Reproducibilidad y Comunicación de Resultados

Diplomado Data Science para Ciencias Sociales
(DTA00005 CA01)

Andrés González Santa Cruz
René Lagos Barrios
Amaru Aguero Jiménez

Presentación y Aspectos Formales

Acuerdos para la clase

- Parte el profesor
- Nombre, experiencia en R, por qué tomó el Diplomado y expectativas del curso
- Requisitos, ver problemas particulares
- Explicación del curso y aspectos formales:
  - Programa y Contenidos
  - Programa y Evaluaciones
  - Programa y Cronograma

¿Qué entienden por reproducibilidad y replicabilidad?, ¿Por qué son importantes cuando hacemos ciencia?

Discutan entre ustedes (de a 2 o 3) y Tomen nota de sus argumentos

Un poco de contexto

Contexto

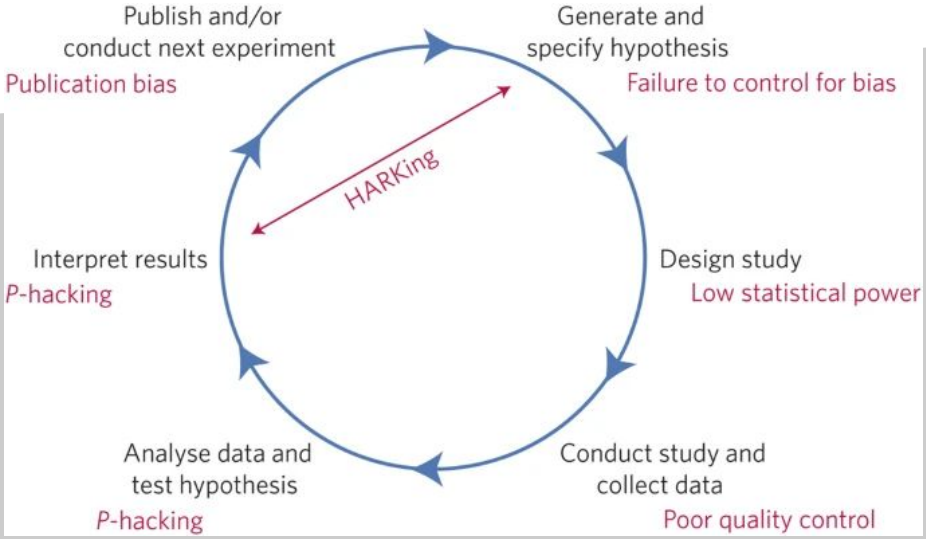
- Apuntamos a producir conocimiento robusto y confiable
- Ensayo y error inherente a la ciencia
  - Callejones sin salida
- PROBLEMA:
  - Problemas complejos
  - Difícil inspección/reutilización
  - “Credibilidad y progreso”

Adam, D. (2023). What reproducibility crisis? New research protocol yields ultra-high replication rate. 10.1038/d41586-023-03486-5

Modelo idealizado del método científico y amenazas

Otros los llaman 4 jinetes del apocalipsis (Bishop, 2024)

- Sesgo de publicación
- Baja potencia estadística
- P-hacking
- HARKing



Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021

Volvamos a las preguntas

- ¿Podría detallar cómo llegó a una respuesta consensuada?
- ¿Existe la posibilidad de que a través de los pasos que siguieron, lleguen a otras conclusiones?
- Si sus compañeros quisieran avanzar en sus argumentos, ¿les servirían realmente sus anotaciones?
- ¿Son generalizables?, ¿Son consistentes y robustas a distintos ejemplos? (ej., replicar en filosofía, historia)
- Si les pidiera definir “ciencia abierta” o “transparencia” y argumentar su importancia, ¿podrían seguir los mismos pasos que siguieron con el ejemplo anterior?
- Comparta sus notas con el otro grupo y pida a cada grupo que argumente únicamente basado en las anotaciones del otro grupo

Definiciones

"*Los datos y el código utilizados para un hallazgo están disponibles y permiten a otro investigador recrearlo*." (Gandrud, 2020)

Replicabilidad vs. Reproducibilidad (ASA, 2017)

- Repetir un estudio con nuevos datos y obtener el mismo resultado.
- Recrear resultados numéricos usando los datos y el código originales.

Ciencia abierta (UNESCO, 2021)
"Un constructo inclusivo que combina movimientos y prácticas para que los conocimientos científicos sean:

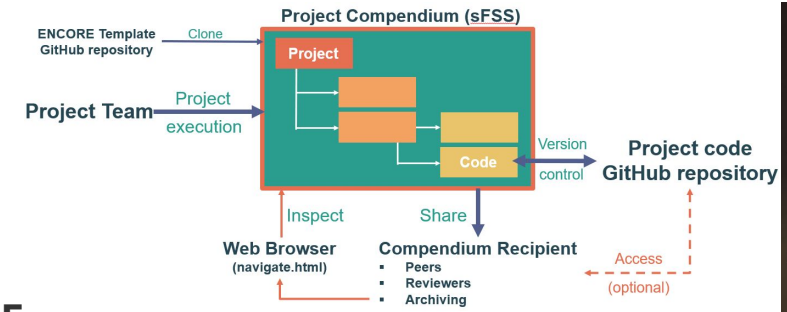
- *Abiertamente disponibles*
- *Accesibles para todos*
- *Reutilizables por todos*"

Implicancias

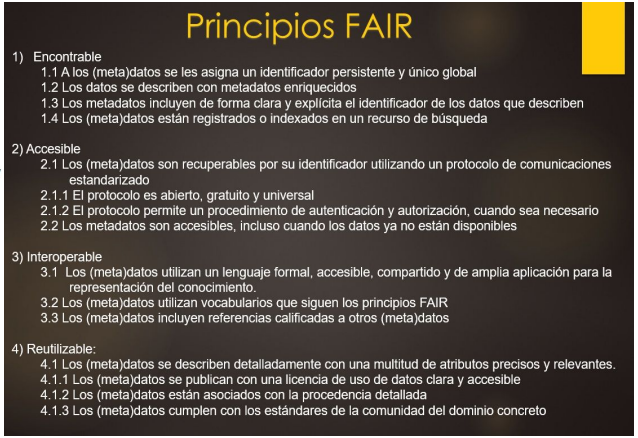
- Open Access es caro y reproduce disparidades económicas
- Según la UNESCO (2018), el marco de legalidad de propiedad intelectual debe ser actualizado y adaptado a condiciones digitales abiertas
- Obstáculos, como la falta de tiempo, la incompatibilidad con los cargos (Rubilar, 2023).

Lineamientos

- **principios FAIR (Findable, Accessible, Interoperable, Reproducible)**
  - Encontrable
  - Accesible
  - Interoperable
  - Reutilizable
- **TOP guidelines**
  - Citación
  - Transparencia de datos
  - Plan analítico
  - Material de inv.
  - Prerregistro
  - Replicabilidad
- **ENCORE**
  - Subcarpetas y sistema de estructura de archivos estandarizada (sFSS)



Fuente: Van Kampen AHC, Mahamune U, Jongejan, A (2023) The standardized file system structure (FSS) navigator. Zenodo. DOI: 10.5281/zenodo.7985655



Fuente: https://biblioguias.cepal.org/c.php?g=495473&p=8022713

	Not Implemented	Level I	Level II	Level III
Citation Standards	No mention of data citation.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used consistent with journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Journal encourages data sharing, or says nothing.	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Journal encourages code sharing, or says nothing.	Article states whether code is available, and, if so, where to access it.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research Materials Transparency	Journal encourages materials sharing, or says nothing.	Article states whether materials are available, and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal encourages design and analysis transparency, or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Study Preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration of study exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis Plan Preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration with analysis plan exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies, or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts results blind review.	Journal uses Registered Reports as a submission option for replication studies with peer review prior to observing the study outcomes.

Fuente: https://docs.google.com/presentation/d/1IThAtkV18kvTiBbskL5JoeIw-xsdRcwUr8lxobbMvn0/edit#slide=id.g1250006a4ed_0_18

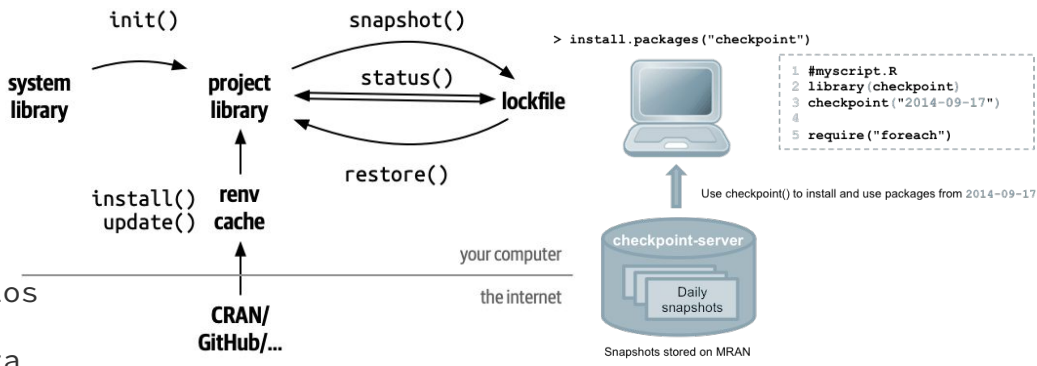
Suficientemente buenas prácticas para computación científica

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

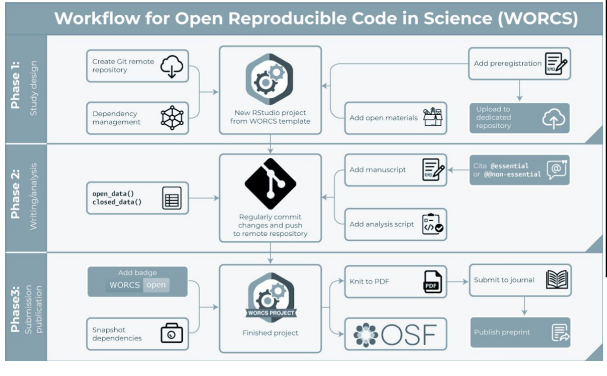
Elementos relevantes en R

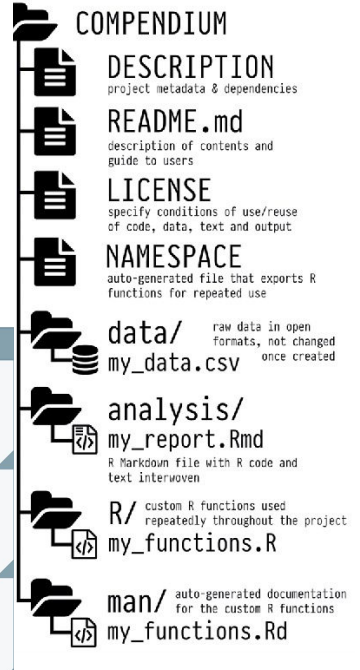
- data.tree= Permite presentar una estructura de los documentos de un proyecto
- renv= Permite aislar bibliotecas de paquetes para cada proyecto, instalar paquetes y versiones específicas para recrearlas en otros entornos
- rig= Permite manejar distintas versiones de R y bibliotecas de paquetes, además permite instalar Rtools
- usethis= Permite automatizar tareas y proyectos e integraciones con otras plataformas (ej., GitHub)
- codebook= Permite incrustar metadatos a objetos como bases de datos.
- dateback= Obtiene paquetes en determinadas fechas, junto con sus dependencias (paquetes asociados que permiten instalarlo)
- checkpoint= Permite instalar paquetes conforme fijados en una fecha en un servidor externo (controlado por Microsoft)
- sessionInfo()= Permite entender el entorno de ejecución de R

Más información en:
https://cran.r-project.org/web/views/ReproducibleResearch.html




```
analysis/
|
├── paper/
|   ├── paper.qmd       # this is the main document to edit
|   └── references.bib  # this contains the reference list information
|
├── figures/            # location of the figures produced by the qmd
|
├── data/
|   ├── raw_data/       # data obtained from elsewhere
|   └── derived_data/   # data generated during the analysis
|
└── templates
    ├── journal-of-archaeological-science.csl
    |                   # this sets the style of citations & reference list
    ├── template.docx   # used to style the output of the paper.qmd
    └── template.Rmd
```





Ejercicio

1) Crear un proyecto en R
2) Generar una estructura del sistema de archivos que distinga bibliografía (_bib), datos de estilo de la carpeta (_style), figuras (_figs), principales paquetes estadísticos (requirements.txt), datos (_data) y salidas (_output)
3) Acceda a esta pagina y baje el archivo correspondiente al **año que le indique el profesor** https://datos.gob.cl/dataset/precios-al-consumidor
4) Obtenga el promedio del precio promedio por grupo y mes. Utilice *tidyverse*
5) Utilice alguna herramienta que contribuya a la reproducibilidad del ejercicio
6) Generar data.tree
7) Comprimir archivo en formato .zip
8) Compartir con un compañer@
9) Enviar captura de pantalla del data.tree, e incluya en el correo una reflexión concienzuda de las debilidades, fortalezas y amenazas a la reproducibilidad y replicabilidad del ejercicio proporcionado por su compañer@ (al menos dos de cada una)

Estructura curso

1- Esquema conceptual de registro y datos

2- Reportes estáticos y dinámicos

3- Repositorios, control de versiones, containers, virtual machines

Fuentes

- Marwick, B. Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. J Archaeol Method Theory 24, 424-450 (2017). https://doi.org/10.1007/s10816-015-9272-9
- Subbaraman, Nidhi. "These Researchers Critique Bad Science. Now Their Own Paper Has Been Retracted." The Wall Street Journal, October 30, 2024. Available at: https://www.wsj.com/science/nature-human-behavior-study-retracted-standards-b589dbe6?utm_source=chatgpt.com (accessed November 25, 2024).
- Gandrud, Christopher. Reproducible Research with R and RStudio (Third Edition). CRC Press/Chapman & Hall, 2020. Source code and files available at: https://github.com/christophergandrud/Reproducible-Research-with-R-and-RStudio (accessed November 25, 2024).
- Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017).
- Chen, X., Dasler, R., Feger, S., Fokianos, P., Gonzalez, J. B., Hirvonsalo, H., Kousidis, D., Lavasa, A., Mele, S., Rodriguez, D. R., Šimko, T., Smith, T., Trisovic, A., Trzcinska, A., Tsanaktsidis, I., Zimmermann, M., Cranmer, K., Heinrich, L., Watts, G., . . . Neubert, S. (2019). Open is not enough. Nature Physics, 15(2), 113-119.
- Figueiredo Filho D, Lins R, Domingos A, Janz N, Silva L. Seven Reasons Why: A User's Guide to Transparency and Reproducibility. Bras political sci rev [Internet]. 2019;13(2):e0001.
- Gobierno de Chile. Directrices de metadatos y mecanismos de interoperabilidad: Proyecto Nodo Nacional de Acceso [Internet]. Santiago: Agencia Nacional de Investigación y Desarrollo (ANID); 2024 [citado 2024 nov 25]. Disponible en: https://acceso-abierto.anid.cl/wp-content/uploads/sites/4/2024/05/Metadatos_para_la_Interoperabilidad_de_los_Repositorios_2024.pdf
- Nosek et al. Guidelines for Transparency and Openness Promotion(TOP)in Journal Policies and Practices (Version1.0.1). 2014. https://osf.io/ud578
- van Kampen, A.H.C., Mahamune, U., Jongejan, A. et al. ENCORE: a practical implementation to improve reproducibility and transparency of computational research. Nat Commun 15, 8117 (2024). https://doi.org/10.1038/s41467-024-52446-8
- Registered Reports at Nature Methods. Nat Methods 19, 131 (2022).
- Fundación Carolina, Secretaría General Iberoamericana (SEGIB). Ciencia abierta: Retos y oportunidades para Iberoamérica. Madrid: Fundación Carolina; 2020 [citado el 26 de noviembre de 2024]. Disponible en: https://www.fundacioncarolina.es/wp-content/uploads/2020/12/Ciencia-Abierta.pdf
- Ross-Hellauer T. Open science, done wrong, will compound inequities. Nature. 2022 Mar 14
- Rubilar A. Diagnóstico y Planificación de gestión del cambio institucional en Ciencia Abierta [presentación]. Semana de Acceso Abierto 2023. Prácticas de Ciencia Abierta en instituciones de Educación Superior. 3 al 5 Octubre; 2023 [citado 2024 Nov 28]. Disponible desde: https://semanaaccesoabierto.cl/exposiciones_2023/dia_2/Diagnostico_y_planificacion_de_gestion_del_cambio_institucional.pdf
- Bishop, D. "Rein in the Four Horsemen of Irreproducibility." (2019). Accessed November 26, 2024. 2.
- Van Lissa CJ, Brandmaier AM, Brinkman L, Lamprecht A, Peikert A, Struiksma ME, Vreede B. WORCS: A Workflow for Open Reproducible Code in Science. Data Science. 2021;4(1):29-49. doi:10.3233/DS-210031.
- Marwick B, Boettiger C, Mullen L. Packaging Data Analytical Work Reproducibly Using R (and Friends). The American Statistician. 2018;72(1):80-8.