Data Falsificada: El Cas Gino.

Per Lorenz Goette i Guillem Riambau

[Aquesta entrada va ser escrita originalment en anglès amb la intenció de ser publicada al blog sobre economia i investigació Nada es Gratis].

Entre el 17 i el 30 de juny de 2023, l'equip de Data Colada va publicar una sèrie de quatre entrades en les quals van mostrar evidència convincent de frau en articles publicats recentment per la professora d'Administració d'Empreses Francesca Gino (Harvard Business School, d'ara endavant HBS). La recerca de Data Colada es va dur a terme el 2021. Basant-se en les seves troballes, HBS va iniciar una investigació interna que va resultar la suspensió "administrativa" durant dos anys sense sou de la Prof. Gino, tal com anuncia la seva pàgina oficial de Harvard.

Les notícies han tingut un impacte més enllà dels cercles acadèmics habituals i de Twitter, sobretot potser perquè el frau es va realitzar de manera irònica en estudis sobre honestedat. The Guardian, The New York Times, The Boston Globe, The Washington Post, The New Yok Post, Business Insider, NPR, i, de manera menys sorprenent, The Chronicle of the Higher Education (també aquí), VOX (el blog, no el partit), o The Atlantic han informat sobre aquesta qüestió, entre molts altres. Pel que sabem, entre els mitjans de comunicació generals espanyols, només Expansión ha destacat les notícies sobre el frau, tot i que La Vanguardia va publicar un article el 2017 sobre la recerca de Gino titulat "Si te ha engañado una vez, lo volverá a hacer, según Harvard".

En un gir inesperat dels esdeveniments, Francesca Gino va anunciar al seu compte de Linkedin a principis d'agost que "no tenia cap altra opció que presentar una demanda contra la Universitat de Harvard i els membres del grup Data Colada, que van treballar junts per destruir la meva carrera i reputació", ja que "[mai], en cap moment, he falsificat dades o comès irregularitats en la recerca de cap mena". La demanda sol·licita ni més ni menys que 25 milions de dòlars com a compensació. (Per als lectors curiosos: és interessant comprovar la gran majoria de respostes al seu post de Linkedin són de suport, en comparació amb les reaccions que hem observat en el món acadèmic).

A continuació, donades les limitacions d'espai, ens centrarem en dues de les acusacions de frau i les respostes de Francesca Gino en la seva demanda. La lògica dels esdeveniments en les altres dues entrades és molt similar i es poden trobar aquí (Data Falsificada (Part 3): "The Cheaters Are Out of Order") i aquí (Data Falsificada (Part 4): "Forgetting The Words").

Data Falsificada (Part 1): "Clusterfake" (publicat el dia 17 de juny de 2023).

En aquesta entrada, els autors aborden l'Estudi 1 de l'article de PNAS del 2012 de Shu, Mazar, Gino i Ariely, "Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end" (espoiler: l'article va ser retirat el 13 de setembre de 2021). L'Estudi 1 es va realitzar a la Universitat de Carolina del Nord (UNC) el 2010. Gino, que era professora a la UNC abans d'unir-se a Harvard el 2010, va ser l'única autora involucrada en la recopilació i l'anàlisi de les dades de l'Estudi 1.

Les dades estaven disponible a Open Science Framework. Hi ha 101 observacions i estan gairebé (però no del tot) ordenades primer per l'assignació de la condició (0 = control, 1 = signar a dalt i 2 = signar a baix) i segon (dins de cada assignació de condició) per un identificador de participant anomenat "P#". Hi ha 8 identificadors de participants que estan duplicats o fora de seqüència d'una manera sospitosa. L'equip de Data Colada argumenta que "[n]o hi ha cap manera, fins allí on podem arribar, d'ordenar les dades per aconseguir aquest ordre. Això vol dir que aquestes files de dades van ser mogudes manualment o que els P# van ser modificats manualment. Veurem que la raó correcta és la primera."

Data Colada assenyala que les dades també inclouen un fitxer d'Excel amb les mateixes dades que conté fórmules. Un fitxer subsidiari que utilitza l'Excel per produir la fulla de càlcul és el calcChain.xml. CalcChain "conserva l'ordre en què les fórmules van ser introduïdes inicialment a la fulla de càlcul", independentment d'on se situïn finalment les cel·les. Utilitzant CalcChain, Data Colada mostra que 6 observacions que apareixen una sobre l'altra en les dades estan fora de seqüència. A més, els P# de les files que envolten els llocs on CalcChain identifica les posicions inicials ometen la posició exacta que hauria estat moguda, reforçant la idea que les observacions van ser mogudes manualment.

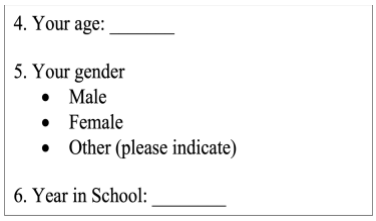
Les 8 observacions són fonamentals pels resultats de l'article, ja que, com assenyala Data Colada, "totes elles estan entre les observacions més extremes dins de la seva condició i totes elles van en la direcció predita". Tot això "suggereix de manera contundent" (argumenten) que les observacions es van modificar per obtenir uns resultats concrets. Per ser precisos, "[a]mb només n = 8, obtenen t(6) = 21.92, amb un valor p minúscul".

Els punts 234-247 (pàgines 52-55) de la demanda de Gino aborden aquesta entrada del blog. La majoria dels seus arguments es basen en el fet que les dades originals es van recopilar en paper (per exemple, 247: "Data Colada sabia que l'Estudi 1 es va realitzar en paper, amb dades recopilades en paper el 2010. Data Colada també sabia que el fet que l'estudi s'hagués realitzat en paper proporcionava una raó raonable i plausible pel fet que les dades (...) no fossin ordenades en cap ordre concret"). Tot i que aquests arguments són correctes, no aborden la qüestió clau assenyalada per Data Colada: que les dades es van reorganitzar dins de la fulla de càlcul d'Excel *després* de la introducció inicial de les observacions.

Data Falsificada (Part 2): "My Class Year Is Harvard" (publicat el 20 de juny de 2023).

En aquesta entrada, els autors discuteixen l'Estudi 4 de l'article de 2015 de Psychological Science "The Moral Virtue of Authenticity: How Inauthenticity Produces Feelings of Immorality and Impurity" (Gino, Kouchaki i Galinsky).

Tots els participants eren estudiants de Harvard. En recopilar informació sociodemogràfica, als participants se'ls va demanar que proporcionessin el seu any a la universitat (Q6, vegeu la captura de pantalla del material original publicat).


4. Your age: _____

5. Your gender
- Male
- Female
- Other (please indicate)

6. Year in School: _____

Les respostes raonables a la pregunta són "Junior", "junior", "3", "class of 2016", "'16", etc. El que sembla menys raonable com a resposta és "Harvard", que es troba fins a 20 vegades a les dades. Com assenyalen els autors de Data Colada, "és difícil imaginar que tants estudiants facin aquest error altament idiosincràtic de manera independent (…) A més, i afegint a la peculiaritat, les respostes d'aquests 20 estudiants es troben totes en un interval de 35 files (de la 450 a la 484) del conjunt de dades publicat".

Totes aquestes observacions proporcionen resultats que coincideixen amb les prediccions dels autors: aquells que, mitjançant assignació aleatòria al tractament, es preveu que donin respostes "altes" ho fan, i aquells que s'assignen a la condició associada a l'expectativa de respostes "baixes" donen respostes "baixes". Com assenyala l'entrada, "l'efecte per a les observacions 'Harvard' és significativament més gran que l'efecte per a les observacions no-Harvard ($p < .000001$). Això suggereix de manera sòlida que aquestes observacions 'Harvard' es van alterar per aconseguir l'efecte desitjat", o, més com a mínim, això sembla una coincidència molt poc probable.

Què diu Gino sobre tot això? Els punts 248–253 (pàgines 55-56) de la seva denúncia aborden aquesta entrada de blog. El seu punt més rellevant és el 250: "Data Colada, com a experimentats científics del comportament, sabien que els participants sovint responen a una enquesta per aconseguir el pagament que els correspon per la seva participació (com a participants de l'estudi) i poden precipitar-se en les respostes, a vegades més d'una vegada per cobrar, i utilitzar valors extrems com a respostes. És àmpliament conegut en la ciència del comportament que, de vegades, els participants en estudis en línia proporcionen dades de baixa qualitat responent enquestes sense l'atenció que requereixen". Cap dels altres 5 punts de Gino en resposta a aquesta entrada de Data Colada aborda per què 20 participants aleatoris que van omplir les respostes virtualment de manera aleatòria ho van fer tots en la mateixa direcció, quan la falta d'atenció faria pensar en tota mena d'errors per a aquestes observacions.

La denúncia de Gino té exactament 100 pàgines. Acaba amb la "petició de reparació" (pàgines 95 en endavant), en la qual demana (pàgina 97): "En la setena causa d'acció per difamació contra els demandats Simonsohn, Nelson i Simmons, [l'equip de Data Colada] danys per almenys 25 milions de dòlars, en una quantitat que es determinarà en el judici, incloent pèrdues econòmiques, oportunitats professionals perdudes, dany a la reputació, angoixa emocional i danys punitius, costos i honoraris d'advocats (...).".

Considerem que la decisió de Gino de portar el tema als tribunals és desafortunada. El millor curs d'acció per a la comunitat acadèmica seria un debat obert sobre què va passar exactament amb tots aquests estudis. Com que aquest debat ara té lloc als jutjats, imposa un gran cost personal a les persones que van plantejar dubtes certament vàlids. Això és particularment

problemàtic perquè té un efecte inhibidor en el futur escrutini de recerca publicada. Per aquesta raó, recolzar el fons de defensa legal de Data Colada és un bé públic important, al qual tots hauríem de contribuir. Si esteu d'acord amb nosaltres, podeu contribuir a l'equip legal de Data Colada a través d'aquest enllaç. No només ells, sinó també segurament l'economia i la ciència en general en sortiran beneficiades.

[Volem destacar que, més enllà de la denúncia de Gino i les entrades citades de Data Colada, hem recopilat gran part de la informació de l'entrada d'Andrew Ganato del 4 d'agost titulada "Addressing the Data Analysis in Francesca Gino's Data Colada Lawsuit".]