ACADEMIA DE CIENCIAS SOCIALES Y DEL MEDIO AMBIENTE DE ANDALUCÍA

Reproductividad en la investigación. Aportes de la Estadística Computacional

Discurso de ingreso del

Ilmo. Sr. D. Andrés González Carmona

Granada 11 de octubre de 2024

Agradecimientos

Excmo. Sr. D. José Carlos Gómez Villamandos, Consejero de Universidad, Investigación e Innovación de la Junta de Andalucía

Excmo. Sr. D. Antonio Pascual Acosta, Presidente de la Academia de Ciencias Sociales y del Medio Ambiente de Andalucía,

Sr. D. Enrique Herrera Viedma, Rector sustituto de la Universidad de Granada.

Excmo. Sr. D. Francisco González Lodeiro, Presidente de la Academia de Ciencias Matemáticas, Físico-Químicas y Naturales de Granada,

Ilmo. Sr. D. Antonio Miguel Posadas Chinchilla, Secretario General de Investigación e Innovación

Ilmo. Sr. D. Santiago Herrero León, Secretario Canciller de la Academia

Sra. D.ª María José Martín Gómez, Delegada Territorial de Desarrollo Educativo y Formación Profesional y de Universidad, Investigación e Innovación en Granada.

Sra. D.ª María Rocío Linares Morera, Secretaria General de Universidad, Investigación e Innovación en Granada.

Ilmas. Sras. Académicas, Ilmos. Sres. Académicos,

Señoras y señores, amigos todos

Debo agradecer en primer lugar a la Academia de Ciencias Sociales y del Medio Ambiente de Andalucía, el honor de haberme elegido como académico de número de la misma, agradecimiento que dedico especialmente a su presidente, Antonio Pascual Acosta, y a mi madrina de ingreso, Pilar Aranda Ramírez.

Por supuesto, hago extensivo este agradecimiento a las personas que hoy presiden este acto. Y con especial cariño al Rector de la Universidad de Granada, Pedro Mercado Pacheco, que no puede acompañarnos hoy como era su deseo y el mío, deseándole su pronta recuperación.

Si tuviera que agradecer una a una a todas las personas que me han permitido desarrollarme como persona, investigador y docente, la lista sería muy, muy larga, excediendo con mucho la duración del propio discurso. Permítanme que lo haga solo a algunos, y entiendan el resto que esta duración es la razón de que no los nombre.

En primer lugar, la fortuna me ha acompañado al haberme puesto excelentes jefes y maestros que me han permitido avanzar por el camino correcto: D. Alfonso Guiraúm Martín, mi querido y admirado maestro D. Ramón Gutiérrez Jáimez, D. Antonio Pascual Acosta, D. Gerardo Pardo Sánchez, D. Francisco González Lodeiro, D. Enrique Hita Villaverde y D.ª Pilar Aranda Ramírez. Tengo con ellos la ventaja de que, al no poder alcanzarlos, me sirven siempre de faro y guía.

Y también todos aquellos con los que más directamente he trabajado, Juan Antonio Maldonado, Pedro Antonio García, Fernando Vera, Lola Martínez, Ramón Gutiérrez, Yolanda Román, Mariluz Gámiz y Rocío Raya, de los que también tanto he aprendido.

Y no olvido a aquellos que han viajado desde otros lugares para acompañarme, Juan Antonio Caballero y Rafaela Osuna.

Aunque los he dejado casi para el final, tengo que agradecer a mi mujer, Yolanda, y a mis hijos, Andrés, Silvia, Juan,

Laura, Claudia y Antonio que siempre son mi apoyo en todas las circunstancias de la vida.

Estoy además agradecido al Presidente de la Academia por haber nombrado, por primera vez en la historia de la misma, a una mujer, la que ha sido primera rectora de la universidad de Granada, para ejercer de madrina en este acto.

Y como dijo ella en el acto de su ingreso en esta Academia, (Aranda Ramírez, 2023) *Espero poder estar a la altura de los valores que representan y poder responder adecuadamente a todo lo que se me demande*.

> *Especialmente porque este es el modus operandi de la ciencia: nada se da por bueno hasta ser validado por otros; toda verdad es provisional, un delicado equilibrio basado en el consenso, también provisional, de toda una comunidad dedicada a analizar los mejores datos de cada momento. Ante resultados especialmente novedosos y rompedores, como los que parecían avalar tan rotundamente la teoría de Linde, la reacción lógica era la sospecha.*
>
> La lira desafinada de Pitágoras (Martín Castro, 2022)

El problema

Existe una condición que debe cumplir cualquier investigación para que puedan aceptarse las conclusiones de la misma. Esta condición es que pueda reproducirse, habitualmente por otros investigadores, o incluso por los mismos.

Ha sido común que investigaciones consideradas válidas con determinadas herramientas estadísticas, al ser analizadas con técnicas estadísticas posteriores hayan hecho sospechar de esas conclusiones.

Hace dos años fue el bicentenario del nacimiento de Gregor Mendel. Todos conocemos los resultados de sus investigaciones con guisantes que condujeron a los conceptos de fenotipo y genotipo. Su trabajo no fue valorado cuando lo presentó en 1865 ni cuando lo publicó al año siguiente como (Mendel, 1866) *Experimentos sobre hibridación de plantas*,

y fue completamente ignorado. Puede encontrarse la traducción al inglés realizada por Bateson en 1901 (Bateson, 1868, 1901, 1966) que para mí es más conveniente de usar dado mi desconocimiento del alemán. En el trabajo, se incluyen unas pocas, concretamente tres, tablas y algunas citas con porcentajes, pero no los datos originales. Se incluyen frases como *From a larger number of plants of the same variety only the most vigorous were chosen for fertilization* (De un mayor número de plantas de la misma variedad, solo se eligieron las más vigorosas para la fertilización) que no se corresponden con una selección al azar.

Se *volvieron a descubrir,* independientemente, por tres investigadores al final del siglo, en 1900, que impulsaron el reconocimiento a Mendel.

El mérito de Mendel fue darse cuenta de que en sus experimentos las variantes de los fenotipos, lo que él llamó caracteres, aparecían en proporciones simples. Y hoy en día sus conclusiones se aceptan sin dudarlo.

Pero sus datos parecieron, después, demasiado buenos. Sir R. A. Fisher, en su trabajo clásico del año 1936, *Has Mendel's work been rediscovered?*, (Fisher, 1936) analizó dichos resultados y cuestionó que fuesen reales. Para ello, formuló el análisis del trabajo como una reconstrucción de cómo se pudieron realizar los experimentos.

This objection is not so strong as it might seem, since it can be shown that Mendel left uncounted, or at least unpublished, far more material than appears in his paper (Esta objeción no es tan fuerte como podría parecer, ya que se puede demostrar que Mendel dejó sin contabilizar, o al menos sin publicar, mucho más material del que aparece en su artículo.)

Mendel was an experienced and successful teacher, and might well have adopted a style of presentation suitable for the lecture-room without feeling under any obligation to complicate his story by unessential details. (Mendel era un profesor experimentado y exitoso, y bien podría haber adop-

tado un estilo de presentación adecuado para el aula sin sentirse obligado a complicar su historia con detalles no esenciales.)

The style of didactic presentation, with its conventional simplifications, represents, as is well known, a tradition far more ancient among scientific writers than the more literal narratives in which experiments are now habitually presented.(El estilo de presentación didáctica, con sus simplificaciones convencionales, representa, como es bien sabido, una tradición mucho más antigua entre los escritores científicos que las narraciones más literales en las que ahora se presentan habitualmente los experimentos.)

En varios apartados especifica que en los datos *… the deviation is less than the standard error of random sampling* (…la desviación es menor que el error estándar del muestreo aleatorio.)

El análisis de Fisher fue implacable. Si hubiera sido el *referee* del trabajo, posiblemente las leyes de la herencia hubiesen tardado bastante en descubrirse. Aunque él también cometió fallos en su análisis. En 1955 él mismo indicó que todas las fechas que proponía en la reconstrucción, debían ser un año menos.

Pero hay casos más claros, sin tener que recurrir a los datos originales, como la conocida Ley de Boyle-Mariotte, llamada así por los dos investigadores que la publicaron en 1662 y 1679 respectivamente —en realidad Mariotte amplió la de Boyle puesto que usó distintas temperaturas—, con solo 17 años de diferencia, que relaciona Presión, Volumen y Temperatura absoluta de un gas *ideal*,

$$\frac{PV}{T} = K$$

Si vamos a la obra original del primero, (Boyle, 1662), disponible en la Biblioteca Virtual del Patrimonio Bibliográfico, https://bvpb.mcu.es/en/consulta/registro.cmd?id=406806, podemos comprobar que, en las 144 páginas del libro, solo aparecen dos pequeñas tablas de media página cada una. El

resto se dedica a defender la propuesta contra las opiniones reinantes en aquel momento.

Es de justicia reconocer el mérito de derivar una regla tan general basándose en tan pocos datos, sin que aún se hubieran descubierto las técnicas de regresión.

Posteriormente, con nuevos datos, el que llegó a ser Premio Nobel, el matemático y físico J. Van der Waals, en su tesis de 1873, llegó a una ecuación donde eliminó del volumen total el que ocupan las propias partículas y sumó a la presión la fuerza de atracción de dichas partículas,

$$\left(P + \frac{n^2 a}{V^2}\right)(V - nb) = nRT$$

Ecuación que se llamó de los gases *reales*. ¿Tal vez, al repetir la experiencia, los datos no eran *ideales*? A partir de entonces se llamó *ideales* a los gases en condiciones de baja densidad.

Sin menoscabar la importancia de estos investigadores, es bastante probable que estuvieran convencidos de lo que querían demostrar y *forzaron la mano* un poco en los datos.

Y si los datos no son accesibles ¿podemos creerlos? ¿Creeríamos en la veracidad de los datos del CERN sobre el bosón de Higgs y sus métodos de análisis, y por tanto de sus conclusiones? ¿Haríamos lo mismo con las conclusiones de las encuestas de opinión?

En el primer caso hay **muchas** personas de prestigio participando en el experimento y todo el proceso es público, y además prácticamente irrepetible.

Del segundo caso no suele existir información que garantice la exactitud de las conclusiones, ni que estas puedan aplicarse posteriormente en casos similares; son informaciones del momento. No puedo resistirme a señalar como ejemplo de encuestas en cuanto a su credibilidad las que realiza actualmente el CIS.

Antonio Pascual me enseñó que, si los resultados de una encuesta no incluyen, al menos, la ficha técnica de la misma, creerlos es tanto o más arriesgado que jugar a la lotería.

¿Y qué decir si los datos han sido directamente inventados, como puede realizarse fácilmente, sin más que pedírselos a un programa de Inteligencia Artificial? En Nature (Naddaf, 2023) se publicó el 22 de noviembre de 2023, un análisis de un artículo publicado en JAMA Ofthalmology (Andrea Taloni; Vincenzo Scorcia; Giuseppe Giannaccare, 2023) de 9 de noviembre en que los autores utilizaron GPT-4 para generar un conjunto de datos e, incluso, **analizarlo** mediante *Advanced Data Analysis* (ADA), un modelo que ejecuta *Python* dentro de GPT-4, para contrastar una determinada hipótesis. Procedieron a crear con GPT-4 un conjunto de datos falsos y utilizarlo como si fuesen reales. Como indica uno de los autores *Our aim was to highlight that, in a few minutes, you can create a data set that is not supported by real original data* (Nuestro objetivo era destacar que, en pocos minutos, se puede crear un conjunto de datos que no está respaldado por datos originales reales).

Los autores pidieron a GPT-4 que fabricara unos datos que dieran soporte a que un procedimiento de técnica de operación ocular era mejor que otro en cuanto a visión tras la operación. La IA generó unos datos con 160 hombres y 140 mujeres y los analizó, con resultados que indicaban que era cierta la hipótesis planteada.

Al analizar los datos por un equipo independiente, se encontraron errores en los mismos, como nombres que no se correspondían con el sexo, falta de correlación entre los valores de visión pre y post operatorias, agrupación inusual de los datos de edades y un número insólito de participantes con edades terminadas en 7 y 8.

Estos *fallos* en los datos, posiblemente se deben a que, en el mundo real, donde se entrena la IA, no suelen encontrarse descripciones detalladas de cómo generar e identificar los datos falsos, y posiblemente la IA generó los datos por agregación de generaciones unidimensionales de cada variable

implicada. Pero si aprende algo más, será capaz de crear unos datos *correctos*, sin *fallos*.

ChatGPT

El 13 de noviembre de 2023 pedí a ChatGPT (versión gratuita) **escribe un artículo sobre reproducibilidad.** El 25 de agosto de 2024 volví a darle esta orden y me respondió con una versión muy mejorada de la anterior que reproduzco a continuación:

La Importancia de la Reproducibilidad en la Ciencia Moderna

La reproducibilidad es un pilar fundamental de la ciencia moderna. Se refiere a la capacidad de repetir un experimento o estudio bajo las mismas condiciones originales y obtener resultados consistentes. Este concepto no solo es crucial para validar los hallazgos científicos, sino que también es esencial para el progreso acumulativo del conocimiento. En este artículo, exploraremos la importancia de la reproducibilidad, los desafíos actuales que enfrenta y las posibles soluciones para mejorar la calidad de la investigación científica.

¿Por Qué Es Importante la Reproducibilidad?

La reproducibilidad es esencial porque garantiza que los resultados científicos no son producto del azar, error o sesgo, sino que reflejan verdaderamente la realidad. Cuando un experimento es reproducible, otros científicos pueden confirmar la validez de los resultados, lo que fortalece la confianza en las conclusiones extraídas. Además, la reproducibilidad permite que otros investigadores puedan construir sobre el trabajo existente, avanzando en el conocimiento científico de manera acumulativa.

Desafíos en la Reproducibilidad

A pesar de su importancia, la reproducibilidad ha enfrentado serios desafíos en la investigación científica moderna. Algunos de los problemas más comunes incluyen:

1. **Falta de Transparencia**: Muchas veces, los detalles completos de los métodos y datos utilizados en un estudio no se comparten adecuadamente. Esto dificulta que otros investigadores puedan replicar los experimentos.
2. **Sesgo de Publicación**: Existe una tendencia en las revistas científicas a publicar solo resultados positivos o novedosos, dejando de lado estudios que no muestran efectos significativos o que contradicen hallazgos previos. Este sesgo crea una visión distorsionada del conocimiento científico.
3. **Errores Metodológicos**: A veces, los estudios no son reproducibles debido a errores en el diseño experimental, análisis estadístico o interpretación de los datos. Estos errores pueden pasar desapercibidos en la revisión por pares, especialmente si los revisores no tienen acceso a los datos originales.

4. **Recursos Limitados**: La reproducción de estudios puede ser costosa y requerir mucho tiempo, lo que desincentiva a muchos investigadores a realizar estos esfuerzos.

Soluciones para Mejorar la Reproducibilidad

A fin de abordar los desafíos mencionados, la comunidad científica ha propuesto y adoptado diversas estrategias:

1. **Acceso Abierto a Datos y Métodos**: Compartir abiertamente los datos y los procedimientos experimentales permite a otros investigadores replicar los estudios con mayor precisión. Plataformas como Open Science Framework (OSF) y arXiv facilitan esta práctica.
2. **Prerregistro de Estudios**: El prerregistro consiste en detallar el diseño del estudio, las hipótesis y los métodos antes de realizar el experimento. Esto ayuda a reducir el sesgo de publicación y mejora la transparencia en la investigación.
3. **Fomento de la Replicación**: Las instituciones y revistas científicas deben valorar y promover los estudios de replicación, reconociendo su importancia para la validación científica.
4. **Revisión por Pares Más Rigurosa**: Mejorar los procesos de revisión por pares, incluyendo la verificación de datos y análisis, puede ayudar a identificar errores metodológicos antes de la publicación.
5. **Educación y Capacitación**: Formar a los científicos en buenas prácticas de investigación, incluyendo estadística y diseño experimental robusto, es esencial para mejorar la reproducibilidad.

Conclusión

La reproducibilidad es una piedra angular de la ciencia, esencial para asegurar que el conocimiento generado sea confiable y útil. Aunque enfrenta numerosos desafíos, la comunidad científica está tomando medidas para mejorar la reproducibilidad a través de una mayor transparencia, el fomento de la replicación y la mejora de los procesos de revisión. Estas acciones no solo fortalecerán la integridad científica, sino que también garantizarán que la ciencia continúe avanzando de manera robusta y confiable. En última instancia, mejorar la reproducibilidad no solo beneficia a la comunidad científica, sino que también refuerza la confianza del público en la ciencia.

Y esta mejora nos indica el problema subyacente: cada vez la IA tiene acceso a más datos y posiblemente también llegará a aprender cómo falsificar los datos de forma más eficiente, igual que ahora es capaz de transformar un texto para que no se encuentren señales de plagio por los programas detectores del mismo.

Propuestas de solución

Modern Data Science with R

En el capítulo 8, *Data science ethics*, de *Modern Data Science with R* (Baumer, Kaplan, & Horton, 2021) se realiza un análisis muy completo la reproductividad y de los problemas que su falta produce. Se indican varios casos de análisis de datos con propuestas erróneas, y en el apartado 8.8 se especifican muy bien algunas de las causas que pueden llevar a análisis erróneos y los criterios que deben cumplirse para intentar evitarlas.

Indican los autores que incluso los investigadores originales son incapaces de reproducir sus propios resultados con posterioridad. Este fallo se produce a menudo cuando se usan programas gestionados mediante menú, que no dejan constancia del proceso. Si además se usan procesadores de texto para describir los resultados, reescribiendo los datos e insertando gráficos, la probabilidad tanto de cometer errores como de no poder reproducir los resultados aumenta considerablemente.

Y hay razones externas –como las que he indicado antes– pero también internas para que las sigamos. En multitud de ocasiones, los propios autores de una investigación son incapaces de repetir la investigación. Que esto ocurriera en publicaciones de hace 30 años es explicable, ya que los procedimientos eran de *corta y pega* o incluso de volver a escribir los datos.

Y me incluyo en los que trabajábamos así porque, por ejemplo, yo colaboré en un equipo dirigido por el Dr. Vázquez Mata realizando determinados análisis estadísticos que se publicaron posteriormente (Vázquez Mata, 1992), que han sido bastante citados en su campo, y que si hoy hubiera de explicar a alguien, me sería imposible hacerlo. No dispongo de los datos (y nunca se publicaron), los análisis se hicieron con BMDP (que ya no existe) pero además ni se publicaron los programas ni los procedimientos para llegar a los resultados, y no los conservo. Era lo habitual, debo reconocerlo.

Describen en ese apartado como *análisis reproducible* la práctica de grabar todos y cada uno de los pasos, aun cuando parezcan triviales, y que incluye:

- Datos: Todos los archivos de datos en la forma en que fueron creados
- Metadatos: Información necesaria para comprender los datos
- Órdenes: Programas necesarios para extraer, transformar (piénsese en el *Big Data* y la necesidad de depurar los datos), ejecutar los análisis, ajustar modelos, generar gráficos, etc.
- Índice: Un índice que relacione las salidas y los resultados del informe

Pero no solo se plantea este problema como la verificación por personas externas, sino que los análisis que **nosotros** realizamos en un trabajo real, conllevan una serie de procesos de depuración, de análisis y de uso de medios que si no están debidamente documentados no seríamos capaces de repetir.

Proyecto TIER

El proyecto TIER, https://www.projecttier.org/, siglas de Teaching Integrity in Empirical Research, surgió en Reino Unido con el objetivo de promover y conseguir un cambio en las normas profesionales que tienen que ver con la transparencia y la reproductividad de la investigación empírica en las ciencias sociales, marcando el principio de que en la investigación que contiene datos y análisis estadísticos de los mismos, la transparencia y la reproductividad deben considerarse imprescindibles y realizarse de modo automático, y que nadie debe considerar que una investigación es creíble si no los incluye.

Los socios del proyecto son UK Reproducibility Network, Atlanta University Center (AUC) Data Science Initiative, FORRT (Framework for Open and Reproducible Research Training), ReplicationWiki una base de datos de estudios empíricos en ciencias sociales de los cuales están disponibles los materiales para la replicación, Berkeley Initiative for Transparency in the Social Sciences, Society for the Improvement of Psychological Science, Center for Open Science, y Collaborative Replications and Education Project.

TIER propone un protocolo, a día de hoy el 4.0, véase https://www.projecttier.org/tier-protocol/protocol-4-0/,aplicable a los proyectos que contienen datos experimentales, de tal modo que se exige que la documentación contenga todos los datos, los programas para analizarlos y toda la información necesaria para hacerlo, de modo que puedan reproducirse los cálculos para obtener los resultados que se muestren en el informe del proyecto.

Es un protocolo muy detallado que sirve como guía para los investigadores. Para el análisis computacional, se pide que los programas que se usen puedan admitir programación escrita.

Tal como allí se describe, se intenta conseguir las siguientes cualidades de reproductividad:

Suficiencia: La documentación de un proyecto debe contener todo lo necesario para reproducir los resultados.

De principio a fin: la documentación debe incluir programas que reproduzcan todos los cálculos necesarios para (i) los pasos de procesamiento necesarios para preparar los datos para el análisis, y (ii) el análisis o los procedimientos realizados con los datos procesados que generan los resultados.

Portabilidad: cualquier usuario debería poder ejecutar los programas en su propio ordenador o espacio de trabajo.

Reproducibilidad (casi) con un solo clic: reproducir todos los cálculos necesarios para generar los resultados de

un estudio no debería requerir (casi) nada más que ejecutar un único programa maestro. Decimos "casi" porque hay un paso preliminar que el usuario debe realizar antes de ejecutar el programa maestro, es decir, configurar el directorio de trabajo en la carpeta elegida.

American Statistical Association

La *American Statistical Association*, ASA, advierte de la necesidad de educar en el uso de análisis reproducibles y la necesidad del uso de herramientas como R,y *knitr* y similares. Además, por ejemplo, marcan las directrices para la enseñanza de la Estadística en EE.UU. (GAISE College Report ASA Revision Committee, 2017).

El *Journal of the American Statistical Association*, JASA, es el lugar de publicación de artículos científicos concretos. En la Guía de estilo (JASA, 2024) se indica que para producir un artículo con R/LaTeX, se utiliza una plantilla, jss-article-rnw.zip, y todo el código fuente se incluye en un archivo de la misma, article.Rnw, que tras ser procesado en R con la función knit del libro knitr producirá el archivo article.tex que puede tratarse con LaTeX.

Por otra parte, las opiniones de ASA suelen ser publicadas en la revista *The American Statistician*, que reciben todos los miembros de la asociación, y que por tanto gozan de gran difusión directa.

Marwick, Boettiger y Mullen, en su artículo *Packaging Data Analytical Work Reproducibly Using R (and Friends)* (Marwick, Boettiger, & Mullen, 2018) marcan el papel fundamental que desempeñan los ordenadores en el proceso de investigación, así como la necesidad de que la investigación sea reproducible, especialmente tras los fallos en conseguir esa reproducción en estudios biomédicos y en otros psicológicos (los autores pertenecen a una Facultad de Medicina). Confiesan que el propósito fundamental de este trabajo es

mostrar cómo el programa R es el apropiado para conseguir ese objetivo y lo hacen recurriendo a diversos ejemplos del mundo real. Destacan cómo R fomenta la creación de *documentos dinámicos*, y con knitr se consigue que los resultados sean reproducibles. Si, además de disponer de los datos públicamente, las aportaciones de programación en R se estructuran en forma de un libro de R y este se publica en CRAN, se beneficia de los mecanismos de control de calidad que allí se imponen. Abundan además en la necesidad de aplicar una estructura lógica al uso de subdirectorios dentro del proyecto. En conclusión, no solo muestran las ventajas, sino que marcan cómo llevarlas a la práctica.

Hofmann, Wickham y Cook, en su artículo *The 2013 Data Expo of the American Statistical Association* (Hofmann, Wickham, & Cook, 2019), analizan la competición del año 2013 que organiza ASA anualmente, con un paréntesis entre 1997 y 2006) desde 1983. En ellas se suministra a los participantes un conjunto de datos con el objetivo de que procedan a realizar un análisis estadístico del mismo, al que pueden incorporar otros datos adicionales.
La principal diferencia con ediciones anteriores es que los análisis presentados, algunos de los cuales fueron publicados en *The American Statistician*, llevaban asociados los códigos de ordenador para su realización, que fueron publicados añadiendo a los artículos la dirección del repositorio en que pueden aún verse: https://github.com/COSTDataExpo2013, constituyendo el primer experimento en investigación reproducible. La idea subyacente fue que, a diferencia de las ocasiones anteriores, los lectores de los trabajos no vieran solamente los resultados, sino que fueran capaces de comprender cómo fueron creados. Y ya, en tres de ellos se utilizó knitr.

Xiong y Cribben, en su artículo *The State of Play of Reproducibility in Statistics: An Empirical Analysis*, (Xiong & Cribben, 2022), destacan que los estudios cuyos resultados no pueden reproducirse o replicarse por la comunidad científica deben considerarse con precaución. Entiende por replicar, el hecho de poder repetir todo el experimento, en tanto que la

reproductividad parte de los datos aportados. Puesto que es muy difícil, por diferentes cuestiones, la replicación, entiende como estándar la reproductividad. Para evaluar la reproductividad en Estadística intentó reproducir los resultados de artículos en siete revistas de prestigio: *Annals of Applied Statistics, Biometrics, Bio-statistics, the Journal of Computational and Graphical Statistics, the Journal of the American Statistical Association, the Journal of the Royal Statistical Society: Series C, y Statistics in Medicine* durante el periodo 2010-2021. Los artículos revisados fueron aquellos cuyos datos correspondían a Resonancia magnética funcional, en total 93 y solo pudo reproducirlos en 14 de ellos, y ello por diferentes causas. Ello lleva a la conclusión de que, sin los datos brutos, las instrucciones de limpieza de los mismos, sin instrucciones detalladas, etc. no es posible la reproducción.
Muchas de estas revistas están en proceso de cambio, o han cambiado, las instrucciones a los autores en cuanto a la disponibilidad de los datos y de los programas de ordenador.
El autor se extiende además en otros campos de ciencias experimentales en los que encuentra fallos similares, y destaca el uso de R y de knitr.
Finaliza haciendo propuestas a las diferentes revistas de qué condiciones deben exigir a los autores para poder realizar las publicaciones, así como que las propias revistas las cumplan.
Además, como conclusión indica que, en el campo de la Estadística, aunque algunas revistas tienen buenas políticas en cuanto a reproductividad, en general el tratamiento es pobre y está necesitado de mejoras.

Entre las muchas citas que realiza, hay una muy breve de *Rigor mortis. Cómo la ciencia chapucera crea curas inútiles, aplasta la esperanza y desperdicia miles de millones* (Harris, 2017) que ha sido un superventas y que trata, ampliamente, diez casos de esas *chapuzas*. Su lectura es muy recomendable. Y aunque allí no lo cita, existe un libro de características antagónicas, en este caso con falsos artículos, que también es de lectura muy recomendable: Demoliendo papers (Golombek, 2005, 2014).

National Academies of Sciences, Engineering, and Medicine

El tema queda reflejado también en varias publicaciones de la *National Academies of Sciences, Engineering, and Medicine* de Estados Unidos.

En el informe *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific* (National Academies of Sciences, Engineering, and Medicine, 2016) ya menciona un estudio de 2011 en el que encontraban que el 65% de los estudios médicos eran inconsistentes cuando se intentaban reproducir. También, un análisis publicado en *Nature* en 2012 encontró que, de 53 artículos sobre cáncer, 47 no se podían reproducir. Así, la poca reproductividad de los resultados publicados por laboratorios académicos, limita la capacidad de transferir los resultados a la práctica médica.
Se hacen eco de la confusión en cuanto al uso de la palabra reproductividad, proponiendo diversos conceptos:

- Reproductividad, como la capacidad de un investigador de duplicar los resultados de un estudio utilizando los mismos materiales.
- Repetibilidad, como la capacidad de ver los datos, ejecutar los programas y seguir todos los pasos descritos en la publicación.
- Replicabilidad, como la capacidad de duplicar los resultados, pero con nuevos datos. Suele fallar porque lo que en un estudio fue estadísticamente significativo, en otro puede no serlo.
- Robustez, como la resistencia del experimento ante (pequeños) cambios en las condiciones y supuestos.
- Reproductividad estadística, en cuanto a cómo las técnicas y métodos estadísticos contribuyen a la poder creer que un resultado es reproducible y medir la reproductividad.
- Reproductividad computacional, producido por las diferencias computacionales, desde usar una hoja de cálculo en un ordenador personal hasta el uso de sistemas avanzados con grandes cantidades de datos y programas.

Como consecuencia proponen compartir mucho más los materiales de investigación, y potenciar tanto la formación como el entrenamiento en reproductividad, en todos los niveles, desde estudiantes hasta investigadores senior.

En el informe *Reproducibility and Replicability in Science* (National Academies of Sciences, Engineering, and Medicine, 2019) indican que si se encuentran resultados similares, tienden a confirmar la veracidad de la conclusión de un trabajo, y, por la misma regla, los fallos repetidos ponen las conclusiones originales en duda.
Cuando la NSF estaba en trámites para que NASEM realizara un estudio sobre reproductividad y replicabilidad en las Ciencias Sociales, el Congreso de EE.EE. promulgó una disposición legal para expandir el objeto del estudio a toda la ciencia e ingeniería.
En este informe se definen de nuevo los conceptos:

- **Reproductividad**, para indicar reproductividad computacional, esto es, si se obtienen resultados computacionales consistentes utilizando los mismos datos, pasos en ordenador, métodos y condiciones de análisis
- **Replicabilidad**, si se obtienen resultados consistentes en otros experimentos similares

Y justifican el nuevo concepto de reproductividad en el cambio de los soportes informáticos y la extensión de su uso, especialmente en este siglo.
Ahora bien, el uso de *software* propietario y que los datos no sean públicos, va en contra de la reproductividad. Destacan que, aunque se están desarrollando herramientas que capturan todas las decisiones que toman los analistas y las incluyen en el informe, estas solo las usa una minoría de investigadores. Además, no existe un procedimiento homogéneo para que las revistas, las instituciones académicas y las gubernamentales conserven los datos y los pongan a disposición de otros investigadores.
Destacan que un tipo de herramienta de investigación, la Inferencia Estadística, ha tenido un papel muy destacado en las sesiones de trabajo, debido a su uso incorrecto, como por ejemplo el conocido uso del Valor P (*P-value*) como umbral

para discriminar una publicación, con el consiguiente sesgo en las publicaciones.

En el informe *Enhancing Scientific Reproducibility in Biomedical Research Through Transparent Reporting: Proceedings of a Workshop* (National Academies of Sciences, Engineering, and Medicine, 2020) ponen claro que para que los logros experimentales puedan ser aceptados por la comunidad científica, deben ser verificados por otros investigadores. También manifiestan que ha habido un número creciente de estudios en biomedicina que no han podido ser replicados y que las técnicas de meta análisis muestran la misma tendencia. Admiten como definiciones de Reproductividad y Replicabilidad las establecidas en el informe antes citado, y señalan que para mejorar en la Reproductividad deben formar parte muchos actores, no solo los investigadores, sino también los que proporcionan fondos, los que dictan las políticas, los editores de las revistas, los organizadores de conferencias, los centros de enseñanza, las sociedades científicas, e incluso los periodistas.

En el informe *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies* (National Academies of Sciences, Engineering, and Medicine, 2022) —La transparencia en la información estadística del Centro Nacional de Estadística de Ciencias e Ingeniería y de todas las agencias estadísticas federales— se plantea el problema de que en EE.UU. hay trece agencias estadísticas principales, así como unidades y programas en muchas otras agencias, que producen estadísticas que son clave en muchas áreas. Pero las estadísticas oficiales suelen ser el resultado de métodos complejos de recopilación, procesamiento y estimación de datos. Estos métodos pueden resultar difíciles de documentar para los organismos que las generan y de comprender por los usuarios, ya que no son intrínsecamente transparentes. Los usuarios de los datos no tienen recursos para verificar la precisión de la información estadística, y por tanto deben confiar en la reputación de las agencias. Las agencias crean y mantienen el respeto y la confianza mediante compromisos públicos claros con la práctica profesional y la

transparencia en todo lo que hacen, incluyendo la información sobre las fortalezas y debilidades de los datos que ofrecen.
Por tanto, destacan la importancia de la transparencia y el concepto estrechamente relacionado de reproducibilidad, y Pide a las agencias que se esfuercen por lograr transparencia en su trabajo.
Al igual que en lo que se pide en los informes científicos, se marcan cuatro preguntas a las que responder:

- ¿Los datos y el análisis se presentan con suficiente transparencia y claridad para que se puedan comprobar los resultados?
- Si se marca esta casilla, ¿los datos y el análisis ofrecidos para respaldar el resultado realmente respaldan ese resultado?
- Si se demuestra que los datos y el análisis respaldan el resultado original, ¿es posible encontrar el resultado informado nuevamente en el contexto de estudio específico investigado?
- Finalmente, ¿puede el resultado informado o la inferencia extraída encontrarse nuevamente en un conjunto más amplio de contextos de estudio?

Además, los usuarios de los datos deben poder localizar fácilmente información suficiente sobre programas estadísticos actuales y pasados, conjuntos de estimaciones y conjuntos de datos de entrada, para estar al tanto de su existencia y evaluar su calidad y su idoneidad para su uso.

Recomendaciones similares existen en Eurostat.

En Psicología

El interés en este campo también es intenso. En (Obels, Lakens, Coles, Gottfried, & Green, 2020) ponen de manifiesto que existe en la actualidad soporte informático que permite con gran facilidad compartir los datos de una investigación y los programas utilizados para su análisis, lo que

permite que a posteriori puedan ser de nuevo utilizados. Destacan que, en su campo, la mayoría de investigadores no tienen formación en reproductividad por lo que, para evaluar el estado de la cuestión en Psicología, evaluaron estas capacidades en los artículos de su campo, desde 2014 a 2018 e intentaron reproducir los resultados computacionales de cada artículo. Se centraron en los *Registered Reports*, en los cuales es necesario acompañar los datos y los programas, o sea, que están en el nivel más alto para la reproductividad, y encontraron 62 artículos que se ajustaban a su campo, de ellos solo 36 tenían completos los datos y los programas, pero solo en 31 funcionaban los programas y solo lograron realizar la reproducción de 21 artículos. Nunca recurrieron a pedir información complementaria a los autores, ya que deseaban la evaluación de lo depositado en los registros.
Destacan en su estudio la diferencia entre usar *software* libre o propietario. Cuando es libre, es inmediato poder ponerse manos a la obra con el análisis, como ocurre con R. Cuando es propietario, hay que adquirir ese *software* o trabajar para poder realizar los análisis con *software* libre exportando los datos, aunque puede que los resultados sean diferentes. Asimismo, en cuanto a los datos, destacan la importancia fundamental de disponer de una descripción completa de los datos: Nombres de las variables, valores de codificación y nombres, etc.

Retraction Watch

Retraction Watch, https://retractionwatch.com/, forma parte de *The Center For Scientific Integrity*, y su finalidad es mantener información sobre artículos que han sido publicados y posteriormente *retirados* por cualquier causa, fundamentalmente por ser erróneos. Dispone de una base de datos, que puede ser utilizada para buscar esos artículos por diferentes conceptos, que en la actualidad contiene más de 50.000 entradas (artículos todos que superaron los criterios de selección de las revistas). Entre otros tiene 93 artículos que han sido escritos por ChatGPT, y 440 retirados sobre COVID-19. Como curiosidad, dispone de un TOP 10 de los artículos más

citados antes de ser retirados. Citaré solamente la revista, el año y el número de esas citas antes y después de ser retirados:

Revista	Año	Antes	Después
Nature	2022	4482	0
Lancet	2021	2250	448
New England Journal of Medicine	2013	1695	869
Nature	2008	2348	0
The Journal of Clinical Hypertension	2008	1929	99
Gynecologic Oncology	2008	1865	56
Lancet	1998	642	1052
Progress in Materials Science	2004	1550	70
Science	2005	232	1303
Plant Journal	2003	895	473

Adviértase el daño que produce en la ciencia el efecto de la **buena reputación** de una revista, viendo la cantidad de citas que se realizan de artículos pese a que han sido retirados de la revista.

Mi opinión

Yo me permito incluir un apartado más: En experimentos clásicos, debe incluir el diseño del experimento y el modo en que se llevó a cabo.

Una de las razones de los fallos es debida al **uso incorrecto** de la Estadística. En los contrastes de hipótesis, Neyman y Pearson (Neyman & Pearson, 1933) establecieron que si hemos de decidir entre dos hipótesis alternativas, de las que no se puede conocer cuál es la cierta, $\mathrm{H_a}$ y H_b, debe establecerse que una de ellas será H_0, la que consideramos que es cierta por nuestro conocimiento previo, y la otra, H_1, que será la elegida si descartamos la que considerábamos cierta. Aquí

hay que destacar que no son intercambiables ambas hipótesis, no es lo mismo tomar como cierto que *todos son inocentes hasta que se demuestre lo contrario* que tomar como cierto que *todos son culpables hasta que se demuestre lo contrario*.

Por tanto, cuando en multitud de ocasiones, encontramos contrastes para decidir si una muestra de una población que sigue la distribución normal, suele darse el error de utilizarla para comenzar a creerlo, cuando en realidad es una prueba para decidir si debemos dejar de creer en dicha hipótesis. Si es que estamos convencidos de que es normal, sin datos previos, debería ser raro que la observación de datos nos decida a dejar de creerlo.

Otra fuente común de error es desconocer que un contraste de hipótesis depende totalmente de las hipótesis del mismo. En pruebas basadas en métodos *empíricos*, como es el caso de la t de Student, la prueba tiene *sentido* incluso aunque no se cumplan las hipótesis, son *robustas* frente a la pérdida de esas condiciones. Pero, cuando se extiende a pruebas como el análisis de la varianza, donde para comprobar la igualdad de medias se usa el cociente de dos varianzas, si no se cumplen las condiciones de los teoremas, se pierde todo el sentido. La presión para realizar pruebas cada vez más complejas, que dan apariencia de mayor rigor científico, hace que se utilicen, sin atender a las condiciones necesarias.

Aún más, unos solos datos, sirven para realizar un contraste único sobre los mismos. Cuando se usan los mismos datos para contrastar la hipótesis de normalidad, y también hipótesis sobre la media, y también sobre la varianza, etc. Los valores del contraste NO SE CONSERVAN. Es necesario recurrir a contraste MÚLTIPLES, que a menudo no están disponibles o requieren de hipótesis diferentes.

La presión por *publicar* fuerza a los investigadores a cometer errores, a modificar *levemente* los datos, y fuerza a los editores que no pueden ya dar largos plazos y encargan la revisión de los trabajos a personas, a menudo recomendadas por los propios autores, que deben realizar sus revisiones en

tiempos muy escasos, de una o dos semanas como mucho, lo que hace que se den por buenos los trabajos sin realizar una revisión concienzuda.

En definitiva, el tema de la reproductividad queda recogido como uno de los aspectos a tener en cuenta en el ejercicio profesional. A partir de ahora me centraré en el **cómo** llevarlo a efecto en el trabajo estadístico.

Aportaciones de la Estadística Computacional a la solución

En el capítulo 4 de *The New Statistics with R. An Introduction for Biologists* (Hector, 2021) se recalca que la investigación debe ser reproducible y lo mismo los análisis estadísticos. Así, se pregunta *Point-and-click statistical software may be quick to start working with, but could you ever repeat the exact sequence of clicks you have made?* (Comenzar a trabajar con un programa estadístico de menú puede ser muy rápido, pero ¿podría alguna vez repetir la secuencia exacta de selecciones que ha realizado?)

Esta pregunta apunta, evidentemente, contra **SPSS** y similares, pero debo destacar que no es totalmente cierto lo que en ella se indica, ya que en cada resultado de **SPSS** se especifica la *SINTAXIS* con la que se ha construido y que, incluso, es posible insertar directamente el resultado en un documento externo, aunque el método no es trivial ni desglosa los apartados.

En el capítulo 7 de *Doing Data Science in R. An Introduction for Social Scientists* (Andrews, 2021) se indica que el fin último es evitar el *corta y pega* en el post-análisis, eliminando así una alta probabilidad de introducir errores.

También indica que **SAS** tiene una posibilidad de realizar análisis reproducibles pero que es muy limitado, como en **SPSS**.

Y ¿de dónde viene la limitación? Especialmente de que la programación que ambos incluyen se limita a la selección de variables, de datos y de técnicas, pero sin la posibilidad de desarrollar nuevos análisis estadísticos. Y, además, que fomentan la utilización mediante el menú. Es algo similar a lo que ocurre entre usar **Word** o usar **LaTeX**. **Word** puede utilizarse introduciendo en el mismo estructuras, pero lo que se fomenta es no usarlas. En **LaTeX** el uso de estructuras es obligatorio.

R (Y Python)

En 1988 se publicó *The New S Language*, llamado coloquialmente *libro azul*, (Becker, Chambers, & Wilks, 1988) que recogía la definición del entorno de programación estadística **S** desarrollado en los laboratorios AT&T que, al ser AT&T un monopolio, por ley era de libre uso y distribución. Al desaparecer el monopolio, el lenguaje fue modificado para adaptarlo a lo que entonces era considerado *de éxito*, el uso mediante menús, surgiendo **S+**, lo que finalmente condujo a su desaparición.

En la solapa del libro *Modelos aditivos generalizados* (Hastie & Tibshirani, 1990) los autores indicaban que habían desarrollado los conceptos allí recogidos utilizando este entorno, que permitía pensar en Estadística, y luego los habían traducido a Fortran para aumentar su rapidez.

Al desaparecer **S+** su nicho fue ocupado por **R** (Ihaka & Gentleman, 1993) que, a partir de un lenguaje experimental no pensado para la Estadística, que volvió a ser de uso y modificación libres, y su desarrollo fue espectacular, posiblemente por ese motivo.

El lenguaje **R** (The R Project for Statistical Computing, s.f.) no adolece de los *defectos* antes mencionados. No existe un menú, salvo libros –lo que en inglés se denominan *packages*– que, so pretexto de facilitar su uso, los introducen; y que me perdonen los responsables de *R Commander* (J, MM, & M, 2024). Recordemos el destino de S+ por introducir menús para elegir opciones. Los libros, en **R**, se utilizan incluyéndolos en un directorio llamado *library* (biblioteca) y por eso los denomino de ese modo, por coherencia.

- Todas las cosas se deben escribir, datos, órdenes, etc.
- Cada objeto es manipulable de modo independiente.
- Cada salida de una orden puede redirigirse a otra orden (*pipe*)
- Existen multitud de libros que permiten realizar una cantidad enorme de tareas distintas. Todo lo que existe

en Estadística susceptible de ser tratado en un ordenador, está implementado en **R**.
- La enorme mayoría (abrumadora) está escrita en código accesible, por lo que los que trabajan en un tema concreto pueden revisar cómo actúa y corregirlo, por lo que a los más torpes nos queda la seguridad de que lo que se produce es correcto (Recordemos aquí algunos casos de **SPSS**, como las varianzas negativas, las correlaciones erróneas, y algunas más en modelos complejos; que no pueden detectarse revisando el código porque esa acción es un delito, al menos en España).
- Y, no es lo menos importante, R es totalmente gratuito e incluso su licencia permite su modificación. El hecho de su gratuidad hace que diferentes organismos lo recomienden ya que permite a investigadores sin recursos para adquirir otros programas realizar los análisis más complejos.

¿Y **Python**? Goza de bastantes virtudes similares a **R**, aunque **R** lo supera en el análisis estadístico. Pero goza de una ventaja respecto de **R** es que su nombre, más largo, permite búsquedas eficaces en Internet. También se distingue de **R** porque dispone de un amplio abanico de aplicaciones a otros tipos de análisis, no estadísticos, que **R** no posee. Probablemente quien se dedique profesionalmente a la Estadística, debería utilizar ambos.

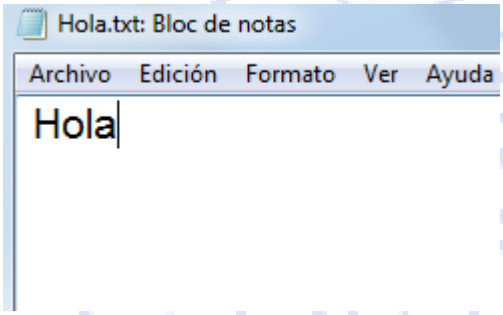
Ambos lenguajes poseen un amplio *ecosistema* de ampliaciones, libros en R y bibliotecas –*Libraries*- en Python, aunque ese es a su vez uno de sus inconvenientes, ya que pueden incluirse errores (Kasimir Schulz; Kieran Evans, 2024) (CERT Coordination Center, 2024).

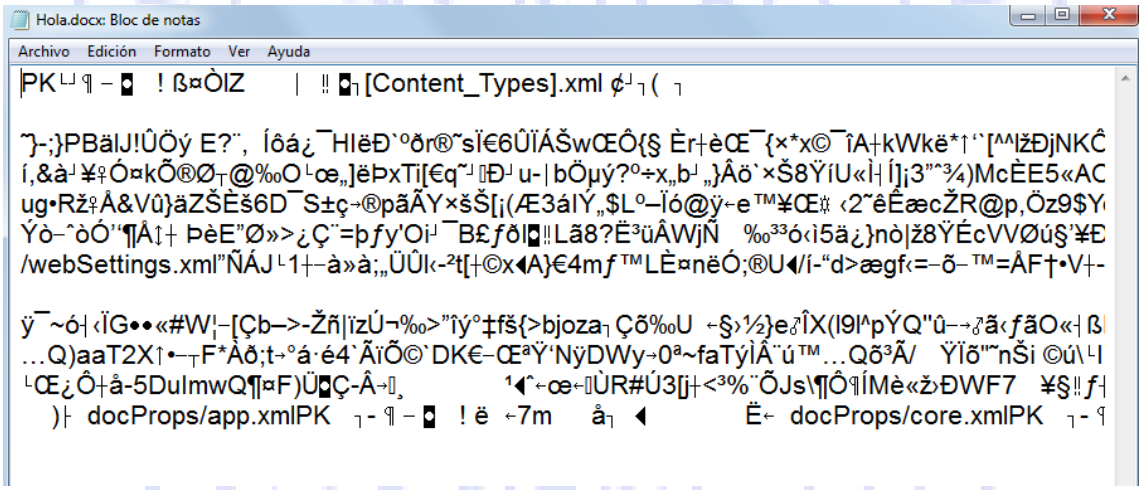
Reproductibilidad

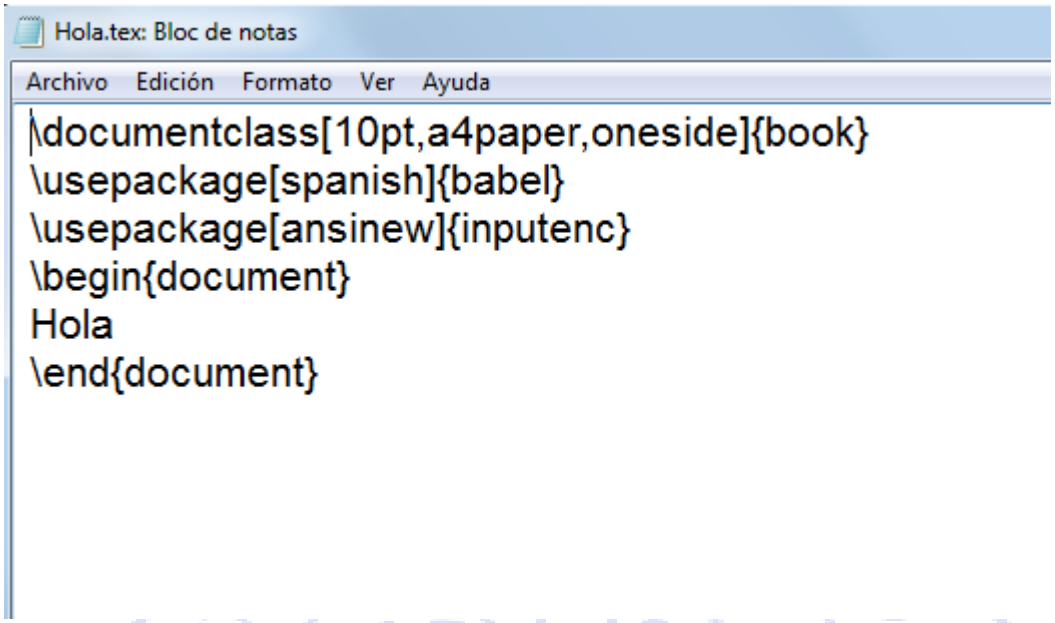
Ahora que hemos visto estos temas, tratemos la **reproductibilidad**, como una medida de cuánto facilita e incluso fomenta un programa concreto el uso de análisis reproducibles.

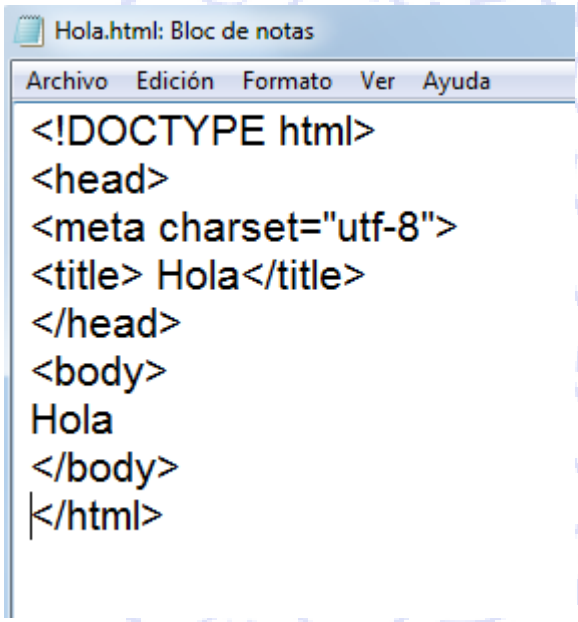
En todo lo que sigue utilizaré en los ejemplos salvo indicación en contra el sistema operativo **Windows** y el lenguaje **R**. Voy a considerar cuatro archivos triviales para ilustrar el procedimiento: Hola.txt, Hola.docx, Hola.HTML y Hola.tex, que corresponden a archivos de texto, Word, HTML y LaTeX.

Todos ellos van a mostrar la palabra Hola. Pero, internamente, no lo hacen del mismo modo. Si los visualizamos en un editor de texto como el Bloc de notas, tendremos lo siguiente:









Cuando cada uno de ellos se abre con un visor apropiado, los cuatro presentan la misma información: La palabra **Hola**.

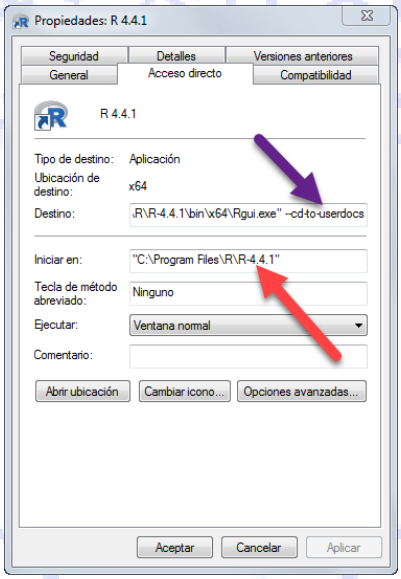
El primero, es el más sencillo, pero carece de toda estructura. En el de **Word** (que no se presenta completo) apreciamos que es imposible leer cuál es la estructura que el programa ha añadido y por tanto no existe una regla clara para modificarla. Los archivos de **LaTeX** y **HTML** tienen estructuras explícitas que debemos escribir lo que hace que parezcan *más difíciles*, pero al ser explícitas, pueden modificarse de modo inmediato. De hecho, el menos riguroso de los dos es

HTML, ya que, si no utilizamos completas las estructuras, los visores intentarán completarlas (en este ejemplo, ni siquiera se ha incluido el archivo **CSS** asociado).

Del mismo modo, si comparamos **R**, **SAS** y **SPSS**, los tres tienen la capacidad de realizar análisis **reproducibles**, pero no con la misma intensidad ni lo fomentan de igual modo. Veremos ahora que la **reproductibilidad** es mucho mayor en **R** que en **SAS** o **SPSS**.

Proyectos en R

Cuando se instala R, tras descargarlo, por ejemplo, de *CRAN* (The R Core Team, 2024) se crea un icono con un enlace al programa. Pero no es un simple enlace. También añade una serie de parámetros a la ejecución del programa.



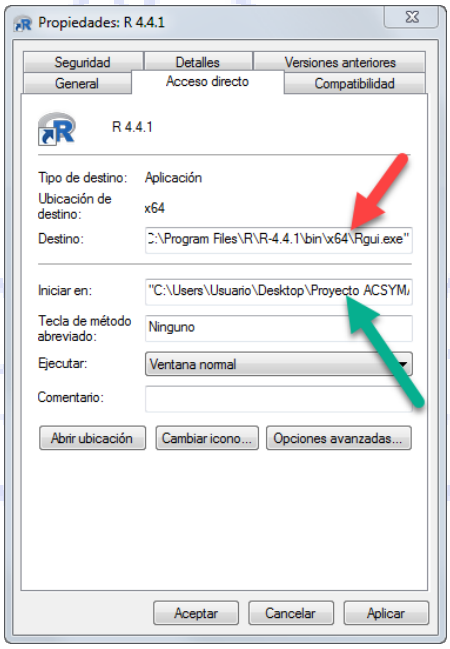
La primera flecha señala a un parámetro que hace que la ejecución del programa (tenga en cuenta que el programa hace referencia a una versión, en este caso la 4.3.2. R nunca borra una instalación anterior, al contrario que Word, por ejemplo. Esto permite mejorar la reproductividad.) modifique el lugar de trabajo que corresponde a *userdocs* –que es

C:/Users/Usuario/Documents– y ello independientemente de dónde esté situado el icono.

La segunda flecha señala a un parámetro que marca el directorio de trabajo, que es cambiado inmediatamente por el parámetro anterior.

En ese directorio es donde, predeterminadamente, se almacenará todo lo que R utilice. Y eso ¿qué significa? Que estará mezclada toda la información de todos los trabajos que se realicen con R, el espacio de trabajo contendrá todas las órdenes de todos los trabajos, con lo que se saturará e irá más lento, y si se crean objetos o funciones es posible que se solapen con otros anteriores ocultándolos.

Afortunadamente hay una solución trivial a este problema: personalizar los proyectos. Para ello, basta con que, para cada proyecto creemos un directorio, en este caso lo he creado en el escritorio con el nombre *Proyecto ACSYMA*. En esa carpeta he puesto una copia del icono y he cambiado sus propiedades como aquí se indica



Con este procedimiento, TODA la información referente al proyecto se guarda y se recupera desde ese directorio y el entorno global solo contiene la información del trabajo de ese proyecto.

Documentar

Es fundamental documentar todo lo que se va haciendo cuando se realiza un trabajo. Absolutamente todo.

Ello permite que la persona que lee la documentación del trabajo (incluso el mismo autor) pueda seguir el hilo lógico del desarrollo y saber por qué se han ido realizando las diferentes tareas.

En R, el uso del símbolo # (almohadilla), indica que el resto de la línea es un comentario, por lo que es muy sencillo proceder a la documentación.

Por supuesto, si el trabajo es complejo deberá realizarse simultáneamente una documentación externa en un documento que pueda consultarse sin tener que leer el contenido del programa que se ha realizado.

Integrar

Integrar es el antónimo de corta y pega. Y es una parte fundamental de la reproductibilidad.

En SPSS, por ejemplo, podemos realizar análisis estadísticos mediante selección de parámetros en un menú, pero a la hora de presentar los resultados, habremos de ir escribiendo el texto deseado y, posteriormente, añadirle los trocitos de información que copiamos de la salida de SPSS, sean tablas, gráficos, etc. Este sistema es altamente propenso a causar errores de copiado y conlleva mucho tiempo de trabajo.

Si se realiza un trabajo repetitivo diario, en el que, cambiando los datos, se realizan informes complejos, el procedimiento es insatisfactorio.

En R existen varias soluciones para resolver este problema y para mí, personalmente, la mejor solución es el uso del libro *knitr*. El autor del mismo, Xie, debió ponerle el nombre con la finalidad de reírse de los que intentáramos pronunciarlo, aunque él alega que lo puso como combinación de *knit* (en inglés, tejer, tricotar) y R.

Este libro permite construir un archivo que una vez compilado produce el resultado definitivo, sin necesidad de cortar y pegar.

Además de favorecer la reproductividad es un recurso perfecto para los trabajos repetitivos, como control de calidad diario en una fábrica, por ejemplo.

El libro knitr

Introducción

El libro **knitr** (Xie et al.;, 2024) permite construir dinámicamente un archivo HTML, PDF, etc. que integra resultados de R.

Son recomendables (Xie, Dynamic Documents with R and knitr, 2nd ed. , 2015) (Xie, Bookdown. Authoring Books and Technical Documents with R Markdown, 2016), (Xie; Thomas; Hill, 2018) (Xie; Allaire; Grolemund, 2019), (Xie; Dervieux; Riederer, 2021) y (Xie, Knitr Manual, 2024) para lograr una visión total del libro *knitr* y sus posibilidades.

Evidentemente, el libro con todas sus dependencias debe estar en la biblioteca, en caso contrario hay que usar la orden install.packages("knitr", dependencies = TRUE) para que lo incluya en la biblioteca. Para crear un archivo HTML, una vez cargado con la orden library(knitr), basta con crear un archivo con la extensión Rhtml que incluya un texto en formato

HTML y que además incluya unos *comentarios especiales* que serán ejecutados en R y cuyos resultados, mezclados con el texto original, constituirán un archivo con el mismo nombre que el archivo suministrado y con extensión html.

Esto es, si el archivo se llama 1.Rhtml y se encuentra en el directorio de trabajo, se ejecuta la orden knit("1.Rhtml") y se obtendrá el archivo 1.html y varios archivos adicionales en el directorio *figure* en caso de que se incluyan gráficos. Debe tenerse en cuenta que si los gráficos no tienen nombre explícito se guardan con nombres predeterminados en el directorio *figure* borrando a los que hubiese del mismo nombre. Así que no se debe ejecutar dos archivos con gráficos en el mismo directorio si no se asignan nombres a los gráficos.

Comentarios especiales

Los comentarios especiales son de la forma

```
<!--begin.rcode [metaórdenes]
        Órdenes de R
end.rcode-->
```

dónde las metaórdenes afectan al tipo de salida de R.

Así, para generar de 20 números pseudoaleatorios de la normal 0, 1 basta escribir

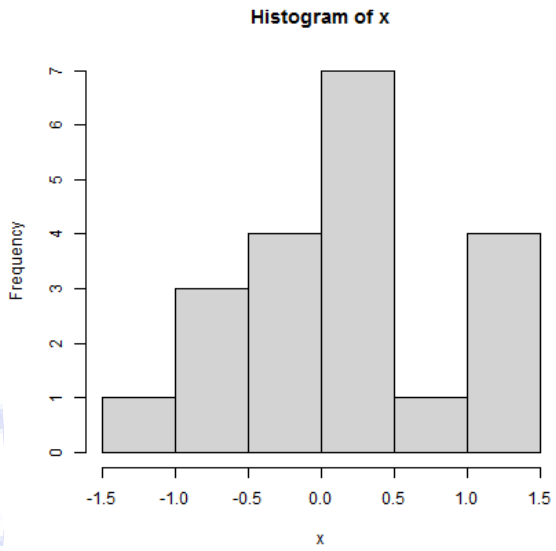
```
<!--begin.rcode
   ## Esta línea es un comentario que no afecta al resultado final
   x=rnorm(20)
end.rcode-->
```

Si no se quiere que aparezcan las órdenes que producen un resultado, basta con que use la metaorden echo=FALSE. En el siguiente bloque se presentan los valores anteriormente generados sin que aparezca la orden con la que se consigue.

```
<!--begin.rcode     echo=FALSE
x
end.rcode-->
```

La inclusión de gráficos es igual de sencilla. Así, el histograma de esos datos (centrado mediante la opción fig.align='center' y de tamaño 6x6 mediante las opciones fig.width=6, fig.height=6, la unidad de medida predeterminada son 72 puntos), y que se almacena con un nombre que comienza con **figura1**:

```
<!--begin.rcode        fig.align='center',        fig.width=6,
fig.height=6, 'figura1'
        ## @knitr histograma
    hist(x)
end.rcode-->
```



Resultados *inline*

Se pueden incluir expresiones en el texto con la opción **<!--rinline expresión -->** como por ejemplo,

```
 La media de x es <!--rinline mean(x) --> que está comprendida entre 0 y 1, aproximadamente en el centro del intervalo.
```

La media de x es 0.2085034 que está comprendida entre 0 y 1, aproximadamente en el centro del intervalo.

Esto elimina el problema del *corta-pega*.

El libro datacleanr

El libro datacleanr (Hurley; Peters; Pappas, 2021) realiza la limpieza de datos interactivamente, pero, pensando en la reproductividad, traslada todas las operaciones realizadas a código R, por lo que puede repetirse el proceso fácilmente. Un ejemplo de su uso en Ecología se encuentra en este artículo (Hurley; Peters; Pappas; Steger; Heinrich, 2022) en el que demuestran la utilidad del libro abordando dos problemas que aparecen frecuentemente en el análisis de datos:

- La identificación de estructuras problemáticas en datos anidados jerárquicamente, y
- La prevención de la pérdida excesiva de datos.

Además, en consonancia con la reproductividad, todos los datos y programas están en repositorios públicos.

El libro papaja

El libro papaja (Aust & Barth, 2023-09-29), acrónimo de *Preparing APA Journal Articles*, es un libro de R que permite crear artículos reproducibles, siguiendo el estilo APA. El estilo APA (APA, 2020) es uno de los principales estilos para artículos científicos, especialmente en Ciencias Sociales, aunque se usa mucho en otros muchos ámbitos. Es el estilo definido por *American Psychological Association*. Puesto que seguir las instrucciones del estilo puede ser muy complejo, son más de 250 páginas, este libro resulta fundamental. Desde la séptima edición, además recomienda abandonar el uso de procesadores de texto y utilizar lo que denominan informes científicos dinámicos, esto es, R y cualquiera de los

libros, como knitr, haciendo referencia a informe de reproductividad en la ciencia de *U.S. National Science Foundation* (Cacioppo, Kaplan, Krosnick, Olds, & Dean, 2015).

Gestión de la reproductividad

Existen varios programas que ayudan a gestionar la reproductividad, porque en análisis complejos existe propensión a cometer fallos.

En R hay varias posibilidades como son

- el libro *renv* (Ushey; Wickham, 2024) que ayuda a crear entornos reproducibles para los proyectos de R. Con su ayuda se lleva un control de todos los libros que se usan, crea y administra bibliotecas R locales del proyecto, guarda el estado de estas bibliotecas en un archivo y luego restaura su biblioteca según sea necesario, permite reinstalarlos en otro ordenador, y permite volver a instalar todo el proyecto en cualquier momento;
- el libro *groundhog* (Simonsohn; Gruson, 2024) que hace que los programas en R sean reproducibles, garantizando que cada vez que se ejecute un programa determinado, se cargue la misma versión de los libros utilizados para lo que utiliza la orden *groundhog.library* en lugar de *library* e incluye una fecha en la llamada. La fecha se utiliza para llamar a la misma versión del libro cada vez;
- el libro *Require* (McIntire; Chubaty , 2024) que enfatiza las funciones que devuelven el mismo resultado ya sea la primera vez que se ejecuta la función o sucesivas;
- el libro *miniCRAN* (de Vries; Chubaty, 2024) que es desde mi punto de vista el mejor, prácticamente automatizando lo que haríamos manualmente: Permite crear un repositorio interno, similar a *CRAN*, solamente con los libros utilizados. El usuario especifica un conjunto de libros deseados y se lee recursivamente el árbol de dependencias de estos libros y descarga este

subconjunto. El usuario puede instalar libros directamente desde este repositorio local, en lugar de hacerlo desde *CRAN*;

En realidad, R tiene un sistema de instalación en el que cada versión se instala en un directorio distinto, sin afectar a ninguna instalación anterior. Puesto que el tamaño total de R y los libros que se utilicen en un proyecto concreto es bastante pequeño (en las capacidades actuales de almacenamiento) es sencillo y conveniente no desinstalar nunca ninguna versión antigua. Si además los proyectos se desarrollan en directorios distintos como hemos indicado en la personalización de proyectos, basta con conservar todos estos directorios y será posible reproducir los análisis en cualquier momento (Mientras el sistema operativo continúe manteniendo la compatibilidad con las versiones anteriores).

En definitiva, hay que calificar como genial el diseño de R, que supongo que no se realizó de modo expreso, que permite desarrollar a la perfección un análisis estadístico totalmente reproducible, y ello, además, manteniéndolo a lo largo del tiempo y de las diferentes versiones.

Reproductibilidad

Ahora que hemos visto estos temas, tratemos la reproductibilidad, como una medida de cuánto facilita e incluso fomenta un programa concreto el uso de análisis reproducibles.

Si comparamos R, SAS y SPSS, los tres tienen la capacidad de realizar análisis reproducibles, pero no con la misma intensidad ni lo fomentan de igual modo.

Creo haber dejado claro que la reproductibilidad es mucho mayor en R que en los otros, y siendo la reproductividad una cualidad deseable, R debería ser nuestra elección primera.

Conclusión

Quienes vayan a realizar una investigación y deseen publicarla deberán, cada vez más, realizarla de acuerdo con los criterios, cada vez más extendidos, de reproductividad o incluso de replicabilidad. Hay un aforismo de la Medicina muy repetido al público no especialista: *no se automedique.* En la investigación hay que pensar en uno similar: *no se autoestadistique.* Debe siempre contar con una persona formada en Estadística que le ayude y que deje constancia de todos los pasos: diseño del experimento, recogida de los datos, depuración de los mismos, análisis estadístico y obtención de conclusiones, y se preocupe de poner todo en repositorios públicos.

Circula una frase que se atribuye a Mario Moreno, Cantinflas, aunque no he logrado encontrar ningún documento que así lo pruebe, y si es de una de sus películas es posible que no fuera suya sino de algún guionista, que me parece de aplicación en este momento:

> «*Estamos peor, pero estamos mejor. Porque antes estábamos bien, pero era mentira. No como ahora que estamos mal, pero es verdad*.»

Y es que, aunque pueda parecer que tantas cosas que antes nos han parecido correctas, las vemos desde un punto de vista que muestra sus flaquezas, sin embargo, hemos aprendido de esos errores y ahora estamos en el camino de mejorarlas.

Muchas gracias.

Bibliografía

Andrea Taloni; Vincenzo Scorcia; Giuseppe Giannaccare. (2023). Large Language Model Advanced Data Analysis Abuse to Create a Fake Data Set in Medical Research. *JAMA Ophthalmology*, 1174-1175.

Andrews. (2021). *Doing Data Science in R. An Introduction for Social Scientists.* London: SAGE.

APA. (2020). Publication Manual of the American Psychological Association, Seventh Edition. *https://apastyle.apa.org/products/publication-manual-7th-edition*.

Aranda Ramírez, P. (2023). Dieta mediterránea: Una dieta saludable y sostenible. *Anales de la Academia de Ciencias Sociales y del Medio Ambiente de Andalucía*.

Aust, F., & Barth, M. (2023-09-29). papaja: Prepare American Psychological Association Journal Articles with R Markdown. *CRAN https://cran.r-project.org/web/packages/papaja/index.html*.

Bateson, G. M. (1868, 1901, 1966). Experiments in Plant Hybridization. *The Electronic Scholarly Publishing Project*.

Baumer, Kaplan, & Horton. (2021). *Modern Data Science with R, 2nd ed.* Boca Raton, FL: CRC Press.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language A Programming Environment for data analysis and graphics.* Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Boyle, R. (1662). *A Defence of the Doctrine Touching the Spring and Weight of the Air.* London: Thomas Robinson.

Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.* National Science Foundation.

CERT Coordination Center. (29 de 04 de 2024). *Software Engineering Institute.* Obtenido de R Programming Language implementations are vulnerable to arbitrary code execution during deserialization of .rds and .rdx files: https://kb.cert.org/vuls/id/238194

de Vries; Chubaty. (2024). miniCRAN: Create a Mini Version of CRAN Containing Only Selected Packages. *CRAN https://cran.r-project.org/web/packages/miniCRAN/index.html*.

Fisher, R. (1936). Has Mendel's work been rediscovered? *Annals of Science*.

GAISE College Report ASA Revision Committee. (2017). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016.* ASA https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

Golombek, D. (2005, 2014). *Demoliendo papers. La trastienda de las publicaciones científicas.* Siglo XXI.

Harris, R. F. (2017). *Rigor mortis. How sloppy science creates worthless cures, crushes hope, and wastes billions.* New York: Basic Books.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models.* Suffolk: Chapman and Hall.

Hector. (2021). *The New Statistics with R. An Introduction for Biologists, 2nd ed.* Oxford: Oxford University Press.

Hofmann, Wickham, & Cook. (2019). The 2013 Data Expo of the American Statistical Association. *Computational Statistics*, 1443-1447.

Hurley; Peters; Pappas. (2021). datacleanr: Interactive and Reproducible Data Cleaning. *CRAN*.

Hurley; Peters; Pappas; Steger; Heinrich. (2022). Addressing the need for interactive, efficient, and reproducible data processing in ecology with the datacleanr R package. *PLoS ONE 17 (5)*.

Ihaka, R., & Gentleman, R. (1993). *R*. Obtenido de The R Project for Statistical Computing: https://www.r-project.org/

J, F., MM, M., & M, B.-V. (2024). *Rcmdr: R Commander. R package version 2.9-2*. Obtenido de https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/

JASA, E. T. (2024). Style Guide, https://www.jstatsoft.org/style. *Journal of Statistical Software*.

Kasimir Schulz; Kieran Evans. (29 de 04 de 2024). *HiddenLayer.* Obtenido de R-bitrary Code Execution: Vulnerability in R's Deserialization: https://hiddenlayer.com/research/r-bitrary-code-execution/

Márquez-Sánchez, S.-C. (2005). ¿Fueron manipulados los datos de Mendel? (argumentos en contra). *Agricultura, Sociedad y Desarrollo*, 39-46.

Martín Castro, A. (2022). *La lira desafinada de Pitágoras.* Madrid: HarperCollins Ibérica.

Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, 80-88.

McIntire; Chubaty . (2024). Require: Installing and Loading R Packages for Reproducible Workflows. *CRAN https://cran.r-project.org/web/packages/Require/index.html*.

Mendel, G. (1866). Versuche über Pflanzenhybriden. *Abhandlungen (Universität Regensburg UNIVERSITÄTBIBLIOTHEK)*.

Naddaf, M. (2023). ChatGPT generates fake data set to support scientific hypothesis. *Nature*, 895-896.

National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific.* Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science .* Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2020). *Enhancing Scientific Reproducibility in Biomedical Research Through Transparent Reporting: Proceedings of a Workshop.* Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2022). *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies.* Washington, DC: The National Academies Press.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical transactions of the Royal Society A*, 694-706.

Obels, Lakens, Coles, Gottfried, & Green. (2020). Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. *Advances in Methods and Practices in Psychological Science*, 229-237.

Simonsohn; Gruson. (2024). groundhog: Version-Control for CRAN, GitHub, and GitLab Packages. *CRAN*

https://cran.r-project.org/web/packages/groundhog/index.html.

The R Core Team. (2024). *The Comprehensive R Archive Network*. Obtenido de https://cran.r-project.org/

The R Project for Statistical Computing. (s.f.). Obtenido de https://www.r-project.org/

Ushey; Wickham. (2024). renv: Project Environments. *CRAN https://cran.r-project.org/web/packages/renv/index.html*.

Vázquez Mata, G. (1992). Factors related to quality of life 12 months after discharge from an intensive care unit. *Critical Care Medicine*.

Xie. (2015). *Dynamic Documents with R and knitr, 2nd ed. .* Boca Raton FL: 2015.

Xie. (2016). *Bookdown. Authoring Books and Technical Documents with R Markdown.* Boca Raton FL: CRC Press.

Xie. (2024). Knitr Manual. *https://yihui.org/knitr/demo/manual/*.

Xie et al.;. (2024). knitr: A General-Purpose Package for Dynamic Report Generation in R. *CRAN https://cran.r-project.org/web/packages/knitr/index.html*.

Xie; Allaire; Grolemund. (2019). *R Markdown. The Definitive Guide.* Boca Raton FL: CRC Press.

Xie; Dervieux; Riederer. (2021). *R Markdown Cookbook.* Boca Raton FL: CRC Press.

Xie; Thomas; Hill. (2018). *Blogdown. Creating Sites with R Markdown.* Boca Raton FL: CRC Press.

Xiong, X., & Cribben, I. (2022). The State of Play of Reproducibility in Statistics: An Empirical Analysis. *The American Statistician, 77(2)*, 115–126.

ACADEMIA DE CIENCIAS SOCIALES Y DEL MEDIO AMBIENTE DE ANDALUCIA

Índice

Agradecimientos 2
El problema 5
ChatGPT 10
Propuestas de solución 12
Modern Data Science with R 12
Proyecto TIER 13
American Statistical Association 15
National Academies of Sciences, Engineering, and Medicine 18
En Psicología 21
Retraction Watch 22
Mi opinión 23
Aportaciones de la Estadística Computacional a la solución 26
R (Y Python) 27
Reproductibilidad 28
Proyectos en R 31
Documentar 33
Integrar 33
El libro knitr 34
Introducción 34
Comentarios especiales 35
Resultados *inline* 36
El libro datacleanr 37
El libro papaja 37
Gestión de la reproductividad 38
Reproductibilidad 39
Conclusión 40
Bibliografía 41