recuperado de: http://www. academia. edu/download/35987432/ESTADISTICA_DESCRIPTIVA_E_INFERENCIAL. pdf* [Preprint]. Available at: http://franciscojaviercruzariza.com/wp-content/uploads/2014/04/Medidas-Descriptivas-COBACH.pdf.

PEÑA, D. (2002) *Análisis de datos multivariantes*. Madrid: McGraw-hill Madrid. et al: https://www.researchgate.net/profile/Daniel-Pena/publication/40944325_Analisis_de_Datos_Multivariantes/links/549154880cf214269f27ffae/Analisis-de-Datos-Multivariantes.pdf.

PEREZ, V.E.M. *ET AL.* (2017) 'Estrategia de imputación con la media bajo el uso de árboles de regresión', *Comunicaciones en Estadística*, 10(1), pp. 9–40.

PÉRTEGAS DÍAZ, S. AND PITA FERNÁNDEZ, S. (2001) 'La distribución normal', *Cad Aten Primaria*, 8, pp. 268–274.

RUBIN, D.B. (1976) 'Datos pérdidos e inferencia, *Biometrika*, 63(3), pp. 581–592.

SERRANO CARVAJAL, A.G. (2018) *Método de remuestreo BOOTSTRAP: aplicaciones en correlación y control de calidad*.

DE TIRATEL, S.R. (2000) *Guía de fuentes de información especializadas*. Grebyd.

VAN BUUREN, S. (2018) *Flexible imputation of missing data*. CRC press.

WARE, J.H. *ET AL.* (2012) 'Missing Data', *New England Journal of Medicine*, 367(14), pp. 1353–1354. doi:10.1056/NEJMsm1210043.

WILKS, S.S. (1932) 'Momentos y distribuciones de parámetros de población a partir de muestras fragmentarias', *Estadística matemática*, 3(3), pp. 163–195.

ZÚÑIGA MALDONADO, C.A., HERNÁNDEZ RIPALDA, M.D. AND JIMÉNEZ GARCÍA, J.A. (2018) 'Comparación de técnicas de imputación para tratar respuestas censuradas en un diseño de experimentos bivariado', *Nova scientia*, 10(20), pp. 190–212.

LEONARDO FABIO MEDINA NUSTE

Firmado digitalmente por LEONARDO FABIO MEDINA NUSTE
Fecha: 2021.10.20 09:14:03 -05'00'

ANEXOS

ANEXO A: FUNCIÓN LOGISTIC EN EL SOFTWARE R

```
library(MASS)
logistic <- function(x, p = 0.5) {
  m <- p / 0.5
  m * (exp(x) / (1 + exp(x)))
}
```

ANEXO B: FUNCIÓN GENERACIÓN DE ESCENARIOS EN EL SOFTWARE R

```
naGenerator <- function(n, mec, prob, mu = c(0,0), Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2)){
 y <- mvrnorm(n = n, mu = mu, Sigma = Sigma)
 yobs <- y[, 2]
 switch(mec,
       mcar = {
        r2.mcar <- 1 - rbinom(n = n, size = 1, prob = prob)
        ymis_mcar <- y[, 2]
        ymis_mcar[r2.mcar == 0] <- NA
        data.frame('Y1' = y[, 1], 'Y2_obs' = yobs, 'Y2_mcar' = ymis_mcar)
       },
       mar = {
        r2.mar <- 1 - rbinom(n = n, size = 1, prob = logistic(y[, 1], p = prob))
        ymis_mar <- y[, 2]
        ymis_mar[r2.mar == 0] <- NA
        data.frame('Y1' = y[, 1], 'Y2_obs' = yobs, 'Y2_mar' = ymis_mar)
       },
       mnar = {
        r2.mnar <- 1 - rbinom(n = n, size = 1, prob = logistic(y[, 2], p = prob))
        ymis_mnar <- y[, 2]
        ymis_mnar[r2.mnar == 0] <- NA
        data.frame('Y1' = y[, 1], 'Y2_obs' = yobs, 'Y2_mnar' = ymis_mnar)
       })
}

data <- naGenerator(n = 5, mec = 'mar', prob = 0.01)
data
data[, 3] <- NA
```

ANEXO C: FUNCIÓN IMPUTACIÓN EN EL SOFTWARE R

```
imputation <- function(data, technique){
  switch(technique,
        eliminate = {
          na.omit(data[, 3])
        },
        means = {
          if(length(na.omit(data[,3])) >= 2 & length(na.omit(data[,3])) <= (nrow(data) - 1)){
            data[, 3][is.na(data[, 3])] <- mean(data[, 3], na.rm = T)
          } else{
            data[, 3][!is.na(data[, 3])] <- NA
          }
          return(data[, 3])
        },
        medians = {
          if(length(na.omit(data[,3])) >= 2 & length(na.omit(data[,3])) <= (nrow(data) - 1)){
            data[, 3][is.na(data[, 3])] <- median(data[, 3], na.rm = T)
          } else{
            data[, 3][!is.na(data[, 3])] <- NA
          }
          return(data[, 3])
        },
        regression = {
          if(length(na.omit(data[,3])) >= 2 & length(na.omit(data[,3])) <= (nrow(data) - 1)){
            lr <- lm(data[, 3] ~ data[, 1])
            data[, 3][is.na(data[, 3])] <- lr$coefficients[1] + (data[, 1][is.na(data[, 3])] *
lr$coefficients[2])
            attr(data[, 3], "settings") <- c('coeff' = lr$coefficients, 'pval_coef' =
summary(lr)[4][[1]][2, 4],
                                             'R^2' = summary(lr)[8])
          } else{
            data[, 3][!is.na(data[, 3])] <- NA
          }
          return(data[,3])
        })
}
```

ANEXO D: FUNCIÓN SIMULACIÓN EN EL SOFTWARE R

```
simulations <- function(nSim, n, mec, prob, technique, mu = c(0,0), alpha = 0.05,
                        Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2)){
  sims <- replicate(nSim, {
    data <- naGenerator(n = n, mec = mec, prob = prob)
    imp <- imputation(data = data, technique = technique)
    if(sum(is.na(imp)) == nrow(data)){
      meanIMp <- varImp <- mp_a <- NA
      impVal <- 0
    } else{
      meanIMp <- mean(imp)
      varImp <- var(imp)
      mp_a <- (1 / sum(is.na(data[, 3]), na.rm = T)) * sum((data[, 2][is.na(data[,3])] -
imp[(is.na(data[, 3]))])^2, na.rm = T)
      impVal <- 1
    }
    list('Mean Imputation' = meanIMp, 'Var Imputation' = varImp,
         'Precission' = mp_a, "Validate Imputation" = impVal)
  })
  nSimValidate = sum(unlist(sims[4,]))
  # P. Ajuste
  if(technique == "eliminate"){
    prA <- NA
  } else{
    prA <- c('Mean_Pre_ajuste' = mean(unlist(sims[3,]), na.rm = T),
             'Var_Pre_ajuste' = var(unlist(sims[3,]), na.rm = T),
             'Coef_Variacion' = sd(unlist(sims[3,]), na.rm = T) / mean(unlist(sims[3,]), na.rm = T))
  }
```

ANEXO E: MANUAL DEL PAQUETE IDF EN R-STUDIO

1. **Presentación del paquete:**

Package: IDF
Type: Package
Title: Funciones de simulación de datos normales y técnicas de imputación
Versión: 0.1.0
Author: Jamilton Vinueza Chalco y Galo Masaquiza Aragon
Maintainer: Jamilton Vinueza jamilton.vinuezac@espoch.edu.ec
Galo Masaquiza galo.masquiza@espoch.edu.ec
Descripción: El paquete IDF replica n veces un conjunto de p variables de naturaleza cuantitativa que sigue una distribución normal, a cada conjunto simula datos faltantes dependiendo de la proporción que el usuario quiera trabajar. Una vez que genera los conjuntos con los datos faltantes, el paquete usa técnicas de imputación como la eliminación, la media, mediana y regresión lineal para imputar esos valores que faltan. Finalmente el paquete obtiene una medida de precisión de ajuste que sirve para comparar cada técnica y obtener la mejor, también comprueba si cumplen con las propiedades de insesgadez y mínima varianza para los estimadores de la media y varianza.

License: GPL-2
Encoding: UTF-8
LazyData: True
RoxygenNote: 6.1.1

2. **Funciones**

Antes de empezar se debe instalar y activar el paquete:
Es necesario instalar el paquete “MASS” para que funcione correctamente el paquete “IDF”.

Código en R:

```
install.packages(“IDF)
install.packages(“MASS”)
library(“IDF”)
library(“MASS”)
```

2.1. Función: naGenerator

Descripción

Esta función realiza dos pasos:

1. Genera una matriz de datos multivariada (bivariante por defecto) que sigue una distribución normal.
2. Una vez generada la matriz crea datos faltantes mediante los tres mecanismos de respuesta: MCAR (la probabilidad de faltar es la misma para todos los casos), MAR (Si la probabilidad de faltar es la misma solo dentro de los grupos definidos por los datos observados) y MNAR (La probabilidad de faltar varía por razones que se desconocen)

Uso

naGenerator(n, mec,prob, mu = c(0, 0), Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))

Argumentos

mec	indica el mecanismo de respuesta a simular mec = 'mcar', 'mar', 'mnar'
prob	indica la proporción de pérdida de datos a generar
mu	indica la media poblacional bivariante
Sigma	indica las varianzas y covarianzas, al ser bivariante se trabaja con una matrix de 2x2

Ejemplo

naGenerator(n = 5, mec = 'mar', prob = 0.1, mu = c(0,0), Sigma = matrix(c(1,0.5,0.5,1), nrow = 2))

2.2. Función: Imputación

Descripción

Esta función, a partir de la base de datos generada con datos faltantes imputa mediante técnicas de imputación. Existen cuatro técnicas:

1. Eliminación por listas, se encarga de eliminar los datos perdidos y trabaja sin ellos

2. Imputación de la media, esta técnica busca los datos perdidos y los reemplaza por la media aritmética

3. Imputación de la mediana, esta técnica busca los datos perdidos y los reemplaza por la mediana

4. Regresión Lineal, esta técnica construye el modelo lineal solo de los datos observados, y se predijo los datos faltantes mediante el modelo.

Uso

imputation(data, technique)

Argumentos

data	Indica la base de datos donde se va imputar los datos
technique	Indica la técnica de imputación a imputar, puede ser 'eliminate', 'means', 'medians' y 'regression'

Ejemplo

imputation(data, technique = 'medians')

2.3. Función: Simulación

Descripción

Esta función genera n simulaciones de la función "*naGenerator*", es decir, realiza la simulación de Montecarlo y da como resultado la medida de precisión de ajuste y la comprobación de las propiedades de Insesgadez y mínima varianza.

Uso

simulations(nSim, n, mec, prob, technique, mu = c(0, 0), alpha = 0.05, Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))

Argumentos:

nSim	indica el número de simulaciones
n	indica el tamaño de la muestra de cada matriz a simular

mec	indica el mecanismo de respuesta
prob	indica la proporción de datos faltantes a simular
technique	indica la técnica de imputación a imputar
mu	indica la media poblacional
alpha	indica el nivel de significancia

Ejemplo:

```
simulations(nSim = 10000, n = 5, prob = 0.25, technique = 'means', alpha = 0.05)
```

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS DEL APRENDIZAJE

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: *29 / 09 / 2021*

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: *Galo Alexander Masquiza Aragón* *Jamilton Daniel Vinueza Chalco*
INFORMACIÓN INSTITUCIONAL
Facultad: *Ciencias*
Carrera: *Ingeniería en Estadística Informática*
Título a optar: *Ingeniero en Estadística Informática*
f. Analista de Biblioteca responsable: *Ing. Leonardo Medina Ñuste MSc.*

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS DEL APRENDIZAJE

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: *29 / 09 / 2021*

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: *Galo Alexander Masquiza Aragón* *Jamilton Daniel Vinueza Chalco*
INFORMACIÓN INSTITUCIONAL
Facultad: *Ciencias*
Carrera: *Ingeniería en Estadística Informática*
Título a optar: *Ingeniero en Estadística Informática*
f. Analista de Biblioteca responsable: *Ing. Leonardo Medina Ñuste MSc.*

LEONARDO FABIO MEDINA NUSTE

Firmado digitalmente por LEONARDO FABIO MEDINA NUSTE
Fecha: 2021.09.29 13:50:40 -05'00'

1813-DBRA-UTP-2021