Junio 2024 • e-ISSN: 2448-5365

SahuarUS

Revista electrónica de Matemáticas

Vol. 8
Número 1

UNIVERSIDAD DE SONORA

Sahuarus. Revista Electrónica del Departamento de Matemáticas

SAHUARUS. REVISTA ELECTRÓNICA DE MATEMÁTICAS, número 1, volumen 8, junio 2024 - diciembre 2024, es una publicación semestral, editada por la Universidad de Sonora, a través del Departamento de Matemáticas. Blvd. Luis Encinas y Rosales S/N, colonia Centro, Hermosillo, Sonora, México. C.P. 83000. Tel. (662) 2592155. Página web: sahuarus.unison.mx. Correo electrónico: sahuarus@unison.mx. Editor responsable: Misael Avendaño Camacho. Reserva de Derechos al Uso Exclusivo No. **04-2023-032214541100-102**, e-ISSN: 2448-5365, ambos otorgados por el Instituto Nacional de Derechos de Autor.

Sahuarus. Revista Electrónica del Departamento de Matemáticas

Volumen 8, número 1, Junio 2024- Diciembre 2024

"El saber de mis hijos
hará mi grandeza"

ÍNDICE

ARTÍCULOS

La replicabilidad en la ciencia y el papel transformador de la metodología estadística de knockoffs
Alejandro Román Vásquez, Gabriel Escarela Pérez, Gabriel Núñez Antonio,
José Ulises Márquez Urbina 1-22

Distribuciones de máxima entropía, incremento de aleatoriedad y teorema límite en probabilidad
Evgueni I. Gordienko, Adolfo Minjárez-Sosa 23-44

Criptografía de los cifrados de bloque
Eduardo Velasco-Barreras. 45-82

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.

Vol. 8 / No. 1 / Junio 2024, pp. 23-44
SahuarUS

Distribuciones de máxima entropía, incremento de aleatoriedad y teoremas límite en probabilidad

Evgueni I. Gordienko[1] y Adolfo Minjárez-Sosa[2]
[1] Departamento de Matemáticas, Universidad Autónoma Metropolitana, Iztapalapa
[2] Departamento de Matemáticas, Universidad de Sonora
[1] gord@xanum.uam.mx, [2] adolfo.minjarez@unison.mx

Resumen

En este artículo se explora el concepto de distribuciones de máxima entropía en el contexto de la teoría de la probabilidad. Se analiza la relación entre incremento en la aleatoriedad, máxima entropía y los teoremas límite en probabilidad, proporcionando un enlace entre la incertidumbre estadística y los principios termodinámicos, en particular con la segunda ley de la termodinámica. En términos generales, veremos que el aumento en la entropía va acompañado con el aumento de la incertidumbre o aleatoriedad en un sistema, lo cual puede ser interpretado como una transición a un estado de equilibro termodinámico. Finalmente algunos de estos resultados se presentan en el marco de la teoría de la información.

Palabras Clave: Entropía; Incertidumbre; Teoremas Límete en Probabilidad; Termodinámica; Teoría de la Información.

DOI: 10.36788/sah.v8i1.146
Recibido: 30 de enero de 2024
Aceptado: 26 de junio de 2024

1. Introducción

La palabra "entropía" viene de griego "$\varepsilon\nu\tau\rho o\pi\acute{\iota}\alpha$" y significa evolución o transformación. Dicho término fue acuñado en 1865 por el físico alemán Rudolf Clausius en el contexto de la termodinámica clásica, una rama de la física dedicada al estudio de las relaciones entre las "propiedades macroscópicas" observables de gases, líquidos, etc., tales como presión, temperatura, energía interna, conversión de calor en trabajo mecánico, entre otras. Es por esta razón que la entropía de Clausius se conoce actualmente como entropía termodinámica la cual, en términos generales, puede entenderse como "la porción de la energía térmica de un sistema que no puede convertirse en trabajo mecánico" .

Al mismo tiempo, a través de numerosos estudios, y gracias a una extensa base experimental, se ha establecido firmemente que en sistemas aislados físicos y químicos la entropía aumenta a medida que los sistemas transitan de "estados de transición" a "estados de equilibrio termodinámico" estos últimos caracterizados por una entropía máxima. Por lo tanto, en

ISSN 2448 5365
Universidad de Sonora

sistemas aislados, la entropía siempre aumenta o se mantiene constante. El siguiente ejemplo ilustra esta idea.

Supongamos que la mitad izquierda de un recipiente aislado contiene gas a temperatura constante, mientras que la mitad derecha está vacía (ver Figura 1.1).

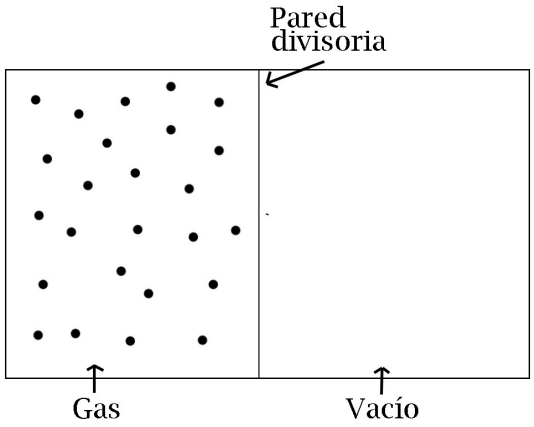

Figura 1:

Si la pared divisoria es removible y, en algún momento, se desliza fuera del recipiente sin romper la hermeticidad del aire, entonces el sistema se encontrará en un estado de desequilibrio, y para restablecer el equilibrio el gas se distribuirá uniformemente en todo el volumen del recipiente en cuestión de milisegundos (ver Figura 1.2).

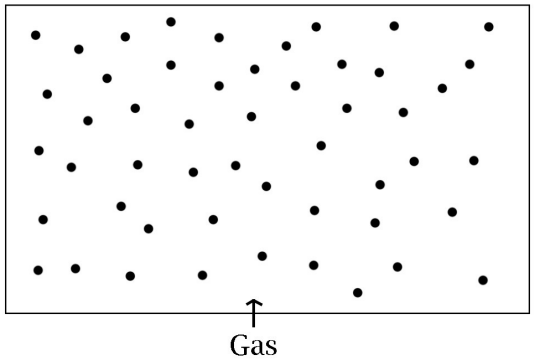

Figura 2:

Posteriormente, veremos que la transición al estado del sistema en la Figura 1.2 está acompañada por un aumento en la entropía, y dicho estado constituye el estado de entropía máxima del gas en el recipiente sin pared divisoria. Es importante destacar que tal transición es irreversible en el tiempo: el gas "nunca" pasará espontáneamente del estado de equilibrio (Figura 1.2) al estado presentado en la Figura 1.3.

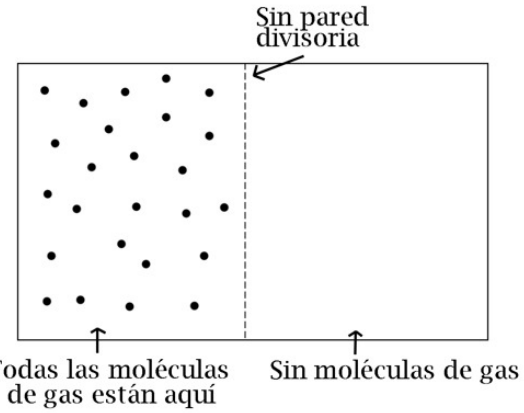

Figura 3:

En resumen, un hecho experimental y parcialmente demostrado matemáticamente es que en sistemas termodinámicos aislados, la entropía no puede disminuir, y dichos sistemas se desplazan de manera irreversible hacia el equilibrio termodinámico, que por lo general se caracteriza por una entropía máxima. Esta esencia constituye la famosa segunda ley de la termodinámica, proporcionando "la flecha del tiempo" es decir, la dirección inevitable del tiempo desde el pasado hacia el futuro.

Años después de la conceptualización de Clausius, el científico austriaco Ludwig Boltzmann introdujo en 1877 la formulación matemática actual de la entropía, y luego J. Willard Gibbs propuso una breve extensión de esta definición. Desde la perspectiva de la teoría de la probabilidad moderna, ambas definiciones utilizan la densidad de probabilidad f_t (en cada tiempo t) de posiciones y velocidades de moléculas de un sistema termodinámico (gas en nuestro caso), definida sobre el espacio fase M, concluyendo que la entropía es definida por el siguiente número:

$$H(f_t) = -\int_M f_t(x) \log [f_t(x)]\, dx, \tag{1}$$

donde x representa las posiciones y velocidades de todas las partículas. Tomando como base esta definición y utilizando ecuaciones de la mecánica clásica para el movimiento de partículas, Boltzmann se dio a la tarea de justificar matemáticamente la segunda ley de la termodinámica. Es decir, demostró que para un "gas ideal" en un recipiente aislado, la entropía $H(f_t)$ no disminuye en función del tiempo, y bajo algunas restricciones, la distribución límite de posiciones es uniforme en el contenedor y la de velocidades es normal. Inicialmente Boltzmann pensó que esta demostración del aumento de la entropía se fundamentaba exclusivamente en las ecuaciones de la mecánica clásica reversibles en el tiempo. Sin embargo, se hizo evidente posteriormente que, de hecho, estaba empleando una suposición probabilística implícita que conduce a una dirección unidireccional del tiempo. En términos probabilísticos, este aumento en la entropía se relaciona con el crecimiento de la incertidumbre o "aleatoriedad" en el sistema, que puede interpretarse como una transición hacia un "estado de equilibrio": a mayor entropía, mayor es la incertidumbre asociada a la variable aleatoria.

Es en este punto donde se relaciona la teoría de la probabilidad y la entropía. En efecto, una parte importante de la teoría de la probabilidad moderna la constituyen los llamados teoremas límite que estudian el comportamiento asintótico de las distribuciones de sumas y productos de variables aleatorias cuando el número de términos crece a infinito. Estos teoremas juegan un papel esencial en la teoría de procesos estocásticos, estadística matemática y, por supuesto, en termodinámica estadística, facilitando la comprensión de por qué las distribuciones de probabilidad como la normal, uniforme y de Poisson, son comunes en diversos fenómenos naturales, ámbitos económicos y contextos científicos. Por ejemplo, el teorema límite mas conocido, el llamado Teorema Central de Límite, establece que la distribución límite de la suma normalizada de variables aleatorias independientes e idénticamente distribuidas con media y varianza finitas, converge a la distribución normal estándar. El punto importante en este teorema es que la distribución límite, la normal, no depende de una distribución particular de las variables aleatorias. En este sentido la distribución límite es universal.

En el ámbito de la entropía, hacia finales del siglo XX, distintos estudios se centraron en el hecho de que las distribuciones límite, en varios teoremas límite, exhiben la máxima entropía dentro de ciertas clases de distribuciones. Un ejemplo de esto es que la densidad normal estándar posee la entropía máxima entre las densidades de variables aleatorias con media cero y varianza unitaria. Mas aún, en el contexto del teorema central del límite, se ha demostrado que las entropías de las sumas de variables aleatorias aumentan a medida que se incrementa el número de términos, llegando su límite a la entropía de la densidad normal estándar, como es de esperarse. En consecuencia, se observa un fenómeno análogo a la segunda ley de la termodinámica. Este resultado no sorprende, dado que las distribuciones de entropía máxima, como la normal o la uniforme, implican mayores niveles de incertidumbre y deben "acumular" las incertidumbres de todas las variables aleatorias hacia las distribuciones límite.

El objetivo del presente artículo es exponer, desde el punto de vista matemático, los aspectos antes descritos, poniendo énfasis en los aspectos intuitivos y de interpretación.

Por otro lado, aun cuando la conceptualización de la entropía tiene sus orígenes en la termodinámica, en 1948 el matemático e ingeniero electricista americano Claude Shannon involucró este concepto en la *Teoría de la Información*. Esta teoría tiene su fundamento en la cuantificación de la información contenida en mensajes que se transmiten através de canales donde está presente algún ruido aleatorio. Básicamente, el trabajo de Shannon consiste en relacionar la entropía termodinámica con la cantidad promedio de información en los mensajes. En este sentido, concluiremos el artículo presentando el papel que juega la entropía en la teoría de información con un ejemplo específico.

2. Entropía como una medida de incertidumbre ("aleatoriedad")

Con el fin de ilustrar el papel que juega la entropía como una medida de la incertidumbre, presentamos algunos ejemplos considerando variables aleatorias discretas y absolutamente continuas. Para este fin, observe que la versión discreta de (1) es la siguiente. Sea X una variable aleatoria con distribución $\mathcal{D}_X$ que toma valores $x_1, x_2, \ldots$. Entonces

$$H(\mathcal{D}_X) = -\sum_k p_k \log(p_k), \tag{2}$$

donde $p_k = P\left[X = x_k\right]$, $k = 1, 2, \ldots$.

Ahora, sea X la variable aleatoria de Bernoulli con parámetro $p \in [0, 1]$. En este caso, X asume solo los valores 0 y 1, y su distribución $\mathcal{D}_X$ es: $P(X = 1) = p$; $P(X = 0) = 1 - p$. Entonces, por (2)

$$H(\mathcal{D}_X) \equiv H(p) = -[p \log(p) + (1 - p) \log(1 - p)]. \tag{3}$$

En (3), $0 \log(0) \stackrel{\text{def}}{=} 0$. La gráfica de la función $H(p)$, $p \in [0, 1]$, está dada en la Figura 4. Observe que el máximo de la entropía es alcanzado en $p = 0.5$, i.e. cuando $P(X = 1) = P(X = 0) = 0.5$ y su valor es

$$H(0.5) \approx 0.69315. \tag{4}$$

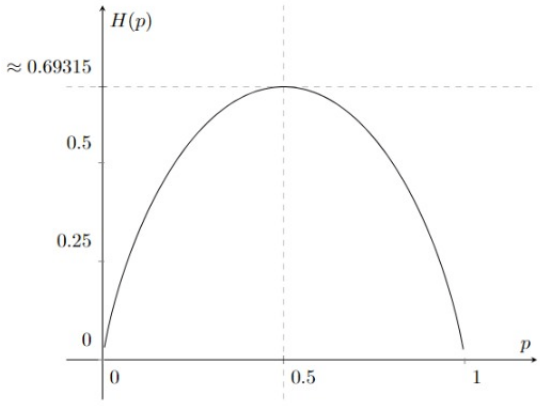

Figura 4:

Este es el caso de *máxima incertidumbre* sobre el posible valor de X. Por ejemplo cuando se lanza una moneda justa, difícilmente se puede predecir el resultado.

Sea ahora $p = 0.99999$, i.e. $P(X = 1) = 0.99999$, $P(X = 0) = 0.00001$. Uno puede "casi seguramente" predecir el valor 1, con lo cual podemos decir que estamos en un caso de *baja incertidumbre*. La entropía en este caso es $H(0.99999) \approx 0.000125129$.

En términos generales, la variable aleatoria con distribución de Bernoulli con $p = 0.99999$ es "menos aleatoria" que la variable aleatoria con distribución de Bernoulli con $p = 0.5$.

Consideremos otro ejemplo. Supongamos que la variable aleatoria X tiene la densidad uniforme f_X en el intervalo (0,1) cuya gráfica se muestra en la Figura 5.

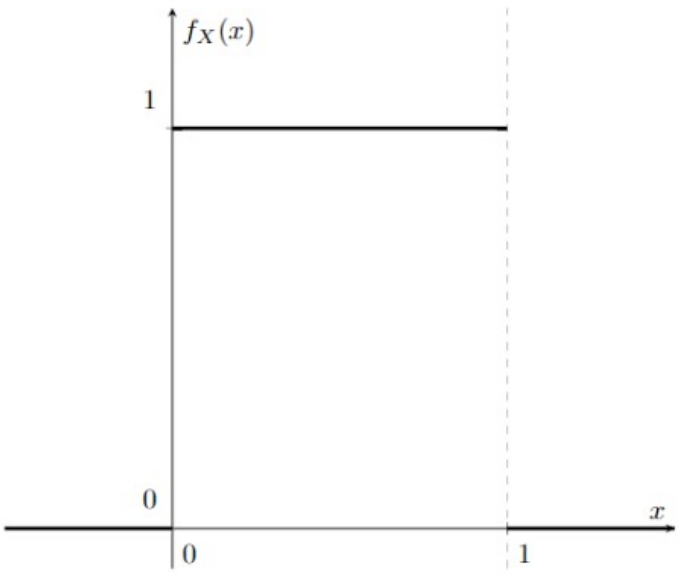

Figura 5:

Entonces, aplicando la definición (1), la entropía de f_X es

$$H(f_X) = -\int_0^1 1\log(1)\,dx = 0.$$

La densidad uniforme en la Figura 5 se acopla a nuestro sentimiento intuitivo de *completa incertidumbre* al "escoger un punto *al azar*" del intervalo $(0, 1)$. En efecto, consideremos la variable aleatoria Y con densidad uniforme f_Y en el intervalo (0.45,0.55) (ver Figura 6). En este caso, estamos bastante seguros de que Y tomará algún valor cercano a 0.5. Por otro lado, la entropía es

$$H(f_Y) = -\int_{0.45}^{0.55} 10\log(10) \approx -2.30258,$$

la cual es mucho menor que $H(f_X) = 0$. Ahora, sea f_Z la densidad uniforme en el intervalo $[-100, 100]$ (ver Figura 7). Entonces, el grado de incertidumbre sobre los posibles valores de Z es bastante alto, y la entropía $H(f_Z) \approx 5.29832$ es muy grande.

A partir de estos ejemplos, una primera conclusión es que entre mas grande es el valor de la entropía, mayor es la incertidumbre asociada a la variable aleatoria.

Otra situación donde se presenta un incremento de la entropía, es decir mayor incertidumbre, es en la suma de variables aleatorias, lo cual toma importancia en algunos teoremas

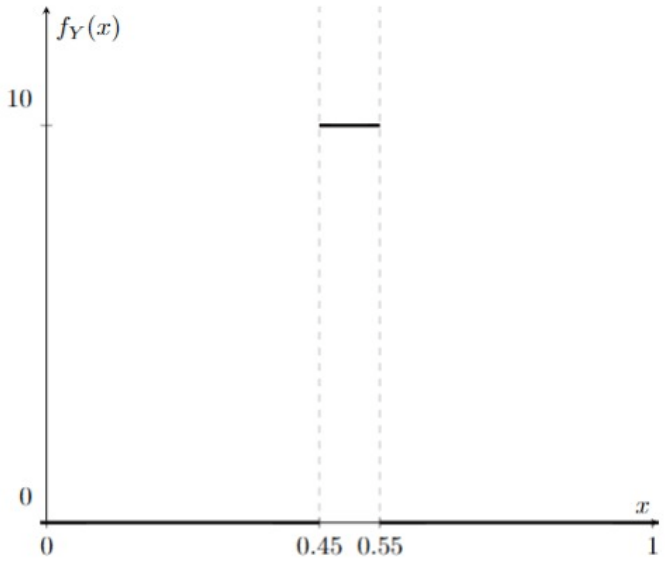

Figura 6:

límite de la Teoría de la Probabilidad. En efecto, como veremos en la Sección 4, similarmente a la segunda ley de la termodinámica, en algunos teoremas límite, la entropía crece conforme aumenta el número de sumandos hasta alcanzar la entropía máxima de ciertas distribuciones.

Vale la pena ilustrar mediante el siguiente ejemplo el por qué la entropía (y la incertidumbre) pueden crecer al aumentar el número de sumandos de variables aleatorias independientes.

Ejemplo 1. *Sean $X_1, X_2, \ldots, X_5$ variables aleatorias independientes e idénticamente distribuidas (i.i.d.), y sea X_1 la variable aleatoria de Bernoulli con parámetro $p = 0.95$. Entonces $P(X_1 = 1) = 0.95$, $P(X_1 = 0) = 0.05$, y realmente esperaríamos la aparición del valor 1 (digamos, "en algún experimento"). Aplicando (3),*

$$H(\mathcal{D}_{X_1}) = -[0.95\log(0.95) + 0.05\log(0.05)] \approx 0.19852, \tag{5}$$

lo cual es significativamente menor que la entropía en (4).

Por otro lado, la suma $S_5 = X_1 + \cdots + X_5$ tiene la distribución binomial $\mathcal{D}_{S_5}$ con parámetros $n = 5$ y $p = 0.95$. Usando la fórmula binomial, encontramos que:

$$\begin{aligned}
P(S_5 = 0) &\approx 0.0000003125;\\
P(S_5 = 1) &\approx 0.000029687;\\
P(S_5 = 2) &\approx 0.0011281;\\
P(S_5 = 3) &\approx 0.021434;\\
P(S_5 = 4) &\approx 0.20363;\\
P(S_5 = 5) &\approx 0.77378.
\end{aligned}$$

Entonces la incertidumbre ha aumentado. En efecto:

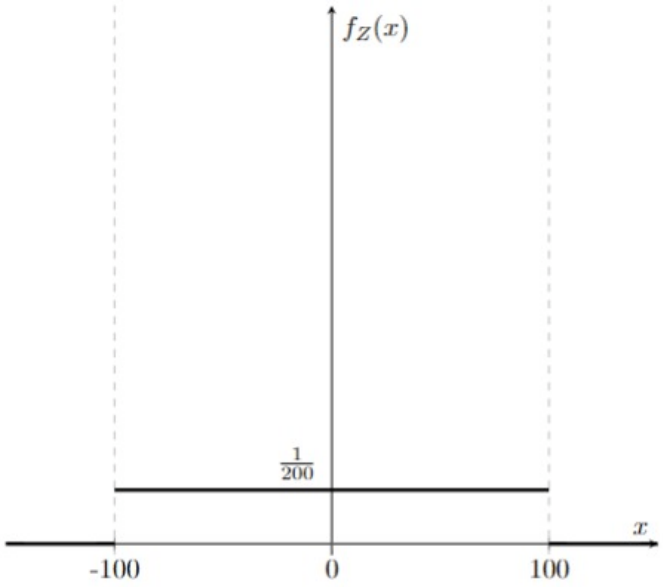

Figura 7:

(a) *ahora la distribución está "dispersa" en el conjunto* $\{0, 1, \ldots, 5\}$*;*

(b) *a pesar de que hay solamente dos valores "realmente probables",* $S_5 = 4$ *y* $S_5 = 5$*, no estamos muy seguros de cuál podría aparecer.*

Lo anterior se refleja en la entropía ya que en este caso, aplicando (2), fácilmente obtenemos:

$$H(\mathcal{D}_{S_5}) \approx 0.61285,$$

lo cual es mayor que la entropía en (5).

3. Densidades de máxima entropía

Para un entero dado $m \geq 1$, sea $\mathbb{R}^m$ el espacio euclidiano m-dimensional, y sea $\mathcal{B}_m$ la σ-álgebra de Borel de subconjuntos de $\mathbb{R}^m$. Sea también $f : \mathbb{R}^m \longrightarrow [0, \infty)$ una densidad (de algún vector aleatorio absolutamente continuo $\bar{X}$).

Definición 2. *De acuerdo con (1), la* ***entropía*** *de* f *se define como el siguiente número (finito o no)*

$$H(f) \stackrel{def}{=} -\int_{\mathbb{R}^m} f(\bar{x}) \log[f(\bar{x})]\, d\bar{x}, \tag{6}$$

siempre que la integral en (6) esté bien definida.

Como siempre, en (6) $0 \log(0) \stackrel{\text{def}}{=} 0$. Las demostración de todas las afirmaciones en esta sección pueden encontrarse, por ejemplo, en el trabajo [5].

3.1. Caso I: Densidades en un conjunto acotado

Sea $B \in \mathcal{B}_m$ un subconjunto acotado de $\mathbb{R}^m$ dado, con volumen (medida de Lebesgue) $\ell(B)$.

Definición 3. *Se dice que una densidad f_{unif} es **uniforme** en B si*

$$f_{unif}(\bar{x}) = \begin{cases} 1/\ell(b), & si\ \bar{x} \in B, \\ 0, & de\ otro\ modo. \end{cases}$$

En la Figura 8 se muestra la gráfica de f_{unif} para el caso $m = 2$.

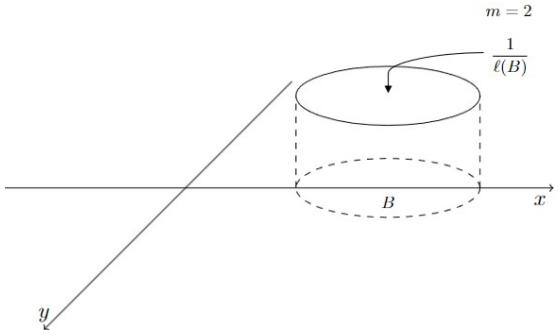

Figura 8:

Por (6), la entropía es

$$H(f_{unif}) = -\int_B 1/\ell(B) \log[1/\ell(B)]\, d\bar{x} = \log[\ell(B)],$$

i.e.

$$H(f_{unif}) = \log[\ell(B)]. \tag{7}$$

Particularmente, en la recta real, si $B = (a, b)$, entonces

$$H(f_{unif}) = \log(b - a). \tag{8}$$

Consideremos la familia $\mathcal{D}_B$ de todas las densidades $f : B \longrightarrow [0, \infty)$ con soporte el conjunto B, i.e. $f(x) = 0$ para cada $x \notin B$.

Proposición 4. *Para cada $f \in \mathcal{D}_B$, $H(f) \leq H(f_{unif}) = \log[\ell(B)]$. Además, para $f \in \mathcal{D}_B$, $H(f) = \log[\ell(B)]$ si y solo si $f = f_{unif}$ (en "casi todos los puntos de B").*

Esta proposición afirma que entre todas las densidades con soporte en un conjunto acotado fijo, la densidad uniforme tiene entropía máxima.

Observacion 5. *La Proposición 4 y (7) dan argumentos adicionales a favor de la segunda ley de la termodinámica, y explican el cambio en la distribución del gas, por ejemplo, en las Figuras 1.1 y 1.2 en la Sección 1. Mas aún, de (7) vemos que la distribución uniforme en un volumen grande tiene mayor entropía.*

3.2. Caso II: Densidades en $\mathbb{R}^m$ con promedio y una matriz de covarianza fijas

Sea $\bar{X} = (X_1, \ldots, X_m)$ un vector aleatorio (con valores en $\mathbb{R}^m$) tal que $E\,||\bar{X}||^2 < \infty$. Entonces, el promedio $\bar{a} = (EX_1, \ldots, EX_m)$ y la matriz de covarianza Σ con elementos $Cov(X_i, X_j) \stackrel{\text{def}}{=} E[(X_i - EX_i)(X_j - EX_j)]$, $i, j = 1, 2, \ldots, n$ están bien definidos.

Fijemos un vector arbitrario $\bar{a} \in \mathbb{R}^m$ y una matriz de $m \times m$ simétrica y positiva definida Σ, y denotemos por $\mathcal{D}_{\bar{a},\Sigma}$ a la familia de todas las densidades de vectores aleatorios de cuadrado integrables continuos en $\mathbb{R}^m$ con promedio $\bar{a}$ y matriz de covarianza Σ. Sea también

$$f_N(\bar{x}) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x} - \bar{a})^T \Sigma^{-1}(\bar{x} - \bar{a})\right\} \tag{9}$$

la densidad Normal del vector aleatorio Normal $\bar{X}$ con promedio $\bar{a}$ y matriz de covarianza Σ.

En (9), $|\Sigma|$ es el *determinante* de Σ y Σ^{-1} es la matriz inversa. Aplicando (6), no es difícil calcular la entropía de f_N lo cual se establece en el siguiente resultado.

Proposición 6. *La entropía de la densidad f_N está dada por*

$$H(f_N) = \frac{1}{2}[m + \log(2\pi|\Sigma|)]. \tag{10}$$

Proposición 7. *Para cada densidad $f \in \mathcal{D}_{\bar{a},\Sigma}$, $H(f) \leq H(f_N)$. Además, para $f \in \mathcal{D}_{\bar{a},\Sigma}$, $H(f) = H(f_N)$ (dado en (10)) si y solo si $f = f_N$ (ver (9)).*

Por lo tanto, de nuevo, en la clase de todas las densidades con promedio y matriz de covarianza fijos, la **densidad Normal** tiene **entropía máxima**.

El caso uno-dimensional es el más importante para nosotros. La fórmula (9) resulta ser:

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}},\ x \in \mathbb{R}. \tag{11}$$

Si $X \sim Norm(a, \sigma)$ es la variable aleatoria Normal con promedio $a = EX$ y varianza $Var(X) = \sigma^2$, i.e. $f_X = f_N$, entonces podemos decir que dadas las restricciones anteriores, la variable aleatoria X se interpreta como la variable aleatoria "más incierta" (o, informalmente, "la más aleatoria").

En vista de (10), la entropía de $f_N = f_X$ con $X \sim Norm(a, \sigma)$ es

$$H(f_N) = \frac{1}{2} + \log(\sqrt{2\pi}\sigma), \tag{12}$$

y para la variable aleatoria Normal estándar: $\eta \sim Norm(0, 1)$ (i.e. con $a = 0$ y $\sigma = 1$)

$$H(f_\eta) = \frac{1}{2} + \log(\sqrt{2\pi}) \approx 1.41894. \tag{13}$$

Observacion 8. *(a) La Proposición 7 arroja algo de luz sobre la presencia de la densidad Normal (de una velocidad) ya que se sostiene el principio general de que en la termodinámica los estados de equilibrio tienen entropía máxima.*

(b) Quizá, es la propiedad de maximalidad de entropía de las densidades Normales lo que puede explicar (al menos parcialmente) la prevalencia de cantidades normalmente distribuidas en la naturaleza y la ciencia. (Ver también la Sección 4.1 sobre la conexión entre el teorema central del límite y el crecimiento de entropía.)

3.3. Caso III: Densidades en $(0, \infty)$ con un promedio fijo

Sea $\lambda > 0$ un número dado, y sea $\mathcal{D}_\lambda$ la familia de todas las densidades f_X de variables aleatorias continuas **positivas e integrables** tales que $EX = 1/\lambda$. Uno de los miembros de $\mathcal{D}_\lambda$ es la densidad de la **variable aleatoria exponencial** $X_{Exp} \sim Exp(\lambda)$ con el parámetro $\lambda > 0$. Esto es:

$$f_{Exp}(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

cuya gráfica se muestra en la Figura 9.

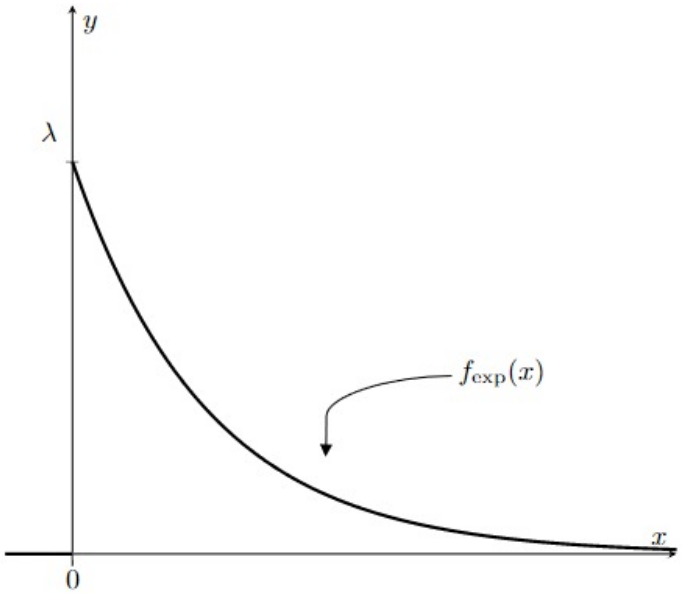

Figura 9:

Proposición 9. *Para cada $f \in \mathcal{D}_\lambda$, $H(f) \leq H(f_{Exp})$. Además, para $f \in \mathcal{D}_\lambda$, $H(f) = H(f_{Exp})$ si y solo si $f = f_{Exp}$.*

Por lo tanto, la *densidad exponencial* es de *máxima entropía* en la clase $\mathcal{D}_\lambda$, la cual, por

(6), está dada como:

$$H(f_{Exp}) = -\int_0^\infty \lambda e^{-\lambda x} \log(\lambda e^{-\lambda x})\, dx = -\log\lambda + 1, \text{ i.e.}$$
$$H(f_{Exp}) = 1 - \log\lambda. \tag{14}$$

Ejemplo 10. *Sea $X \sim Exp(\lambda = 1)$ (ver Figura 10),*

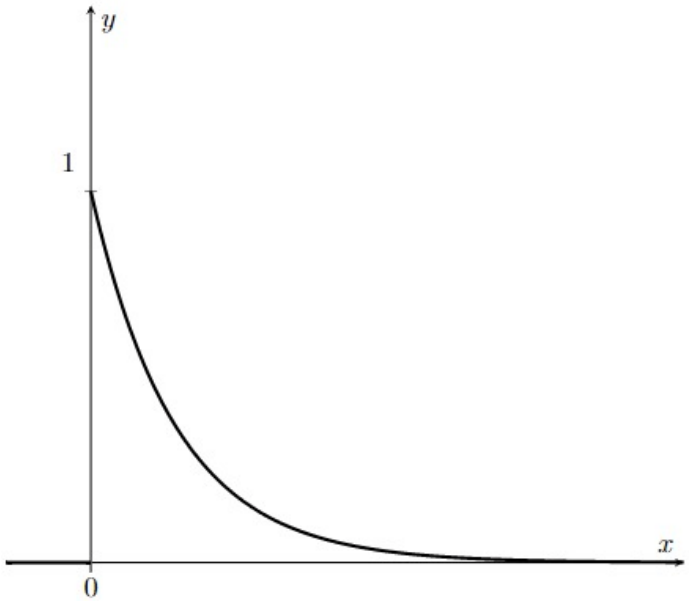

Figura 10:

y supóngase que la variable aleatoria Y tiene densidad $f_Y(x) = 0.5e^{-|x|}$, $x \in \mathbb{R}$. Observe que $f_Y \notin \mathcal{D}_\lambda$ (ver Figura 11).

Por (14), $H(f_X) = 1$ y con cálculos simples, $H(f_Y) \approx 1.6932$. Se puede observar de las Figuras 10 y 11 que comparado con Y, la variable aleatoria X es "menos incierta".

4. Densidades de máxima entropía y algunos teoremas límite

Los teoremas del límite central y otros teoremas límite asociados con ciertos esquemas de suma de variables (o vectores) aleatorias juegan un papel prominente no solo en la Teoría de Probabilidad moderna, si no también en estadística matemática, la teoría de procesos estocásticos y sus numerosas aplicaciones.

Como sabemos de cálculo, la suma de una serie convergente puede ser cualquier número. Cuando se suman (y normalizan apropiadamente) variables aleatorias independientes, las distribuciones límite comúnmente resultan ser **universales** (i.e. las mismas para diferentes

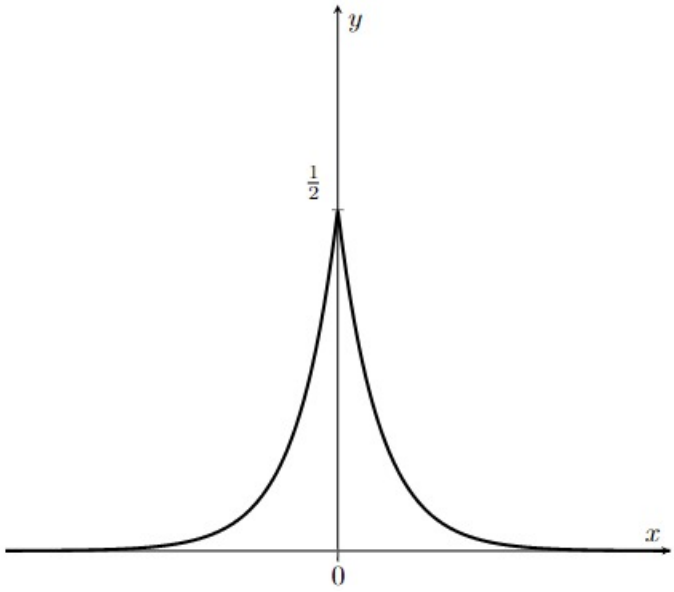

Figura 11:

distribuciones de sumandos). Es incluso más curioso que esas distribuciones (densidades) límite tienen entropía máxima en clases específicas de densidades. Podría decirse que la suma "acumula (en cierto sentido) incertidumbres" y las lleva a la densidad del límite. Esto se compara con la segunda ley de la termodinámica.

4.1. Teorema Central del Límite y crecimiento de entropía

Una de las formas más simples del Teorema Central del Límite (TCL) es la siguiente.

Sean $X_1, X_2, \dots$ variables aleatorias i.i.d. tales que $\sigma^2 = Var(X_i) \in (0, \infty)$. Denotando $a = EX_1 (= EX_2 = EX_3 = \dots)$, sean $S_n = X_1 + X_2 + \cdots + X_n$, $n = 1, 2, \dots$, y

$$Y_n = \frac{S_n - na}{\sigma\sqrt{n}},\ n \geq 1, \tag{15}$$

éstas últimas llamadas sumas normalizadas. Sea también $\eta \sim N(0,1)$ la variable aleatoria Normal estándar. Entonces la "versión canónica" del TCL dice:

Para cada $x \in \mathbb{R}$, cuando $n \to \infty$, $F_{Y_n}(x) \to F_\eta(x)$, donde F_{Y_n} y $F_\eta(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$ son las funciones de distribución correspondientes.

Es bien sabido que suponiendo que X_1 tiene una densidad, bajo ciertas hipótesis adicionales obtenemos la *convergencia de densidades* (ver e.g. [10]), esto es:

- ya sea

$$f_{Y_n}(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}, \text{ cuando } n \to \infty, \tag{16}$$

- o (bajo otras hipótesis complementarias) tenemos la convergencia en la llamada norma L_1:

$$\int_{-\infty}^{\infty} |f_{Y_n}(x) - f_\eta(x)|\, dx \to 0 \text{ cuando } n \to \infty. \tag{17}$$

En (16) y (17), f_{Y_n} es *la densidad* de la variable aleatoria Y_n y

$$f_\eta(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{18}$$

es la densidad de la variable aleatoria Normal estándar.

Para simplificar la notación (y, de hecho, sin pérdida de generalidad), a partir de ahora supondremos que $a = 0$ y $\sigma^2 = 1$. Entonces las sumas normalizadas en (15) tomarán la forma:

$$Y_n = \frac{S_n}{\sqrt{n}},\ n = 1, 2, \ldots. \tag{19}$$

Además, supongamos que la variable aleatoria X_1 tiene una densidad. Como antes, la ***entropía*** de Y_n es

$$H(f_{Y_n}) \stackrel{\text{def}}{=} -\int_{-\infty}^{\infty} f_{Y_n}(x) \log[f_{Y_n}(x)]\, dx. \tag{20}$$

Para establecer el siguiente resultado usaremos la siguiente definición (ver, e.g., [9]).

Definición 11. *Sean f y g densidades definidas en $\mathbb{R}$ tal que si $f > 0$ entonces $g > 0$. Definimos la entropía relativa de f con respecto a g como*

$$H(f|g) = \int_{\mathbb{R}} f(x) \log\left[\frac{f(x)}{g(x)}\right] dx,$$

cuando la integral existe.

Un resultado que relaciona la entropía relativa con las densidades es la llamada desigualdad de Pinsker-Csiszár-Kullback (ver [11]):

$$\int_{\mathbb{R}} |f(x) - g(x)|\, dx \le \sqrt{2H(f|g)}. \tag{21}$$

Ahora, obeserve que la entropía relativa de Y_n con respecto a la densidad Normal f_η en (18) es

$$H(f_{Y_n}|f_\eta) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f_{Y_n}(x) \log\left[\frac{f_{Y_n}(x)}{f_\eta(x)}\right] dx. \tag{22}$$

En 1986, A. Barron demostró la siguiente "versión entrópica" del TCL.

Teorema 12. *(Ver [2])*

*(a) Si para algún $n_0 \geq 1$, $H(f_{Y_{n_0}}|f_\eta)$ es **finito**, entonces, cuando $n \to \infty$*

$$H(f_{Y_n}|f_\eta) \to H(f_\eta|f_\eta) = 0. \tag{23}$$

*(b) Si para algún $n_0 \geq 1$, $H(f_{Y_{n_0}})$ es **finito**, entonces, cuando $n \to \infty$,*

$$H(f_{Y_n}) \to H(f_\eta) = 0.5 + \log(\sqrt{2\pi}). \tag{24}$$

Observacion 13. *En (15), $EY_n = \frac{1}{\sqrt{n}}E(X_1 + \cdots + X_n) = 0$, $n = 1, 2, \ldots$, y $Var(Y_n) = \frac{1}{n}Var(X_1 + \cdots + X_n) = 1$, $n = 1, 2, \ldots$. Además, $E\eta = 0$ y $Var(\eta) = 1$. Por lo tanto (ver Proposición 7), (24) afirma que cuando $n \to \infty$, la entropía de f_{Y_n} se acerca a la **entropía máxima** en la clase de densidades de variables aleatorias con media cero y varianza unitaria.*

De (23) y la desigualdad (21) se obtiene la L_1-convergencia de las densidades como se establece en el siguiente resultado.

Corolario 14. *Bajo las condiciones de la parte (a) del Teorema 12,*

$$\int_{-\infty}^{\infty} |f_{Y_n}(x) - f_\eta(x)|\, dx \to 0 \text{ cuando } n \to \infty. \tag{25}$$

Observacion 15. *(Ver [2]) (a) Las condiciones del Teorema 12(a) y el Corolario 14 se satisfacen si para algún n la densidad f_{Y_n} es acotada.*

(b) Existe una variable aleatoria X_1 tal que f_{Y_n} es no acotada para cada $n \geq 1$, pero (24) y (25) son verdaderas, a pesar de que la densidad f_η es acotada.

La siguiente pregunta interesante es si la convergencia de entropía en (24) es *monótona*. La respuesta positiva fue dada en el año 2004 por S. Arstein, K. Ball, F. Barthe y A. Naor (ver [1]).

Teorema 16. *[1] Para cada $n = 1, 2, \ldots$*

$$H(Y_{n+1}) \geq H(Y_n) \tag{26}$$

o

$$H\left(\frac{S_{n+1}}{\sqrt{n+1}}\right) \geq H\left(\frac{S_n}{\sqrt{n}}\right).$$

De (26) vemos que en (24) la entropía de las densidades de sumas normalizadas converge monótonamente (i.e. no decrecientemente) a su máximo valor $H(f_\eta)$.

Ejemplo 17. *Sea $X_1 \sim Unif(-\sqrt{3}, \sqrt{3})$ (la variable aleatoria uniforme en el intervalo $(-\sqrt{3}, \sqrt{3})$). Entonces $a = EX_1 = 0$; $\sigma^2 = Var(X_1) = 1$, y en (15) $Y_1 = \frac{X_1}{\sqrt{1}} = X_1$, con $H(Y_1) \approx 1.2425$ (ver (8)). En vista de los Teoremas 12 y 16, podemos ver que dado el "valor inicial" 1.2425, la entropía $H\left(\frac{S_n}{\sqrt{n}}\right)$ converge monótonamente a $H(f_\eta) \approx 1.41894$ (ver (13)).*

Bajo las condiciones del Teorema 12 la entropía relativa $H(f_{Y_n}|f_\eta)$ tiende a cero, cuando $n \to \infty$. En un trabajo relativamente reciente [4], S. Bobkov, G. Chistyakov y F. Götze establecieron la tasa de convergencia. En efecto, si $EX_1^4 < \infty$ y la sucesión $\{H(f_{Y_n}|f_\eta)\}$ está acotada, entonces para alguna constante C se cumple

$$H(f_{Y_n}|f_\eta) \leq \frac{C}{n},\ n = 1, 2, \ldots, \tag{27}$$

y, aplicando la desigualdad (21),

$$\int_{-\infty}^{\infty} |f_{Y_n}(x) - f_\eta(x)|\, dx \leq \frac{\sqrt{2C}}{\sqrt{n}},\ n = 1, 2, \ldots. \tag{28}$$

Nótese que la cota en (28) es la típica tasa de convergencia en el TCL.

4.2. Suma de variables aleatorias independientes en grupos. Convergencia a las distribuciones uniformes y crecimiento de entropía

En la Sección 3 vimos que una densidad uniforme tiene la entropía máxima entre las densidades con soporte un conjunto acotado. Hay una amplia clase de teoremas del límite que tratan con la suma de variables aleatorias en grupos compactos. Este es el caso cuando las densidades límite son uniformes. Solo consideraremos un esquema particular simple: *adición módulo 1* en el grupo (intervalo) $[0, 1) \subset \mathbb{R}$.

Consideremos al grupo (intervalo) $G = [0, 1)$ equipado con la operación $\oplus$ que representa la adición módulo 1, i.e.

$$x \oplus y = \begin{cases} x + y & \text{if } x + y < 1, \\ x + y - 1 & \text{if } x + y \geq 1. \end{cases} \tag{29}$$

Entonces $(G, \oplus)$ es un ejemplo simple de un *grupo conmutativo compacto*.

En este caso, la distribución, la densidad (si existe) y la independencia de variables aleatorias con valores en G se definen de manera usual, considerando a $G = [0, 1)$ como un subconjunto de Borel de $\mathbb{R}$. En este sentido, una función (medible) $f_X : [0, 1) \to [0, \infty)$ es *la densidad* de una variable aleatoria X en G si $P(X \in B) = \int_B f_X(x)\, dx$, para cada subconjunto de Borel $B \subset [0, 1)$. Particularmente, $f_{unif}(x) = 1$, $x \in [0, 1)$ es llamada la *densidad uniforme*.

Sean $X_1, X_2, \ldots$ variables aleatorias con valores en G. Usando la operación (29) se define la variable aleatoria $X_1 \oplus X_2$ en G. Luego, por inducción, la suma $S_n \stackrel{\text{def}}{=} X_1 \oplus X_2 \oplus \cdots \oplus X_n$ se define para cada entero $n = 1, 2, \ldots$.

No es difícil demostrar que si las variables aleatorias $X_1, X_2, \ldots$ son i.i.d. y X_1 tiene una densidad (i.e. X_1 es "absolutamente continua"), entonces para cada $n \geq 1$, la variable aleatoria S_n tiene una densidad, que se denota por f_{S_n}.

El siguiente teorema fue demostrado en 1972 por R. N. Bhattacharya (ver [3]).

Teorema 18. *Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. en G con una densidad común f_X. Asúmase que hay una constante $\alpha > 0$ tal que*

$$\inf_{x\in[0,1)} f_X(x) \geq \alpha. \tag{30}$$

Entonces:

$$\int_0^1 |f_{S_n}(x) - f_{unif}(x)|\, dx \equiv \int_0^1 |f_{S_n}(x) - 1|\, dx \leq 2(1-\alpha)^n, \quad n = 1, 2, \ldots. \tag{31}$$

Observe que la condición (30), aunque es muy fuerte, garantiza que la tasa de convergencia en (31) de las densidades f_{S_n} a la uniforme, $f_{unif} \equiv 1$, sea geométrica, la cual es mas rápida en comparación con la tasa en (28).

Ejemplo 19. *Sea* $f_X(x) = \begin{cases} \frac{1}{2\sqrt{x}}, & x \in (0,1) \\ 1/2, & x = 0. \end{cases}$

Entonces en (30), $\alpha = 0.5$ y, en efecto, $\int_0^1 |f_{S_{30}}(x) - 1|\, dx \leq 0.00000000187$.

En los Teoremas 12 y 16 notamos que la convergencia de las densidades de las sumas normalizadas $Y_n = \frac{S_n}{\sqrt{n}}$ a la densidad Normal estándar $f_\eta(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ (de máxima entropía) está acompañada por un *incremento monótono* de la *entropía* $H(Y_n)$ hacia su máximo valor $H(f_\eta)$. A partir de lo anterior, surge la siguiente pregunta: ¿Qué pasa con el crecimiento de entropía en el Teorema 18?. Se tiene una respuesta a partir de la desigualdad inversa en (21). En efecto, en el artículo [11], bajo ciertas condiciones adicionales, la desigualdad contraria a (21) fue demostrada. Esto es, para alguna constante *positiva* γ:

$$\left\{ \int_{\mathbb{R}} |f(x) - g(x)|\, dx \right\}^2 \geq \gamma H(f|g). \tag{32}$$

Puede demostrarse que bajo las hipótesis del Teorema 18, las condiciones mencionadas para (32) se satisfacen. Por lo tanto, combinando (31) y (32), y observando que $H(f_{S_n}|f_{unif}) = -H(f_{S_n})$ obtenemos el siguiente resultado.

Proposición 20. *Bajo las hipótesis del Teorema 18,*

$$0 \geq H(f_{S_n}) \to H(f_{unif}) = 0, \ \textit{cuando } n \to \infty. \tag{33}$$

Vale la pena enfatizar nuevamente que en el lado derecho de (33) tenemos la *entropía máxima* de densidades en $[0, 1)$.

4.3. El Teorema de Rényi y la convergencia a la densidad exponencial

Observe que la densidad exponencial entra en una de las clases de densidades de máxima entropía que examinamos en la Sección 3. Nos preguntamos ahora si con esta densidad existen

resultados análogos a los teoremas límite analizados anteriormente. Para responder este punto presentaremos el bien conocido hecho de que una distribución exponencial es el límite débil de las llamadas *sumas geométricas*. Esto es parte del Teorema de Rényi publicado por el matemático Alfréd Rényi en 1956.

Sean $X_1, X_2, \ldots$ variables aleatorias no negativas i.i.d. tales que $EX_1^2 < \infty$ y $a \stackrel{\text{def}}{=} EX_1 > 0$. Sea también N una variable aleatoria *geométrica* con parámetro $p \in (0,1)$. Asúmase que N es *independiente* de $X_1, X_2, \ldots$. Recordemos que la variable aleatoria N es geométrica si $P(N = k) = p(1-p)^{k-1}$, $k = 1, 2, \ldots$.

Consideremos las siguientes *sumas geométricas*:

$$S_N \stackrel{\text{def}}{=} X_1 + X_2 + \ldots X_N. \tag{34}$$

Nótese que el número de términos en (34) es aleatorio y sigue la distribución geométrica.

Teorema 21 (Rényi)**.** *Cuando* $p \to 0$*,*

$$pS_N \Rightarrow Y, \tag{35}$$

donde $Y \sim Exp(\lambda = 1/a)$ *es la variable aleatoria con la densidad exponencial con parámetro* $\lambda = 1/a$*.*

El símbolo "$\Rightarrow$" en (35) significa la convergencia *débil*, i.e. la *convergencia en distribución*. En el libro [8], Capítulo 3, V. Kalashnikov demostró un resultado más fuerte.

Teorema 22. *Bajo las hipótesis del Teorema 21 existe una constante* $\beta < \infty$ *tal que*

$$\sup_{x \in \mathbb{R}} |P(pS_N \le x) - P(Y \le x)| \le \beta p,\ p \in (0,1). \tag{36}$$

En particular, el lado izquierdo de (36) tiende a cero cuando $p \to 0$*.*

Si adicionalmente suponemos que X_1 tiene una densidad f_{X_1}, entonces la variable aleatoria pS_N también tiene una densidad f_{pS_N}. Desafortunadamente, escapa a nuestro conocimiento bajo qué condiciones se puede afirmar que, cuando $p \to 0$,

$$\int_0^\infty |f_{pS_N}(x) - \frac{1}{a} e^{-x/a}|\, dx \to 0. \tag{37}$$

La convergencia en (36) sugiere que $P(pS_N \in I) \to P(Y \in I)$ uniformemente sobre **todos los intervalos** $I \subset \mathbb{R}$, mientras que (37) es equivalente a $P(pS_N \in B) \to P(Y \in B)$ uniformemente sobre todos los **subconjuntos de Borel** de $\mathbb{R}$. La última convergencia es referida como *convergencia fuerte*.

En el Teorema 12 y el Corolario 14 observamos que la convergencia fuerte está estrechamente relacionada con el *incremento de la entropía* de sumas de variables aleatorias a

la entropía máxima en las clases apropiadas. Ahora, en el contexto del Teorema de Rényi, podemos notar que por la igualdad de Wald

$$E(pS_N) = pEX_1EN = EX_1 = a.$$

Recordemos que (ver Sección 3) en la clase de todas las densidades en $(0,\infty)$ con un promedio fijo a, la densidad exponencial f_{exp} tiene la entropía máxima $H(f_{exp}) = 1 + \log(a)$ (ver (14)), obtenemos que para cada $p \in (0,1)$, $H(f_{pS_n}) \leq H(f_{exp}) = 1+\log(a)$. Sin embargo, no tenemos información sobre el siguiente hecho: $H(f_{pS_N}) \to H(f_{exp})$ cuando $p \to 0$. Tal vez este es un problema abierto.

5. Crecimiento de entropía y teoría de la información

La entropía en la teoría de la información guarda una conexión con la entropía termodinámica. Como lo hemos comentado anteriormente, en el contexto de la termodinámica, la entropía se basa en el análisis de un sistema de partículas con estados que representan posición y velocidad, y los cuales poseen una distribución de probabilidad específica. En la teoría de la información clásica (ver, e.g., [6]), para una distribición discreta $\mathcal{D}_X$, la *entropía de Shannon* $H_{Sh} = H$ se define mediante la misma ecuación (2), y eso estima la cantidad promedio de información contenida en los valores de la variable aleatoria X con distribución $\mathcal{D}_X$. Por ejemplo, como vimos en la Sección 2, si $\mathcal{D}_X$ es la distribución de Bernoulli con $p = 0.5$, entonces al observar la realización del valor $X = 1$, hemos obtenido la información $H_{Sh} \approx 0.69315$. Pero, si $p = 0.99999$ entonces $H_{Sh} \approx 0.000125129$ (un "poco de información" ya que el valor $X = 1$, en este caso, es muy posible.)

Ilustraremos la observación anterior mediante el siguiente ejemplo (ver también [7]).

Sea $p \in (0,1)$ un número dado y $X_1, X_2, ...$ variables aleatorias i.i.d. tomando valores en $\{-1,1\}$, con distribución común

$$P[X_1 = 1] = p \quad \text{y} \quad P[X_1 = -1] = 1 - p.$$

Además, definimos la variable aleatoria

$$Y_n := X_0 \cdot X_1 \cdot ... \cdot X_n, \tag{38}$$

donde $X_0 = 1$. Es claro que la variable aleatoria Y_n también toma valores en $\{-1,1\}$.

En determinado contexto, las variables aleatorias $X_0, X_1, ...$ y Y_n podrían tener la siguiente interpretación en los términos de la retransmisión múltiple de algún mensaje.

Supongamos que el evento $[X_0 = 1]$ representa que alguna persona ha recibido un mensaje completamente veráz que solo acepta significados de "si" o "no". Este mensaje es transmitido a otro sitio, ya sea correctamente, lo cual se representa con el evento $[X_1 = 1]$, con probabilidad p, o cambiando su contenido al lado opuesto representado por el evento $[X_1 = -1]$, con probabilidad $1 - p$. Ahora, este primer receptor retransmite el mensaje bajo el mismo procedimiemto, esto es,

- con probabilidad p el mensaje es retransmitido en la forma en la cual ha sido recibido, lo cual se representa por el evento $[X_2 = 1]$;
- con probabilidad $1 - p$ el mensaje es retransmitido cambiando su contenido a lo contrario, representado por el evento $[X_2 = -1]$.

Una vez que el segundo receptor recibe el mensaje, este procedimiento se repite una y otra vez.

Bajo este escenario, observe que $[Y_n = 1]$ representa el evento de que el mensaje ha sido recibido correctamente despues de n retransmisiones. Entonces es importante analizar el comportamiento de Y_n para n suficientemente grande y la relación con su entropía. Para este fin, primero calculemos la distribución de Y_n.

Observe que

$$EX_1 = 1 \cdot p + (-1)(1 - p) = 2p - 1.$$

Además, por la independencia de las variables aleatorias $X_1, X_2, ...,$

$$EY_n = EX_1 EX_2 \cdots EX_n = (2p - 1)^n. \tag{39}$$

Por otro lado, sea $p_n = P[Y_n = 1]$ y $1 - p_n = P[Y_n = -1]$. Entonces

$$EY_n = 1 \cdot p_n + (-1)(1 - p_n) = 2p_n - 1. \tag{40}$$

Combinando (39) y (40) tenemos

$$2p_n - 1 = (2p - 1)^n,$$

con lo cual obtenemos la distribución de Y_n dada por

$$P[Y_n = 1] = p_n = \frac{(2p - 1)^n + 1}{2} \tag{41}$$

y

$$P[Y_n = -1] = 1 - p_n = \frac{1 - (2p - 1)^n}{2}, \quad n = 1, 2, ... \tag{42}$$

Ahora, como $p \in (0, 1)$, se cumple $|2p - 1| < 1$. Entonces, de (41) y (42), cuando $n \to \infty$,

$$P[Y_n = 1] \to \frac{1}{2} \quad \text{y} \quad P[Y_n = -1] \to \frac{1}{2}, \tag{43}$$

con una tasa de convergencia exponencial.

Observe que este comportamiento es el mismo para cada valor de $p \in (0, 1)$. Entonces, el $n-$ésimo receptor, i.e., el último receptor, para n suficientemente grande, conocerá la

veracidad o la falsedad del mensaje con la misma probabilidad 1/2, es decir con completa incertidumbre.

Analicemos esto último desde el punto de vista de la entropía considerando el siguiente caso particular con $p = P[X_1 = 1] = 0.9999$. Observe (ver (2)) que la entropía de la variable aleatoria $Y_1 = X_1$ es aproximadamente 0.00082108, muy cercana a cero. De acuerdo a (43), cuando $n \to \infty$, la entropía de Y_n crece a su máximo valor $-\log(0.5) \approx 0.69315$, que es la entropía de la variable aleatoria

$$Y_\infty = \begin{cases} 1, & \text{con probabilidad } 1/2; \\ -1 & \text{con probabilidad } 1/2, \end{cases}$$

la cual, como era de esperarse, expresa completa incertidumbre.

A partir de los resultados presentados en el ejemplo previo, resulta relevante compararlos con los resultados de las secciones anteriores. Primero, la situación descrita presenta una clara analogía en el ámbito de la física. En efecto, cuando un sistema de gas aislado pasa a un estado de equilibrio termodinámico, la distribución límite de moléculas, uniforme en posición y normal en velocidad, no tiene memoria de la distribución original. Observe que en los teoremas límite, como el teorema del límite central, también se evidencia cierta pérdida de información. Es decir, la distribución límite normal estándar carece de información sobre las distribuciones específicas de las variables aleatorias involucradas en la suma, a excepción de la media y la varianza. Además, las distribuciones límite alcanzan entropía máxima en determinadas clases, y el proceso de transición a dicho límite conlleva un aumento en la entropía. Todos estos aspectos se reflejan en el ejemplo a partir de que independientemente del valor de p, la distribución límite es la misma, y es de entropía máxima.

Agradecimientos. Este trabajo fue apoyado por el Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT - MÉXICO) bajo el proyecto Ciencia de Frontera CF-2019 87787.

Referencias

[1] Sh. Artstein, K. Ball, F. Barthe and A. Naor, Solution of Shannon's problem on the monotonicity of entropy, Journal of the American Mathematical Society, 17, 975-982 (2004).

[2] A. R. Barron, Entropy and the central limit theorem, The Annals of Probability, 14, 336-342 (1986).

[3] R. N. Bhattacharya, Speed of convergence of n-th fold convolution of a probability measure on a compact group, Z. Wahrscheinlichkeitstheorie verw. Geb., 25, 1-10 (1972).

[4] S. G. Bobkov, G.P. Chistyakov and F. Götze, Berry-Essen bounds in the entropic central limit theorem, Probab. Theory Relat Fields, 435-478 (2014).

[5] K. Conrad, Probability distributions and maximal entropy, Expository paper (2013) URL: https: //kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf

[6] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Edition, Willey, 2006.

[7] E. Gordienko and X. I. Popoca Jimenez, Introducción a la Teoría de Probabilidad y Métricas Probabilísticas con Aplicaciones en Seguros y Finanzas, Instituto de Matemáticas UNAM, México, 2018.

[8] V. V. Kalashnikov, Geometric Sums: Bounds for Rare Events with Applications, Kluwer Academic Publishers, 1997.

[9] M. C. Mackey, Time's Arrow: The Origins of Thermodynamic Behavior, Springer-Verlag, New York, 1992.

[10] V. V. Petrov, Sums of Independent Random Variables, Springer-Verlag, Berlin, 1975.

[11] I. Sason, On reverse Pinsker inequalities, IEEE. Trans. on information Theory, 61 (2015).

Como citar este artículo: E. I. Gordienko y J. A. Minjárez-Sosa, "Distribuciones de máxima entropía, incremento de aleatoriedad y teoremas límite en probabilidad", Sahuarus. Revista Electrónica de Matemáticas, vol. 8, no. 1, pp. 23-44, 2024. https://doi.org/10.36788/sah.v8i1.146.

Vol.8 / No.1 / Junio 2024, pp. 45-82
SahuarUS

Criptografía de los cifrados de bloque

Eduardo Velasco-Barreras
Departamento de Matemáticas, Universidad de Sonora.
eduardo.velasco@unison.mx

Resumen

En este trabajo discutiremos la importancia y el funcionamiento de los cifrados de bloque en criptografía. Particularmente, presentaremos algunas herramientas matemáticas en las que dichos cifrados están basados, y cómo la complejidad de las mismas ha logrado asegurar el nivel de seguridad que exigían las aplicaciones de su época. Entre los cifrados de bloque que se discutirán, se encuentra el *Data Ecryption Standard*, el cual fue el cifrado de bloque más utilizado desde la década de los ochenta hasta finales de los noventa.

Palabras Clave: Criptografía; Cifrados de bloque; Data Encryption Standard; Aritmética modular

DOI: 10.36788/sah.v8i1.100
Recibido: 23 de abril de 2019
Aceptado: 29 de junio de 2024

1. Motivación

La criptografía es la ciencia de la escritura secreta. Su objetivo es ocultar el significado de un mensaje, de manera que sólo la persona a quien dicho mensaje es enviado sea capaz de leerlo. Actualmente, la criptografía tiene muchísimas aplicaciones relacionadas con la vida cotidiana, como son el acceso seguro a páginas web, cifrado de correos electrónicos, sistemas de respaldo de archivos, firmas digitales, trasacciones finacieras, entre muchas otras.

A pesar de que el desarrollo que en las últimas décadas ha tenido la criptografía está relacionado con los saltos agigantados de la revolución tecnológica, su historia se remonta incluso hacia las civilizaciones más antiguas. En efecto, prácticamente cada civilización que ha desarrollado una forma de comunicación escrita ha utilizado también formas secretas de comunicación. Por ejemplo, en el antiguo pueblo egipcio de Menat Jufu, se han encontrado *jeroglíficos no estándares* escritos en la tumba del nomarca de Orix llamado Jnumhotep II. Dicho esquema de cifrado es lo que hoy en día llamaríamos *cifrado por sustitución*.

Otro ejemplo muy conocido de criptografía es el de la escítala, la cual era utilizada por los éforos (magistrados) de la ciudad de Esparta. La escítala consistía de una vara cilíndrica alrededor de la cual se enrollaba una tira de pergamino, de manera que al escribir sobre ella y posteriormente desenrollarla, se obtenía un mensaje ininteligible para un tercero que intentara leerlo. Para descifrar dicho mensaje, era necesario enrollar la tira de pergamino

ISSN 2448 5365
Universidad de Sonora

nuevamente en algún cilindro del mismo diámetro. Desde tiempos del historiador y filósofo griego Plutarco se creía que el propósito de la escítala era impedir a terceros conocer el contenido del mensaje. De esta manera, hablaríamos de un *cifrado por trasposición*, pues cada caracter (o bloques de caracteres) son desplazados siguiendo un patrón bien definido. Sin embargo, la revisión de autores más tempranos ha permitido concluir que la escítala no era utilizda como método de cifrado [8], sino que quizás era un *método de autenticación*, es decir, que su fin era validar la identidad del emisor del mensaje [13].

Figura 1: A la izquierda, una escítala en forma de cilindro hexagonal (creada por Eivind Lindbråten `https://commons.wikimedia.org/wiki/File:Skytale.png`). A la derecha, una forma simple de hacerla (`http://unmuseum.mus.pa.us/excoded.htm`).

Para cerrar esta parte introductoria, presentaremos un cifrado clásico que también es bastante conocido, pero que además será útil para los propósitos de esta exposición. Se dice que el político y militar romano Julio César se comunicaba con sus tropas usando un corrimiento de todas las letras del alfabeto por medio de un número fijo de pasos. Supongamos que, en lugar de utilizar como alfabeto las letras A, B, C, D, ..., W, X, Y, Z en el orden usual, tomamos el alfabeto empezando en alguna otra letra, digamos, l, m, n, ñ, ... , h, i, j, k. Es decir, que al escribir un mensaje, en lugar de utilizar la letra "A", usaríamos la letra "l"; en vez de escribir la letra "B", pondríamos "m" y así sucesivamente. En particular, al llegar al final del alfabeto se continúa desde el principio (la "O" se reemplaza por la "z" y la siguiente letra que es la "P" se reemplaza por la "a"). La tabla 1 muestra la clave completa de este ejemplo. Si el mensaje que queremos cifrar es "este texto es indescifrable", entonces, de acuerdo a dicha tabla, escribiríamos "ODEO EOIEZ OD SXÑODNSPCLMVO".

a	b	c	d	e	f	g	h	i	j	k	l	m	n	ñ	o	p	q	r	s	t	u	v	w	x	y	z
L	M	N	Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K

Tabla 1: La clave completa de nuestro ejemplo de cifrado de Julio César.

Como un ejercicio para el lector, ¿pudiera descifrar el siguiente mensaje, sabiendo que fue cifrado con la clave anterior? Disculpen la falta de ortografía, pero nuestra clave no incluye caracteres con acento.

> AZÑOWZD LPSCWLC NLEOQZCSNLWOXEO BFO OV SXQOXSZ RFWLXZ XZ OD NLALK ÑO SXGOXELC XSXQFX NZÑSQZ DONCOEZ BFO OV ACZASZ SXQOXSZ RFWLXZ XZ AFOÑL ÑODNSPCLC. OÑQLC LVVLX AZO.

2. Aritmética modular

De los ejemplos presentados en la introducción pudiera tenerse la impresión de que las matemáticas no juegan un rol importante en la criptografía, o al menos no es muy transparente el cómo las matemáticas pueden ayudar a describir los cifrados presentados arriba o bien, generar otros más complejos. Resulta ser que la teoría de números, así como estructuras algebraicas como los grupos, anillos y campos, son esenciales en muchos de los cifrados que se usan hoy día.

Una de las herramientas fundamentales para la criptografía es la aritmética modular. Para el lector que no esté familiarizado con este tema, daremos aquí un pequeño repaso. Sin embargo, es recomendable que el lector pueda profundizar en un tema tan bonito, para lo cual sugerimos el libro [11, Capítulo 2]. En cambio, si el lector considera estar suficientemente familiarizado con las nociones básicas de aritmética modular y la noción de inversos multiplicativos, puede omitir el resto de esta sección y continuar la lectura en la sección 3.

2.1. Suma y multiplicación

Para un primer ejemplo de cómo funciona la aritmética modular, fijemos un número, por ejemplo, 6. Ahora supongamos que vamos a ordenar todos los números enteros en seis grupos, llamados *clases*, usando lo que se conoce por *relaciones de equivalencia*: Decimos que los enteros a y b pertenecen a la misma clase si $b - a$ es múltiplo de 6. Por ejemplo, 17 y 53 pertenecen a la misma clase, pues su diferencia es $53 - 17 = 36$, que es múltiplo de 6. Asimismo, 17 y -1 también pertenecen a la misma clase, pues $-1 - 17 = -18$, que es múltiplo de 6. No es difícil observar que todos estos números, -1, 17 y 53, pertenecen todos a la misma clase que el 5: las respectivas diferencias con 5 son -6, 12 y 48, los cuales son todos múltiplos de 6.

Una manera un poco más eficiente de proceder sería respondiendo a lo siguiente: ¿cuáles son todos los números enteros que pertenecen a la misma clase que el 5? Puesto que, por definición, hemos definido que todos los números de una misma clase tengan diferencia múltiplo de 6, podemos ir contando de 6 en 6 a partir de 5. De esta manera, recorreríamos hacia adelante los números 5, 11, 17, 23, 29, 35, 41, 47, 53, 59, 65, ... Similarmente, si nos vamos de 6 en 6 a partir de 5 pero hacia atrás, recorreríamos los números 5, -1, -7, -13, -19, ... En la tabla 2, los números de la clase de equivalencia del 5 se han colocado en el renglón superior. Asimismo, si empezamos desde el 0 y contamos de 6 en 6 hacia delante y hacia atrás, recorreremos los números que pertenecen al reglón inferior. De manera similar, podemos partir de los números 1, 2, 3 o 4 y contar de 6 en 6 hacia adelante y hacia atrás, obteniendo cada uno de los otros renglones de la tabla 2.

Ahora observemos la siguiente propiedad: si tomamos el 52 y hacemos su división entre 6, el resultado no es entero, sino que deja residuo 4, $52 = 6 \times 8 + 4$. Esto nos dice que el 52 pertenece a la misma clase que el 4, y que para llegar del 4 al 52 es hay que avanzar 8 lugares a la derecha. Similarmente, como 31 deja residuo 1 al dividirse entre 6, se tiene que 31 y 1 pertenecen a la misma clase.

Por otra parte, este acomodo tan sencillo de los números en renglones que van de 6 en

...	-19	-13	-7	-1	5	11	17	23	29	35	41	47	53	59	65	...
...	-20	-14	-8	-2	4	10	16	22	28	34	40	46	52	58	64	...
...	-21	-15	-9	-3	3	9	15	21	27	33	39	45	51	57	63	...
...	-22	-16	-10	-4	2	8	14	20	26	32	38	44	50	56	62	...
...	-23	-17	-11	-5	1	7	13	19	25	31	37	43	49	55	61	...
...	-24	-18	-12	-6	0	6	12	18	24	30	36	42	48	54	60	...

Tabla 2: Cada uno de los seis renglones es una clase de equivalencia módulo 6.

6 tiene varias cualidades interesantes. Tomemos un número del renglón verde, por ejemplo, el 44, y tomemos un número que pertenezca al reglón rojo, por ejemplo, el -13. Si sumamos ambos números obtenemos $44 + (-13) = 31$, que pertenece al renglón azul. Resulta que si tomamos cualquier otro par de números que también pertenezcan a los renglones verde y rojo, su suma siempre va a caer en el renglón azul (por ejemplo, 2+5=7). Esta propiedad no es exclusiva de los renglones verde y rojo. Si elegimos cualquier otro par de renglones, por ejemplo, azul y amarillo, las sumas entre elementos de dichos renglones siempre caerán en el renglón naranja.

Una manera de sintetizar lo anterior es la siguiente: denotemos por **0** a la clase del cero (al conjunto de todos los elementos del renglón blanco, o sea los múltiplos de 6). Asimismo, denotemos por **1** a la clase del 1 (el conjunto de todos los elementos del renglón azul). Similarmente, **2**, **3**, **4** y **5** denotarán a los renglones verde, amarillo, naranja y rojo. Puesto que la suma de cualquier elemento de la clase del 2 (renglón verde) con cualquier elemento de la clase del 5 (renglón rojo) pertenece a la clase del 1 (renglón azul), escribimos $\mathbf{2}+\mathbf{5}=\mathbf{1}$. Por otra parte, también tenemos que las sumas de elementos de la clase del 1 con elementos de la clase del 3 (azul y amarillo) da elementos de la clase del 4 (naranja), podemos escribir $\mathbf{1}+\mathbf{3}=\mathbf{4}$. Todas las posibles relaciones de sumas entre las clases **1**, **2**, **3**, **4**, **5** y **6** se resumen en la tabla de la suma que se presenta a continuación. Más aún, con la multiplicación también se tiene una operación bien definida, como se exhibe en la segunda tabla.

$+$	**0**	**1**	**2**	**3**	**4**	**5**
0	**0**	**1**	**2**	**3**	**4**	**5**
1	**1**	**2**	**3**	**4**	**5**	**0**
2	**2**	**3**	**4**	**5**	**0**	**1**
3	**3**	**4**	**5**	**0**	**1**	**2**
4	**4**	**5**	**0**	**1**	**2**	**3**
5	**5**	**0**	**1**	**2**	**3**	**4**

$\times$	**0**	**1**	**2**	**3**	**4**	**5**
0	**0**	**0**	**0**	**0**	**0**	**0**
1	**0**	**1**	**2**	**3**	**4**	**5**
2	**0**	**2**	**4**	**0**	**2**	**4**
3	**0**	**3**	**0**	**3**	**0**	**3**
4	**0**	**4**	**2**	**0**	**4**	**2**
5	**0**	**5**	**4**	**3**	**2**	**1**

Estas tablas dan lugar a la *aritmética módulo 6*. Sin embargo, esto puede hacerse con cualquier otro número. Un ejemplo cotidiano lo tenemos en las horas del día, así como en los días de la semana.

Ejemplo 2.1 *Un joven emocionado por su próximo cumpleaños le dice a sus compañeros: "Faltan 531 horas para mi cumpleaños". Si esto ocurrió a las 7 de la tarde, ¿a qué hora nació el joven?*

Solución. En este ejemplo, aplicaremos aritmética módulo 24, ya que es el número de horas que tiene un día. A las 7 de la tarde son las **19** horas. Si agregamos **531** horas, tenemos que $\mathbf{19 + 531 = 540 = 12}$, pues el residuo que deja 540 al dividirse entre 24 es 12. Por lo tanto, el joven nació a las 12 del día. ■

Ejemplo 2.2 *Supongamos que un turista decide visitar nueve ciudades el mismo número de días cada una. Se sabe que a la primera ciudad llegó un viernes y se fue un miércoles, y que en el mismo día que se iba de una ciudad llegaba a la otra. ¿Qué día se fue el turista de la última ciudad?*

Solución. En este ejemplo carecemos de suficiente información para conocer cuántos días estuvo exactamente el turista en cada ciudad. Sin embargo, sí podemos dar respuesta a la pregunta anterior utilizando *aritmética módulo 7*. Como el turista llegó en viernes a la primera ciudad y se fue en miércoles, el turista pudo haber estado 5 días si no permaneció alguna semana completa, pero también pudo haber estado 12 días si se estuvo una semana completa, o bien 19 días si estuvo dos semanas completas, etc. En general, en número de días que el turista permaneció en la primera ciudad fue $7k + 5$, donde k es el número de semanas completas que el turista permaneció en dicho lugar. En otras palabras, el número de días que el turista permaneció en la primera ciudad es algún elemento de **5**, la clase del 5 en módulo 7. Dado que cada ciudad permaneció el mismo número de días, en total estuvo $\mathbf{5{\times}9{=}45{=}3}$ días sin contar las semanas. A partir del viernes, contando 3 días llegamos al lunes. Puesto que el viaje inició en viernes, el turista tuvo que irse un lunes de la última ciudad. ■

Estos dos ejemplos ilustran la manera intuitiva en que podemos pensar en la aritmética modular. Así como las horas del día y los días de la semana se repiten cíclicamente, también las operaciones en aritmética modular se repiten en ciclo.

Ejemplo 2.3 *Otro ejemplo cotidiano de aritmética modular es cuando trabajamos con números pares e impares. Sabemos que la suma de dos números pares da como resultado un número par, al igual que la suma de dos números impares. En cambio, cuando sumamos dos números de distinta paridad el resultado es impar. En cuanto a la multiplicación, el producto de dos números impares es impar, mientras que el producto con un número par siempre es par. Estos hechos pueden resumirse en las tablas de suma y multiplicación módulo 2 que presentamos a continuación. La clase* **0** *en módulo 2 consiste de todos los números pares, mientras que* **1** *consiste de todos los números impares.*

$+$	**0**	**1**
0	**0**	**1**
1	**1**	**0**

$\times$	**0**	**1**
0	**0**	**0**
1	**0**	**1**

La suma módulo 2 puede extenderse de manera natural a bloques de igual longitud. Por ejemplo, si $A = 1111010010101$ *y* $B = 0001101001110$*, entonces*

$$A \oplus B := 1111010010101 \oplus 0001101001110 = 1110111011011.$$

Esta suma, llamada suma exclusiva *o XOR, lo que hace es sumar módulo 2 cada una de las componentes.*

En general, dado un entero $n \in \mathbb{Z}$, consideremos el conjunto $n\mathbb{Z}$ de los enteros múltiplos de n. Luego, podemos dividir al anillo de los enteros $\mathbb{Z}$ en n clases de equivalencia dadas por la siguiente relación: $a \sim b$ si $b - a \in n\mathbb{Z}$. Las n clases de equivalencia son $\mathbf{0}, \mathbf{1}, \ldots, \boldsymbol{n-1}$ y denotaremos al conjunto de dichas clases de equivalencia por $\mathbb{Z}/n\mathbb{Z}$. Asimismo, usaremos la notación $a \mod n$ para referirnos al número entero entre 0 y $n-1$ que pertenece a la misma clase de equivalencia que a. Ahora bien, el hecho de que las operaciones de suma y multiplicación de clases de equivalencia estén bien definidas recae fundamentalmente en que el conjunto $n\mathbb{Z}$ es un *ideal* en el anillo $\mathbb{Z}$. Esto quiere decir que:

1. La suma de dos múltiplos de n es nuevamente un múltiplo de n: $n\mathbb{Z} + n\mathbb{Z} \subseteq n\mathbb{Z}$.
2. Al multiplicar un múltiplo de n por *cualquier otro entero* el resultado es múltiplo de n: $n\mathbb{Z} \times \mathbb{Z} \subseteq n\mathbb{Z}$.

En estos términos, el conjunto de clases de equivalencia $\mathbb{Z}/n\mathbb{Z}$ adquiere estructura de anillo con la suma y multiplicación de clases de equivalencia. La teoría de ideales dentro de las estructuras algebraicas son de gran importancia, no sólo en aritmética modular, sino para la construcción de otros ejemplos que se usan en criptografía. Por ello, sugerimos al lector interesado el profundizar en esta teoría, por ejemplo, en la referencia [6].

2.2. Inversos módulo n y el algoritmo euclidiano extendido

Consideremos el conjunto $\mathbb{Z}/n\mathbb{Z}$ de los enteros módulo n, y sea $\boldsymbol{a} \in \mathbb{Z}/n\mathbb{Z}$ la clase de equivalencia de algún entero $a \in \mathbb{Z}$. Una pregunta natural es cúando existe un entero $b \in \mathbb{Z}$ tal que $\boldsymbol{b} \in \mathbb{Z}/n\mathbb{Z}$ sea el inverso multiplicativo de $\boldsymbol{a}$. Por *inverso multiplicativo* queremos decir que la multiplicación de ambos sea $\mathbf{1}$, $\boldsymbol{ab} = \mathbf{1}$. En el lenguaje de anillos, se dice que $\boldsymbol{a}$ es una *unidad* en el anillo $\mathbb{Z}/n\mathbb{Z}$.

Por ejemplo, observemos la tabla de multiplicación módulo 6 que presentamos en la subsección anterior. Podemos ver que, como resultado de una multiplicación, el $\mathbf{1}$ solamente se obtiene de $\mathbf{1} \times \mathbf{1} = \mathbf{1}$ y $\mathbf{5} \times \mathbf{5} = \mathbf{1}$. En ese sentido, solamente $\mathbf{1}$ y $\mathbf{5}$ son unidades en $\mathbb{Z}/6\mathbb{Z}$. Observemos ahora la tabla de multiplicación en módulo 7. En este caso, podemos ver que en cada renglón, exceptuando el primero, aparece un $\mathbf{1}$, lo que nos dice que *todos los elementos distintos de cero de* $\mathbb{Z}/7\mathbb{Z}$ *son unidades*. Más precisamente, como $\mathbf{1} \times \mathbf{1} = \mathbf{1}$, $\mathbf{2} \times \mathbf{4} = \mathbf{1}$, $\mathbf{3} \times \mathbf{5} = \mathbf{1}$, $\mathbf{4} \times \mathbf{2} = \mathbf{1}$, $\mathbf{5} \times \mathbf{3} = \mathbf{1}$ y $\mathbf{6} \times \mathbf{6} = \mathbf{1}$, tenemos que $\mathbf{1}$ es su propio inverso, $\mathbf{2}$ y $\mathbf{4}$ son inversos uno del otro, $\mathbf{3}$ y $\mathbf{5}$ son inversos uno del otro y $\mathbf{6}$ es su propio inverso.

El hecho de que alguna clase residual $\boldsymbol{a}$ admita un inverso $\boldsymbol{b}$ en $\mathbb{Z}/n\mathbb{Z}$ depende tanto de $\boldsymbol{a}$ como de n. En términos de representantes de clase, es decir, trabajando con números enteros $a \in \boldsymbol{a}$ y $b \in \boldsymbol{b}$ en lugar de con sus clases de equivalencia, el que $\boldsymbol{a} \times \boldsymbol{b} = \mathbf{1}$ significa que $ab - 1$ sea múltiplo de n, es decir, $ab - 1 = nk$ para algún k. Esto último equivale a $ab - nk = 1$, es decir, a expresar al 1 como *combinación lineal* de a y n. Se puede probar que esto es posible siempre y cuando a y n sean *primos relativos* [11], es decir, que no tengan divisores comunes aparte de ± 1. Más precisamente, que el máximo común divisor de a y n sea 1, $(a, n) = 1$.

×	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
1	**0**	**1**	**2**	**3**	**4**	**5**	**6**
2	**0**	**2**	**4**	**6**	**1**	**3**	**5**
3	**0**	**3**	**6**	**2**	**5**	**1**	**4**
4	**0**	**4**	**1**	**5**	**2**	**6**	**3**
5	**0**	**5**	**3**	**1**	**6**	**4**	**2**
6	**0**	**6**	**5**	**4**	**3**	**2**	**1**

Tabla 3: Tabla de multiplicación módulo 7.

Proposición 2.1 *Un entero a admite un inverso multiplicativo en módulo n si y sólo si a y n son primos relativos. Dicho de otra manera, la clase $\boldsymbol{a} \in \mathbb{Z}/n\mathbb{Z}$ es una unidad si y sólo si $(a, n) = 1$ para algún representante $a \in \boldsymbol{a}$.*

Utilizando el criterio anterior, podemos explicar por qué en la multiplicación módulo 6 solamente 1 y 5 tienen inverso: resulta que 1 y 5 no tienen factores en común con 6, mientras que 0, 2, 3, y 4 sí lo tienen. En el caso de la multiplicación módulo 7, todos los elementos distintos de cero tienen inverso porque 7 es número primo, y por lo tanto, los únicos enteros que tienen factores en común con 7 son los múltiplos de 7 (la clase $\mathbf{0}$).

El algoritmo euclidiano. De acuerdo a la discusión anterior, la construcción de una clase $\boldsymbol{b}$ tal que $\boldsymbol{a} \times \boldsymbol{b} = \mathbf{1}$ equivale a encontrar un entero b tal que $ab - nk = 1$ para alguna k, es decir, a expresar 1 como combinación lineal de a y n. Este procedimiento puede hacerse mediante el *algoritmo euclidiano extendido*.

Fijemos dos enteros n y a y supongamos $n > a > 0$. Recordemos que el *algoritmo de la división* nos dice que existen únicos enteros q y r tales que

$$n = aq + r, \qquad 0 \leq r < a.$$

Resulta que el máximo común divisor (n, a) coincide con el de (a, r), con la ventaja de que ahora los enteros involucrados son más pequeños: $n > a > r$. De esta manera, podemos repetir el algoritmo de la división para generar enteros cada vez más pequeños cuyo máximo común divisor sea siempre el mismo:

$$\begin{aligned}
n &= aq + r, & (a, r) &= (n, a);\\
a &= rq_1 + r_1, & (r, r_1) &= (a, r);\\
r &= r_1q_2 + r_2, & (r_1, r_2) &= (r, r_1);\\
r_1 &= r_2q_3 + r_3, & (r_2, r_3) &= (r_1, r_2);\\
r_2 &= r_3q_4 + r_4, & (r_3, r_4) &= (r_2, r_3);\\
&\vdots & &\vdots
\end{aligned}$$

Recordemos que en cada etapa hemos generado un residuo que es más pequeño tal que el máximo común divisor entre cualesquiera dos consecutivos es el mismo, $n > a > r > r_1 >$

$r_2 > \ldots$, $(r_i, r_{i+1}) = (n, a)$. Puesto que en cada etapa obtenemos un residuo más pequeño, eventualmente debemos obtener residuo cero:

$$r_{m-2} = r_{m-1}q_m + r_m,$$
$$r_{m-1} = r_m q_{m+1},$$

es decir, r_m es el último residuo no cero, y el siguiente residuo ya es cero, $r_{m+1} = 0$. Como

$$(n, a) = (r_m, r_{m+1}) = (r_m, 0) = r_m,$$

concluimos que el último residuo no cero, es decir r_m, es el máximo común divisor.

Proposición 2.2 (Algoritmo euclidiano) *Sean a y n enteros positivos. Al aplicar sucesivamente el algoritmo euclidiano, el último residuo no cero es el máximo común divisor (a, n).*

Ejemplo 2.4 *Consideremos $n = 2019$ y $a = 1115$. Aplicando el algoritmo euclidiano, obtenemos*

$$\begin{aligned}
2019 &= 1115 \times 1 + 904,\\
1115 &= 904 \times 1 + 211,\\
904 &= 211 \times 4 + 60,\\
211 &= 60 \times 3 + 31,\\
60 &= 31 \times 1 + 29,\\
31 &= 29 \times 1 + 2,\\
29 &= 2 \times 14 + 1,\\
2 &= 1 \times 2.
\end{aligned}$$

Por lo tanto, el máximo común divisor de 2019 y 1115 es $(2019, 1115) = 1$.

De las ecuaciones obtenidas arriba, es fácil ver que r_i es combinación lineal de los dos anteriores, $r_i = r_{i-2} - q_i r_{i-1}$. Más aún, de dichas relaciones se puede expresar a r_m como combinación lineal de cada par de residuos consecutivos r_{i-1} y r_i. Veámoslo como continuación del ejemplo anterior.

Ejemplo 2.5 *Del ejemplo 2.4, tenemos que*

$$\begin{aligned}
1 &= 29 - 2 \times 14,\\
2 &= 31 - 29 \times 1,\\
29 &= 60 - 31 \times 1,\\
31 &= 211 - 60 \times 3,\\
60 &= 904 - 211 \times 4,\\
211 &= 1115 - 904 \times 1,\\
904 &= 2019 - 1115 \times 1,
\end{aligned}$$

donde cada uno de los residuos ha sido expresado como combinación lineal de los anteriores. Usaremos dichas relaciones para expresar el último residuo, es decir, el 1, como combinación lineal de 2019 y 1115. En las siguientes igualdades,

$$\begin{aligned}
1 &= 29 \times 1 - 2 \times 14 = 29 - (31 - 29 \times 1) \times 14 \\
&= 29 \times 15 - 31 \times 14 = (60 - 31 \times 1) \times 15 - 31 \times 14 \\
&= 60 \times 15 - 31 \times 29 = 60 \times 15 - (211 - 60 \times 3) \times 29 \\
&= 60 \times 102 - 211 \times 29 = (904 - 211 \times 4) \times 102 - 211 \times 29 \\
&= 904 \times 102 - 211 \times 437 = 904 \times 102 - (1115 - 904 \times 1) \times 437 \\
&= 904 \times 539 - 1115 \times 437 = (2019 - 1115 \times 1) \times 539 - 1115 \times 437 \\
&= 2019 \times 539 - 1115 \times 976,
\end{aligned}$$

la columna de la izquierda presenta al 1 como combinación lineal de los otros residuos hasta llegar a 2019 y 1115, mientras que la columna de la derecha muestra el procedimiento que lleva a la siguiente combinación a partir de las expresiones de arriba. El lector puede verificar que, efectivamente, $2019 \times 539 - 1115 \times 976$ *es igual a 1.*

Proposición 2.3 (Algoritmo euclidiano extendido) *Sean a y n enteros positivos. El máximo común divisor $d = (a, n)$ puede expresarse como combinación lineal de a y n, es decir, $d = a \times b + n \times k$ para algunos enteros b y k. Más aún, los enteros b y k son únicos módulo n/d y a/d, respectivamente. En particular, si $d = 1$, entonces b es el inverso de a módulo n.*

Ejemplo 2.6 *Tomemos $n = 2019$ y $a = 1115$. Puesto que* $2019 \times 539 - 1115 \times 976 = 1$*, se concluye que $b = -976$ es el inverso multiplicativo de* 1115 *módulo 2019.*

Hemos visto que el inverso multiplicativo de $\mathbf{a} \in \mathbb{Z}/n\mathbb{Z}$ puede construirse si $(a, n) = 1$, aplicando el algoritmo euclidiano extendido para expresar a 1 como combinación lineal de a y n. Es importante mencionar que este procedimiento no sólo es válido para los enteros módulo n, es decir, al trabajar con clases de equivalencia del anillo de los enteros $\mathbb{Z}$, sino que también aplica en anillos más generales, llamados *euclidianos*, en los cuales se puede llevar a cabo una versión del algoritmo euclidiano arriba descrito. Esto es significativo, por ejemplo, para entender el trasfondo matemático del *Advanced Encryption Standard (AES)*, que es uno de los métodos de cifrado más utilizados actualmente, y parte del cual está basado en la construcción de inversos de clases de equivalencia de polinomios [12, Sección 4.3].

3. Cifrados de bloque: algunos ejemplos

3.1. ¿Por qué utilizar cifrados de bloque?

Antes de entrar de lleno con los cifrados de bloque, analicemos el cifrado de Julio César desde la perspectiva de la aritmética modular. Recordemos que este cifrado lo que hace es

recorrer cíclicamente el orden del alfabeto. Si a cada una de las letras a, b, c, ..., y, z le hacemos corresponder los números 0, 1, 2, ..., 25, 26, entonces la clave de la tabla 1 lo que hace es reemplazar el 0 por el 11, el 1 por el 12, el 2 por el 13, ..., el 15 por el 26, el 16 por el 0, el 17 por el 1, ... y el 26 por el 10.

Desde el punto de vista de la aritmética modular, el cifrado lo que hace es reemplazar la letra asociada con el número a con la letra asociada al número $a + 11 \mod 27$, es decir, al representante de clase módulo 27 de $a + 11$ que está entre 0 y 26. Esto lo podemos ver en la tabla 4, donde cada entrada del tercer renglón es el resultado de sumarle 11 en módulo 27 a la entrada que está arriba de ella.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	ñ	o	p	q	r	s	t	u	v	w	x	y	z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	0	1	2	3	4	5	6	7	8	9	10
L	M	N	Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K

Tabla 4: La clave de la tabla 1 es simplemente sumar 11 en la artimética módulo 27.

Ejemplo 3.1 *El mensaje "este texto es indescifrable" usando números, y omitiendo los espacios en blanco, se expresaría por*

$$x = 4\ 19\ 20\ 4\ 20\ 4\ 24\ 20\ 15\ 4\ 19\ 8\ 13\ 3\ 4\ 19\ 2\ 8\ 5\ 18\ 0\ 1\ 11\ 4.$$

Sumando 11 $\mod 27$*, obtenemos*

$$y = 15\ 3\ 4\ 15\ 4\ 15\ 8\ 4\ 26\ 15\ 3\ 19\ 24\ 14\ 15\ 3\ 13\ 19\ 16\ 2\ 11\ 12\ 22\ 15.$$

Dichos números corresponden a las letras "ODEOEOIEZODSXÑODNSPCLMVO", que es el mensaje cifrado.

Con un alfabeto de 27 letras, solamente hay 27 opciones para crear un cifrado de Julio César: simplemente elegimos cuál número sumar módulo 27, que en este caso fue 11. Esto se expresa diciendo que el *espacio de claves* del cifrado de Julio César tiene solamente 27 elementos. Con la tecnología actual, este tamaño del espacio de claves no ofrece ninguna seguridad, tan sólo le serviría a los niños para enviar mensajes secretos a los amigos de la escuela.

El cifrado de Julio César es simplemente una *traslación* en módulo 27, pues cada uno de los números se suma por un elemento fijo. Un primer intento para construir un cifrado tal que el tamaño del espacio de claves ofrezca un mayor nivel de seguridad sería usando *transformaciones afines* módulo 27, es decir, combinar una transformación lineal con una traslación. En este caso, para el cifrado elegiríamos dos enteros a y b entre 0 y 26, tal que $(a, 27) = 1$. La cualidad de que a sea primo relativo con 27 es para que a tenga inverso, y que el mensaje cifrado pueda ser leído (descifrado). Luego, una vez que nuestro mensaje se convierte en una serie de números $x = x_1x_2 \dots x_n$ módulo 27, cada uno de los números x_i es transformado en $y_i = ax_i+b$. La persona que recibe el mensaje cifrado, para descifrarlo deberá

aplicar la transformación afín inversa $x_i := cy_i + d \mod 27$, donde $c := a^{-1}$ y $d := -a^{-1}b$. Es un ejercicio sencillo para el lector verificar que esto efectivamente define la transformación inversa del cifrado.

Ejemplo 3.2 *Sean $a = 2$ y $b = 3$. En este caso, el cifrado consiste en multiplicar por $a = 2$ y al resultado sumarle $b = 3$. Luego, la relación entre cada caracter con su cifrado sería la dada por la tabla 5. Nuevamente, si queremos cifrar el mensaje "este texto es indescifrable", pero ahora usando nuestra transformación afín $y = 2x + 3 \mod 27$, el resultado es "JMÑJÑJVÑEJMQAHJMFQLKBDWJ". La persona que recibe este mensaje tendría que usar la clave anterior pero en sentido inverso. Para encontrarla, la transformación afín a usar debe ser la inversa a $y = 2x + 3 \mod 27$. Puesto que $14 \times 2 = 28 \equiv 1 \mod 27$, tenemos que 14 es el inverso de 2. Como $14 \times 3 \equiv 15 \mod 27$, resulta que la transformación afín inversa es $x = 14y - 15 \mod 27$.*

A	B	C	D	E	F	G	H	I	J	K	L	M	N	Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
3	5	7	9	11	13	15	17	19	21	23	25	0	2	4	6	8	10	12	14	16	18	20	22	24	26	1
d	f	h	j	l	n	o	q	s	u	w	y	a	c	e	g	i	k	m	ñ	p	r	t	v	x	z	b

Tabla 5: La clave de la transformación afin $y_i = ax_i + b \mod 27$, con $a = 2$, $b = 3$.

Hemos pasado de usar traslaciones módulo 27 a transformaciones afines. En este caso, el espacio de clave consiste de $18 \times 27 = 486$ claves distintas, pues tenemos 18 opciones para elegir a tal que $(a, 27) = 1$ y 27 opciones para elegir b. Ahora el tamaño de clave es mucho más grande que el original de 27 que teníamos con el cifrado de Julio César, pero nuevamente este tamaño de clave es muy pequeño y no resistirá un ataque por *fuerza bruta* con las computadoras actuales. Un ataque por fuerza bruta consiste en probar cada una de las 486 posibles claves hasta encontrar la que funciona.

De la discusión anterior, vemos que es indispensable tener un espacio de claves bastante grande para poder resistir el ataque más elemental, que es el de prueba y error. Sin embargo, *no es suficiente* que el espacio de clave sea muy grande para asegurar que el cifrado sea seguro, como veremos en el siguiente ejemplo.

Las traslaciones y las transformaciones afines de caracter a caracter ofrecen un espacio de clave bastante pequeño. Una generalización es considerar *todas las posibles permutaciones* de nuestro alfabeto de 27 caracteres. Es decir, asociar a cada letra del alfabeto alguna otra sin algún orden aparente. Por ejemplo, la a con la R, la b con la X, la c con la H, la d con la E, etc. En este caso, el espacio de clave será de tamaño 27!, es decir, $27 \times 26 \times 25 \times \ldots \times 3 \times 2 \times 1 = 10888869450418352160768000000 \approx 1 \times 10^{28}$. Si contáramos con una potencia de cómputo tal que pudiéramos hacer un ataque por fuerza bruta capaz de probar un trillón de claves por segundo, nos podría tomar la edad actual del universo el encontrar la permutación que se utilizó para cifrar el mensaje. En este sentido, esta clase más amplia de cifrados es bastante resistente a ataques por fuerza bruta. Sin embargo, existen maneras más inteligentes de atacar a este cifrado, como veremos en el siguiente ejemplo.

Ejemplo 3.3 *Consideremos el mensaje cifrado*

```
ZKQ GDHGOTRQR KTETWQDOQ FZT ZK EQBGH XTEAHDOQY

RTST ATKTD GQDQ WTD UDQROTKAT TW FZT WZ

DHAQEOHKQY WTQ ETDH, GTDH TY DTEOGDHEH WHYQBTKAT

WT AOTKT TK DTUOHKTW WOBGYTBTKAT EHKTVQW.
```

en el que cada caracter se representa por una única letra. Los espacios entre palabras se han conservado, al igual que los signos de puntuación. Los acentos no se han tomado en cuenta.

T	Q	D	K	H	W	O	E	G	A	Z	Y	R	B	F	U	X	S	V
31	14	13	12	12	10	9	8	7	7	5	5	4	4	2	2	1	1	1

Tabla 6: Tabla de frecuencias del ejemplo 3.3.

Obsérvese en la tabla de frecuencias 6 que los caracteres más utilizados fueron `T`*,* `Q`*,* `D`*,* `K`*,* `H` *y* `W` *en ese orden. Tomando cuenta que en la lengua castellana los caracteres más utilizados en el lenguaje escrito son las letras "e", "a", "o", "s", "r" "n"(en ese orden), podemos empezar haciendo prueba y error cambiando los caracteres más utilizados del texto cifrado por algunos de estos últimos. Por ejemplo, cambiando* `T` *por* `e` *y* `Q` *por* `a`*, obtenemos*

```
ZKa GDHGOeRaR KeEeWaDOa FZe ZK EaBGH XeEAHDOaY

ReSe AeKeD GaDa WeD UDaROeKAe eW FZe WZ

DHAaEOHKaY Wea EeDH, GeDH eY DeEOGDHEH WHYaBeKAe

We AOeKe eK DeUOHKeW WOBGYeBeKAe EHKeVaW.
```

Pensando en que también la letra `W` *fue una de las más utilizadas, podemos suponer que se trate de la* `o`*, la* `s`*, la* `r` *o la* `n`*. Haciendo la prueba con cada una de ellas, vemos que lo que mejor corresponde a lo que ya hemos descifrado es sustituirla por la* `s`*:*

```
ZKa GDHGOeRaR KeEesaDOa FZe ZK EaBGH XeEAHDOaY

ReSe AeKeD GaDa seD UDaROeKAe es FZe sZ
```

```
DHAaEOHKaY sea EeDH, GeDH eY DeEOGDHEH sHYaBeKAe

se AOeKe eK DeUOHKes sOBGYeBeKAe EHKeVas.
```

Tratando de reemplazar `D`*,* `K` *y* `H` *por* `o`*,* `r` *y* `n` *en algún orden, puede verse que la* `D` *y la* `K` *no corresponden a la* `o` *y que probablemente* `D` *corresponda a* `r`*. Así, reemplazando* `H` *por* `o`*,* `D` *por* `r` *y* `K` *por* `n`*, obtenemos*

```
Zna GroGOeRaR neEesarOa FZe Zn EaBGo XeEAorOaY

ReSe Aener Gara ser UraROenAe es FZe sZ

roAaEOonaY sea Eero, Gero eY reEOGroEo soYaBenAe

se AOene en reUOones sOBGYeBenAe EoneVas.
```

De este mensaje se puede deducir que la `Z` *corresponde a* `u` *y que la* `E` *representa* `c`*. Haciendo ese cambio, se puede deducir que* `O` *representa* `i` *y después que* `A` *representa* `t`*. Después de unas cuantas pruebas y errores, se puede descifrar el mensaje:*

```
una propiedad necesaria que un campo vectorial

debe tener para ser gradiente es que su

rotacional sea cero, pero el reciproco solamente

se tiene en regiones simplemente conexas.
```

En el ejemplo anterior se explotó la siguiente debilidad del cifrado: el mensaje cifrado conserva las *propiedades estadísticas* del mensaje original [12, Subsección 1.2.2]. Esto permitió que con un poco de deducción lógica y basándonos en reglas de ortografía elementales [4], se pudiera deducir el contenido del mensaje en un tiempo relativamente corto (menor que la edad del universo).

Una lección que nos brinda el ejemplo anterior es que aunque el tener un espacio de claves grande es una condición necesaria para que un cifrado resista ataques por fuerza bruta, esto no es suficiente para garantizar que el cifrado sea seguro. En virtud de ello, conviene presentar un par de propiedades adicionales que se necesitan tener para que un cifrado sea seguro [12, Subsección 3.1.1]:

1. **Confusión:** La relación entre el mensaje cifrado y la clave utilizada para cifrarlo no es inmediata.

2. **Difusión:** El cambiar un símbolo del mensaje original influye en muchos de los símbolos del texto cifrado.

El cifrado por sustitución que presentamos arriba posee la propiedad de confusión, pero no la de difusión. Es importante mencionar que para que un cifrado sea seguro, se necesita que se tengan ambas propiedades al mismo tiempo. En la práctica, los cifrados de bloque aplican alternadamente operaciones de confusión y difusión para una mayor seguridad, es el caso del DES y el AES. En la siguiente parte presentaremos un primer ejemplo de cifrado de bloque, el cual lo podemos pensar como un procedimiento para obtener confusión y difusión (pero que tampoco es lo bastante seguro por sí solo, como veremos en su momento).

Para evitar que un cifrado conserve las propiedades estadísticas del mensaje original, debemos evitar operar caracter a caracter. Una manera de hacerlo es utilizando *cifrados de bloque*, como veremos a continuación.

3.2. Transformaciones afines

El cifrado que presentamos en este apartado consiste en aplicar transformaciones afines, pero a diferencia de las que presentamos arriba y que sólo nos permitían transformar un caracter a la vez, éstas utilizan multiplicación de matrices y suma de vectores, de una cierta longitud fija. Esto nos permitirá cifrar *bloques completos* de dicha longitud fija de una sola vez.

En este procedimiento, nuevamente pensaremos en las 27 letras del abecedario, junto con el espacio en blanco, como números de 0 a 27, o más precisamente, como *clases de residuos módulo 28* (ver tabla 7). Naturalmente, este procedimiento puede adaptarse para admitir un mayor número de caracteres como letras con acentos, espacios y signos de puntuación. Una manera de hacer eso es mediante el *código ASCII*[1], que permite manejar 256 caracteres distintos.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	ñ	o	p	q	r	s	t	u	v	w	x	y	z	
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

Tabla 7: La relación que permite trabajar numéricamente con letras.

Cifrando bloques de longitud 3. Para fines ilustrativos, vamos a presentar una manera de cifrar bloques de longitud 3, para lo cual utilizaremos la siguiente matriz de tamaño 3×3 y el vector de longitud 3:

[1]El código ASCII puede consultarse en `https://elcodigoascii.com.ar/`.

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

En general, para cifrar bloques de longitud n se tomaría una matriz $n \times n$ y un vector de tamaño n. Lo único que debemos de cuidar es que la matriz A que elijamos sea invertible módulo 28, para poder descifrar correctamente el mensaje. Esto quiere decir que el *determinante* de A sea primo relativo con 28, $(\det A, 28) = 1$.

Utilizando A y b podemos definir una transformación afín $y = Ax + b$, donde x y y son bloques de longitud 3. Por ejemplo, supongamos que queremos cifrar el mensaje "ejemplos de cifrados de bloque". Para lograrlo, necesitamos dividir nuestro mensaje en bloques verticales de tamaño 3, incluyendo los espacios:

$$\begin{pmatrix} e & m & o & d & c & r & o & d & b & q \\ j & p & s & e & i & a & s & e & l & u \\ e & l & & & f & d & & & o & e \end{pmatrix}.$$

Utilizando la relación de la tabla 7, podemos convertir nuestro arreglo de letras en una matriz X de tamaño 3×10 con entradas en $\mathbb{Z}/28\mathbb{Z}$,

$$X = \begin{pmatrix} 4 & 12 & 15 & 3 & 2 & 18 & 15 & 3 & 1 & 17 \\ 9 & 16 & 19 & 4 & 8 & 0 & 19 & 4 & 11 & 21 \\ 4 & 11 & 27 & 27 & 5 & 3 & 27 & 27 & 15 & 4 \end{pmatrix}.$$

Ahora bien, para obtener el mensaje cifrado, a cada columna x_i le vamos a aplicar la transformación afín mencionada arriba para obtener el bloque cifrado $y_i = Ax_i + b$. Es decir, multiplicamos a la matriz del mensaje X por la matriz A, y a cada columna de la matriz resultante le sumamos el vector b. Cabe mencionar que las operaciones de suma y multiplicación se efectúan módulo 28:

$$\begin{pmatrix} 0 & 2 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 4 & 12 & 15 & 3 & 2 & 18 & 15 & 3 & 1 & 17 \\ 9 & 16 & 19 & 4 & 8 & 0 & 19 & 4 & 11 & 21 \\ 4 & 11 & 27 & 27 & 5 & 3 & 27 & 27 & 15 & 4 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 22 & 15 & 9 & 7 & 21 & 3 & 9 & 7 & 9 & 18 \\ 12 & 6 & 13 & 1 & 12 & 24 & 13 & 1 & 3 & 25 \\ 17 & 12 & 21 & 10 & 12 & 8 & 21 & 10 & 13 & 27 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 23 & 16 & 10 & 8 & 22 & 4 & 10 & 8 & 10 & 19 \\ 11 & 5 & 12 & 0 & 11 & 23 & 12 & 0 & 2 & 24 \\ 17 & 12 & 21 & 10 & 12 & 8 & 21 & 10 & 13 & 27 \end{pmatrix}$$

Esta matriz resultante la volvemos a converitr a letras, si así se prefiere, de acuerdo a la tabla 7. Así, obtenemos el mensaje cifrado "WLQPFMKMUIAKVLMEWIKMUIAKKCNSX ", donde al final hay un espacio.

Una de las ventajas que tiene el cifrar de esta manera es que un mismo caracter no corresponde siempre al mismo caracter. En efecto, la palabra *ejemplos* se convirtió en $WLQPFMKM$ donde las letras e del mensaje original fueron a parar a letras distintas, W y Q. Por otra parte, dos letras distintas l y s fueron a parar al mismo caracter. Lo que de fondo está ocurriendo es que nuestro cifrado toma bloques completos de longitud 3 y cifra su contenido. Por ejemplo, en nuestro mensaje original aparece dos veces el bloque "os " (con espacio al final), el cual se cifra como "KMU" las dos veces que aparece. Lo mismo ocurre con el bloque "de ".

Descifrando el mensaje. Para que el receptor del mensaje cifrado pueda leerlo, necesitará descifrarlo. En este caso, hay que usar una *transformación afín inversa*. Recordemos que cada bloque del mensaje original x de longitud 3 fue cifrado con una transformación afín $y = Ax+b$. Para recuperar x a partir de y, usaremos la transformación afín $x = Cy + d$, donde $C = A^{-1}$ y $d = -A^{-1}b$. En nuestro caso,

$$C = A^{-1} = \begin{pmatrix} 6 & 25 & 16 \\ 16 & 6 & 25 \\ 25 & 16 & 6 \end{pmatrix} \qquad d = -A^{-1}b = \begin{pmatrix} 19 \\ 18 \\ 19 \end{pmatrix}.$$

El lector puede verificar que al tomar cada bloque y, de longitud 3 en el mensaje cifrado, y aplicarle la transformación afín inversa $x = Cy + d$, recuperamos la matriz del mensaje original.

La debilidad de la linealidad. Notemos que cada uno de los tres caracteres que conforman a un bloque cifrado y_i dependen de los tres caracteres del bloque original x_i. En este sentido, nuestro cifrado de bloque tiene buena difusión dentro del bloque de tamaño 3. A pesar de esto, este modo en que estamos implementando las transformaciones afines *sí conserva algunas propiedades estadísticas del mensaje original* y en principio sí pudiera explotarse dicha debilidad para descifrar el contenido del mensaje. Sin embargo, la mayor debilidad que presenta este cifrado se relaciona con la inherente linealidad que lo define. Esto hace que *con muy poca información que se conozca, se pueda descubrir la clave* con que se cifra un mensaje. En efecto, puesto que nuestro cifrado utiliza transformaciones afines de orden 3, a un atacante le puede bastar conocer cuatro parejas de bloques original-cifrado (x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3) para descubrir la clave. Más precisamente, es suficiente que $v_1 := x_1 - x_0$, $v_2 := x_2 - x_0$ y $v_3 := x_3 - x_0$ sean linealmente independientes para recuperar la transformación afín. Si denotamos $w_1 := y_1 - y_0$, $w_2 := y_2 - y_0$ y $w_3 := y_3 - y_0$, entonces

$$w_i = y_i - y_0 = (Ax_i + b) - (Ax_0 + b) = A(x_i - x_0) = Av_i.$$

Por lo tanto, para conocer A, basta resolver el sistema lineal $Av_1 = w_1$, $Av_2 = w_2$ y $Av_3 = w_3$.

Supongamos que un atacante ha logrado interceptar parte del mensaje anterior, por ejemplo, que sabe que la palabra "*ejemplos de* "(con espacio al final) ha sido cifrado en "WLQPFMKMUIAK". Con esta información, podemos recuperar la matriz A utilizada

arriba. En nuestro caso, tenemos que

$$x_0 = \begin{pmatrix} e \\ j \\ e \end{pmatrix} = \begin{pmatrix} 4 \\ 9 \\ 4 \end{pmatrix}, \quad x_1 = \begin{pmatrix} m \\ p \\ l \end{pmatrix} = \begin{pmatrix} 12 \\ 16 \\ 11 \end{pmatrix}, \quad x_2 = \begin{pmatrix} o \\ s \end{pmatrix} = \begin{pmatrix} 15 \\ 19 \\ 27 \end{pmatrix}, \quad x_3 = \begin{pmatrix} d \\ e \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 27 \end{pmatrix},$$

por lo cual

$$v_1 = x_1 - x_0 = \begin{pmatrix} 8 \\ 7 \\ 7 \end{pmatrix}, \qquad v_2 = x_2 - x_0 = \begin{pmatrix} 11 \\ 10 \\ 23 \end{pmatrix}, \qquad v_3 = x_3 - x_0 = \begin{pmatrix} 27 \\ 23 \\ 23 \end{pmatrix}.$$

Similarmente,

$$y_0 = \begin{pmatrix} W \\ L \\ Q \end{pmatrix} = \begin{pmatrix} 23 \\ 11 \\ 17 \end{pmatrix}, \quad y_1 = \begin{pmatrix} P \\ F \\ M \end{pmatrix} = \begin{pmatrix} 16 \\ 5 \\ 12 \end{pmatrix}, \quad y_2 = \begin{pmatrix} K \\ M \\ U \end{pmatrix} = \begin{pmatrix} 10 \\ 12 \\ 21 \end{pmatrix}, \quad y_3 = \begin{pmatrix} I \\ A \\ K \end{pmatrix} = \begin{pmatrix} 8 \\ 0 \\ 10 \end{pmatrix},$$

por lo cual

$$w_1 = y_1 - y_0 = \begin{pmatrix} 21 \\ 22 \\ 23 \end{pmatrix}, \qquad w_2 = y_2 - y_0 = \begin{pmatrix} 15 \\ 1 \\ 4 \end{pmatrix}, \qquad w_3 = y_3 - y_0 = \begin{pmatrix} 13 \\ 17 \\ 21 \end{pmatrix}.$$

Luego, aplicando Gauss-Jordan con los vectores traspuestos de v_1, v_2, v_3 y w_1, w_2, w_3,

$$\left(\begin{array}{ccc|ccc} 8 & 7 & 7 & 21 & 22 & 23 \\ 11 & 10 & 23 & 15 & 1 & 4 \\ 27 & 23 & 23 & 13 & 17 & 21 \end{array}\right),$$

podemos recuperar la traspuesta de la matriz A. En efecto, primero multiplicamos al primer renglón por 23 y al resultado le restamos 7 veces el tercero:

$$\left(\begin{array}{ccc|ccc} 23 & 0 & 0 & 0 & 23 & 18 \\ 11 & 10 & 23 & 15 & 1 & 4 \\ 27 & 23 & 23 & 13 & 17 & 21 \end{array}\right).$$

Observemos que el inverso de 23 módulo 28 es 11. Luego, multiplicando al primer renglón por 11, obtenemos

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 2 \\ 11 & 10 & 23 & 15 & 1 & 4 \\ 27 & 23 & 23 & 13 & 17 & 21 \end{array}\right)$$

En este paso hemos obtenido que (0 1 2) es el primer renglón de la matriz traspuesta de A. Procediendo de manera similar, recuperamos la matriz A que usamos para cifrar el mensaje. Finalmente, el vector b se recupera haciendo

$$b = y_0 - Ax_0 = \begin{pmatrix} 23 \\ 11 \\ 17 \end{pmatrix} - \begin{pmatrix} 0 & 2 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 4 \\ 9 \\ 4 \end{pmatrix} = \begin{pmatrix} 23 \\ 11 \\ 17 \end{pmatrix} - \begin{pmatrix} 22 \\ 12 \\ 17 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

En general, cuando al cifrar se usa una transformación afín de orden n, es suficiente tener $n+1$ parejas *genéricas* $(x_0, y_0), \ldots, (x_n, y_n)$ para poder descifrarlo, lo cual en términos prácticos representa un nivel de seguridad demasiado bajo.

3.3. El telegrama Zimmermann

Hace poco más de un siglo, Europa estaba en medio del peor conflicto armado hasta entonces: la Primera Guerra Mundial. Hacia inicios de 1917, Estados Unidos aún se mantenía neutral. Sin embargo, mantenía fuertes relaciones comerciales con Francia y el Imperio Británico. Particularmente, los banqueros estadounidenses habían realizado fuertes préstamos a dichos países para que pudieran sostenerse durante la guerra. En consecuencia, Estados Unidos tenía interés en que estos países resultaran vencedores en este conflicto.

En el bando contrario, el Imperio Alemán utilizaba sus submarinos para atacar los barcos comerciales de Estados Unidos que viajaban rumbo a Inglaterra. Aún así, buscaban que los Estados Unidos no se involucraran directamente dentro del conflicto armado. Sin embargo, la entrada de los Estados Unidos en la primera guerra mundial era un hecho cada vez más inevitable. En este conexto, el ministro Alemán de Asuntos Exteriores Arthur Zimmermann envió un telegrama el 16 de enero de 1917 a su embajador en México, el conde Heinrich von Eckardt. En dicho telegrama, le daba instrucciones de que, en caso de que Estados Unidos decidiera entrar al conflicto armado, formara una alianza militar con México, a cambio de la cual se le ofrecería devolverle una parte de los territorios que México había perdido 69 años antes cuando los Estados Unidos invadieron el país.

Por su parte, el gobierno de México, encabezado entonces por Venustiano Carranza, declinó la oferta. De acuerdo con [9], cuando el contenido del telegrama fue hecho público, el gobierno mexicano negó haberlo recibido, aunque hubo testigos asegurando que Carranza y sus colaboradores cercanos recibieron y rechazaron inmediatamente la propuesta, mientras que también hay quienes afirman que la propuesta no fue entregada a Carranza por lo peligroso de la misma. En cualquier caso, la propuesta en sí misma era inviable desde el punto de vista mexicano. En esa época, el gobierno estaba ocupado socabando las revueltas de Pancho Villa en el norte y de Emiliano Zapata en el sur. Asimismo, la fuerza militar de los Estados Unidos era ya bastante más poderosa que la de México, como se había exhibido pocos años antes en la ocupación estadounidense de Veracruz en 1914.

El telegrama Zimmermann fue enviado de forma cifrada, tal como se muestra en la figura 2. Resulta que el mismo día en que el telegrama fue enviado, los ingleses lo interceptaron y descifraron, gracias a que ya conocían parte del cifrado utilizado. El método de cifrado utilizado era una especie de diccionario, es decir, cada uno de los bloques numéricos del telegrama representa una palabra. Por ejemplo, algunas de las palabras que aparecen en dicho telegrama son [5, Capítulo I]:

14936	eingeschränkten	22049	sich
15021	einzeln	22200	stop
15099	Empfang	22295	sofortiger

Al descifrado del telegrama Zimmermann, y su revelación al gobierno estadounidense por parte de los ingleses, se le ha llamado “el mayor golpe de inteligencia de todos los tiempos” [5, Capítulo I]. Esto en parte se debe a que, al descifrar y publicar su contenido, se logró cambiar la postura antibelicista de muchos estadounidenses, siendo el último empujón para la entrada de los Estados Unidos en la guerra, favoreciendo decisivamente el fin del conflicto.

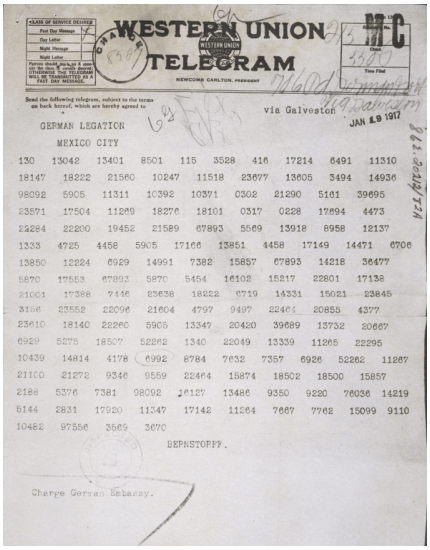

Figura 2: El telegrama Zimmermann.

4. Data Encryption Standard

Los ejemplos de los apartados anteriores nos han permitido ilustrar algunas de las cualidades que son deseables en los cifrados, así como ciertas debilidades que conviene evitar. Por un lado, la propiedad de confusión nos ayuda a que no sea fácil identificar la clave con la que se ha cifrado el mensaje. La propiedad de difusión, por otro lado, ayuda a que pequeños cambios en el mensaje original se traduzcan en varios cambios al mensaje cifrado. Finalmente, hemos visto que la linealidad en los cifrados brinda difusión, pero es muy fácil de descifrar conociendo poca información.

A continuación, presentaremos el *Data Encryption Standard (DES)*, el cual cifra bloques de 64 bits. Quiere decir que, en lugar de trabajar con bloques formados por enteros del 0 al 27 como en la correspondencia de la tabla 7, *los bloques a cifrar consisten únicamente de ceros y unos*. En resumen, el *DES* cifra bloques de ceros y unos de longitud 64.

Históricamente, los cifrados utilizados por los gobiernos se habían manejado de manera secreta. O sea que no se conocía tanto el procedimiento de *cifrado* en sí como la *clave* usada por dicho cifrado. Hoy en día, los esquemas de cifrado utilizados son conocidos. Es decir, se sabe cuál es el cifrado utilizado por las instituciones gubernamentales, financieras o de comunicación, manteniendo oculta la clave que se utiliza, por supuesto. Esto tiene sus grandes ventajas, pues al ser conocido o público el esquema de cifrado que se utiliza, hay muchos más criptoanalistas trabajando para estudiar y descubrir las posibles vulnerabilidades del cifrado. De esta manera, se puede confiar más en la seguridad propia del cifrado que en la secrecía del mismo.

Este paso de mantener los cifrados en secreto a hacerlos públicos se dio justamente con el *DES* [12, Capítulo 3]. En 1972 el US National Bureau of Standards (NBS), hoy llamado *National Institute of Standards and Technology*, realizó solicitudes para un cifrado de uso estandarizado en los Estados Unidos. Se buscaba que dicho cifrado pudiera utilizarse en

diferentes aplicaciones, tanto de gobierno, como financieras y comerciales. Dos años más tarde, recibieron una propuesta de un grupo de criptógrafos de IBM. Dicha propuesta está basada en un algoritmo conocido como *red de Feistel*, que describiremos en su momento. El cifrado propuesto se llamaba *Lucifer* (en inglés, *-cifer* se pronuncia igual que la palabra *cipher*, que significa "cifrado"), el cual tenía la capacidad de cifrar bloques de 64 bits usando una clave que consistía de 128 bits.

La NBS remitió la examinación de la seguridad del cifrado propuesto a la *National Security Agency* (NSA), cuya existencia no era admitida en aquella época. Dicha agencia de seguridad, además de cambiar el nombre del cifrado a *Data Ecryption Standard*, decidió reducir el tamaño de clave de 128 a 56 bits, haciéndolo más vulnerable a ataques por fuerza bruta. Debido a esto, se temía que dicha agencia hubiera encontrado alguna vulnerabilidad matemática sólo conocida por ellos que les permitiera romper el cifrado a voluntad. A pesar de esas inquietudes, las especificaciones del cifrado fueron puestas a disposición del público en 1977. El haber hecho público el algoritmo de dicho cifrado, junto con el rápido crecimiento en el uso de computadoras a principos de los ochenta, permitió que la comunidad de investigadores pudiera analizar a profundidad al *DES*.

4.1. Descripción general del cifrado

Como mencionamos arriba, el *DES* cifra bloques de 64 bits usando claves de 56 bits basado en una *red de Feistel*. Esto significa que al principio se aplica al bloque de 64 bits una permutación inicial; después se aplica un algoritmo iterativo de 16 rondas, en cada una de las cuales se realiza prácticamente el mismo procedimiento. Por último, se aplica una permutación final. A partir de la clave original k se derivan 16 subclaves $k_1, k_2, \ldots, k_{16}$, cada una de las cuales se utiliza en cada ronda. Este procedimiento puede visualizarse esquemáticamente en el diagrama de la figura 3.

Una descripción un poco más detallada del procedimiento es la siguiente. Primeramente, al mensaje original x, que es un bloque de 64 bits, se le aplica una *permutación inicial* $IP(x)$. El bloque resultante es dividido en dos bloques L_0 y R_0 (izquierdo y derecho) de 32 bits cada uno. Ambas mitades entran como argumento de la red de Feistel de 16 rondas. Para cada $i = 1, \ldots, 16$, al bloque (L_{i-1}, R_{i-1}) se le aplica el procedimiento de la i-ésima ronda, dando como resultado el bloque (L_i, R_i) por la fórmula siguiente:

$$L_i := R_{i-1}, \qquad R_i := L_{i-1} \oplus f(R_{i-1}, k_i).$$

Aquí, f es una función que toma el bloque derecho anterior R_{i-1} y la subclave de la ronda k_i y devuelve un bloque de longitud de 32 bits. La operación $\oplus$ es la disyunción exclusiva o XOR, que fue descrita en el ejemplo 2.3. Después de la ronda 16, se aplica la permutación final, que consiste en intercambiar las dos mitades del bloque (L_{16}, R_{16}) y aplicarle la inversa de la permutación inicial. El resultado es el mensaje cifrado,

$$y = \mathrm{DES}_k(x) = IP^{-1}(R_{16}, L_{16}).$$

En este procedimiento iterativo, las propiedades de confusión y difusión se dan en cada ronda. La propiedad de confusión es asegurada por la estructura de la función f: su im-

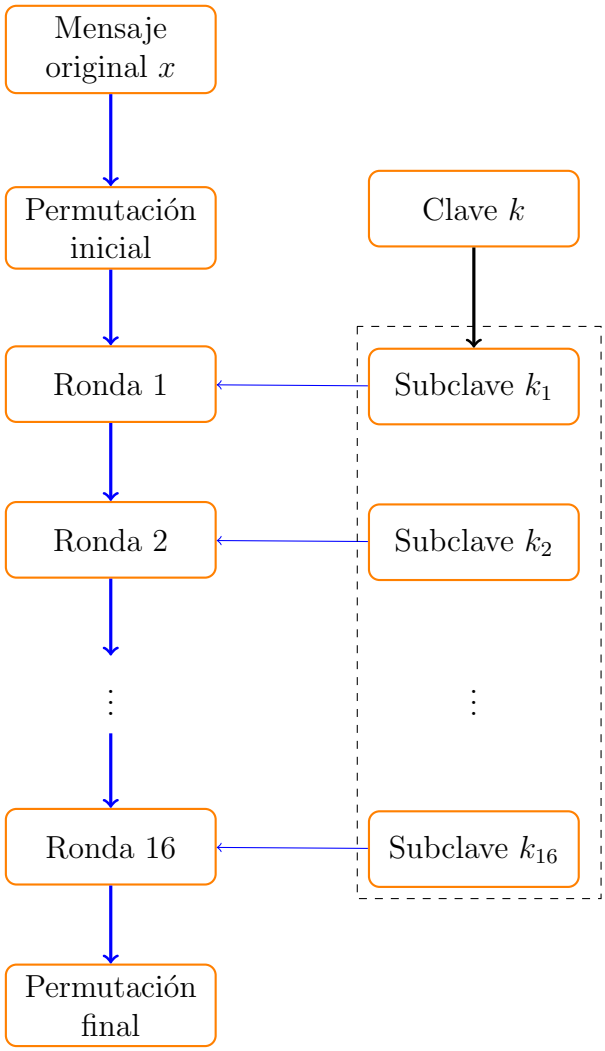

Figura 3: Diagrama de flujo del cifrado DES.

predictibilidad asegura que no exista una relación inmediata entre el mensaje cifrado y la clave original. Sin embargo, como puede observarse, de una ronda a otra solamente se cifra la mitad izquierda L_{i-1} cuando se suma con el resultado de la función f, mientras que la mitad derecha R_{i-1} pasa sin cifrarse a ser la nueva parte izquierda. Esto exige que para tener una alta difusión se deban de utilizar 16 rondas. El procedimiento que se aplica en cada ronda viene esquemáticamente descrito en la figura 4.

Para comprender a detalle el cifrado DES, necesitamos describir las permutaciones IP e IP^{-1}, explicar cómo opera la función f y cómo se generan las subclaves k_i a partir de la clave k.

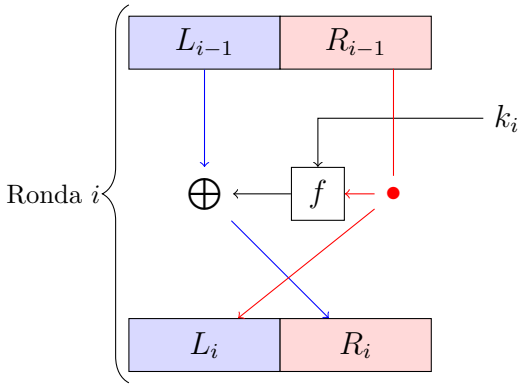

Figura 4: En cada ronda, el texto de 64 bits se divide en dos partes izquierda y derecha de 32 bits. La mitad derecha entra como argumento a la función f, pero pasa sin cifrarse a la parte izquierda. La mitad izquierda es cifrada cuando se suma con el resultado de la función f.

4.2. La permutación inicial y su inversa

La manera más concisa para describir la permutación inicial es la siguiente: Si $a = a_1a_2\dots a_{64}$ es un bloque de 64 bits, entonces su imagen bajo la permutación inicial IP es $b = IP(a) := b_1b_2\dots b_{64}$, donde para cada $i = 1,\dots,64$, se define $b_n := a_{p(n)}$, con

$$p(n) := \begin{cases} 58n \mod 66 & \text{si } n \leq 32, \\ 58n - 9 \mod 66 & \text{si } n \geq 33. \end{cases}$$

Esto quiere decir que el primer bit del bloque resultante será el que estaba en la posición $p(1) = 58$, el segundo bit del bloque resultante será el que estaba en la posición $p(2) = 50$, y así sucesivamente. Si bien ésta es una forma concisa de describirla, para implementarla es más eficiente presentar todos los valores como en la tabla de la figura 5.

IP							
58	50	42	34	26	18	10	2
60	52	44	36	28	20	12	4
62	54	46	38	30	22	14	6
64	56	48	40	32	24	16	8
57	49	41	33	25	17	9	1
59	51	43	35	27	19	11	3
61	53	45	37	29	21	13	5
63	55	47	39	31	23	15	7

Figura 5: La permutación inicial del DES. Tomada de [12, Tabla 3.1].

Similarmente, la permutación inversa IP^{-1} puede describirse de manera concisa como sigue: Si $b = b_1b_2\dots b_{64}$ es un bloque de 64 bits, su imagen es $a = IP^{-1}(b) := a_1a_2\dots a_{64}$, donde $a_m := b_{q(m)}$, con

$$q(m) := \begin{cases} 4m \mod 33 & \text{si } m \text{ es par}, \\ (4m+3 \mod 33) + 33 & \text{si } m \text{ es impar}. \end{cases}$$

Por ejemplo, el primer bit del bloque resultante será aquél que estaba en la posición $q(1) = 40$, el segundo bit resultante es el que estaba en la posición $q(2) = 8$, etcétera. Dicha permutación puede describirse por la figura 6.

IP^{-1}							
40	8	48	16	56	24	64	32
39	7	47	15	55	23	63	31
38	6	46	14	54	22	62	30
37	5	45	13	53	21	61	29
36	4	44	12	52	20	60	28
35	3	43	11	51	19	59	27
34	2	42	10	50	18	58	26
33	1	41	9	49	17	57	25

Figura 6: La permutación inversa a la inicial en el DES. Tomada de [12, Tabla 3.2].

4.3. La función *f*

La función f que aparece en el algoritmo del DES es el elemento esencial de este cifrado. Es la función f la que le otorga la propiedad de confusión. Asimismo, es el único elemento no lineal en el cifrado.

En cada ronda, esta función toma como argumentos a la subclave de la ronda y a la mitad derecha del bloque obtenido en la ronda anterior, de longitudes 48 y 32 bits, respectivamente, y devuelve un bloque de 32 bits. Más precisamente, en la ronda i la función f toma el bloque R_{i-1} y le aplica una *expansión* para convertirlo en un bloque $E(R_{i-1})$ de 48 bits. Posteriormente, el bloque expandido se suma, con la operación XOR, con la subclave k_i. El bloque resultante, de 48 bits, es dividido en ocho sub-bloques de 6 bits, cada uno de los cuales pasa por unas *cajas de sustitución*, que devuelven en total ocho bloques de 4 bits. Finalmente, el bloque completo de 32 bits es permutado, y el bloque obtenido es el resultado de la función, $f(R_{i-1}, k_i)$ (ver figura 7).

A continuación, describimos cada paso con detalle.

La expansión. Como comentábamos arriba, el primer paso que realiza la función es expandir el bloque R_{i-1}, de 32 bits, a otro bloque $E(R_{i-1})$ de 48 bits. Básicamente, lo que hace la función de expansión es repetir los bits que se ubican en las posiciones 0 y 1 módulo 4, es decir, los que están en las posiciones 4, 5, 8, 9, 12, 13, 16, 17, 20, 21, 24, 25, 28, 29, 32 y 1. Esta expansión puede describirse por la figura 8. Dicha expansión, también puede describirse de manera concisa usando aritmética modular, tal como lo hicimos con la permutación incial y su inversa, sólo que usando una expresión un poco más complicada. Si $R = r_1 r_2 \dots r_{32}$ es el bloque de 32 bits a expandir, el resultado de la expansión es $E(r) := e_1 e_2 \dots e_{48}$, donde $e_n := r_{g(n)}$, con

$$g(n) := ((6n + 25 - 5(n - 1 \mod 6)) \mod 32) + 1, \qquad n = 1, 2, \dots, 48.$$

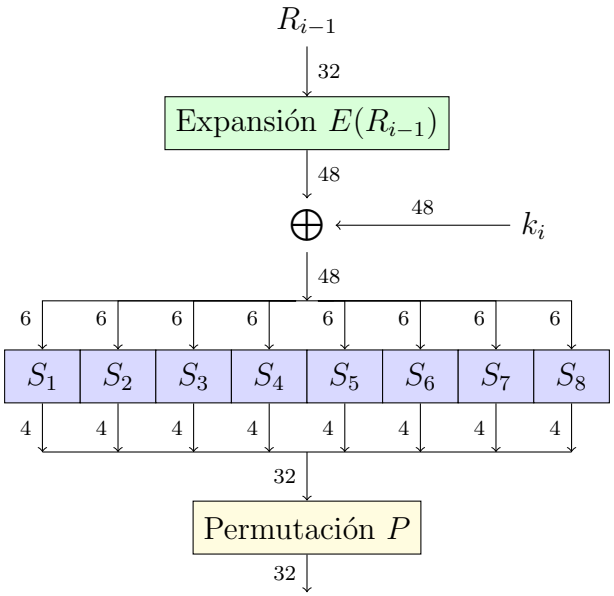

Figura 7: Descripción esquemática de la función f. A lado de cada flecha hemos indicado el tamaño de bloque correspondiente.

E					
32	1	2	3	4	5
4	5	6	7	8	9
8	9	10	11	12	13
12	13	14	15	16	17
16	17	18	19	20	21
20	21	22	23	24	25
24	25	26	27	28	29
28	29	30	31	32	1

Figura 8: La la expansión en la función f del DES. Tomada de [12, Tabla 3.3].

Las cajas de sustitución. Como se ha indicado en el diagrama de la figura 7, el bloque obtenido después de la expansión se suma con la subclave de la ronda. El bloque resultante, de 48 bits, es dividido en ocho sub-bloques de 6 bits cada uno, los cuales pasan a una de las ocho *cajas de sustitución*, o más brevemente, *S-cajas*.

Cada una de las S-cajas recibe un bloque de 6 bits y devuelve uno de 4 bits. Las S-cajas son las operaciones más importante del DES, en cuanto a que en ellas recae la seguridad del cifrado [3]. En efecto, las S-cajas son el único elemento no lineal de dicho algoritmo [5, Sección 2.2] y le brindan la propiedad de confusión [12, Subsección 3.3.2], evitando explotar en el DES las debilidades de la linealidad discutidas en la sección anterior.

A continuación, explicamos cómo es que dichas S-cajas están definidas y posteriormente haremos comentarios al respecto de dicha definición y su importancia.

Una manera de pensar en una S-caja es como una colección de cuatro permutaciones σ_0,

σ_1, σ_2, σ_3 del conjunto $\{0, 1, 2, \ldots, 14, 15\}$, las cuales difieren de una S-caja a otra. Cuando una S-caja recibe como argumento un bloque de 6 bits $b = b_5b_4b_3b_2b_1b_0$, el primer y último bit se piensan como la representación binaria de un entero x entre 0 y 3, $x := (b_5b_0)_2$. Similarmente, los bits restantes se piensan como la representación binaria de un entero y entre 0 y 15, $y := (b_4b_3b_2b_1)_2$. Los cuatro bits que la S-caja arroja como resultado son la representación binaria del resultado de aplicar la permutación σ_x de la S-caja al entero y: $S(b) := \sigma_x(y)$.

Para ilustrar la descripción general anterior, consideremos la primera S-caja del DES, la cual consiste de las permutaciones

$$\begin{aligned}
\sigma_0 &= (0\ 14)(1\ 4\ 2\ 13\ 9\ 10\ 6\ 11\ 12\ 5\ 15\ 7\ 8\ 3),\\
\sigma_1 &= (0)(1\ 15\ 8\ 10\ 12\ 9\ 6\ 13\ 5\ 2\ 7)(3\ 4\ 14)(11),\\
\sigma_2 &= (0\ 4\ 13\ 10\ 9\ 12\ 3\ 8\ 15)(1)(2\ 14\ 5\ 6)(7\ 11),\\
\sigma_3 &= (0\ 15\ 13)(1\ 12\ 10\ 3\ 2\ 8\ 5\ 9\ 11\ 14\ 6)(4)(7).
\end{aligned}$$

Hemos presentado cada permutación en términos de su descomposición en producto de ciclos ajenos. Esta notación significa, por ejemplo, que la permutación σ_3 envía el 0 al 15, el 15 al 13 y éste 13 de vuelta al 0; envía el 1 al 12, éste al 10, éste al 3, éste al 2, éste al 8, éste al 5, éste al 9, éste al 11, éste al 14, éste al 6 y éste de vuelta al 1; finalmente, el 4 es enviado en sí mismo, al igual que el 7.

Supongamos que la S-caja S_1 recibe por argumento al bloque de seis bits $b = 111010$. El entero x es aquél cuya representación binaria es 10, ya que estos son el primer y último bits de b. En otras palabras, es $x = 10_2 = 2$. Similarmente, es $y = 13$, pues los cuatro bits centrales de b, a saber, 1101, representan al número 13. Por tanto, debemos aplicar la permutación indicada por $x = 2$ al entero $y = 13$, es decir, $\sigma_2(13) = 10$. Finalmente, el resultado obtenido de aplicar la S-caja S_1 al bit b es la representación en cuatro bits (binaria) de 10, que es 1010:

$$S_1(111010) = \sigma_2(13) = 10 \equiv 1010_2.$$

Cada una de las ocho cajas de sustitución operan de manera análoga que S_1, sólo que las cuatro permutaciones de cada S-caja son diferentes:

- Permutaciones de la caja de sustitución S_2:

$$\begin{aligned}
\sigma_0 &= (0\ 15\ 10\ 2\ 8\ 9\ 7\ 4\ 6\ 3\ 14\ 5\ 11\ 13)(12),\\
\sigma_1 &= (0\ 3\ 7\ 14\ 11\ 10\ 1\ 13\ 9)(2\ 4\ 15\ 5)(6\ 8\ 12),\\
\sigma_2 &= (0)(1\ 14\ 2\ 7)(3\ 11\ 6\ 13)(4\ 10\ 12\ 9\ 8\ 5)(15),\\
\sigma_3 &= (0\ 13\ 5\ 15\ 9\ 6\ 4\ 3\ 1\ 8\ 11\ 12)(2\ 10\ 7)(14).
\end{aligned}$$

- Permutaciones de la caja de sustitución S_3:

$$\begin{aligned}
\sigma_0 &= (0\ 10\ 12\ 11\ 7\ 5\ 3\ 14\ 2\ 9\ 13\ 4\ 6\ 15\ 8\ 1),\\
\sigma_1 &= (0\ 13\ 11\ 14\ 15\ 1\ 7\ 10\ 5\ 4\ 3\ 9\ 8\ 2)(6)(12),\\
\sigma_2 &= (0\ 13\ 10\ 2\ 4\ 8\ 11\ 12\ 5\ 15\ 7)(1\ 6\ 3\ 9)(14),\\
\sigma_3 &= (0\ 1\ 10\ 14\ 2\ 13\ 5\ 9\ 15\ 12\ 11\ 3)(4\ 6\ 8)(7).
\end{aligned}$$

- Permutaciones de la caja de sustitución S_4:

$$\begin{aligned}\sigma_0 &= (0\ 7\ 10\ 8\ 1\ 13\ 12\ 11\ 5\ 6\ 9\ 2\ 14\ 4)(3)(15),\\ \sigma_1 &= (0\ 13\ 10\ 2\ 11\ 12\ 1\ 8\ 4\ 6)(3\ 5\ 15\ 9\ 7)(14),\\ \sigma_2 &= (0\ 10\ 3)(1\ 6\ 7\ 13\ 2\ 9)(4\ 12\ 5\ 11\ 14\ 8\ 15),\\ \sigma_3 &= (0\ 3\ 6\ 13\ 7\ 8\ 9\ 4\ 10\ 5\ 1\ 15\ 14\ 2)(11)(12).\end{aligned}$$

- Permutaciones de la caja de sustitución S_5:

$$\begin{aligned}\sigma_0 &= (0\ 2\ 4\ 7\ 6\ 11\ 15\ 9\ 5\ 10\ 3\ 1\ 12\ 13)(8)(14),\\ \sigma_1 &= (0\ 14\ 8\ 5\ 7\ 1\ 11\ 10\ 15\ 6\ 13\ 9)(2)(3\ 12)(4),\\ \sigma_2 &= (0\ 4\ 10\ 12\ 6\ 7\ 8\ 15\ 14)(1\ 2)(3\ 11\ 5\ 13)(9),\\ \sigma_3 &= (0\ 11\ 9\ 15\ 3\ 7\ 13\ 4\ 1\ 8\ 6\ 2\ 12\ 10)(5\ 14).\end{aligned}$$

- Permutaciones de la caja de sustitución S_6:

$$\begin{aligned}\sigma_0 &= (0\ 12\ 14\ 5\ 2\ 10\ 3\ 15\ 11\ 4\ 9\ 13\ 7\ 8)(1)(6),\\ \sigma_1 &= (0\ 10\ 13\ 11\ 14\ 3\ 2\ 4\ 7\ 5\ 12)(1\ 15\ 8\ 6\ 9),\\ \sigma_2 &= (0\ 9)(1\ 14\ 11\ 10\ 4\ 2\ 15\ 6\ 12)(3\ 5\ 8\ 7)(13),\\ \sigma_3 &= (0\ 4\ 9\ 14\ 8\ 11\ 7\ 10\ 1\ 3\ 12\ 6\ 15\ 13)(2)(5).\end{aligned}$$

- Permutaciones de la caja de sustitución S_7:

$$\begin{aligned}\sigma_0 &= (0\ 4\ 15\ 1\ 11\ 7\ 13\ 10\ 9\ 12\ 5)(2)(3\ 14\ 6\ 8),\\ \sigma_1 &= (0\ 13\ 15\ 6\ 1)(2\ 11\ 12)(3\ 7\ 10\ 5\ 9)(4)(8\ 14),\\ \sigma_2 &= (0\ 1\ 4\ 12)(2\ 11\ 8\ 10\ 6\ 7\ 14\ 9\ 15)(3\ 13\ 5),\\ \sigma_3 &= (0\ 6\ 10)(1\ 11\ 15\ 12\ 14\ 3\ 8\ 9\ 5\ 4)(2\ 13)(7).\end{aligned}$$

- Permutaciones de la caja de sustitución S_8:

$$\begin{aligned}\sigma_0 &= (0\ 13)(1\ 2\ 8\ 10\ 3\ 4\ 6\ 11\ 14\ 12\ 5\ 15\ 7)(9),\\ \sigma_1 &= (0\ 1\ 15\ 2\ 13\ 14\ 9\ 5\ 3\ 8\ 12)(4\ 10\ 6\ 7)(11),\\ \sigma_2 &= (0\ 7\ 2\ 4\ 9\ 6\ 14\ 5\ 12\ 15\ 8)(1\ 11\ 13\ 3)(10),\\ \sigma_3 &= (0\ 2\ 14\ 6\ 8\ 15\ 11)(1)(3\ 7\ 13\ 5\ 10\ 9\ 12)(4).\end{aligned}$$

Ya habíamos comentado arriba, y esto puede entreverse del procedimiento general de operación de las S-cajas, que en las cajas de sustitución recae la parte esencial de la seguridad del cifrado. Más aún, la elección específica de estas S-cajas permite que el cifrado sea resistente a un ataque criptográfico conocido como *criptoanálsis diferencial*. A grandes rasgos, el criptoanálisis diferencial busca romper la seguridad de un cifrado a partir de estudiar cómo pequeños cambios en el mensaje original se traducen al mensaje cifrado. En la época en que

el DES fue presentado al público, el criptoanálisis diferencial no había sido descubierto por la comunidad científica. Sin embargo, en el año de 1990, que fue cuando el criptoanálisis diferencial fue descubierto, el equipo de criptógrafos de IBM declaró que *ellos ya conocían* de ese tipo de ataque, que no lo habían revelado a la comunidad por considerarlo un tipo de ataque bastante avanzado y que *las S-cajas del DES fueron específicamente diseñadas para resistir al criptoanálisis diferencial* [3]. Sobre el criterio de diseño de las S-cajas, lo único que se sabe a ciencia cierta es que fueron diseñadas para satisfacer las siguientes características [12, Subsección 3.3.2]:

1. Cada S-caja recibe bloques de 6 bits y devuelve bloques de 4 bits.
2. Ninguno de los bits resultantes debe ser cercano a una combinación lineal de los bits ingresados.
3. Si el primer y el último bits a ingresar son fijos, y variamos los 4 bits de enmedio, todos los posibles resultados de 4 bits deben poder obtenerse. En otras palabras, las S-cajas consisten efectivamente de permutaciones.
4. Si dos entradas difieren **en un solo bit**, los bloques resultantes difieren **en al menos dos bits**.
5. Si dos entradas difieren **en los dos bits centrales**, sus resultados difieren **en al menos dos bits**.
6. Si dos entradas difieren en los primeros bits y son idénticos en los últimos dos, ambos resultados son diferentes.
7. Dada una diferencia no nula en 6 bits entre entradas, a lo más 8 de las 32 parejas de bits exhibiendo dicha diferencia pueden dar como resultado la misma diferencia.
8. Una **colisión** (diferencia cero en la salida) sólo es posible para tres S-cajas adyacentes.

La permutación P. Después de que las ocho S-cajas producen cada una un bloque de 4 bits, se le aplica una permutación P al bloque resultante de 32 bits. El objetivo de dicha permutación es brindarle difusión al cifrado, ya que ésta hace que los bits que resultan de una S-caja particular en una ronda concreta, en la siguiente ronda entren como argumento a S-cajas diferentes. Más aún, esto produce un *efecto avalancha*, ya que **a partir de la quinta ronda, cada bit resultante depende de todos los bits iniciales** [12, Subsección 3.3.2]. Esta permutación de los bloques de 32 bits, descrita en producto de ciclos, es

$$P = (1\ 16\ 10\ 15\ 31\ 4\ 21\ 32\ 25\ 19\ 24\ 9)(2\ 7\ 28\ 6\ 12\ 26\ 13\ 5\ 29\ 22\ 27\ 30\ 11\ 23\ 3\ 20\ 14\ 18\ 8\ 17).$$

4.4. Esquema de generación de las subclaves

Para terminar de entender cómo opera el cifrado DES, resta explicar cómo se generan las subclaves $k_1, k_2, \ldots, k_{16}$ que entran como argumento a la función f en la ronda correspondiente. Esencialmente, cada una de ellas es una permutación (de algunos) de los bits de la clave original k.

Formalmente, la clave original k del cifrado DES consiste de 64 bits, por lo que en principio existen 2^{64} claves diferentes. Sin embargo, el primer paso del esquema de generación de las subclaves *ignora los bits que se ubican en las posiciones múltiplo de ocho*. De esta manera, solamente 56 de los 64 bits originales intervienen en el cifrado. Por ello, podemos decir con toda justicia que *el tamaño del espacio de claves del DES es de solamente* 2^{56} *bits*.

Elección permutada 1. Una descripción más precisa de cómo pasamos de 64 a 56 bits es diciendo que se aplica una *elección permutada* a los 64 bits. Esta primera elección permutada puede describirse por la tabla de la figura 9, es decir: si $k = k_1 k_2 \ldots k_{64}$ es la clave original de 64 bits, entonces el resultado de $PC_1(k)$ es el bloque de 56 bits $l_1 l_2 \ldots l_{56}$ dado por

$$l_n := k_{F(n)},$$

donde $F(n)$ es el elemento en la entrada n de dicha tabla ($F(1) = 57, F(2) = 49, \ldots$). La función $F : \{1, \ldots, 56\} \to \{1, \ldots, 64\}$ puede describirse analíticamente por $F(n) :=$

$PC-1$							
57	49	41	33	25	17	9	1
58	50	42	34	26	18	10	2
59	51	43	35	27	19	11	3
60	52	44	36	63	55	47	39
31	23	15	7	62	54	46	38
30	22	14	6	61	53	45	37
29	21	13	5	28	20	12	4

Figura 9: La primera elección permutada del esquema de generación de subclaves del DES. Tomada de [12, Tabla 3.13].

$f(g(h(n)))$, donde

$$f(n) := 57n \mod 65, \qquad g(n) := \begin{cases} n+20 & \text{si } 29 \leq n \leq 32, \\ n-20 & \text{si } 49 \leq n \leq 52, \\ n+8 & \text{si } 33 \leq n \leq 36, \\ n-8 & \text{si } 41 \leq n \leq 44, \\ n+16 & \text{si } 37 \leq n \leq 40, \\ n-16 & \text{si } 53 \leq n \leq 56, \end{cases}$$

$$h(n) := n - (n-1 \mod 8) + ((n \mod 8 + 3) \mod 8) \qquad \text{si } n \geq 33.$$

Construcción recursiva de las subclaves. Una vez que de la clave k se ha extraído el bloque $PC_1(k)$ de 56 bits, éste se divide en dos mitades C_0 y D_0 de 28 bits cada una. A partir de ellos, se van construyendo los bloques $C_1, D_1, C_2, D_2, \ldots, C_{16}, D_{16}$ para después extraer de C_i y D_i la subclave k_i.

El procedimiento de construcción de C_i, D_i y la clave k_i a partir de C_{i-1} y D_{i-1} viene descrito en el diagrama de la figura 10. Cada una de las mitades C_{i-1} y D_{i-1} es permutada cíclicamente uno o dos lugares hacia la izquierda (*left shifting* LS_i), según la ronda de la que se trate. En las rondas $i = 1, 2, 9$ y 16, el left shifting LS_i consiste de permutar hacia la izquierda en un lugar, mientras que en el resto de las rondas, LS_i permuta dos lugares hacia la izquierda. El resultado de dichas permutaciones son los nuevos bloques C_i y D_i:

$$C_i := LS_i(C_{i-1}), \qquad D_i := LS_i(D_{i-1}).$$

La clave k_i de 48 bits es extraída del bloque de 56 bits formado por las dos mitades C_i y D_i. De entre todos los 56 bits, se toman 48 de acuerdo a una *segunda elección permutada* PC_2, descrita en la tabla de la figura 11,

$$k_i := PC_2(C_i, D_i).$$

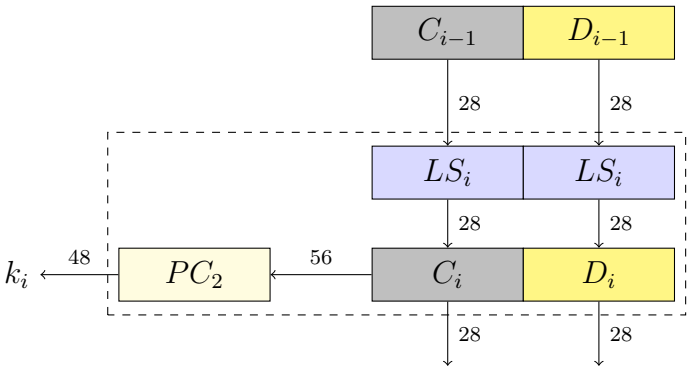

Figura 10: Descripción esquemática de la construcción de la clave k_i de la ronda i.

$PC-2$							
14	17	11	24	1	5	3	28
15	6	21	10	23	19	12	4
26	8	16	7	27	20	13	2
41	52	31	37	47	55	30	40
51	45	33	48	44	49	39	56
34	53	46	42	50	36	29	32

Figura 11: La segunda elección permutada del esquema de generación de subclaves del DES. Tomada de [12, Tabla 3.14].

4.5. Descifrando con el DES

Ahora procederemos a explicar cómo es el proceso de descifrado con el *DES*. Un aspecto soprendente de este cifrado es que *las operaciones para cifrar y descifrar son exactamente las mismas*, cambiando únicamente (y sólo un poco) el esquema de generación de claves. Para poder ver esta propiedad, presentamos primero un resumen del cifrado DES:

1. Al bloque original x, de 56 bits, se le aplica la permutación inicial IP (ver subsección 4.2).
2. El resultado de aplicar la permutación se divide en dos bloques L_0 y R_0 de 28 bits cada uno: $(L_0, R_0) := IP(x)$.
3. Para cada $i = 1, 2, \ldots, 16$, se definen $L_i := R_{i-1}$ y $R_i := L_{i-1} \oplus f(R_{i-1}, k_i)$, donde f es la función descrita en la subsección 4.3 y k_i es la subclave de la ronda i.
4. Las mitades L_{16} y R_{16} se intercambian y al bloque resultante se le aplica la inversa de la permutación inicial (ver subsección 4.2),
$$y = IP^{-1}(R_{16}, L_{16}).$$
5. El resultado es el mensaje cifrado $y = \mathrm{DES}_k(x)$.

Ahora bien, el hecho de que el proceso de cifrado y descifrado sean el mismo, salvo por el esquema de generación de las subclaves, se debe a que el DES está basado en una red de Feistel. Vamos a mostrar que efectivamente, podemos recuperar el mensaje original x a partir del mensaje cifrado y aplicando el mismo procedimiento.

Consideremos el mensaje cifrado $y = \mathrm{DES}_k(x)$. Apliquémosle la permutación inicial $IP(y)$ y al resultado dividámoslo en dos bloques izquierdo y derecho L_0^d y R_0^d (el superíndice d indica que estamos en modo de *descifrado*). Luego, para cada $i = 1, 2, \ldots, 16$ defínanse

$$L_i^d := R_{i-1}^d \qquad \text{y} \qquad R_i^d := L_{i-1}^d \oplus f(R_{i-1}^d, k_{17-i}).$$

Escencialmente estamos definiendo las mitades L_i^d y R_i^d tal como lo hicimos en el modo de cifrado, sólo que las subclaves se usan en el orden contrario: la primera subclave es k_{16}, la segunda es k_{15}, etcétera. Finalmente, sea $\widetilde{x} := IP^{-1}(R_{16}^d, L_{16}^d)$. Nuestro objetivo es probar que $\widetilde{x} = x$, es decir, que después de aplicarle al mensaje cifrado $y = \mathrm{DES}_k(x)$ el mismo procedimiento, pero usando las subclaves en el orden contrario, recuperamos el mensaje original x.

Para empezar, vamos a probar por inducción sobre el número de rondas que $L_i^d = R_{16-i}$ y que $R_i^d = L_{16-i}$ para todo $i = 0, 1, 2, \ldots, 16$. Para $i = 0$, tenemos que

$$(L_0^d, R_0^d) = IP(y) = IP(IP^{-1}(R_{16}, L_{16})) = (R_{16}, L_{16}).$$

Supongamos que para $i = 0, \ldots, j$ se cumple que $L_i^d = R_{16-i}$ y que $R_i^d = L_{16-i}$. Para $i = j+1$, tenemos por definición de R y L y por hipótesis de inducción que

$$L_i^d = L_{j+1}^d = R_j^d = L_{16-j} = R_{16-(j+1)} = R_{16-i}.$$

La primera y quinta igualdades son por $i = j + 1$, la segunda y cuarta igualdades son por la definición recursiva de L y la tercera igualdad es hipótesis de inducción. Similarmente,

$$\begin{aligned} R_i^d &= L_{i-1}^d \oplus f(R_{i-1}^d, k_{17-i}) = L_j^d \oplus f(R_j^d, k_{16-j}) \\ &= R_{16-j} \oplus f(L_{16-j}, k_{16-j}) = R_{16-j} \oplus f(R_{16-(j+1)}, k_{16-j}) \\ &= (L_{16-(j+1)} \oplus f(R_{16-(j+1)}, k_{16-j})) \oplus f(R_{16-(j+1)}, k_{16-j}) \\ &= L_{16-i} \oplus (f(R_{16-(j+1)}, k_{16-j}) \oplus f(R_{16-(j+1)}, k_{16-j})) \\ &= L_{16-i}. \end{aligned}$$

Aquí hemos aplicado definición de R^d, $i = j + 1$, hipótesis de inducción, definición de L, definición de R, asociatividad, $i = j + 1$ y el hecho de que $a \oplus a = 0$ si a es un bloque de bits, pues trabajamos módulo 2.

En particular, tenemos que $L_{16}^d = R_0$ y $R_{16}^d = L_0$. Por lo tanto, tomando en cuenta la definición de L_0 y R_0, tenemos

$$\widetilde{x} = IP^{-1}(R_{16}^d, L_{16}^d) = IP^{-1}(L_0, R_0) = IP^{-1}(IP(x)) = x.$$

Esto demuestra que **el descifrado en el DES es el mismo procedimiento que el cifrado, usando las subclaves en orden contrario**.

Subclaves en modo de descifrado. En principio uno pudiera suponer que para aplicar el proceso de descifrado en el DES es necesario generar primero todas las subclaves $k_1, k_2, \ldots, k_{16}$ según se describió en la subsección 4.4. Sin embargo, dicho esquema de generación de las subclaves *permite generar las subclaves en el orden inverso*, es decir, generar primero k_{16}, después k_{15}, etc.

Recordemos que las subclaves de cada ronda se generaban de la siguiente manera:

1. Aplicar la primera elección permutada a la clave k.
2. El resultado se divide en dos bloques de 28 bits cada uno: $(C_0, D_0) := PC_1(k)$.
3. Para cada $i = 1, 2, \ldots, 16$, hacer

$$C_i := LS_i(C_{i-1}), \qquad D_i := LS_i(D_{i-1}), \qquad k_i := PC_2(C_i, D_i),$$

donde LS_i es un *left shifting* sencillo cuando $i = 1, 2, 9$ y 16, y doble en los demás casos y PC_2 es la segunda elección permutada.

Puesto que $k_i := PC_2(C_i, D_i)$, el problema de construir las subclaves en el orden inverso se reduce a obtener C_i y D_i en el orden inverso también. Más aún, puesto que los bloques C_i y D_i se obtuvieron a partir de los anteriores haciendo corrimientos sencillos y dobles hacia la izquierda, según la ronda, podemos generar dichos bloques en el orden inverso *haciendo corrimientos sencillos y dobles hacia la derecha*, siempre y cuando generemos primero los bloques C_{16} y D_{16}.

Vamos a mostrar la siguiente afirmación, sencilla, pero a su vez poco evidente: **los bloques C_{16} y D_{16} son iguales a C_0 y D_0, respectivamente**. En efecto, recordemos que C_{16} se obtiene de aplicar LS_{16} a C_{15}, que a su vez se obtuvo de aplicar LS_{15} a C_{14}, etcétera. Por tanto, C_{16} se obtiene de aplicar $LS_{16}, LS_{15}, \ldots, LS_2, LS_1$ a C_0. Dado que LS_1, LS_2, LS_9 y LS_{16} son corrimientos sencillos y los demás son dobles, en total tenemos 4 corrimientos sencillos y 12 corrimientos dobles, dando un total de 28 corrimientos a la izquierda. En otras palabras, C_{16} se obtiene de C_0 haciendo 28 corrimientos a la izquierda. Recordando que C_0 es un bloque de 28 bits, el aplicarle 28 corrimientos a la izquierda lo deja igual. Por tanto $C_{16} = C_0$. De la misma manera, $D_{16} = D_0$.

Resumiendo la discusión anterior, para generar las subclaves en el orden inverso se procede de la siguiente manera:

1. Aplicar la primera elección permutada a la clave k.
2. El resultado se divide en dos bloques de 28 bits cada uno: $(C_{16}, D_{16}) := PC_1(k)$.
3. Para cada $j = 1, 2, \ldots, 16$, hacer

$$C_{16-j} := RS_j(C_{17-j}), \qquad D_{16-j} := RS_j(D_{17-j}), \qquad k_{17-j} := PC_2(C_{17-j}, D_{17-j}),$$

   donde RS_j es un *right shifting* sencillo cuando $j = 1, 2, 9$ y 16, y doble en los demás casos y PC_2 es la segunda elección permutada.

De esta manera, se generan las subclaves en el orden contrario, y en dicho orden es que se utilizan para descifrar en el DES.

4.6. Ataques al cifrado y alternativas

Para la época en que el DES se utilizó de manera estandarizada, es decir, de 1977 a 1999, el cifrado DES era suficientemente resistente a ataques por fuerza bruta. Esto se debe a que el espacio de claves del DES es de 2^{56}, que para la tecnología de esa época era bastante grande. A lo largo de la década de los 90 se propusieron diferentes ataques al DES con máquinas costosas y especialmente diseñadas para ello, logrando romper el cifrado en cuestión de horas [12, Sección 3.5]. Posteriormente, en el año de 2006, las universidades de Bochum y de Kiel construyeron la máquina llamada COPACOBANA (Cost-Optimized Parallel Code-Breaker), que tuvo un costo aproximado de $10,000 dólares. Dicha máquina en promedio es capaz de descubrir la clave usada en un cifrado DES en solamente 7 días en promedio. Por supuesto, quienes tengan suficientes recursos para ello, digamos gobiernos y grandes empresas, son capaces de invertir recursos para romper el cifrado con ataques por fuerza bruta.

Por otra parte, los ataques analíticos pueden ayudar a romper un cifrado. En 1990 el *criptoanálisis diferencial* fue dado a conocer a la comunidad científica, y tres años después el *criptoanálisis lineal*. Ambos métodos permiten descubrir una clave utilizada en cualquier cifrado de bloque siempre que se conozcan una cierta cantidad de parejas (x, y), donde y es el mensaje cifrado de x. Para el DES, el criptoanálisis diferencial permite descubrir la clave si se conocen 2^{47} parejas, mientras que el criptnoanálisis lineal baja este número a 2^{43}. Estos

números, aunque siguen siendo bastante grandes, son significativamente más pequeños que el espacio de clave del DES, que, aunado al desarrollo tecnólogico dado desde principios de la década de los 90, cada vez era más viable romper el cifrado DES por fuerza bruta o por cualquier otro método.

Hoy en día, un tamaño de clave de 56 bits es considerado pequeño para cuestiones de seguridad, ya que la capacidad de cómputo actual permite romper cifrados de este tamaño de clave mediante un ataque por fuerza bruta relativamente rápido. Por este motivo, se han llegado a utilizar algunas variantes del DES en las que el tamaño de clave es mayor.

El DES triple. Como su nombre lo indica, el DES triple lo que hace es aplicar el DES tres veces, usando tres claves diferentes a la vez. Más precisamente, si κ_1, κ_2 y κ_3 son claves de 56 bits, el triple DES (denotado 3DES o TDEA) aplicado a un mensaje x es

$$y = 3\mathrm{DES}_{\kappa_1,\kappa_2,\kappa_3}(x) := \mathrm{DES}_{\kappa_3}(\mathrm{DES}_{\kappa_2}(\mathrm{DES}_{\kappa_1}(x))).$$

Por supuesto, este procedimiento es aproximadamente tres veces más lento que el DES simple. Sin embargo, esto permite que el cifrado sea mucho más resistente a ataques por fuerza bruta, pues el tamaño de clave se triplica y el espacio de clave consiste de 2^{168} elementos. Otra variante similar a la anterior es

$$y = 3\mathrm{DES}_{\kappa_1,\kappa_2,\kappa_3}(x) := \mathrm{DES}_{\kappa_3}(\mathrm{DES}^{-1}_{\kappa_2}(\mathrm{DES}_{\kappa_1}(x))),$$

donde en el segundo paso se aplica el proceso del DES inverso, es decir, en modo de descifrado. Esta variante tiene la cualidad de que si $\kappa_1 = \kappa_2 = \kappa_3$, el procedimiento coincide con el cifrado DES simple, lo cual es requerido en ciertas aplicaciones. El 3DES es eficiente en hardware pero no en software, y ha sido muy popular en aplicaciones financieras y protección biométrica de información en pasaportes electrónicos.

El DES con blanqueamiento de clave. Otra variante del DES triple utiliza la técnica de *blanqueamiento de clave*. Es un procedimiento muy sencillo en el que se elige una clave k para el DES y dos claves adicionales κ_1, κ_2. Entonces, se utiliza el siguiente esquema de cifrado,

$$y = \mathrm{DES}_{k,\kappa_1,\kappa_2}(x) = \mathrm{DES}_k(x \oplus \kappa_1) \oplus \kappa_2$$

que prácticamente posee la misma velocidad que el DES pero que aumenta significativamente la seguridad del cifrado.

Hay un aspecto criptoanalítico del DES que puede incluso llegar a comprometer la seguridad del 3DES y es la existencia de *claves débiles*, *claves semi-débiles* y *claves posiblemente semi-débiles*. Recordemos que tanto en el proceso de cifrado como el de descifrado del DES se llevan a cabo dieciséis rondas, en las cuales se utiliza una subclave k_i generada a partir de la elección permutada $PC_1(k)$ de una clave original k de 64 bits. En ese sentido, una clave k se dice ser *débil* si resulta que las dieciséis subclaves de rondas $k_1, \ldots, k_{16}$ son todas iguales entre sí. Recordando que las subclaves se obtenen al permutar cíclicamente las mitades de $PC_1(k)$ en una cierta cantidad de lugares según la ronda, la igualdad de todas las subclaves

ocurre cuando los bits de cada mitad de $PC_1(k)$ son todos iguales, es decir, en cada mitad todos 0 o todos 1. Así, las siguientes cuatro son las claves débiles para el DES, escritas en formato hexadecimal:[2]

$$\begin{array}{rr} 0101010101010101 & \text{FEFEFEFEFEFEFEFE} \\ \text{E0E0E0E0F1F1F1F1} & \text{1F1F1F1F0E0E0E0E} \end{array}$$

Por otro lado, como el 3DES es una composición del DES consigo mismo tres veces, es necesario tomar en cuenta que hay ciertas elecciones de claves κ_1 y κ_2 para las cuales DES_{κ_1} y DES_{κ_2} cifran igual. Esto provoca que en el cifrado del 3DES

$$y = \text{3DES}_{\kappa_1,\kappa_2,\kappa_3}(x) = \text{DES}_{\kappa_3}(\text{DES}^{-1}_{\kappa_2}(\text{DES}_{\kappa_1}(x)))$$

la operación $\text{DES}^{-1}_{\kappa_2}$ revierte lo hecho por DES_{κ_1}. Luego, ese tipo de par de claves, llamadas *semi-débiles*, debe de evitarse en el 3DES debido a que su uso reduce su nivel de seguridad al del DES. Dichos pares de claves semi-débiles son

$$\begin{gathered} \text{011F011F010E010E y 1F011F010E010E01,} \\ \text{01E001E001F101F1 y E001E001F101F101,} \\ \text{01FE01FE01FE01FE y FE01FE01FE01FE01,} \\ \text{1FE01FE00EF10EF1 y E01FE01FF10EF10E,} \\ \text{1FFE1FFE0EFE0EFE y FE1FFE1FFE0EFE0E,} \\ \text{E0FEE0FEF1FEF1FE y FEE0FEE0FEF1FEF1.} \end{gathered}$$

Además, para el DES también hay 48 claves *posiblemente débiles*, que pueden consultarse en [1, p. 12], con la cualidad de que, de las dieciséis subclaves $k_1, \ldots, k_{16}$, sólo hay cuatro distintas entre sí.

Finalmente, es importante señalar que los cifrados de bloques de 64 bits, como el DES y sus variantes, tienen otro problema importante de seguridad: la alta frecuencia de *colisiones*. Una colisión es simplemente que se obtengan dos textos cifrados iguales, lo cual, cuando se encuentra, permite obtener información significativa acerca del texto original. En el caso de los cifrados de bloque de 64 bits, la probabilidad de obtener una colisión después de cifrar $2^{32} = 4294967296$ bloques es muy alta. Pudiera parecer que son muchos bloques de bits, pero tómese en cuenta que esta cantidad equivale a medio gigabyte de información. Para evitar estos problemas de seguridad, en 2017 el US National Institute of Standards and Technology (NIST) recomendó no utilizar una misma clave en el 3DES para cifrar más de 2^{20} bloques de 64 bits [1, Subsección 3.4] y desde 2024 ha desautorizado su uso en nuevas aplicaciones [2, Sección 2].

[2] En el formato hexadecimal, cada símbolo $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \text{A}, \text{B}, \text{C}, \text{D}, \text{E}$ y F representa un número entre 0 y 15, los cuales en binario se expresan con cuatro dígitos 0 o 1. Particularmente, se tiene que $0 = 0000$, $1 = 0001$, $\text{E} = 1110$ y $\text{F} = 1111$. De esta manera, las claves del DES, que constan de 64 bits, se expresan más brevemente con 16 símbolos hexadecimales.

5. Comentarios finales

En esta exposición empezamos describiendo algunos cifrados sencillos, que aunque ya no son de utilidad práctica, nos han permitido entender algunas de las propiedades deseables para los cifrados, así como algunas debilidades típicas que se deben evitar. También hemos presentado el cifrado DES, el cual consiste de un algoritmo más complejo cuya estructura es la de una red de Feistel y que posee fuertemente las propiedades de confusión, difusión, no linealidad y resistencia al criptoanálisis diferencial. Además, posee un espacio de claves de tamaño 2^{56}, el cual es bastante grande para la época en que fue dominante y por lo tanto resistente a los ataques por fuerza bruta. El DES fue un cifrado estandarizado y ampliamente utilizado de 1977 a 1999. Originalmente había sido pensado para utilizarse por solamente 10 años, hasta 1987. Sin embargo, como no se le encontraron debilidades serias, se extendió su uso hasta 1999.

Este cifrado nos permitió entender la complejidad que está detrás de los cifrados modernos, los cuales realizan operaciones por computadora que difícilmente podríamos realizar a mano. Sin duda alguna, la tecnología ha permitido avanzar en los aspectos de la seguridad, abriendo la puerta a todas las aplicaciones que actualmente forman parte de nuestra cotidianeidad.

En 1997, el US National Institute of Standards and Technology (NIST) abrió una convocatoria para un nuevo cifrado estandarizado, el *Advanced Encryption Standard (AES)* para reemplazar al DES. Esta fue una convocatoria totalmente abierta en la que el NIST fungió como administrador, a diferencia de con el DES. En cada una de las tres rondas de selección del AES, el NIST y la comunidad científica discutieron las ventajas y desventajas de cada cifrado hasta que se seleccionó un ganador.

Los requisitos que el NIST impuso a las propuestas fueron los siguientes:

1. Que fuera un cifrado de bloques de 128 bits.
2. Que sea capaz de funcionar con claves de longitudes de 128, 192 y 256 bits.
3. Eficiencia en software y hardware.

En agosto de 1999, fueron anunciados cinco finalistas: *Mars*, de la IBM; *RC6* de los Laboratorios RSA; *Rijndael*, diseñado por los criptógrafos belgas Joan Daemen y Vincent Rijmen; *Serpent*, diseñado por Ross Anderson, Eli Biham y Lars Knudsen; y *Twofish*, diseñado por criptógrafos de la Counterpane Internet Security, de Princeton y Berkeley. Después de un año de exhaustivos análisis entre los cifrados presentados, se anunció como ganador al cifrado Rijndael, pasando a ser el nuevo *Advanced Encryption Standard*.

Este ejemplo nos enseña que los cifrados no son eternos y que aunque un cifrado sea capaz de responder a las necesidades de una época concreta, eventualmente tanto la tecnología como los avances de la criptografía hacen que los cifrados deban ser reemplazados por mejores algoritmos. Actualmente, la expectativa es que la computación cuántica se desarrolle de manera significativa y sea criptográficamente relevante en la próxima década. De hecho, desde hace tiempo ya existen los llamados *algoritmos cuánticos* que, en principio, podrían llegar a ser implementados en computadoras cuánticas y ser capaces de romper muchos de los cifrados que hoy son considerados seguros, incluyendo algunas versiones del AES.

Por ejemplo, el *algoritmo de Grover* (1996) es un algoritmo cuántico de búsqueda que permite con una alta probabilidad realizar ataques por fuerza bruta exitosos en un tiempo del orden de la raíz cuadrada del tamaño del espacio de claves [7]. En el caso del AES-128, es decir, la versión del AES con claves de 128 bits, el espacio de claves es de tamaño 2^{128}. En teoría, cuando las computadoras cuánticas lleguen a ser capaces de implementar el algoritmo de Grover, el nivel de seguridad del AES-128 bajaría a $\sqrt{2^{128}} = 2^{64}$, volviéndose vulnerable a ataques por fuerza bruta. Similarmente, el nivel de seguridad del AES-196 bajaría de 2^{196} a 2^{98} mientras que el del AES-256 bajaría a 2^{128}, siendo este último tamaño de clave todavía resistente a ataques por fuerza bruta.

Como puede intuirse, existe una preocupación real de que dentro de pocos años la computación cuántica vuelva inseguros los cifrados más utlizados actualmente. Debido a ello, en 2016 el NIST lanzó convocatorias para la adopción de algoritmos estandarizados que sean resistentes a la computación cuántica. En 2022, se seleccionaron cuatro algoritmos [10], de los cuales tres ya han sido estandarizados en 2023 y se espera que se seleccionen y estandaricen más algoritmos para su uso y estudio. De esta manera, estamos viviendo cómo se inaugura una nueva etapa de la criptografía postcuántica.

Agradecimientos

El autor agradece a los tres revisores anónimos por tomarse el tiempo de revisar y señalar varios errores que había en la primera versión del artículo, así como por sugerir la inclusión de referencias adicionales que, sin duda, han dado mayor realce a los temas aquí discutidos.

Referencias

[1] E. Barker and N. Mouha, "Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher," *Special Publication (NIST SP) 800-67 Rev. 2.*, pp. 1–25, 2017. DOI: 10.6028/NIST.SP.800-67r2. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-67r2

[2] E. Barker and A. Roginsky, "Transitioning the Use of Cryptographic Algorithms and Key Lengths," *Special Publication (NIST SP) 800-131A Rev. 2.*, pp. 1–27, 2019. DOI: 10.6028/NIST.SP.800-131Ar2. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-131Ar2

[3] D. Coppersmith, "The Data Encryption Standard (DES) and its strength against attacks," *IBM Journal of Research and Development*, vol. 38, no. 3, pp. 243–250, 1994. DOI: 10.1147/rd.383.0243. [Online]. Available: https://doi.org/10.1147/rd.383.0243

[4] J. A. de la Peña, *Álgebra en todas partes.* Fondo de Cultura Económica, 1999.

[5] J. v. z. Gathen, *CryptoSchool*, 1st ed. Springer Berlin, Heidelberg, 2016. [Online]. Available: https://doi.org/10.1007/978-3-662-48425-8

[6] L. C. Grove, *Algebra.* Academic Press, 1983.

[7] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing.* New York, NY, USA: Association for Computing Machinery, 1996. DOI: 10.1145/237814.237866 p. 212–219. [Online]. Available: https://doi.org/10.1145/237814.237866

[8] T. Kelly, "The myth of the skytale," *Cryptologia*, vol. 22, no. 3, pp. 244–260, 1998. DOI: 10.1080/0161-119891886902. [Online]. Available: https://doi.org/10.1080/0161-119891886902

[9] G. Morales-Luna, "Sobre el telegrama zimmerman," *CINVESTAV-IPN*, 2016, 21 de diciembre. [Online]. Available: https://delta.cs.cinvestav.mx/~gmorales/

[10] NIST, "Nist announces first four quantum-resistant cryptographic algorithms," *NIST news*, 2022. [Online]. Available: https://www.nist.gov/news-events/news/2022/07/nist-announces-first-four-quantum-resistant-cryptographic-algorithms

[11] I. M. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed. Jhon Wiley, 1991.

[12] C. Paar, J. Pelzl, and T. Güneysu, *Understanding Cryptography*, 2nd ed. Springer Berlin, Heidelberg, 2024. [Online]. Available: https://doi.org/10.1007/978-3-662-69007-9

ISSN 2448 5365
Universidad de Sonora

[13] F. Russel, *Information Gathering in Classical Greece*. University of Michigan Press, 1999.

Cómo citar este artículo: E. Velasco-Barrera, "Criptografía de los cifrados de bloque", Sahuarus. Revista Electrónica de Matemática, vol. 8, no. 1, pp. 45 – 82, 2024. https://doi.org/10.36788/sah.v8i1.100