Vol.8 / No.1 / Junio 2024, pp. 1-22
SahuarUS

La replicabilidad en la ciencia y el papel transformador de la metodología estadística de knockoffs

Alejandro Román Vásquez[*,1], **Gabriel Escarela Pérez**[*,2], **Gabriel Núñez Antonio**[*,3] y **José Ulises Márquez Urbina**[**,4],

[*]Departamento de Matemáticas, Universidad Autónoma Metropolitana–Unidad Iztapalapa, Av. San Rafael Atlixco 186, C.P. 09340, Iztapalapa, CDMX. México.

[**]Centro de Investigación en Matemáticas A.C., Unidad Monterrey, 66629 Monterrey, Nuevo León, México.

[**]Consejo Nacional de Humanidades, Ciencia y Tecnología, Av. Insurgentes Sur 1582, Col. Crédito Constructor, Benito Juárez, 03940, CDMX, México.

[1]arv@xanum.uam.mx, [2]ge@xanum.uam.mx, [3]gabnunez@xanum.uam.mx, [4]ulises@cimat.mx.

Resumen

Un aspecto importante en la ciencia es la replicabilidad de los resultados científicos. En este artículo se examinan algunas causas fundamentales que contribuyen a la falta de replicabilidad, centrando el análisis en un componente crucial: la estadística y la inferencia selectiva. Partiendo de los desafíos inherentes a las pruebas de hipótesis múltiples en situaciones de alta dimensionalidad, una estrategia para abordar la problemática de la replicabilidad se basa en la implementación del modelo-X de imitaciones. Esta metodología se destaca por generar variables sintéticas que imitan a las originales, permitiendo diferenciar de manera efectiva entre asociaciones genuinas y espurias, y controlando de manera simultánea la tasa de falsos descubrimientos en entornos de muestras finitas. Los aspectos técnicos del modelo-X de imitaciones se describen en este trabajo, subrayando sus alcances y limitaciones. Se enfatiza la efectividad de esta metodología con casos de éxito, tales como la estimación de la pureza en tumores, el análisis de asociación genómica, la identificación de factores pronósticos en ensayos clínicos, la determinación de factores de riesgo asociados al COVID-19 de larga duración, y la selección de variables en estudios de tasa de criminalidad. Estos ejemplos concretos ilustran la preponderante utilidad práctica y la versatilidad del modelo-X de imitaciones en diversas áreas de investigación. Sin lugar a dudas, este enfoque contribuye de manera original a los desafíos actuales en cuanto a la replicabilidad, marcando un hito significativo en la mejora de la confiabilidad y robustez de la evidencia científica.

Palabras Clave: Crisis de replicabilidad; Hipótesis estadísticas múltiples; Modelo-X de imitaciones.

DOI:10.36788/sah.v8i1.148
Recibido: 9 de febrero de 2024.
Aceptado: 4 de junio de 2024.

ISSN 2448 5365
Universidad de Sonora

1. Introducción

Cuando nos sumergimos en la aplicación de conocimientos científicos, tendemos a presuponer que el fundamento del saber se erige sólidamente sobre la piedra angular de experimentos replicables. Robert Boyle, precursor de la ciencia experimental moderna, resaltó la importancia crucial de la replicabilidad para conferir credibilidad a los descubrimientos científicos (Benjamini, 2020). Sin embargo, al explorar los anales de la historia científica, nos encontramos con una crisis que cuestiona esta base misma: *la crisis de replicabilidad*.

La falta de replicabilidad, evidentemente, genera preocupación entre aquellos dedicados a la investigación científica, ya que socava la credibilidad de esta y pone en tela de juicio tanto su labor, su utilidad y las bondades de sus hallazgos. Esta crisis afecta a todos los consumidores del conocimiento científico, y por ende surge la pregunta: ¿por qué debería importar al resto de la población la falta de replicabilidad? Siendo usuarios de la ciencia y el conocimiento, nos beneficiamos de sus numerosas ventajas, pero también nos enfrentamos a las adversidades que pueden surgir cuando hay reportes científicos que potencialmente no son replicables.

Consideremos, por ejemplo, un nuevo procedimiento médico que se presenta como innovador en el tratamiento de un padecimiento específico. Los individuos que padecen la enfermedad van a consultar los servicios de un médico y se van a someter a un procedimiento, que potencialmente es costoso en términos económicos y que puede implicar cuidados paliativos. Sin embargo, con el tiempo, se descubre que el tratamiento no solo no alivia los síntomas, sino que también conlleva efectos secundarios adversos. Los pacientes sufren las consecuencias de una investigación con hallazgos incorrectos, enfrentando un impacto negativo financiero o daños en su salud. Es entonces comprensible preocuparse de que la confianza en el testimonio científico pueda desviarnos del camino correcto, si muchos de los hallazgos en los que confiamos no pueden replicarse.

El propósito de este trabajo es describir una metodología estadística recientemente desarrollada, la cual busca mejorar la replicabilidad de los hallazgos científicos, y que es conocida como el *modelo-X de imitaciones* (model-X knockoffs, en inglés). Para comprender la relevancia de esta técnica, las siguientes dos sección explican la magnitud de la crisis de replicabilidad, aluden las causas principales que la impulsan y exponen el papel crucial que desempeña la estadística en esta crisis, particularmente en el aspecto de la inferencia selectiva en pruebas de hipótesis múltiples. Tras este preámbulo, las tres secciones subsecuentes abordarán desde una perspectiva matemática y estadística, los aspectos fundamentales del modelo-X de imitaciones, incluyendo métodos para generar variables sintéticas, estadísticos de imitación, alcances, limitaciones y ejemplos exitosos. El documento finaliza con una sección de conclusiones.

2. Replicabilidad en crisis

La preocupación moderna por la replicabilidad en las ciencias encuentra sus raíces en las décadas de 1960 y 1970, como evidencian los trabajos de Ahlgren (1969) y Smith (1970). En respuesta a esta inquietud, se estableció la revista “Replications in Social Psychology” a

finales de los años 70, con el propósito de fomentar y resaltar la importancia de la replicación de estudios (Campbell and Jackson, 1979). Lamentablemente, esta iniciativa cesó su publicación tras solo tres números (Romero, 2019). La denominada *crisis de replicabilidad* tiene aproximadamente 30 años, coincidiendo con la industrialización de los procesos científicos, caracterizada por avances en herramientas genómicas, tecnología de imágenes y análisis de datos, como destaca Benjamini (2020). Mann (1994) y Lander and Kruglyak (1995) advirtieron problemas de replicabilidad en la genética del comportamiento y la genómica, respectivamente, contribuyendo a la creciente inquietud sobre este tema. Ioannidis (2005) amplió la popularidad del tema al afirmar que la mayoría de los hallazgos de investigación publicados son falsos, generando un amplio interés y esfuerzos para abordar el problema.

Como señala Romero (2019), tres casos notorios intensificaron la preocupación por la replicabilidad. El estudio sobre el caminar de personas mayores de Bargh et al. (1996), que fue altamente citado durante años, no se replicó con éxito en intentos posteriores más rigurosos (Doyen et al., 2012; Pashler et al., 2011). La serie de estudios de percepción extrasensorial de Bem (2011), que afirmaba que las personas podían prever el futuro, generó desconfianza en las prácticas experimentales de la psicología debido a un uso incorrecto de métodos y herramientas estadísticas comúnmente empleados. Informes de Amgen y Bayer Healthcare revelaron dificultades para replicar hallazgos biomédicos (Begley and Ellis, 2012; Prinz et al., 2011). Adicionalmente a estos casos, las retractaciones de las obras de Diederik Stapel destacan, debido a falsificaciones y fabricaciones de datos, que conllevaron a la preocupación general (Stroebe et al., 2012).

En paralelo a estos casos de no replicabilidad, un grupo de científicos se unió para llevar a cabo un extenso intento de replicar los hallazgos publicados en tres influyentes revistas de psicología (Open Science Collaboration, 2015). Cada resultado principal de una selección de 100 artículos tomados al azar de estas revistas líderes en el campo fue sometido a pruebas de replicación. Al concluir este esfuerzo, que se inició en 2011 y terminó en 2015, solo el 34 % de los resultados principales de los estudios se lograron replicar.

3. Causas de la falta de replicabilidad y el papel de la estadística

La falta de replicabilidad en la investigación científica puede manifestarse a través de diversas fuentes, muchas de las cuales están vinculadas a errores humanos o elecciones tomadas por los investigadores. Entre los factores más influyentes, destacan el sesgo de publicación, los incentivos de investigación mal alineados, errores, informes incompletos y, lamentablemente, casos de fraude. Además de estos desafíos, la aplicación de técnicas de inferencia estadística inapropiadas, o mal especificadas, también ha contribuido de manera significativa a la falta de replicabilidad en la investigación científica (National Academies of Sciences, Engineering, and Medicine, 2019).

El sesgo de publicación representa una distorsión significativa en la literatura científica, alimentado por la preferencia hacia resultados estadísticamente significativos. La presión para publicar hallazgos positivos conduce a la exclusión sistemática de resultados que no alcanzan significancia estadística. Este fenómeno genera una representación sesgada de la realidad, ya

que sólo a través de la inclusión de efectos significativos y no significativos se puede lograr una estimación precisa del tamaño real del efecto. Los incentivos académicos desalineados representan un desafío considerable para la integridad de la investigación. Elementos como la permanencia y el financiamiento de proyectos pueden comprometer la calidad de los estudios, ya que los investigadores, impulsados por métricas de productividad, pueden sentir la presión de publicar rápidamente, descuidando así los estándares científicos. Un aspecto central de este problema radica en la regla de prioridad, que premia exclusivamente al primer científico que realiza un descubrimiento. Esta práctica, aunque arraigada en el sistema de recompensas académicas, desalienta la replicación de estudios (Romero, 2019).

La presencia de errores en la investigación, ya sea en aspectos metodológicos, computacionales o en la recopilación de datos, constituye un factor crítico que puede dar lugar a la falta de replicabilidad. Detectar estos errores presenta diversos retos, incluyendo la importancia crucial de la transparencia, tanto en la obtención de los datos como en la elección de la metodología y el consecuente procesamiento computacional, para asegurar la reproducibilidad de un estudio. La transparencia no sólo facilita la identificación y corrección de posibles errores, sino que también contribuye en la consolidación de la replicabilidad de la investigación a lo largo del tiempo. En particular, la carencia de información detallada sobre los aspectos fundamentales del estudio puede obstaculizar significativamente los esfuerzos de replicación. Por lo tanto, el compartir de manera exhaustiva y clara los detalles de las metodologías de investigación se convierte en un elemento esencial para facilitar la reproducción de los resultados.

El fraude y la mala conducta representan el extremo más grave de la falta de replicabilidad. Cuando se descubren casos de investigadores que manipulan, fabrican, o falsifican de manera deliberada los datos, se produce un perjuicio al progreso científico. Este comportamiento no solo socava la integridad de la investigación, sino que también mina la confianza pública en el proceso científico. Finalmente, el uso inapropiado de diversas técnicas estadísticas ha sido ampliamente citado como una causa fundamental que ha contribuido significativamente a la crisis de replicabilidad (Colling and Szűcs, 2021). Cuatro grandes aspectos estadísticos han emergido como protagonistas en la acentuación de esta problemática, según lo señalado por algunas fuentes, de las que destacan la de Romero (2019), la de National Academies of Sciences, Engineering, and Medicine (2019) y la de Benjamini (2020).

La primera problemática se vincula con el papel que desempeñan tanto la investigación exploratoria como la confirmatoria. Mientras que la investigación exploratoria genera una hipótesis a partir de los datos recopilados, la investigación confirmatoria implica hipótesis predefinidas y sigue un procedimiento planificado de pruebas estadísticas. El utilizar la investigación exploratoria con objetivos confirmatorios conlleva a una violación del principio de no emplear los datos tanto para la generación de hipótesis como para validar hallazgos, comprometiendo así la integridad de los resultados estadísticos. Este fenómeno se asocia estrechamente con el concepto de HARKing (Hypothesizing After Results are Known, por su acrónimo en inglés), que erróneamente basa la hipótesis en los datos recopilados y luego utiliza esos mismos datos como evidencia para respaldar la hipótesis.

El segundo aspecto crítico se relaciona con las prácticas de investigación cuestionables (QRPs, por sus siglas en inglés). Dado que la significancia estadística juega un papel de-

terminante en la publicación, los científicos enfrentan incentivos para reportar resultados sesgados, a veces de manera inconsciente, con el fin de obtenerla. Una práctica particularmente perniciosa en este contexto es el p-hacking, que implica aprovechar la flexibilidad en la recopilación de datos para obtener significancia estadística. Esto puede incluir acciones como recopilar más datos o excluir selectivamente datos hasta obtener los resultados deseados. Un estudio significativo de simulación por computadora realizado por Simmons et al. (2011) revela que una combinación de técnicas de p-hacking puede aumentar la tasa de falsos descubrimientos (falsos positivos) hasta en un preocupante 61 %.

Un tercer componente en estadística que ha exacerbado la crisis de replicabilidad, son las Pruebas de Significación de la Hipótesis Nula (NHST, por sus siglas en inglés), también conocidas como pruebas de significancia, y sus valores-p asociados (Nuzzo, 2014). Varios estudios indican una tendencia común entre aquellos que utilizan estas pruebas: malinterpretación de los valores-p, no comprensión de la lógica detrás de las técnicas inferenciales y la confusión de la significancia estadística con la importancia científica (Cohen, 1990; Fidler et al., 2006; Ziliak and McCloskey, 2010). En respuesta a estas problemáticas, algunas revistas han prohibido la inclusión de valores-p en sus páginas, mientras que otras sugieren una redefinición de lo que se considera estadísticamente significativo (Trafimow and Marks, 2015; Benjamin et al., 2018).

La crisis de replicabilidad resalta la falta generalizada de comprensión acerca de los valores-p y las bases de las estadísticas frecuentistas, así como las dificultades para justificar inferencias basadas en la significancia estadística. Durante la crisis y los eventuales debates, la Asociación Estadounidense de Estadística (ASA por sus siglas en inglés) emitió dos declaraciones con el objetivo de aclarar el significado y uso de la significancia estadística y los valores-p (Wasserstein and Lazar, 2016; Wasserstein et al., 2019). Estos esfuerzos buscan mejorar la comprensión y el uso responsable de las pruebas de significancia, destacando la importancia de interpretar los resultados en un contexto más amplio y subrayando que la significancia estadística no debe ser la única medida de relevancia científica.

Por último, un cuarto aspecto que impacta la replicabilidad es la denominada inferencia selectiva. Según Benjamini (2020), esta plantea una amenaza sustancial para la replicabilidad y se ha vuelto más desafiante en el contexto de la ciencia industrializada. La inferencia selectiva se vuelve problemática cuando la elección se realiza entre los numerosos resultados evidentes en el trabajo publicado, pero la inferencia estadística no se ajusta para tener en cuenta dicha selección. En tales circunstancias, las garantías estadísticas convencionales ofrecidas por todos los métodos estadísticos se debilitan. Dado que la selección sólo puede ocurrir cuando hay muchas oportunidades, este fenómeno a veces se denomina como el problema de la multiplicidad o pruebas de hipótesis múltiples. Cuando se prueban múltiples hipótesis, la probabilidad de obtener un resultado significativo por casualidad aumenta. Por lo tanto, esta selección debe ajustarse para obtener valores-p e intervalos de confianza válidos; de lo contrario, perderían su función como evaluaciones cuantitativas creíbles de la incertidumbre (Benjamini, 2020).

4. Pruebas de hipótesis múltiples y el modelo-X de imitaciones

En el contexto de la realización de múltiples pruebas de hipótesis independientes, resulta importante comprender el concepto de tasa de error por familia, que representa la probabilidad de experimentar al menos un falso rechazo entre todas las hipótesis. Esta tasa se calcula mediante la expresión $1 - (1 - \alpha)^r$, donde r denota el número de pruebas y α la significancia de cada prueba, es decir, la probabilidad de incurrir en un falso descubrimiento, también conocido como error tipo I (Bretz et al., 2016). Consideremos, por ejemplo, el caso específico de un investigador que lleva a cabo diez pruebas estadísticas independientes. Bajo la suposición de que la hipótesis nula es verdadera para todas las pruebas, la probabilidad de cometer al menos un falso positivo asciende a aproximadamente el 40 %, considerando una $\alpha = 0.05$. Esta cantidad revela que la magnitud del error por familia en este escenario particular resultaría elevada.

La problemática de la inferencia selectiva constituye una situación común y, lamentablemente, suele pasar desapercibida entre los investigadores (Benjamini, 2020). Un claro ejemplo de esta falta de atención se evidencia en una reciente encuesta bibliográfica, la cual acompañó a un artículo que resaltaba la presencia de esta multiplicidad oculta en el análisis de datos. Alarmantemente, sólo alrededor del 1 % de los investigadores, examinando 819 artículos de seis destacadas revistas de psicología, tomaron en consideración este fenómeno al interpretar sus resultados (Cramer et al., 2016). Este revelador hallazgo subraya la extensión de la confusión existente en torno a este tema crucial. Adicionalmente, cabe destacar que incluso expertos reconocidos han manifestado sorpresa al descubrir la prevalencia de esta situación (Bishop, 2014), sugiriendo así que este conocimiento no está ampliamente difundido en la comunidad científica.

Cuando nos encontramos ante un número moderado de comparaciones y la necesidad de obtener resultados estadísticamente robustos, el control de la tasa de error por familia emerge como un procedimiento adecuado, a pesar de ser potencialmente conservador. Este enfoque se emplea con mayor frecuencia en estudios confirmatorios, o en situaciones en las cuales un falso rechazo podría conllevar a consecuencias perjudiciales, como sucede, por ejemplo, en ensayos clínicos (Ren, 2021). Para una familia de inferencias potenciales sobre diferentes parámetros, los métodos tales como el de Bonferroni ofrecen un amplio control simultáneo de la tasa de error por familia (Benjamini, 2020).

El enfoque denominado "en promedio sobre la selección" se presenta como una alternativa más flexible, asegurando que cualquier inferencia de un método estadístico permanezca válida, en promedio, a lo largo de múltiples selecciones (Benjamini, 2020). Este enfoque, orientado al control de la tasa de falsos descubrimientos (TFD o FDR por sus siglas en inglés), fue introducido por Benjamini and Hochberg (1995). Ésta se caracteriza por asociar descubrimientos con hipótesis rechazadas y descubrimientos falsos con errores tipo I. Su objetivo principal es maximizar los descubrimientos, al mismo tiempo que se controla la proporción de descubrimientos falsos en el nivel esperado q. En contraste, con la tasa de error por familia, la TFD se posiciona como una métrica más liberal y, por ende, resulta más adecuada para estudios de carácter exploratorio y en situaciones de alta dimensionalidad; es decir, cuando el número de comparaciones es elevado (Ren, 2021).

La mayoría de los métodos convencionales para llevar a cabo pruebas múltiples, independientemente de las métricas consideradas, tienen como base los valores-p (Bretz et al., 2016). En términos más específicos, estos métodos parten de la premisa de que, dada la información disponible, es posible obtener un valor-p válido, denotado aquí como p_j, para cada característica nula j, de manera que se cumple $P(p_j \leq t) \leq t$, para cualquier $t \in (0, 1)$. En otras palabras, si la hipótesis nula es verdadera, los valores-p deben exhibir una distribución uniforme (Colling and Szűcs, 2021).

A pesar de la extensa literatura en este ámbito, la suposición de contar con valores-p válidos resulta ser notablemente restrictiva. En términos generales, la obtención de valores-p válidos se convierte en una tarea desafiante sin un modelado paramétrico sólido, especialmente en contextos de alta dimensionalidad. Investigadores han subrayado que al emplear métodos de inferencia clásicos en situaciones donde la dimensión p del número de pruebas es del mismo orden que el número de muestras n, los valores-p resultantes no se comportan según lo esperado. Contrariamente, muestran un comportamiento que podría aumentar potencialmente el error tipo I (Sur and Candès, 2019; Candès et al., 2018). Este señalamiento destaca la necesidad de explorar enfoques más avanzados y adaptativos en la inferencia estadística, especialmente en escenarios donde la complejidad dimensional desafía las premisas tradicionales.

El modelo-X de imitaciones, concebido inicialmente por Barber and Candès (2015) y respaldado por las investigaciones de Candès et al. (2018) y las de Barber et al. (2020), representa un enfoque innovador en el ámbito de pruebas múltiples. Este procedimiento supera la dependencia de los valores-p y logra un control de la TFD con resultados no asintóticos; es decir, proporciona garantías de control en entornos de muestras finitas. El modelo-X de imitaciones representa una herramienta eficaz para la selección de variables, centrando su interés en identificar qué factores $X_1, X_2, \ldots, X_p$ están verdaderamente relacionados con una variable de interés Y. La tarea de identificar las características esenciales dentro de un conjunto potencial de candidatos $\boldsymbol{X} = (X_1, \ldots, X_p)$ se enmarca en pruebas de hipótesis múltiples de independencia condicional. Más específicamente, esta condición implica determinar qué variables X_j cumplen con la condición $Y \perp X_j | \boldsymbol{X}_{-j}$, donde $\boldsymbol{X}_{-j}$ representa todas las variables excepto X_j.

En términos generales, el modelo-X de imitaciones implica la generación de un conjunto de copias, denominadas knockoffs y representadas por $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_p)$, que imitan las variables originales y su estructura de dependencia, sin emplear información alguna de la variable de respuesta Y. Este enfoque permite discernir qué variables tienen una verdadera asociación y cuáles variables son espurias, al comparar estadísticas de importancia de las variables originales con las de sus homólogos de imitación. En consecuencia, el modelo-X de imitaciones se destaca como una herramienta valiosa para el análisis de datos en situaciones en las que se busca una mayor robustez en la selección de variables.

5. Aspectos principales del modelo-X de imitaciones

En el contexto de la selección de variables, se parte de la premisa de que la variable respuesta Y está relacionada con un conjunto de variables explicativas $\boldsymbol{X} = (X_1, \ldots, X_p)$. Utilizando un conjunto de observaciones independientes e idénticamente distribuidas (i.i.d.) de la distribución de probabilidad $P_{\boldsymbol{X}Y}$, el objetivo es identificar las variables realmente asociadas con la respuesta. En este escenario, se define una variable X_j como nula si Y es independiente de X_j dado el resto de las variables $\boldsymbol{X}_{-j} = \{X_i : i \neq j\}$; en notación probabilística, $Y \perp X_j | \boldsymbol{X}_{-j}$. En términos de índices, $S_0 \subset \{1, 2, \ldots, p\}$ representa al conjunto de variables nulas, y $S = \{1, 2, \ldots, p\} \setminus S_0$ denota al conjunto de variables no nulas. El propósito principal radica en estimar S, abordando el problema de pruebas de hipótesis múltiples mientras se controla la TFD, la cual se define como el valor esperado de la proporción de falsos descubrimientos (PFD), que es:

$$TFD = E\left[\frac{|\hat{S} \cap S_0|}{|\hat{S}| \vee 1}\right]$$

donde $\hat{S}$ es el conjunto de variables seleccionadas, y $a \vee b = max(a, b)$. Según la descripción de Sechidis et al. (2021), el procedimiento del modelo-X de imitaciones consta de tres pasos fundamentales: (1) construcción de las variables de imitación, conocidas como los knockoffs; (2) estimación del estadístico de imitación y (3) el cálculo del umbral dependiente de datos.

El primer paso, que consiste en la generación de copias o knockoffs, resulta fundamental para mantener el control sobre la TFD. Aquí, las variables sintéticas $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ deben replicar la información de las variables originales, sin establecer ninguna asociación con la respuesta Y; consecuentemente, dichas variables no deben ser determinadas por ningún método de selección de variables. En esta etapa, es imperativo construir los knockoffs $\tilde{\boldsymbol{X}}$, cumpliendo con dos propiedades fundamentales:

1. Independencia condicional: $Y \perp \tilde{\boldsymbol{X}} | \boldsymbol{X}$.

2. Intercambiabilidad: Para cualquier subconjunto $R \subset [p] := \{1, 2, \ldots, p\}$, se cumple que $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{\text{SWAP}(R)} \equiv (\boldsymbol{X}, \tilde{\boldsymbol{X}})$, donde SWAP$(R)$ significa intercambiar X_j con $\tilde{X}_j$ para cada $j \in R$.

La primera condición se satisface de manera sencilla, ya que los procedimientos de muestreo no incorporan información alguna sobre la respuesta Y. La propiedad de intercambiabilidad establece que no podemos distinguir si la columna j corresponde a una variable real o a una imitación únicamente observando los vectores $\boldsymbol{X}$ y $\tilde{\boldsymbol{X}}$.

Una vez generadas las variables de imitación $\tilde{\boldsymbol{X}}$, el segundo paso consiste en construir un vector de estadísticas de imitación $\boldsymbol{W} = (W_1, \ldots, W_p)$ utilizando el conjunto de datos originales y sus respectivas copias. Cada W_j se forma a través de la relación $W_j = f(Z_j, Z_{j+p})$, donde Z_j y Z_{j+p} son estadísticas de importancia que miden la relevancia de la variable original X_j y su contraparte $\tilde{X}_j$. Aquí, f es una función antisimétrica, la cual satisface $f(u, v) = -f(v, u)$. Esta propiedad implica que el intercambio entre la j-ésima variable y

su imitación cambia el signo de W_j, propiedad conocida como signo inverso. Nótese que es posible emplear cualquier función de importancia que posea la propiedad del signo inverso.

En la literatura se han propuesto varias funciones para medir la importancia de las variables explicativas (Candès et al., 2018), siendo comunes los valores absolutos de los coeficientes de regresión en un modelo regularizado de LASSO (Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés). Estas estadísticas de importancia basadas en los coeficientes de regresión conducen a la creación de la estadística de imitación conocida como la estadística de Diferencia de Coeficientes de LASSO (DCL), dada por:

$$W_j = Z_j - Z_{j+p} = |\beta_j| - |\beta_{j+p}|.$$

En tal formulación, un valor grande y positivo de W_j proporciona evidencia de que la distribución de Y depende de la variable original X_j. En contraste, para variables nulas (aquellas no asociadas con la respuesta), W_j exhibe una distribución simétrica; por lo tanto, es igualmente probable que tome valores positivos o negativos Candès et al. (2018).

Si las estadísticas de imitación W_j para las variables nulas exhiben una distribución simétrica, entonces, para cualquier umbral fijo $t > 0$, se verifica que el número de variables para las cuales W_j es menor o igual a $-t$ supera al número de variables nulas cuyo W_j es menor o igual a $-t$, ya que el conjunto de variables nulas es un subconjunto del total de las variables. Por simetría, también se cumple la propiedad de que el número de variables para las cuales W_j es menor o igual a $-t$ es mayor que el número de variables nulas cuyo W_j es mayor o igual a t. Estas relaciones se expresan, respectivamente, como:

$$\begin{aligned} \#\{j : W_j \leq -t\} &\geq \#\{j \text{ nulas} : W_j \leq -t\} \\ \#\{j : W_j \leq -t\} &\geq \#\{j \text{ nulas} : W_j \geq t\}. \end{aligned}$$

Si se seleccionan como variables con asociación aquellas con un valor de W_j suficientemente grande, $W_j \geq t$, la proporción de falsos descubrimientos se calcula como:

$$PFD = \frac{\#\{j \text{ nulas} : W_j \geq t\}}{\#\{j : W_j \geq t\}},$$

lo que representa el cálculo del cociente del número de variables que son realmente nulas entre el número de variables seleccionadas. Sin embargo, dado que no se conoce de antemano cuáles son las variables nulas, la PFD puede estimarse con la siguiente expresión:

$$\widehat{PFD} = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}.$$

Como se mencionó previamente, el numerador $\#\{j : W_j \leq -t\}$ es una estimación sesgada hacia arriba del numerador desconocido $\#\{j \text{ nulas} : W_j \geq t\}$; por lo tanto, la premisa del tercer paso en la metodología de los knockoffs es seleccionar un umbral dependiente de los datos que sea tan liberal como sea posible, que al mismo tiempo estima PFD y controla la TFD. Para un valor dado $q \in (0, 1)$, la elección de las variables $\hat{S} = \{j : W_j \geq T_+\}$ controla la TFD en el valor objetivo q, donde T_+ está determinado por (Candès et al., 2018):

$$T_+ = min\left\{c > 0 : \frac{1 + \#\{j : W_j \leq -c\}}{\#\{j : W_j \geq c\}} \leq q\right\}.$$

Como señala Candès et al. (2018), es importante destacar que estos resultados no son asintóticos, lo que significa que no dependen de tener muestras de tamaño grande.

5.1. Métodos para generar knockoffs

Una propiedad fundamental de la metodología del modelo-X de imitaciones radica en la creación de variables de imitación que sean válidas, en el sentido de cumplir con las dos propiedades esenciales: independencia condicional $Y \perp \tilde{\boldsymbol{X}}|\boldsymbol{X}$ y la intercambiabilidad de pares. En el trabajo de Candès et al. (2018), se introduce un algoritmo secuencial denominado Pares Independientes Condicionales Secuenciales (PICS). Aunque el procedimiento PICS resulta en variables de imitación válidas, los autores destacan que su implementación puede ser bastante complicada, ya que implica recalcular la distribución condicional en cada paso. En el caso muy particular en que el vector aleatorio de covariables $\boldsymbol{X}$ se pueda expresar como una cadena de Markov, el trabajo de Sesia et al. (2019) propone un procedimiento para generar variables de imitación secuenciales que aprovechan este algoritmo. Sin embargo, fuera de este caso específico, obtener un procedimiento práctico a partir de PICS no es trivial.

Ahora, consideremos el caso en el que el vector de covariables sigue una distribución Gaussiana multivariada. Sin pérdida de generalidad, supongamos que cada covariable ha sido trasladada y reescalada para tener media cero y varianza uno; específicamente, $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, donde $\boldsymbol{0}$ es un vector de ceros y $\boldsymbol{\Sigma}$ es la matriz de covarianza que, debido al reescalado, es equivalente a la matriz de correlación, y que debe ser positiva semi-definida; de esta manera, una distribución conjunta que permite generar knockoffs válidos es la siguiente:

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \sim N_{2p}(\boldsymbol{0}, \boldsymbol{G}), \quad \text{donde} \quad \boldsymbol{G} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - diag\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - diag\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{pmatrix}.$$

Aquí, $diag\{\boldsymbol{s}\}$ representa cualquier matriz diagonal seleccionada de manera que la matriz de covarianza conjunta $\boldsymbol{G}$ es positiva semi-definida. Una forma de producir una matriz $diag\{\boldsymbol{s}\}$ que cumple con esta condición se basa en el uso de lo que se conoce como programación semi-definida, cuyo algoritmo consta de una optimización convexa sujeta a ciertas restricciones (ver Sección 3 de Candès et al. (2018)).

Una vez encontrada la matriz $\boldsymbol{G}$, las variables sintéticas pueden generarse a partir de la distribución condicional de $\tilde{\boldsymbol{X}}$ dado $\boldsymbol{X}$. Como la distribución conjunta $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ sigue una distribución $N_{2p}(\boldsymbol{0}, \boldsymbol{G})$, y la distribución normal es cerrada bajo condicionamiento, la distribución condicional se modela como una normal multivariada; a saber, $\tilde{\boldsymbol{X}}|\boldsymbol{X} \sim N_p(\boldsymbol{\mu}', \boldsymbol{V})$, donde $\boldsymbol{\mu}' = \boldsymbol{X} - \boldsymbol{X}\boldsymbol{\Sigma}^{-1} diag\{\boldsymbol{s}\}$ y $\boldsymbol{V} = 2diag\{\boldsymbol{s}\} - diag\{\boldsymbol{s}\}\boldsymbol{\Sigma}^{-1} diag\{\boldsymbol{s}\}$. Esto implica que, dados el vector $\boldsymbol{X}$, la matriz de covarianza $\boldsymbol{\Sigma}$ y la matriz diagonal $diag\{\boldsymbol{s}\}$, se puede construir el vector $\boldsymbol{\mu}'$ y la matriz de covarianza $\boldsymbol{V}$, que a su vez generan variables de imitación tras simular realizaciones de la distribución $\tilde{\boldsymbol{X}}|\boldsymbol{X} \sim N_p(\boldsymbol{\mu}', \boldsymbol{V})$.

El procedimiento de muestreo de knockoffs gaussianos descrito anteriormente minimiza la llamada Correlación Media Absoluta (CMA) entre cada X_j y su contraparte $\tilde{X}_j$, con el objetivo de maximizar la potencia estadística. No obstante, como señalan Spector and Janson (2022), este criterio puede resultar ineficaz en el contexto de datos altamente correlacionados. La explicación radica en que al minimizar la CMA entre cada X_j y su knockoff $\tilde{X}_j$, se pueden

crear fuertes dependencias entre X_j y el resto de las variables de imitación $\tilde{\boldsymbol{X}}_{-j}$, lo que permite a los algoritmos reconstruir el efecto de las variables no nulas utilizando las variables sintéticas restantes, disminuyendo así la potencia estadística del método. Para abordar este problema, se han propuesto enfoques que minimizan la reconstructibilidad de las variables originales, basándose en dos medidas: knockoffs basados en Reconstructibilidad de Mínima Varianza (RMV) y knockoffs de Máxima Entropía (ME) (ver Sección 3 en Spector and Janson (2022)).

Varios estudios han propuesto distintos mecanismos para generar variables de imitación en entornos no gaussianos, cada uno con sus ventajas y limitaciones. Un enfoque destacado es el presentado por Candès et al. (2018), que aborda el problema con el método de knockoffs de segundo orden. Aquí, en lugar de cumplir estrictamente con la propiedad de intercambiabilidad por pares, el procedimiento requiere que $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{\mathrm{SWAP}(R)}$ y $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ coincidan en los dos primeros momentos para cualquier subconjunto $R \subset [p] := \{1, 2, \ldots, p\}$. En otras palabras, al intercambiar variables con sus copias, se busca que al menos las distribuciones coincidan en medias y covarianzas. Aunque este enfoque de segundo orden demuestra robustez en ciertas situaciones prácticas, no produce knockoffs válidos y no garantiza el control de la TFD en diversas circunstancias (Spector and Janson, 2022).

Otro enfoque, propuesto por Bates et al. (2021), introduce un generador de knockoffs basado en el algoritmo de Metropolis-Hastings. Este método permite muestrear knockoffs válidos para distribuciones arbitrarias del vector $\boldsymbol{X}$ bajo la suposición de que al menos su densidad no normalizada es conocida. Aunque este procedimiento de muestreo de knockoffs puede ejecutarse en un tiempo razonable para modelos gráficos, los autores que lo proponen advierten que es computacionalmente prohibitivo sin una estructura gráfica. Otros trabajos han explorado enfoques flexibles para el muestreo de knockoffs. En Jordon et al. (2018), se emplean modelos de redes adversarias generativas, que son potentes modelos generativos de aprendizaje profundo. Sin embargo, estos modelos sufren de inestabilidad en el entrenamiento y de un proceso de estimación complicado, como se ha señalado en estudios como los de Gulrajani et al. (2017) y Mescheder et al. (2017).

Romano et al. (2020) presentan Deep-Knockoffs, que utilizan la Discrepancia Máxima de Medias (DMM) como función de pérdida en un modelo generativo de aprendizaje profundo. Aunque en entornos de alta dimensión, la DMM puede no generar knockoffs confiables Ramdas et al. (2015), y su eficacia a menudo depende de la selección de hiperparámetros que pueden ser computacionalmente costosos de determinar (Sudarshan et al., 2020). El algoritmo de Knockoffs de Verosimilitud Directa Profunda es un procedimiento de dos etapas que utiliza máxima verosimilitud para estimar primero la distribución de $\boldsymbol{X}$ a partir de datos observados, y luego estima la distribución de knockoffs, minimizando la divergencia de Kullback–Leibler (KL) entre la distribución conjunta de $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ y la distribución conjunta de cualquier intercambio de coordenadas entre $\boldsymbol{X}$ y $\tilde{\boldsymbol{X}}$ (Sudarshan et al., 2020). En este caso, una desventaja del uso de la divergencia KL es la falta de sensibilidad a la distancia, lo que resulta en que regiones cercanas de alta masa de probabilidad pero no superpuestas no se consideran como distribuciones similares (Bińkowski et al., 2018).

En el entorno no gaussiano mixto, Kormaksson et al. (2021) proponen un algoritmo secuencial para generar knockoffs cuando los datos subyacentes consisten tanto en variables

continuas como categóricas. Este procedimiento se basa en el algoritmo PICS, introducido por Candès et al. (2018), como se mencionó anteriormente, estimando cada distribución condicional en el proceso iterativo $P(X_j|\boldsymbol{X}_{-j}, \tilde{\boldsymbol{X}}_{1:j-1})$ mediante la regresión de la variable j-ésima X_j sobre las $p-1$ variables restantes $\boldsymbol{X}_{-j}$ y las variables de imitación creadas $\tilde{\boldsymbol{X}}_{1:j-1}$. Según Barber et al. (2020), esta estrategia puede dar lugar a condicionales incompatibles; es decir, distribuciones condicionales que al concatenarse, no representan a la distribución original del vector $\boldsymbol{X}$, y esta discrepancia puede resultar en una inflación de la TFD. En Vásquez et al. (2023) se propone un procedimiento que utiliza la cópula gaussiana latente para modelar el vector de predictores con marginales no gaussianas, pudiendo ser continuas, binarias u ordinales. Sin embargo, la desventaja de este método es que no es posible considerar variables categóricas nominales.

5.2. Algunos estadísticos de imitación W_j

Existe una amplia variedad de estadísticas de imitación, representadas por $\boldsymbol{W} = (W_1, \ldots, W_p)$, que se pueden emplear en el modelo-X de imitaciones. Dado que $W_j = f(Z_j, Z_{j+p})$, la diversidad correspondiente depende del tipo de estadística de importancia Z_j, así como del tipo de función antisimétrica $f(u, v)$ utilizada. Las opciones para Z_j incluyen medidas estadísticas como el coeficiente estimado en un modelo lineal generalizado, pero también pueden abarcar medidas más heurísticas, como la importancia de variables en bosques aleatorios (Candès et al., 2018). En cuanto a las funciones antisimétricas $f(Z_j, Z_{j+p})$, existen varias opciones, entre las que destacan $|Z_j| - |Z_{j+p}|, \text{signo}(|Z_j| - |Z_{j+p}|) \times \max(|Z_j|, |Z_{j+p}|)$ y $log(|Z_j|) - log(|Z_{j+p}|)$.

Cuando los datos respaldan la hipótesis de un comportamiento lineal de los predictores con la respuesta, Barber and Candès (2015) sugieren el uso del Máximo Signo de LASSO (MSL), así como el valor absoluto de los coeficientes de un modelo de regresión LASSO, como estadísticas de importancia. El MSL corresponde al parámetro de penalización λ más grande en el cual la variable j entra en el modelo en la regresión de LASSO; en contraste, cuando existe una asociación no lineal entre la respuesta y los predictores, diversos autores han propuesto estadísticas de importancia asociadas a modelos que se adaptan a estas circunstancias. Por ejemplo, en Jiang et al. (2021) se sugiere el uso de la Explicación Aditiva de Shapley (Lundberg and Lee, 2017), que ha sido consistentemente empleada como medida para interpretar las predicciones, especialmente en modelos de árboles de decisión y sus extensiones, como XGBoost (Chen and Guestrin, 2016).

En Lu et al. (2018), se presenta una arquitectura de aprendizaje profundo denominada DeepPINK (selección de variables profunda utilizando Knockoffs no lineales de entrada pareada), la cual se basa en un perceptrón multicapa con la distinción principal de que incorpora una capa de acoplamiento pareada con p filtros, uno por cada variable de entrada, donde cada filtro conecta la variable original con su contraparte de imitación. Otra fuente muy interesante de estadísticas de importancia para los knockoffs proviene de los procedimientos bayesianos. Como se destaca en Candès et al. (2018), lo que hace esto especialmente atractivo es que se obtiene la ventaja de poder incorporar información previa, manteniendo al mismo tiempo una estricta garantía frecuentista sobre el error tipo I. Los estadísticos de imitación pueden

ser caracterizados al utilizar la diferencia de los coeficientes absolutos medios posteriores, la diferencia, o la razón logarítmica de las probabilidades posteriores de coeficientes no nulos con una distribución a priori dispersa (George and McCulloch, 1997).

De la diversidad de propuestas de estadísticos de imitación $\boldsymbol{W}$, lo destacable es que los knockoffs pueden funcionar como una envoltura adaptable para prácticamente cualquier algoritmo de ajuste o predicción de datos. Independientemente del algoritmo seleccionado, se garantiza un control riguroso del error en el proceso de selección de variables.

5.3. Alcances y limitaciones

La metodología del modelo-X de imitaciones presenta tres beneficios notables. En primer lugar, su aplicación se destaca al no depender de valores-p específicos, permitiendo su implementación efectiva en situaciones de alta dimensionalidad. Esto es especialmente valioso en campos como la genética, donde el número de genes supera considerablemente la cantidad de pacientes disponibles para el estudio. El segundo beneficio se evidencia al comparar el modelo-X de imitaciones con otros métodos, como el procedimiento de Benjamini and Hochberg (1995). Investigaciones empíricas, como las llevadas a cabo por Candès et al. (2018) y Kormaksson et al. (2021) demuestran que el modelo-X de imitaciones posee una mayor potencia estadística y un control más efectivo de la TFD.

El tercer beneficio radica en la necesidad de una modelación precisa de la distribución $P_{\boldsymbol{X}}$ para controlar la TFD a niveles predefinidos. Es fácil visualizar escenarios en los que se cuentan con abundantes muestras no etiquetadas del vector $\boldsymbol{X}$, debido a la dificultad para adquirir datos etiquetados (muestras con un valor específico de la respuesta Y). Esto es evidente en estudios genéticos, donde se dispone de cientos de miles o incluso millones de genotipos en diversas poblaciones, pero reclutar pacientes con un fenotipo particular resulta desafiante. La disponibilidad de más datos no etiquetados facilita la estimación precisa de $P_{\boldsymbol{X}}$, favoreciendo así la aplicabilidad de esta metodología.

A pesar de estos beneficios, es importante destacar dos limitaciones del modelo-X de imitaciones. En primer lugar, el umbral T_+ es dependiente de los datos, lo que puede ser problemático cuando la cantidad de datos es limitada. En tales casos, la elección de un valor específico de q para controlar la TFD puede requerir un ajuste mediante ensayo y error, seleccionando un q mayor para obtener un T_+ adecuado. La segunda limitación recae en la necesidad de conocer la distribución $P_{\boldsymbol{X}}$ para lograr un control exacto de la TFD. En situaciones donde $P_{\boldsymbol{X}}$ es desconocido, se debe modelar con los datos para obtener una aproximación $Q_{\boldsymbol{X}}$. La efectividad de este enfoque depende de la cercanía entre $Q_{\boldsymbol{X}}$ y $P_{\boldsymbol{X}}$. Sin embargo, si $Q_{\boldsymbol{X}}$ no se aproxima lo suficiente a $P_{\boldsymbol{X}}$, existe el riesgo de inflación en la TFD. Barber et al. (2020) proponen una cota basada en la divergencia de Kullback-Leibler como medida para evaluar la proximidad entre distribuciones y cuantificar el impacto potencial de errores en la estimación de $Q_{\boldsymbol{X}}$.

5.4. Implementación computacional del modelo-X de imitaciones

Para la aplicación del modelo-X de imitaciones, se pueden emplear bibliotecas de R y Python que ofrecen funcionalidades específicas y eficientes. En particular, el paquete `knockoff` de `R` (Patterson and Sesia, 2018), se destaca por su capacidad para generar knockoffs Gaussianos y de segundo orden, y por permitir la creación de variables de imitación, usando diversos estadísticos de importancia. Cuando la relación entre la respuesta y los predictores es lineal, se tienen los coeficientes de LASSO y el Máximo Signo de LASSO, mientras que en escenarios de no linealidad, esta paquetería permite aplicar la importancia de variables en bosques aleatorios. La función `knock.filter()` ejecuta el procedimiento Knockoffs de manera integral, requiriendo la especificación de la matriz de predictores, el vector de respuestas, el método para crear knockoffs, el estadístico de importancia y la tasa de falsos descubrimientos. Las guías disponibles, como Advanced Usage of the Knockoff Filter for R, Controlled variable Selection with Fixed-X Knockoffs (Patterson and Sesia, 2022a), y Controlled variable Selection with Model-X knockoffs (Patterson and Sesia, 2022b) proveen orientación valiosa para una implementación efectiva.

Otra biblioteca relevante es `knockpy` de `Python` (Spector and Janson, 2022), que ofrece la capacidad de construir knockoffs Gaussianos utilizando los métodos RMV y ME, minimizando la reconstructibilidad de las variables originales. En entornos no Gaussianos con estructuras gráficas definidas, se puede implementar el algoritmo de Metropolis-Hastings para la generación de knockoffs. Además, se pueden emplear estadísticas de importancia como DeepPINK. Una guía rápida de uso se encuentra disponible en Spector (2020).

La combinación de estas herramientas y bibliotecas proporciona un marco integral para la implementación efectiva del modelo-X de imitaciones en una variedad de escenarios estadísticos. Esto se puede constatar en la implementación computacional del método de creación de knockoffs empleando la cópula Gaussiana latente (Vásquez et al., 2023) donde se combinan estas paqueterías en una Jupyter Notebook. La implementación de este enfoque se encuentra disponible en Vásquez (2022).

6. Casos de éxito

Para comprender plenamente el impacto positivo que ha generado esta metodología, es esencial explorar algunos casos de éxito documentados en la literatura. Un ejemplo destacado es el estudio de Jiang et al. (2021), donde aplicaron la metodología del modelo-X de imitaciones para identificar los genes que mejor estiman la pureza en tumores de carcinoma invasivo de mama (BRCA por su acrónimo en inglés) y melanoma cutáneo (SKCM por su acrónimo en inglés). La pureza, medida en términos del porcentaje de células cancerosas en una muestra de tejido tumoral, se evaluó utilizando el procedimiento KOBT (Knockoff Boosted Trees) sobre datos de expresión génica de RNA obtenidos del proyecto Pan-Cancer Atlas. El algoritmo KOBT demostró su eficacia al detectar con éxito genes cruciales para la estimación de la pureza en los tumores de BRCA y SKCM. Este enfoque no solo validó descubrimientos previos (Li et al., 2019; Yoshihara et al., 2013), sino que también identificó nuevos genes relevantes para este propósito. Este caso ilustra claramente cómo la aplicación

de la metodología de imitaciones ha contribuido de manera significativa a la comprensión y caracterización de la pureza tumoral.

El estudio de Sesia et al. (2021) destaca por su exhaustivo análisis de datos provenientes del biobanco del Reino Unido, específicamente centrado en estudios de asociación genómica (GWAS por sus siglas en inglés). Este enfoque involucra la comparación de miles de variantes genéticas con diversos fenotipos, con el objetivo de identificar asociaciones de interés biológico. El análisis se enfoca tanto en fenotipos de rasgos continuos, como altura, índice de masa corporal, recuento de plaquetas y presión arterial sistólica, como en enfermedades específicas, tales como enfermedad cardiovascular, enfermedad respiratoria, hipertiroidismo y diabetes. El método utilizado fue KnockoffGWAS, el cual se posiciona como una herramienta poderosa al ser comparado con el modelo lineal mixto bayesiano conocido como BOLT-LMM (Loh et al., 2018). Los resultados sugieren que KnockoffGWAS exhibe una mayor capacidad de descubrimiento al identificar sitios adicionales en el genoma (loci) con relevancia biológica. La validación de los descubrimientos mediante recursos externos, como el catálogo GWAS, el proyecto de Biobanco de Japón y el recurso FinnGen, respalda de manera concluyente los resultados obtenidos mediante el método KnockoffGWAS. Estos hallazgos no solo refuerzan la eficacia de la metodología empleada, sino que también sugieren nuevas perspectivas para la interpretación y aplicación de los resultados de los estudios de asociación genómica.

En el ámbito de la medicina clínica, la aplicación de la metodología de knockoffs a cuatro ensayos clínicos de fase III del medicamento Cosentyx (Secukinumab), un anticuerpo monoclonal que inhibe la interleucina 17A, ha arrojado resultados reveladores en la investigación sobre la Artritis Psoriásica (Kormaksson et al., 2021). Este estudio aborda el desafío crucial de identificar factores pronósticos para la respuesta ACR20 a las 16 semanas después de la administración de diferentes dosis de Cosentyx. Cabe destacar que ACR20 es un criterio desarrollado por el Colegio Americano de Reumatología que evalúa la mejora en el número de articulaciones inflamadas y dolorosas en los pacientes. Entre los hallazgos más notables, se destaca la eficacia de las dosis de Cosentyx, superando de manera significativa al placebo. Además, factores demográficos, como la edad y la región, han mostrado asociaciones significativas con la respuesta ACR20. No menos importante, los síntomas iniciales, como la presencia de artritis poliarticular y la intensidad elevada del dolor, se correlacionan positivamente con mayores probabilidades de respuesta al tratamiento. En el ámbito de las variables de laboratorio, aspectos como la "Proteína Total" y la "Creatinina" han destacado en la investigación. La presencia de niveles elevados de proteínas o sus productos de descomposición al inicio del estudio se asocia con mayores probabilidades de respuesta, sugiriendo posibles conexiones con la progresión de la enfermedad. Estos resultados no solo consolidan la eficacia de Cosentyx en el tratamiento de la Artritis Psoriásica, sino que también revelan aspectos clínicos y biomarcadores que podrían ser esenciales para la predicción y la personalización de los enfoques terapéuticos en pacientes afectados.

En el trabajo de Wang et al. (2023), se introduce un algoritmo basado en el modelo-X de imitaciones que destaca por su capacidad para identificar señales a partir de la combinación de pruebas de independencia condicional en múltiples fuentes de información. Esto permite manejar la heterogeneidad tanto de los predictores como de la variable de respuesta presente en las diversas bases de datos. La aplicación de esta metodología se llevó a cabo en infor-

mación proveniente de la Cohorte Nacional de Colaboración COVID (N3C) en los Estados Unidos (Haendel et al., 2021). Esta cohorte abarca registros electrónicos de salud junto con una amplia variedad de datos demográficos, socioeconómicos, comorbilidades y medicamentos de pacientes provenientes de más de 77 sitios. El estudio se centró en la identificación de factores de riesgo asociados con el COVID-19 de larga duración, y los resultados revelaron 17 factores de riesgo significativos. Estos incluyen variables como sexo, edad al inicio del COVID, raza, demencia, obesidad, enfermedad coronaria, uso de corticosteroides sistémicos, depresión, cánceres metastásicos sólidos, enfermedad pulmonar crónica, infarto de miocardio, miocardiopatías, hipertensión, resultado negativo de anticuerpos, número de dosis de la vacuna contra el COVID, visitas a servicios de urgencias relacionadas con el COVID y la enfermedad de células falciformes. Este estudio no sólo resalta la eficacia del modelo-X de imitaciones en la exploración de múltiples fuentes de datos heterogéneas, sino que también proporciona valiosa información sobre una variedad de factores que pueden influir en la persistencia del COVID-19.

En la investigación de Dai and Zheng (2023), se presenta un método de selección de variables basado en knockoffs, diseñado para identificar señales mutuas a partir de múltiples conjuntos de datos independientes. La aplicación de esta metodología se llevó a cabo mediante el análisis de un estudio de tasas de criminalidad, con un enfoque particular en identificar características asociadas con la tasa de criminalidad en la comunidad, independientemente de la distribución racial. Para llevar a cabo este análisis, se utilizaron registros del conjunto de datos de Comunidades y Crimen de la Universidad de California Irvine (UCI), que contienen información sobre tasas de criminalidad y 122 variables adicionales de 1994 comunidades en Estados Unidos con diversas composiciones raciales. Los hallazgos de la investigación resaltan variables específicas que están significativamente vinculadas a las tasas de criminalidad. Entre estas variables se encuentran el "porcentaje de hogares con ingresos de asistencia pública en 1989", el "porcentaje de niños nacidos de padres nunca casados", el "porcentaje de personas en viviendas densas", el "porcentaje de hombres que nunca se han casado" y el "número de hogares vacíos". Estos resultados tienen implicaciones significativas para la comprensión y abordaje de factores criminológicos, permitiendo una aproximación más precisa y fundamentada en la identificación y prevención de delitos.

7. Conclusión

La crisis de replicabilidad en la investigación científica ha sido un desafío persistente que ha comprometido la credibilidad de los hallazgos y sus aplicaciones prácticas. A lo largo de este manuscrito hemos explorado la raíz de esta crisis, identificando causas fundamentales como el sesgo de publicación, incentivos académicos desalineados, errores en la investigación, fraude y el uso inapropiado de técnicas estadísticas. El contexto de pruebas de hipótesis múltiples ha emergido como un componente crítico en la falta de replicabilidad. La exploración de métodos tradicionales y sus limitaciones nos ha llevado a la presentación del modelo-X de imitaciones como una metodología estadística innovadora diseñada para mejorar la replicabilidad en investigaciones científicas. Este enfoque aborda la inferencia selectiva y los desafíos asociados

con pruebas múltiples en contextos de alta dimensionalidad.

Al examinar los aspectos técnicos del modelo-X de imitaciones, hemos destacado su necesidad de generar variables de imitación, la elección de estadísticas de imitación relevantes y la flexibilidad en la aplicación de diferentes métodos para generar knockoffs. A pesar de sus beneficios, también hemos identificado algunas limitaciones, como la dependencia de un umbral supeditado a los datos y la necesidad de un conocimiento preciso de la distribución subyacente. Estas limitaciones marcan un camino para posibles avances y desarrollos futuros en este campo que es un área de investigación activa. Los casos de éxito en la aplicación del modelo-X de imitaciones en diversas áreas subrayan su versatilidad y efectividad. La estimación de pureza en tumores de mama y melanoma cutáneo, los análisis de asociación genómica en grandes biobancos, la identificación de factores pronósticos en ensayos clínicos, la identificación de factores de riesgo asociados con el COVID-19 de larga duración y la selección de variables en estudios de tasas de criminalidad son ejemplos concretos que resaltan la utilidad práctica de esta metodología.

En resumen, el modelo-X de imitaciones emerge como una herramienta prometedora y valiosa para mejorar la replicabilidad en la investigación científica, proporcionando un enfoque robusto y flexible para abordar los desafíos estadísticos asociados con pruebas múltiples en contextos de alta dimensionalidad. Su aplicación exitosa en diversas disciplinas respalda su potencial para impulsar la confianza en los hallazgos científicos y avanzar hacia una investigación más sólida y reproducible. A medida que los científicos buscan fortalecer la base de conocimientos, la implementación de enfoques innovadores como los knockoffs podría ser clave para construir conocimientos más confiables.

8. Agradecimientos

Los autores agradecen el apoyo del CONAHCYT-México a través del Sistema Nacional de Investigadores y el Programa de Investigadoras e Investigadores por México, así como el respaldo del Departamento de Matemáticas de la Universidad Autónoma Metropolitana, Unidad Iztapalapa. Además, expresan su gratitud al revisor anónimo por sus observaciones y sugerencias, las cuales contribuyeron a mejorar este manuscrito.

Bibliografía

A. Ahlgren, "A modest proposal for encouraging replication," *American Psychologist*, vol. 24, no. 4, p. 471, 1969.

R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055 – 2085, 2015.

R. F. Barber, E. J. Candès, and R. J. Samworth, "Robust inference with knockoffs," *The Annals of Statistics*, vol. 48, no. 3, pp. 1409 – 1431, 2020.

J. A. Bargh, M. Chen, and L. Burrows, "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action." *Journal of personality and social psychology*, vol. 71, no. 2, p. 230, 1996.

S. Bates, E. Candès, L. Janson, and W. Wang, "Metropolized knockoff sampling," *Journal of the American Statistical Association*, vol. 116, no. 535, pp. 1413–1427, 2021.

C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, pp. 531–533, 2012.

D. J. Bem, "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of personality and social psychology*, vol. 100, no. 3, p. 407, 2011.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer *et al.*, "Redefine statistical significance," *Nature human behaviour*, vol. 2, no. 1, pp. 6–10, 2018.

Y. Benjamini, "Selective Inference: The Silent Killer of Replicability," *Harvard Data Science Review*, vol. 2, no. 4, dec 16 2020, https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

D. Bishop, "Interpreting unexpected significant findings," 2014.

F. Bretz, T. Hothorn, and P. Westfall, *Multiple comparisons using R*. CRC press, 2016.

K. E. Campbell and T. T. Jackson, "The role of and need for replication research in social psychology," *Replications in social psychology*, vol. 1, no. 1, pp. 3–14, 1979.

E. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold:'Model-X' knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 80, no. 3, pp. 551–577, 2018.

T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

J. Cohen, "Things I have learned (so far)." in *Annual Convention of the American Psychological Association, 98th, Aug, 1990, Boston, MA, US; Presented at the aforementioned conference.* American Psychological Association, 1990.

L. J. Colling and D. Szűcs, "Statistical inference and the replication crisis," *Review of Philosophy and Psychology*, vol. 12, pp. 121–147, 2021.

A. O. Cramer, D. van Ravenzwaaij, D. Matzke, H. Steingroever, R. Wetzels, R. P. Grasman, L. J. Waldorp, and E.-J. Wagenmakers, "Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies," *Psychonomic bulletin & review*, vol. 23, pp. 640–647, 2016.

R. Dai and C. Zheng, "False discovery rate-controlled multiple testing for union null hypotheses: a knockoff-based approach," *Biometrics*, 2023.

S. Doyen, O. Klein, C.-L. Pichon, and A. Cleeremans, "Behavioral priming: It's all in the mind, but whose mind?" *PLOS ONE*, vol. 7, no. 1, 01 2012.

F. Fidler *et al.*, "Should psychology abandon p values and teach cis instead? evidence-based reforms in statistics education," 2006.

E. I. George and R. E. McCulloch, "Approaches for bayesian variable selection," *Statistica Sinica*, pp. 339–373, 1997.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

M. A. Haendel, C. G. Chute, T. D. Bennett, D. A. Eichmann, J. Guinney, W. A. Kibbe, P. R. Payne, E. R. Pfaff, P. N. Robinson, J. H. Saltz *et al.*, "The national covid cohort collaborative (n3c): rationale, design, infrastructure, and deployment," *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 427–443, 2021.

J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, 08 2005.

T. Jiang, Y. Li, and A. A. Motsinger-Reif, "Knockoff boosted tree for model-free variable selection," *Bioinformatics*, vol. 37, no. 7, pp. 976–983, 2021.

J. Jordon, J. Yoon, and M. van der Schaar, "Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks," in *International conference on learning representations*, 2018.

M. Kormaksson, L. J. Kelly, X. Zhu, S. Haemmerle, L. Pricop, and D. Ohlssen, "Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool," *Statistics in Medicine*, vol. 40, no. 14, pp. 3313–3328, 2021.

E. Lander and L. Kruglyak, "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results," *Nature genetics*, vol. 11, no. 3, pp. 241–247, 1995.

Y. Li, D. M. Umbach, A. Bingham, Q.-J. Li, Y. Zhuang, and L. Li, "Putative biomarkers for predicting tumor sample purity based on gene expression data," *BMC genomics*, vol. 20, no. 1, pp. 1–12, 2019.

P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price, "Mixed-model association for biobank-scale datasets," *Nature genetics*, vol. 50, no. 7, pp. 906–908, 2018.

Y. Lu, Y. Fan, J. Lv, and W. Stafford Noble, "DeepPINK: reproducible feature selection in deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

C. C. Mann, "Behavioral genetics in transition: A mass of evidence—animal and human—shows that genes influence behavior. but the attempt to pin down which genes influence which behaviors has proved frustratingly difficult," *Science*, vol. 264, no. 5166, pp. 1686–1689, 1994.

L. Mescheder, S. Nowozin, and A. Geiger, "The numerics of gans," *Advances in neural information processing systems*, vol. 30, 2017.

National Academies of Sciences, Engineering, and Medicine, "Reproducibility and replicability in science," 2019.

R. Nuzzo, "Scientific method: Statistical errors," *Nature*, vol. 506, no. 7487, p. 150, 2014.

Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

H. Pashler, C. Harris, and N. Coburn, "Elderly-related words prime slow walking. psychfiledrawer," 2011.

E. Patterson and M. Sesia, "Advanced usage of the knockoff filter for r," https://cran.r-project.org/web/packages/knockoff/vignettes/advanced.html, 2022, [Online; accessed 03-May-2024].

——, "Controlled variable selection with model-x knockoffs," https://cran.r-project.org/web/packages/knockoff/vignettes/knockoff.html, 2022, [Online; accessed 03-May-2024].

——, "knockoff: The knockoff filter for controlled variable selection," *R package version 0.3*, vol. 2, 2018.

F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?" *Nature reviews Drug discovery*, vol. 10, no. 9, pp. 712–712, 2011.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

Z. Ren, *Model-free Methods For Multiple Testing and Predictive Inference*. Stanford University, 2021.

Y. Romano, M. Sesia, and E. Candès, "Deep knockoffs," *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1861–1872, 2020.

F. Romero, "Philosophy of science and the replicability crisis," *Philosophy Compass*, vol. 14, no. 11, p. e12633, 2019.

K. Sechidis, M. Kormaksson, and D. Ohlssen, "Using knockoffs for controlled predictive biomarker identification," *Statistics in Medicine*, vol. 40, no. 25, pp. 5453–5473, 2021.

M. Sesia, C. Sabatti, and E. J. Candès, "Gene hunting with hidden markov model knockoffs," *Biometrika*, vol. 106, no. 1, pp. 1–18, 2019.

M. Sesia, S. Bates, E. Candès, J. Marchini, and C. Sabatti, "False discovery rate control in genome-wide association studies with population structure," *Proceedings of the National Academy of Sciences*, vol. 118, no. 40, p. e2105841118, 2021.

J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science*, vol. 22, no. 11, pp. 1359–1366, 2011.

N. C. Smith, "Replication studies: A neglected aspect of psychological research." *American Psychologist*, vol. 25, no. 10, p. 970, 1970.

A. Spector, "knockpy," https://amspector100.github.io/knockpy/index.html, 2020, [Online; accessed 03-May-2024].

A. Spector and L. Janson, "Powerful knockoffs via minimizing reconstructability," *The Annals of Statistics*, vol. 50, no. 1, pp. 252–276, 2022.

W. Stroebe, T. Postmes, and R. Spears, "Scientific misconduct and the myth of self-correction in science," *Perspectives on psychological science*, vol. 7, no. 6, pp. 670–688, 2012.

M. Sudarshan, W. Tansey, and R. Ranganath, "Deep direct likelihood knockoffs," *Advances in neural information processing systems*, vol. 33, pp. 5036–5046, 2020.

P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14 516–14 525, 2019.

D. Trafimow and M. Marks, "Editorial," *Basic and Applied Social Psychology*, vol. 37, no. 1, pp. 1–2, 2015.

A. R. Vásquez, J. U. Márquez Urbina, G. González Farías, and G. Escarela, "Controlling the false discovery rate by a latent gaussian copula knockoff procedure," *Computational Statistics*, pp. 1–24, 2023.

A. R. Vásquez, "GitHub: LGCK-LCD," https://github.com/AlejandroRomanVasquez/LGCK-LCD/tree/main, 2022, [Online; accessed 03-May-2024].

R. Wang, R. Dai, and C. Zheng, "Controlling fdr in selecting group-level simultaneous signals from multiple data sources with application to the national covid collaborative cohort data," *arXiv preprint arXiv:2303.01599*, 2023.

R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values: context, process, and purpose," pp. 129–133, 2016.

R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a world beyond $p < 0{,}05$," pp. 1–19, 2019.

K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P. W. Laird, D. A. Levine *et al.*, "Inferring tumour purity and stromal and immune cell admixture from expression data," *Nature communications*, vol. 4, no. 1, p. 2612, 2013.

S. T. Ziliak and D. N. McCloskey, *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2010.

Como citar este artículo: A. R. Vásquez, G. Escarela Pérez, G. Núñez-Antonio, y J. U. Márquez Urbina, " La replicabilidad en la ciencia y el papel transformador de la metodología estadística de knockoffs", Sahuarus. Revista. Electrónica de Matemáticas, vol. 8, no. 1, pp. 1–22, 2024. https://doi.org/10.36788/sah.v8i1.148.