Estadísticas de citas: Un informe de la Unión Matemática Internacional (IMU) en cooperación con el Consejo Internacional de Matemáticas Industriales y Aplicadas (ICIAM) y el Instituto de Estadística Matemática (IMS)*

por

Comité Conjunto de Evaluación Cuantitativa de la Investigación: Robert Adler, John Ewing (presidente) y Peter Taylor

Resumen

Este es un informe sobre el uso y abuso de datos sobre citas en la evaluación de la investigación científica. La idea de que la evaluación científica tiene que hacerse usando métodos «simples y objetivos» es cada vez más frecuente en estos días. Los métodos «simples y objetivos» se interpretan mayoritariamente como *bibliométricos*, esto es, datos sobre citas y estadísticas derivadas de ellos. Hay una creencia en que las estadísticas de citas son inherentemente más exactas porque sustituyen juicios complejos por simples números, y, por tanto, superan la posible subjetividad de la *revisión por pares* (el juicio de compañeros científicos). Pero esta creencia carece de fundamento.

- Confiar en las estadísticas no es más acertado cuando las estadísticas se usan incorrectamente. De hecho, las estadísticas pueden engañar cuando se aplican mal o no se entienden. Buena parte de la bibliométrica moderna parece confiar en la experiencia y la intuición sobre la interpretación y la validez de las estadísticas de citas.
- Mientras que los números parecen «objetivos», su objetividad puede ser ilusoria. El significado de una cita puede ser incluso más subjetivo que la revisión por pares. Como esta subjetividad es menos obvia para las citas, es menos probable que los que usan datos de citas entiendan sus limitaciones.
- La única confianza en datos sobre citas proporciona, en el mejor de los casos, una comprensión incompleta y a menudo superficial de la investigación, una comprensión que es válida solo cuando está reforzada por otros juicios. *Los números no son inherentemente superiores a juicios sensatos.*

*Esta es una traducción del informe *Citation Statistics*, publicado el 11 de junio de 2008 por la Unión Matemática Internacional (`http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf`), realizada por Ramón Esteban Romero con permiso de la Unión Matemática Internacional.

Usar datos sobre citas para evaluar la investigación significa, en última instancia, usar estadísticas basadas en datos para ordenar cosas: revistas, artículos, personas, programas y disciplinas. Las herramientas estadísticas usadas para ordenar estas cosas en muchos casos se entienden y se usan de manera incorrecta.

- En cuanto a las revistas, el factor de impacto se usa a menudo para ordenarlas. Este índice es una simple media derivada de la distribución de citas para una colección de artículos en la revista. La media refleja solo una pequeña cantidad de información sobre esta distribución, y es una estadística algo rudimentaria. Además, hay varios factores desconcertantes cuando se juzgan revistas por citas, y cualquier comparación de revistas requiere precaución cuando se usan factores de impacto. Usar únicamente el factor de impacto para juzgar una revista es como usar únicamente el peso para juzgar la salud de una persona.
- En cuanto a los artículos, en vez de confiar en el número real de citas para comparar artículos individuales, la gente lo sustituye con frecuencia por los factores de impacto de las revistas en que los artículos aparecen. Cree que factores de impacto más altos significan mayores números de citas. ¡Pero con frecuencia este *no* es el caso! Ese es un uso incorrecto de las estadísticas muy extendido, pero que debe ser cuestionado cuando y donde ocurra.
- En cuanto a los científicos, los historiales completos de citas pueden ser difíciles de comparar. Como consecuencia, ha habido intentos de encontrar estadísticas simples que reflejen la complejidad total del historial de citas de un científico con un solo número. El más notable de estos es el índice h, que parece ir ganando popularidad. Pero incluso una inspección superficial del índice h y sus variantes muestra que estos son intentos ingenuos de entender historiales de citas complicados. Mientras que contienen una pequeña cantidad de información sobre la distribución de las citas de un científico, pierden información crucial que es esencial para la evaluación de la investigación.

La validez de las estadísticas como el factor de impacto y el índice h no han sido ni bien entendidas ni bien estudiadas. La conexión de estas estadísticas con la calidad de la investigación se establece a veces sobre la base de la «experiencia». La justificación para confiar en ellas es que están «fácilmente disponibles». Los pocos estudios que se han hecho de estas estadísticas se centran específicamente en mostrar una correlación con algunas otras medidas de calidad, en vez de en determinar cómo se puede derivar información útil de los datos sobre citas de la mejor manera.

Nosotros no rechazamos las estadísticas de citas como una herramienta para evaluar la calidad de la investigación: los datos y las estadísticas de citas pueden proporcionar alguna información valiosa. Reconocemos que la evaluación tiene que ser práctica y, por esta razón, las estadísticas de citación que se obtengan con facilidad serán casi con seguridad parte del proceso. Pero los datos sobre citas proporcionan solo una visión limitada e incompleta de la calidad de la investigación, y las estadísticas derivadas de los datos sobre citas a veces se entienden mal y se usan incorrectamente. La investigación es demasiado importante para medir su valor únicamente con una herramienta tan burda.

Esperamos que aquellos que estén implicados en la evaluación lean tanto el comentario como los detalles de este informe para entender no solo las limitaciones de las estadísticas de citas, sino también cómo usarlas mejor. Si fijamos niveles altos para los usos de la ciencia, por supuesto debemos fijar igualmente niveles altos para evaluar su calidad.

Del encargo del comité:

El impulso hacia una mayor transparencia y responsabilidad en el mundo académico ha creado una «cultura de números» en la que las instituciones y los individuos creen que se pueden alcanzar decisiones justas mediante una evaluación algorítmica de algunos datos estadísticos; incapaces de medir la calidad (la meta final), los que tienen que tomar las decisiones sustituyen la calidad por números que pueden medir. Esta tendencia pide el comentario de aquellos que «se dedican a los números» profesionalmente: matemáticos y estadísticos.

Introducción

La investigación científica es importante. La investigación se halla en la base de buena parte del progreso en nuestro mundo moderno y proporciona la esperanza de que podamos resolver algunos de los problemas a los que se enfrenta la humanidad y que parecen intratables, desde el medio ambiente hasta nuestra población en expansión. Por ello, los gobiernos y las instituciones de todo el mundo proporcionan un apoyo financiero considerable a la investigación científica. Naturalmente, quieren saber que su dinero se está invirtiendo sabiamente; quieren evaluar la calidad de la investigación que pagan y así tomar decisiones bien justificadas sobre inversiones futuras.

Esto no es nada nuevo: la gente ha estado evaluando la investigación durante muchos años. Lo que *es* nuevo, sin embargo, es la noción de que la buena evaluación debe ser «simple y objetiva», y de que esta se puede alcanzar confiando fundamentalmente en métricas (estadísticas) derivadas de los datos sobre citas en vez de en una variedad de métodos, incluyendo juicios por los propios científicos. El primer párrafo de un informe reciente indica crudamente esta visión:

> Es intención del Gobierno que el método actual de determinación de la calidad de la investigación universitaria —la Operación de Evaluación de la Investigación del Reino Unido (RAE)— se sustituya después de que se complete el siguiente ciclo en 2008. Las métricas, en vez de la revisión por pares, serán el centro del nuevo sistema, y se espera que las bibliométricas (uso de números de artículos en revistas y sus citas) sean un índice de calidad central en este sistema.
>
> Evidence Report, [12, p. 3]

Los que argumentan a favor de esta simple objetividad creen que la investigación es demasiado importante como para confiar en juicios subjetivos. Creen que las métricas basadas en citas dan claridad al proceso de ordenación y eliminan ambigüeda-

des inherentes a otras formas de evaluación. Creen que las métricas cuidadosamente escogidas son independientes e imparciales. La mayoría cree que estas métricas permiten comparar todas las partes del proceso de investigación —revistas, artículos, personas, programas, e incluso disciplinas enteras— simple y efectivamente, sin el uso de una revisión por pares subjetiva.

Pero esta fe en la exactitud, independencia y eficacia de las métricas está fuera de lugar.

- En primer lugar, la exactitud de estas métricas es ilusoria. Es una máxima común que las estadísticas pueden mentir cuando se usan incorrectamente. El uso incorrecto de las estadísticas de citas está muy extendido y es notorio. A pesar de los repetidos intentos para alertar contra este uso incorrecto (por ejemplo, el uso incorrecto del factor de impacto), los gobiernos, las instituciones e incluso los mismos científicos continúan obteniendo conclusiones injustificables e incluso falsas de la aplicación incorrecta de estadísticas de citas.
- En segundo lugar, la confianza única en las métricas basadas en citas sustituye un tipo de juicio por otro. En lugar de la revisión por pares subjetiva, uno tiene la interpretación subjetiva del significado de una cita. Los que promueven la confianza exclusiva en las métricas basadas en citas suponen implícitamente que cada cita significa lo mismo sobre la investigación citada: su «impacto». Esta es una suposición que no está demostrada, y muy probablemente es incorrecta.
- En tercer lugar, mientras que las estadísticas son valiosas para entender el mundo en que vivimos, proporcionan solo una comprensión parcial. En nuestro mundo moderno, a veces está de moda asegurar una creencia mística en que las medidas numéricas son superiores a las demás formas de comprensión. Los que promueven el uso de estadísticas de citas como *sustitución* de un entendimiento más completo de la investigación mantienen implícitamente tal creencia. No solo necesitamos usar las estadísticas *correctamente*, también necesitamos usarlas prudentemente.

Nosotros no discutimos el esfuerzo para evaluar la investigación, sino más bien la demanda de que estas evaluaciones confíen predominantemente en métricas basadas en citas que sean «simples y objetivas», una demanda que a menudo se interpreta como necesitada de números fácilmente calculables que ordenen publicaciones, personas o programas. **La investigación usualmente tiene muchos objetivos, tanto a corto como a largo plazo, y por tanto es razonable que su valor deba ser juzgado mediante criterios múltiples**. Los matemáticos saben que hay muchas cosas, tanto reales como abstractas, que no pueden ordenarse simplemente, en el sentido de que dos de ellas puedan ser comparadas. La comparación requiere a menudo de un análisis más complicado, que algunas veces deja a uno sin una decisión sobre cuál de las dos cosas es «mejor». La correcta respuesta a «¿Cuál es mejor?» es a veces: «¡Depende!».

La súplica de usar múltiples métodos para evaluar la calidad de la investigación ya se ha hecho antes (por ejemplo, [25] o [9]). Las publicaciones se pueden juzgar

de muchas maneras, no solo por las citas. Estimaciones tales como invitaciones, pertenencia a comités editoriales y premios a menudo miden la calidad. En algunas disciplinas y en algunos países, la financiación de becas puede desempeñar un papel. Y la revisión por pares es una componente importante de la evaluación. (No debemos descartar la revisión por pares simplemente porque a veces está condicionada por la parcialidad, así como no debemos descartar las estadísticas de citas porque a veces estén devaluadas por el uso incorrecto.) Esta es una pequeña muestra de los múltiples caminos por los que se puede llevar a cabo la evaluación. Hay muchas formas de hacer una buena evaluación, y su importancia relativa varía entre disciplinas. A pesar de esto, las estadísticas «objetivas» basadas en citas repetidamente se convierten en el método preferido para la evaluación. El atractivo de un proceso simple y números simples (preferentemente un único número) parece superar el sentido común y el buen juicio.

Este informe ha sido escrito por matemáticos y trata sobre el uso incorrecto de las estadísticas en la evaluación de la investigación científica. Por supuesto, este uso incorrecto está dirigido a veces hacia la propia disciplina de las matemáticas, y esta es una de las razones para escribir este informe. La especial cultura de citas de las matemáticas, con unos números bajos de citas para revistas, artículos y autores, las hace especialmente vulnerables al abuso de las estadísticas de citaciones. Creemos, sin embargo, que *todos* los científicos, así como el público general, deberían estar ansiosos por usar métodos científicos sensatos para la evaluación de la investigación.

Algunos en la comunidad científica prescindirían totalmente de las estadísticas de citas en una reacción cínica a los abusos del pasado, pero hacer esto significaría descartar una herramienta valiosa. Las estadísticas basadas en citas *pueden* desempeñar un papel en la evaluación de la investigación, siempre que se usen adecuadamente, se interpreten con precaución y compongan solo una parte del proceso. Las citas proporcionan información sobre revistas, artículos y personas. No queremos ocultar esta información, queremos iluminarla.

Este es el propósito de este informe. Las primeras tres secciones tratan los modos en que las citas de datos pueden usarse (correcta e incorrectamente) para evaluar revistas, artículos y personas. La siguiente sección discute los significados variados de las citas y las consecuentes limitaciones de las estadísticas basadas en citas. La última sección aconseja sobre el uso prudente de las estadísticas e insta a que la evaluación atenúe el uso de las estadísticas de citas con otros juicios, incluso aunque haga las evaluaciones menos simples.

«Todo debe hacerse tan simple como sea posible, pero no más simple», dijo una vez Albert Einstein.[1] Este consejo de uno de los más eminentes científicos del mundo es especialmente apto cuando se evalúa la investigación científica.

Ordenación de revistas: El factor de impacto[2]

El factor de impacto se creó en la década de los 60 del pasado siglo como una manera de medir el valor de las revistas calculando el número medio de citas por artículo a lo largo de un período especificado de tiempo [18]. La media se calcula a

partir de datos obtenidos por *Thomson Scientific* (antes llamado *The Institute for Scientific Information*), que publica *Journal Citation Reports*. *Thomson Scientific* extrae referencias de más de 9 000 revistas, añadiendo información sobre cada artículo y sus referencias a su base de datos cada año [35]. Usando esta información, se puede contar con qué frecuencia un artículo es citado por otros artículos que se publican en la colección de revistas indexadas. (Notemos que *Thomson Scientific* indexa menos de la mitad de los artículos cubiertos por *Mathematical Reviews* y *Zentralblatt*, las dos revistas de recensiones de matemáticas más importantes.[3])

Para una revista y un año particulares, el factor de impacto de la revista se obtiene calculando el número medio de citas a artículos en la revista durante los dos años precedentes a partir de todos los artículos publicados en ese año dado (en la colección particular de revistas indexada por *Thomson Scientific*). Si el factor de impacto de una revista es 1,5 en 2007, esto significa que en promedio los artículos publicados durante 2005 y 2006 fueron citados 1,5 veces por los artículos de la colección de todas las revistas indexadas publicadas en 2007.

La propia *Thomson Scientific* usa el factor de impacto como un factor a la hora de seleccionar qué revistas indexar [35]. Por otra parte, Thomson promueve el uso del factor de impacto más generalmente para comparar revistas.

> «Como herramienta para la administración de colecciones de revistas en bibliotecas, el factor de impacto proporciona al administrador de la biblioteca información sobre revistas que ya están en la colección y revistas consideradas para su adquisición. Estos datos deben combinarse con los datos de coste y circulación para tomar decisiones racionales sobre compras de revistas.» Thomson, [36]

Algunos escritores han señalado que no se debe juzgar el valor académico de una revista usando datos sobre citas únicamente, y los presentes autores están muy de acuerdo con esta opinión. Además de esta observación general, el factor de impacto ha sido criticado también por otras razones. (Véase [31], [2], [29], [13], [1] y [20].)

1. La identificación del factor de impacto como una media no es absolutamente correcta. Como muchas revistas publican artículos no sustanciales como cartas y editoriales, que son poco citados, estos artículos no cuentan en el denominador del factor de impacto. Por otro lado, a pesar de que es poco frecuente, estos artículos a veces se citan, y estas citas *sí* se cuentan en el numerador. El factor de impacto no es, por tanto, exactamente el promedio de citas por artículo. Cuando las revistas publican un gran número de estos artículos no sustanciales, la desviación puede ser significativa. En muchas áreas, incluyendo las matemáticas, esta desviación es mínima.

2. El período de dos años usado en la definición del factor de impacto se pensó para que la estadística estuviera actualizada [18]. Para algunos campos, como las ciencias biomédicas, esto es apropiado porque la mayoría de los artículos publicados reciben la mayoría de sus citas poco tiempo después de la publicación. En otros campos, como las matemáticas, la mayoría de las citas aparecen tras el período de dos años. Examinando una colección de más de 3 millones de citas recientes en revistas matemáticas (la base de datos *Math Reviews Citation*), se ve que a grandes

rasgos el 90 % de las citas caen fuera de esta ventana de 2 años (figura 1). En consecuencia, el factor de impacto se basa en un mero 10 % de la actividad de citas y pierde la gran mayoría de las citas.[4]

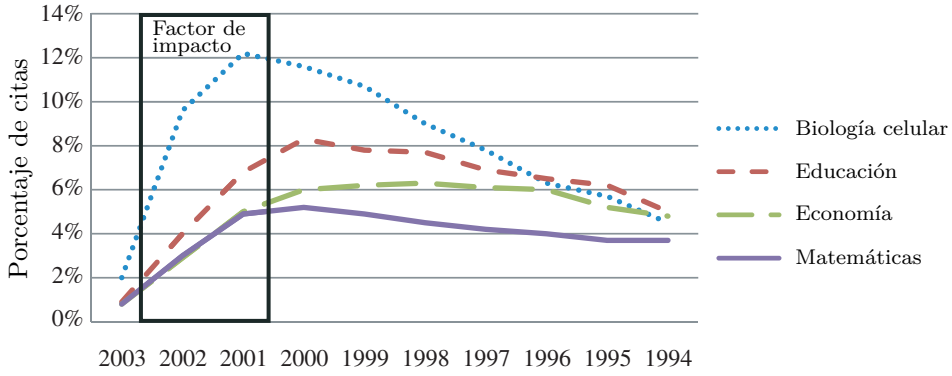

Figura 1: Una gráfica que muestra la edad de las citas de artículos publicados en 2003 que cubre cuatro campos diferentes. Las citas a artículos publicados en 2001–2002 son aquellas que contribuyen al factor de impacto; todas las demás citas son irrelevantes para el factor de impacto. Datos de Thomson Scientific.

¿El intervalo de dos años significa que el factor de impacto es engañoso? Para revistas matemáticas la evidencia es equívoca. *Thomson Scientific* calcula factores de impacto de 5 años, que señalan que se correlacionan bastante bien con los factores de impacto usuales (de 2 años) [17]. Usando la base de datos de citas de Math Reviews, uno puede calcular «factores de impacto» (esto es, citas medias por artículo) para una colección de las 100 revistas matemáticas más citadas usando períodos de 2, 5 y 10 años. La gráfica de la figura 2 muestra que los factores de impacto de 5 y 10 años generalmente siguen la pista del factor de impacto de 2 años.

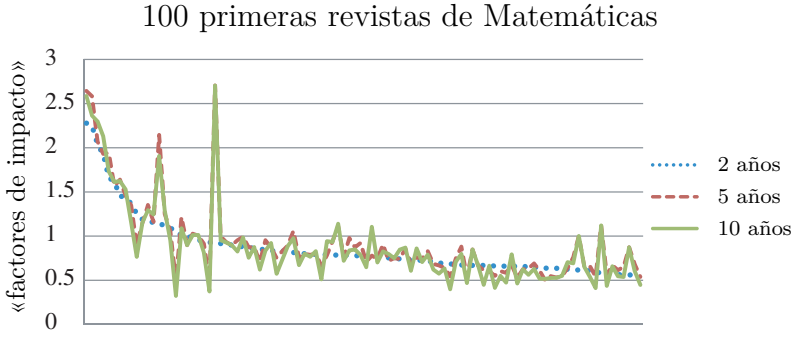

Figura 2: «Factores de impacto» para 2, 5 y 10 años para 100 revistas matemáticas. Datos de la base de datos de citas de Math Reviews.

El pico más grande es una revista que no publicó artículos durante parte de este tiempo; los picos pequeños tienden a ser revistas que publican un número relativamente pequeño de artículos cada año, y la gráfica simplemente refleja la variabilidad normal en factores de impacto para estas revistas. Es evidente que cambiar el número de «años objetivo» al calcular el factor de impacto cambia la ordenación de las revistas, pero los cambios son generalmente modestos, excepto para revistas pequeñas, donde los factores de impacto también varían cuando cambia el «año fuente» (véase más adelante).

3. El factor de impacto varía considerablemente entre disciplinas [2]. Parte de esta diferencia proviene de la observación 2: si en algunas disciplinas muchas citas aparecen fuera de la ventana de dos años, los factores de impacto para sus revistas serán mucho menores. Por otro lado, parte de la diferencia es simplemente que las culturas de citas difieren de disciplina en disciplina, y los científicos citarán artículos a diferentes ritmos y por diferentes razones. (Daremos más detalles después porque el significado de las citas es extremadamente importante.) Se sigue que no se pueden comparar dos revistas en diferentes disciplinas de una manera significativa usando factores de impacto (figura 3).

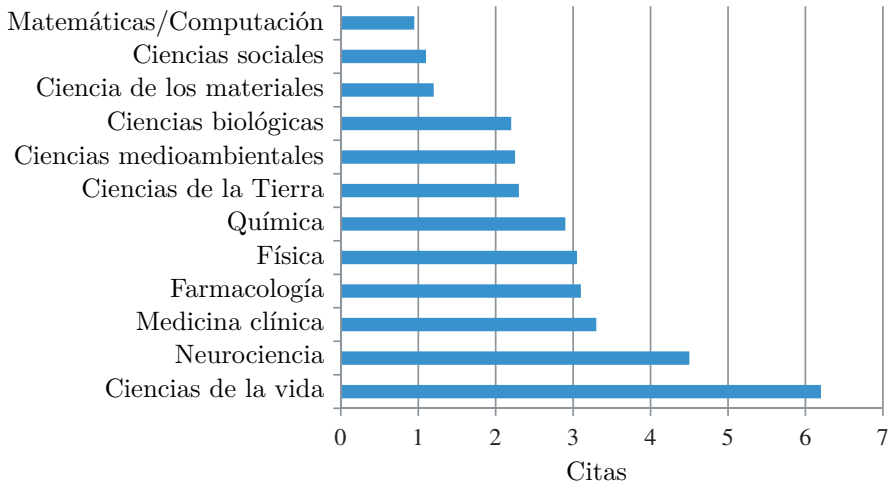

Figura 3: Citas medias por artículo para diferentes disciplinas, que muestran que las prácticas de citas difieren de una manera muy significativa. Datos de Thomson Scientific [2].

4. El factor de impacto puede variar considerablemente de un año a otro, y la variación tiende a ser mayor para revistas pequeñas [2]. Para revistas que publican menos de 50 artículos, por ejemplo, el *cambio* medio en el factor de impacto de 2002 a 2003 fue próximo al 50 %. Esto es completamente esperable, por supuesto, porque el tamaño de la muestra para revistas pequeñas es pequeño. Por otro lado, a veces se comparan revistas para un año fijado, sin tener en cuenta la mayor variación de las revistas pequeñas.

5. Las revistas que publican artículos en idiomas distintos del inglés muy probablemente recibirán menos citaciones porque una gran parte de la comunidad científica no puede leerlas (o no las leen). Y el tipo de revista, más que simplemente la calidad, puede influir en el factor de impacto. Revistas que publican artículos de revisión, por ejemplo, recibirán con frecuencia más citas que revistas que no lo hacen, y por tanto tendrán factores de impacto más altos (a veces, sustancialmente más altos) [2].

6. La crítica más importante del factor de impacto es que no se entiende bien su significado. Cuando se usa el factor de impacto para comparar dos revistas, no hay ningún modelo *a priori* que defina qué significa ser «mejor». El único modelo se deriva del propio factor de impacto: un factor de impacto más grande significa una revista mejor. En el paradigma estadístico clásico, uno define un modelo, formula una hipótesis (de ninguna diferencia) y entonces encuentra un estadístico, que dependiendo de sus valores permite decidir si se acepta o se rechaza la hipótesis. Derivar información (y posiblemente un modelo) de los propios datos es una aproximación legítima al análisis estadístico, pero en este caso no está claro qué información se ha obtenido. ¿Cómo mide la calidad el factor de impacto? ¿Es el mejor estadístico para medir la calidad? ¿Qué es lo que *sí* mide? (Nuestra discusión posterior sobre el significado de las citas es relevante aquí.) Es importante remarcar que se conoce poco sobre un modelo para la calidad de las revistas o cómo podría estar relacionado con el factor de impacto.

Las anteriores seis críticas al factor de impacto son todas válidas, pero solo significan que el factor de impacto es rudimentario, no inútil. Por ejemplo, el factor de impacto puede usarse como un punto de partida en la ordenación de revistas en grupos, usando los factores de impacto inicialmente para definir los grupos, y empleando entonces otros criterios para refinar la ordenación y verificar que los grupos tienen sentido. Pero usar el factor de impacto para evaluar revistas requiere precaución. El factor de impacto no puede usarse para comparar revistas entre disciplinas, por ejemplo, y se debe mirar detenidamente el tipo de revistas cuando se usa el factor de impacto para ordenarlas. También se debe prestar una permanente atención a las variaciones anuales, especialmente para las revistas más pequeñas, y entender que las pequeñas diferencias pueden ser fenómenos puramente aleatorios. Y es importante reconocer que el factor de impacto puede que no refleje exactamente el rango completo de actividad de citas en algunas disciplinas, tanto porque no todas las revistas están indexadas como porque el período es tan corto. Otras estadísticas basadas en períodos de tiempo más largos y más revistas pueden ser mejores indicadores de calidad. Finalmente, las citas son solo una manera de juzgar revistas, y deben ser complementadas con otra información (el mensaje central de este informe).

Todas estas precauciones son semejantes a las que hay que tomar para cualquier ordenación basada en estadísticas. Ordenar sin criterio las revistas según el factor de impacto para un año particular es un uso incorrecto de la estadística. En su favor, *Thomson Scientific* está de acuerdo con esta frase y (suavemente) advierte de todas estas cosas a aquellos que usan el factor de impacto.

> «Thomson Scientific no se basa únicamente en el factor de impacto únicamente en la evaluación de la utilidad de una revista, y tampoco debería

> hacerlo nadie más. El factor de impacto no debería usarse sin prestar cuidadosa atención a los muchos fenómenos que pueden influir en las tasas de citación, como por ejemplo el número medio de referencias citadas en un artículo medio. El factor de impacto debería ser usado junto con una justificada revisión por pares.» Thomson, [36]

Desgraciadamente, esta advertencia no se tiene en cuenta con demasiada frecuencia.

Ordenación de artículos

El factor de impacto y estadísticas similares basadas en citas pueden usarse incorrectamente cuando ordenamos revistas, pero hay un uso incorrecto más fundamental y más insidioso: el uso del factor de impacto para comparar artículos individuales, personas, programas o incluso disciplinas. Este es un problema creciente que se extiende a lo largo de muchos paises y muchas disciplinas, empeorado por las recientes evaluaciones nacionales de la investigación.

En un sentido, este no es un fenómeno nuevo. A veces se apela a científicos para que hagan juicios sobre historiales investigadores, y se oyen comentarios como «Ella publica en buenas revistas» o «La mayoría de sus artículos están en revistas de bajo nivel». Estas pueden ser evaluaciones sensatas: la calidad de las revistas en las que un científico publica generalmente (o consistentemente) es uno de los muchos factores que se pueden usar para evaluar la investigación total de ese científico. El factor de impacto, sin embargo, ha incrementado la tendencia a atribuir las propiedades de una revista a *cada* artículo de esa revista (y a *cada* autor).

Thomson Scientific promueve implícitamente esta práctica:

> «Quizás el uso más importante y reciente del factor está en el proceso de evaluación académica. El factor de impacto puede usarse para dar una burda aproximación del prestigio de las revistas en las que los individuos han publicado.» Thomson, [36]

Aquí hay algunos ejemplos de algunas maneras en las que la gente ha interpretado este consejo, de los que hemos sido informados por matemáticos de todo el mundo:

Ejemplo 1: Mi universidad ha introducido recientemente una nueva clasificación de revistas que usa las revistas del Science Citation Index Core. Las revistas están divididas en tres grupos basados solo en el factor de impacto. Hay 30 revistas en la lista superior, que no contiene ninguna revista matemática. La segunda lista contiene 667, que incluye 21 revistas matemáticas. La publicación en la primera lista hace que el apoyo de la universidad a la investigación sea el triple; la publicación en la segunda lista, el doble. La publicación en la lista principal se recompensa con 15 puntos; la publicación en cualquier revista cubierta por *Thomson Scientific* merece 10. La promoción requiere un número mínimo de puntos fijado.

Ejemplo 2: En mi país, los profesores de universidad con plazas fijas son evaluados cada seis años. Sucesivas evaluaciones positivas son clave para el éxito académico. Además del currículum vítæ, el mayor factor de evaluación se refiere a la

posición de cinco artículos publicados. En los últimos años, cada uno de esos artículos recibe 3 puntos si aparece en revistas del tercio superior de la lista de *Thomson Scientific*, 2 puntos si está en el segundo tercio, y un punto en el último tercio. (Las tres listas han sido creadas usando el factor de impacto.)

Ejemplo 3: En nuestro departamento, cada profesor es evaluado mediante una fórmula en la que aparecen el número de artículos individuales (con sus equivalentes), multiplicado por el factor de impacto de las revistas en las que aparecen. Las promociones y las contrataciones en parte están basadas en esta fórmula.

En estos ejemplos, así como en muchos otros de los que hemos sido informados, el factor de impacto se usa bien explícitamente, bien implícitamente para comparar artículos individuales junto con sus autores: si el factor de impacto de la revista A es mayor que el de la revista B, entonces seguramente un artículo en A debe ser superior a un artículo en B, y el autor A superior al autor B. En algunos casos, este razonamiento se extiende a la ordenación de departamentos o incluso disciplinas enteras.

Se ha sabido desde hace tiempo que la distribución de los números de citas para artículos individuales en una revista es altamente sesgada, aproximándose a lo que se llama una ley potencial ([31], [16]). Esto tiene consecuencias que podemos precisar con un ejemplo.

La distribución de artículos en *Proceedings of the American Mathematical Society* durante el período 2000–2004 puede verse en la figura 4. *Proceedings* publica artículos cortos, normalmente de menos de diez páginas de longitud. Durante este período, ha publicado 2 381 artículos (alrededor de 15 000 páginas). Usando las revistas de 2005 en la base de datos Math Reviews, el número de citas medio por artículo (esto es, el factor de impacto), es 0,434.

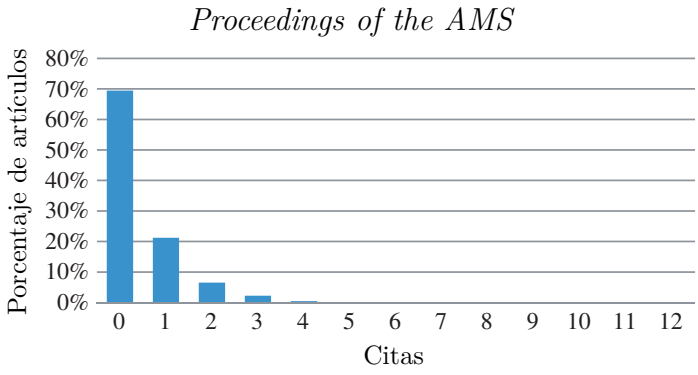

Figura 4: Distribución de citas para artículos en *Proceedings of the American Mathematical Society* durante el período 2000–2004.

Transactions of the AMS publica artículos más largos que habitualmente son más sustanciales, tanto en alcance como en contenido. A lo largo del mismo período de tiempo, las *Transactions* publicaron 1 165 artículos (más de 25 000 páginas), cuyos

números de citas variaban entre 0 y 12 (figura 5). El número medio de citas por artículo fue 0,846, más o menos el doble que el de *Proceedings*.

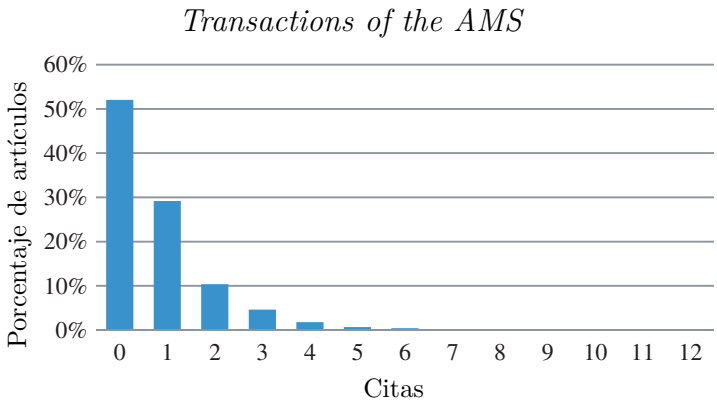

Figura 5: Distribución de citas para artículos en *Transactions of the American Mathematical Society* durante el período 2000–2004

Consideremos ahora a dos matemáticos, uno que publique un artículo en Proceedings y el otro un artículo en Transactions. Usando alguna de las prácticas institucionales citadas anteriormente, el segundo sería juzgado superior al primero, al publicar un artículo en una revista con factor de impacto más alto, de hecho, ¡el doble de alto! ¿Es esta una evaluación válida? ¿Son los artículos de *Transactions of the AMS* el doble de buenos que los de *Proceedings*?

Cuando aseguramos que un artículo individual de *Transactions* es mejor (en el sentido de citas) que un artículo individual de *Proceedings*, no hacemos una pregunta sobre medias, sino más bien una cuestión sobre probabilidades: ¿Cuál es la probabilidad de que estemos equivocados? ¿Cuál es la probabilidad de que un artículo de *Proceedings* seleccionado al azar tenga al menos tantas citas como un artículo de *Transactions* elegido al azar?

Esto es un cálculo elemental, y la respuesta es 62 %. Esto significa que estamos equivocados el 62 % de las veces, y un artículo de *Proceedings* seleccionado al azar será exactamente tan bueno (o mejor) que un artículo de *Transactions* elegido al azar, ¡a pesar de que el factor de impacto de *Proceedings* es sólo la mitad del de *Transactions*! Estamos equivocados más veces de las que tenemos la razón. *La mayoría* de la gente encuentra esto sorprendente, pero es una consecuencia de la distribución altamente sesgada y lo estrecha que es la ventana de tiempo que se usa para calcular el factor de impacto (que es la razón del alto porcentaje de artículos sin citar).[5] Muestra el valor del pensamiento estadístico preciso en lugar de la observación intuitiva.

Este es un comportamiento típico para las revistas, y no hay nada especial en relación con las elecciones de estas dos revistas. (Por ejemplo, *Journal of the AMS* a lo largo del mismo período tiene un factor de impacto de 2,63, seis veces mayor que el de *Proceedings*. Sin embargo, un artículo elegido al azar de *Proceedings* es al

menos tan bueno como un artículo de *Journal*, en el sentido de las citas, el 32 % de las veces.)

Así, **aunque es incorrecto decir que el factor de impacto no proporciona información sobre artículos individuales en una revista, la información es sorprendentemente vaga y puede ser espectacularmente engañosa**.

Se sigue que los tipos de cálculos realizados en los tres ejemplos anteriores —que usan el factor de impacto como representante de los números reales de citas para artículos individuales— tienen poca base racional. Hacer afirmaciones que son incorrectas más de la mitad de las veces (o un tercio de las veces) seguramente no es un buen camino para llevar a cabo una evaluación.

Una vez que se observa que no tiene sentido sustituir el número de citas individual por el factor de impacto, se entiende que no tiene sentido usar el factor de impacto para evaluar a los autores de esos artículos, los proyectos en los que trabajan o (con casi toda seguridad) las disciplinas que representan. El factor de impacto y las medias en general son demasiado rudimentarias para hacer comparaciones sensatas sin más información.

Por supuesto, ordenar a las personas no es lo mismo que ordenar sus artículos. Pero si se quiere ordenar estos usando solo citas para medir la calidad de un determinado artículo, hay que empezar por contar las citas de dicho artículo. El factor de impacto de la revista en la cual aparece el artículo no es un sustituto en el que podamos confiar.

Ordenación de científicos

Mientras que el factor de impacto ha venido siendo la estadística basada en citas más conocida, hay otras estadísticas más recientes cuyo uso se están promoviendo activamente. Aquí tenemos una pequeña muestra de tres de estas estadísticas pensadas para ordenar individuos.

Índice h: El índice h de un científico es el mayor n para el cual ha publicado n artículos, cada uno de ellos con al menos n citas.

Esta es la más popular de las estadísticas mencionadas aquí. Fue propuesta por J. E. Hirsch [21] para medir «la producción científica de un investigador» prestando atención a la larga «cola» de la distribución de citas de una persona. El objetivo era sustituir los números de publicaciones y la distribución de citas por un único número.

Índice m: El índice m de un científico es el índice h dividido por el número de años desde su primer artículo. Fue propuesto también por Hirsch en el artículo antes mencionado. La intención es compensar a los científicos jóvenes porque no han tenido tiempo de publicar artículos o de conseguir muchas citas.

Índice g: El índice g de un científico es el mayor n para el cual los n artículos más citados tienen un total de al menos n^2 citas.

> Esto fue propuesto por Leo Egghe en 2006 [11]. El índice h no tiene en cuenta el hecho de que algunos artículos entre los n primeros pueden tener un número de citas extraordinariamente grande. El índice g pretende compensar esto.

Hay más índices —muchos más índices— que incluyen variantes de los anteriores y que tienen en cuenta la edad de los artículos o el número de autores ([3], [4], [32]).

En su artículo en que definió el índice h, Hirsch escribió que proponía que el índice h fuera «un índice fácilmente calculable, que diera una estimación de la importancia, significación y alto impacto de las contribuciones de investigación acumulada de un científico» [21, p. 5]. Continuó añadiendo que «este índice puede proporcionar un criterio útil para comparar a diferentes individuos que compitan por el mismo recurso cuando un importante criterio de evaluación sea el logro científico».

Ninguna de estas afirmaciones está apoyada por una evidencia convincente. Para apoyar su afirmación de que el índice h mide la importancia y significación de la investigación acumulada de un científico, Hirsch analiza el índice h para una colección de ganadores del Premio Nobel (y, separadamente, miembros de la Academia Nacional). Demuestra que las personas en estos grupos generalmente tienen índices h altos. Se puede concluir que es bastante probable que un científico tenga un índice h alto dado que el científico ha sido galardonado con el Nobel. Pero, sin más información, sabemos bastante poco sobre la probabilidad de que alguien reciba el Premio Nobel o sea un miembro de la Academia Nacional sabiendo que tiene un índice h alto. Este es el tipo de información que se requiere para establecer la validez del índice h.

En su artículo, Hirsch afirma también que se puede usar el índice h para comparar a dos científicos:

> «Argumento que dos individuos con un índice h similar son comparables en términos de su impacto científico global, incluso si su número total de citas es muy diferente. Recíprocamente, que entre dos individuos (de la misma edad científica) con número similar de artículos en total o de número total de citas y muy diferente valor h, el que tiene el mayor h es probablemente el científico más consumado.» Hirsch, [21, p. 1]

Estas aseveraciones parecen ser refutadas por el sentido común. (Pensemos en dos científicos, cada uno de ellos con 10 artículos con 10 citas, pero uno con 90 artículos adicionales con 9 citas cada uno; o supongamos que uno tiene exactamente 10 artículos de 10 citas y el otro exactamente 10 artículos de 100 citas cada uno. ¿Pensaría alguien que son equivalentes?)[6]

Hirsch ensalza las virtudes del índice h afirmando que «h es preferible a otros criterios de un único número comúnmente usados para evaluar la producción científica de un investigador...» [21, p. 1], pero ni define «preferible» ni explica por qué se *quieren* encontrar «criterios de un único número».

Aunque esta aproximación ha sido bastante criticada, ha habido poco análisis serio. La mayoría de este análisis consiste en mostrar «validez convergente», es decir, que el índice h está muy en correlación con métricas de publicación/citas, tales como el número de artículos publicados o el número total de citas. Esta correlación no tiene

nada de especial, ya que estas variables son funciones del mismo fenómeno básico: publicaciones. En un artículo notable sobre el índice h [23], los autores llevan a cabo un análisis más cuidadoso y demuestran que el índice h (en realidad, el índice m) no es tan «bueno» como considerar meramente el número medio de citas por artículo. Incluso aquí, sin embargo, los autores no definen adecuadamente lo que significa el término «bueno». Cuando se aplica el paradigma de la estadística clásica [23], el índice h resulta ser menos fidedigno que otras medidas.

Se han inventado diversas variantes del índice h para comparar la calidad de los investigadores no solo en una disciplina, sino también entre distintas disciplinas ([4], [28]). Otros afirman que el índice h puede usarse para comparar institutos y departamentos [22]. Estos intentos para condensar un historial de citas complejo en un único número son a menudo sorprendentemente ingenuos. De hecho, la principal ventaja de estos nuevos índices sobre simples histogramas de números de citas es que los índices descartan casi todos los detalles de los historiales de citas, y esto hace posible comparar dos científicos cualesquiera. Sin embargo, incluso ejemplos sencillos muestran que la información descartada es necesaria para entender un historial de investigación. Seguramente **la meta cuando se evalúa la investigación tiene que ser la comprensión, no simplemente asegurar que dos personas cualesquiera sean comparables**.

En algunos casos, las agencias de evaluación nacional añaden el índice h o alguna de sus variantes como parte de sus datos. Esto es un uso incorrecto de los datos. Por desgracia, tener un único número para ordenar a cada científico es una idea seductora: puede llegar más fácilmente a un público que a menudo no comprende bien el uso correcto del razonamiento estadístico en situaciones mucho más simples.

El significado de las citas

Los que propugnan las estadísticas de citas como la medida predominante de la calidad científica no responden a la pregunta esencial: ¿Qué significan las citas? Juntan grandes cantidades de datos sobre números de citas, procesan los datos para derivar estadísticas y entonces afirman que el proceso de evaluación resultante es «objetivo». En realidad es la *interpretación* de las citas lo que conduce a la evaluación, y la interpretación se basa en el *significado* de las citas, que es bastante subjetivo.

En la literatura que alienta esta aproximación, es sorprendentemente difícil encontrar afirmaciones claras sobre el significado de las citas.

> «La idea que hay tras la indexación de las citas es fundamentalmente simple. Reconociendo que el valor de la información está determinado por los que la usan, ¿qué mejor manera de medir la calidad del trabajo que midiendo el impacto que produce en la comunidad en general? La población más variada posible en la comunidad académica (o sea, cualquiera que use o cite el material fuente) determina la influencia o impacto de la idea y su creador en nuestro cuerpo de conocimiento.»
>
> Thomson, [37]

> «A pesar de que cuantificar la calidad de los científicos es difícil, la visión general es que es mejor publicar más que menos y que el número de citas de un artículo (en relación con los hábitos de citas en el área) es una medida útil de su calidad.»
>
> Lehmann, Jackson y Lautrup, [23, p. 1003]

> «La frecuencia de las citas refleja el valor de una revista y el uso que de ella se hace... » Garfield, [15, p. 535]

> «Que un médico o un investigador biomédico cite un artículo de una revista, indica que la revista citada lo ha influido de algun modo.»
>
> Garfield, [16, p. 7]

> «Las citas son un reconocimiento de deuda intelectual.» Thomson, [38]

Los términos relevantes son «calidad», «valor», «influencia» y «deuda intelectual». El término «impacto» se ha convertido en la palabra genérica usada para asignar significado a las citas, un término que surgió por primera vez en un artículo corto escrito en 1955 por Eugene Garfield para promover la iniciativa de crear un índice de citas. Escribió:

> «Entonces, en el caso de un artículo altamente significativo, el índice de citas tiene un valor cuantitativo, ya que puede ayudar al historiador a medir la influencia del artículo, esto es, su "factor de impacto".»
>
> Garfield, [14, p. 3]

Está bastante claro que aquí, como en otros sitios, el término «factor de impacto» tiene la intención de sugerir que el artículo que cita ha sido «construido sobre» el trabajo del citado: que las citas son el mecanismo mediante el cual la investigación se propaga y avanza.

Hay mucha documentación sobre el significado real de las citas que sugiere que las citas son más complicadas de lo que estas vagas frases nos inducen a creer. Por ejemplo, en su artículo de 1983 sobre evaluación de la investigación, Martin e Irvine escriben:

> «Subyace a todos estos problemas con el uso de las citas como medida de calidad nuestra ignorancia de las razones de *por qué* los autores citan algunos trabajos y no otros. Los problemas descritos antes... El análisis simple de citas presupone un modelo altamente racional de hacer referencias, en el cual las citas deben reflejar fundamentalmente el reconocimiento científico del trabajo previo de alta calidad o importancia, y los potenciales citadores tienen todos la misma posibilidad de citar artículos particulares... » Martin e Irvine, [26, p. 69]

En su artículo de 1988 sobre el significado de las citas [10], Cozzens asegura que las citas son el resultado de dos sistemas subyacentes a los usos de la publicación científica, uno un sistema de «recompensa» y otro «retórico». El primer tipo tiene el significado más a menudo asociado con una cita: un reconocimiento de que el artículo que cita tiene una «deuda intelectual» hacia el citado. El segundo, sin embargo, tiene

un significado bastante diferente: una referencia a un artículo previo que explica algún resultado, que quizá no es un resultado del autor citado. Estas citas retóricas son simplemente una manera de llevar a cabo una conversación científica, no de establecer una deuda intelectual. Por supuesto, en algunos casos, una cita puede tener ambos significados.

Cozzens hace la observación de que *la mayoría* de las citas son retóricas. Esto se confirma por la experiencia de la mayoría de los matemáticos en ejercicio. (En la base de datos Math Reviews, por ejemplo, alrededor del 30 % de los más de 3 millones de citas son a libros y *no* a artículos de investigación en revistas.) ¿Por qué es esto importante? Porque, al contrario de lo que pasa con las citas de «recompensa», que tienden a referirse a artículos fundamentales, la elección de qué artículo citar retóricamente depende de varios factores: el prestigio del autor citado (el efecto «halo»), la relación entre el artículo que cita y el citado, la disponibilidad de la revista (¿es más fácil que se citen revistas de acceso libre?), la conveniencia de referirse a muchos resultados de un único artículo, y así sucesivamente. Pocos de estos factores están directamente relacionados con la «calidad» del artículo citado.

Incluso cuando las citas son citas de «recompensa», pueden reflejar una variedad de motivos, que incluyen «actualidad, crédito negativo, información operativa, persuasión, crédito positivo, alerta al lector y consenso social» [8]. En la mayoría de los casos, las citas fueron motivadas por más de uno de estos motivos Algunos resultados notables pueden sufrir el efecto de «obliteración», al ser incorporados inmediatamente en el trabajo de otros, que entonces sirve como base para más citas. Otras citas no son recompensas por investigación excelente, sino más bien advertencias sobre resultados o pensamientos defectuosos. El presente informe proporciona muchos ejemplos de estas citas de «advertencia».

La sociología de las citas es un tema complejo, que está lejos del alcance de este informe. Incluso esta somera discusión, sin embargo, muestra que **el significado de las citas no es simple y que las estadísticas basadas en citas no son tan «objetivas» como los proponentes aseguran**.

Algunos podrían argumentar que el significado de las citas es inmaterial porque las estadísticas basadas en citas están altamente en correlación con algunas otras medidas de calidad en la investigación (como la revisión por pares). Por ejemplo, el informe *Evidence* mencionado antes argumenta que las estadísticas de citas pueden (y deben) sustituir otras formas de evaluación debido a esta correlación:

> «Evidence ha argumentado que las técnicas bibliométricas pueden crear indicadores de calidad de la investigación que son congruentes con la percepción del investigador.» Evidence Report, [12, p. 9]

La conclusión parece ser que las estadísticas basadas en citas, a pesar de su significado preciso, deben reemplazar a otros métodos de evaluación, ya que a menudo están de acuerdo con ellos. Aparte de la circularidad de este argumento, la falacia de este razonamiento es fácil de ver.

Uso prudente de las estadísticas

El exceso de confianza entusiasta en métricas (estadísticas) objetivas para evaluar la investigación no es un fenómeno ni nuevo ni aislado. Se describe elocuentemente en el libro popular de 2001 *Condenadas mentiras y estadísticas* (*Damned lies and statistics*) escrito por el sociólogo Joel Best:

> «Hay culturas en las que la gente cree que algunos objetos tiene poderes mágicos; los antropólogos llaman a estos objetos fetiches. En nuestra sociedad, las estadísticas son un tipo de fetiche. Tendemos a mirarlas como si fueran mágicas, como si fueran algo más que simples números. Las tratamos como representaciones potentes de la verdad; actuamos como si destilaran la complejidad y la confusión de la realidad en simples hechos. Usamos las estadísticas para convertir problemas sociales complicados en estimaciones que se entienden más fácilmente, porcentajes y tasas. Las estadísticas dirigen nuestro interés; nos muestran sobre qué tendríamos que preocuparnos y cuánto deberíamos preocuparnos. En cierto sentido, un problema social se convierte en una estadística y, como tratamos las estadísticas como verdaderas e irrefutables, alcanzan un tipo de control mágico y fetichista sobre cómo vemos los problemas sociales. Pensamos en las estadísticas como hechos que descubrimos, no números que creamos.» Best, [6, p. 160]

Esta creencia mística en la magia de las estadísticas de citas puede encontrarse a lo largo de la documentación para procedimientos de evaluación de la investigación, tanto nacionales como institucionales. También puede encontrarse en el trabajo de los que promueven el índice h y sus variantes.

Esta actitud es evidente también en intentos recientes de mejorar el factor de impacto usando algoritmos matemáticos más sofisticados, que incluyen algoritmos de ordenación de páginas, para analizar citas ([5], [33]). Sus proponentes hacen afirmaciones sobre su eficacia que no están justificadas por el análisis y la dificultad para evaluar. Como están basadas en cálculos más complicados, las suposiciones (a menudo ocultas) que hay detrás de ellos no son sencillas de discernir para la mayoría de la gente.[7] Se supone que debemos tratar los números y las ordenaciones con temor reverencial, como verdades en vez de creaciones.

La investigación no es la primera actividad financiada públicamente que está sujeta a examen, y a lo largo de las últimas décadas la gente ha intentado llevar a cabo evaluaciones de rendimiento cuantitativas de todo, desde sistemas educativos (escuelas) hasta la asistencia sanitaria (hospitales e incluso cirujanos individuales). En algunos casos, los estadísticos han intervenido para avisar a los que hacen la medida sobre métricas sensatas y el uso correcto de las estadísticas. **Si se consulta con los médicos al ejercer la medicina, sin duda debería consultarse con los estadísticos (y seguir su consejo) al ejercer la estadística**. Se pueden encontrar dos excelentes ejemplos en [7] y [19]. Mientras que cada uno se ocupa de la evaluación del rendimiento de cosas distintas de la investigación —seguimiento del rendimiento del sector público en el primero y asistencia sanitaria/educación en el

segundo—, cada uno proporciona entendimiento sobre el uso sensato de la estadística en la evaluación académica.

En particular, el artículo de Goldstein y Spiegelhalter trata el uso de las clasificaciones ligueras (ordenaciones, *rankings*) basadas en números simples (por ejemplo, logros de los estudiantes o resultados médicos), y es especialmente relevante en la evaluación de la investigación mediante la ordenación de revistas, artículos o autores usando estadísticas de citas. En su artículo, los autores trazan un esquema de tres partes para cualquier evaluación de rendimiento:

Datos

> «Por elegante que sea, ningún juego de piernas estadístico puede superar insuficiencias básicas ni en la *oportunidad* ni en la *integridad* de los datos recogidos.» Goldstein y Spiegelhalter, [19, p. 389]

Esta es una observación importante para la evaluación del rendimiento basada en citas. El factor de impacto, por ejemplo, está basado en un subconjunto de datos, que incluye solo aquellas revistas seleccionadas por *Thomson Scientific*. (Notamos que el factor de impacto en sí mismo es la parte que más influye en los criterios de selección.) Algunos han cuestionado la integridad de estos datos [30]. Otros apuntan que otros conjuntos de datos podrían ser más completos [27]. Muchos grupos han apoyado la idea de usar Google Académico para implementar estadísticas basadas en citas, como el índice h, pero los datos contenidos en Google Académico son a menudo inexactos (porque cosas como los nombres de los autores son extraídas automáticamente de envíos a la red). Las estadísticas de citas para científicos individuales son a menudo difíciles de obtener porque los autores no son identificados unívocamente y, en algunos escenarios y en algunos países, esto puede ser un enorme impedimento para concertar con exactitud datos sobre citas. En ocasiones, la colección particular de datos que se usa para el análisis de citas se pasa por alto. Es probable sacar conclusiones defectuosas de estadísticas basadas en datos defectuosos.

Análisis estadístico y presentación

> «Prestaremos particular atención a la especificación de un *modelo* estadístico apropiado, la importancia crucial de la *incertidumbre* en la presentación de todos los resultados, técnicas para el *ajuste* de resultados para factores confusos y finalmente hasta dónde se puede confiar en *ordenaciones* explícitas.»

Como hemos escrito previamente, en muchos casos en los que las estadísticas de citas se usan para ordenar artículos, personas y programas, no se especifica ningún modelo por adelantado. En su lugar, los propios datos sugieren un modelo, que es a menudo vago. Un proceso circular parece ordenar objetos más arriba porque están ordenados más arriba (en la base de datos). Frecuentemente se dedica escasa atención a la incertidumbre en *cualquiera* de estas ordenaciones, y poco análisis de cómo esta incertidumbre (por ejemplo, variaciones anuales en el factor de impacto) afectaría a

las ordenaciones. Finalmente, se hace caso omiso de factores confusos (por ejemplo, la disciplina particular, el tipo de artículos que una revista publica, si un científico particular es un experimentalista o un teórico) en estas ordenaciones, especialmente cuando se llevan a cabo en evaluaciones de rendimiento nacionales.

Interpretación e impacto

> «Las comparaciones discutidas en el artículo son de gran interés público, y esta es claramente un área donde la atención cuidadosa a las limitaciones es a la vez vital y proclive a ser desdeñada. Que los resultados ajustados son de alguna manera medidas válidas de "calidad institucional" es una cosa, pero los analistas deberían también tener en cuenta los efectos potenciales de los resultados en términos de cambios de comportamiento futuros de las instituciones y los individuos que tratan de mejorar su subsiguiente "ordenación".»
>
> Goldstein y Spiegelhalter, [19, p. 390]

La evaluación de la investigación *también* es de gran interés público. Para un científico individual, una evaluación puede tener efectos profundos y a largo plazo en su carrera; para un departamento, puede cambiar sus perspectivas de éxito en el futuro lejano; para las disciplinas, una colección de evaluaciones puede marcar la diferencia entre prosperar y languidecer. Para una tarea tan importante, seguramente se debería entender tanto la validez como las limitaciones de las herramientas que se usan para llevarla a cabo. ¿Hasta qué punto miden las citas la calidad de la investigación? Los números de citas parecen estar en correlación con la calidad, y hay una interpretación intuitiva de que los artículos de gran calidad son muy citados. Pero, como se ha explicado antes, algunos artículos, especialmente en algunas disciplinas, son muy citados por razones distintas de la gran calidad, y no se sigue que artículos muy citados sean necesariamente de gran calidad. La precisa interpretación de las ordenaciones basadas en estadísticas de citas necesita ser entendida mejor. Además, si las estadísticas de citas desempeñan un papel central en la evaluación de la investigación, está claro que los autores, editores e incluso las editoriales encontrarán maneras de manipular el sistema en su beneficio [24]. Las implicaciones de esto a largo plazo no son claras y no han sido estudiadas.

Vale la pena leer en estos días el artículo de Goldstein y Spiegelhalter porque deja claro que el exceso de confianza en estadísticas ingenuas para evaluar la investigación no es un problema aislado. Los gobiernos, las instituciones y los individuos han forcejeado con problemas similares en el pasado en otros contextos, y han encontrado maneras de entender mejor las herramientas estadísticas y aumentarlas con otros medios de evaluación. Goldstein y Spiegelhalter acaban su artículo con una declaración positiva de esperanza:

> «Finalmente, a pesar de que generalmente hemos sido críticos con muchos intentos actuales de proporcionar juicios sobre instituciones, no deseamos dar la impresión de que creemos que todas estas comparaciones son necesariamente erróneas. Nos parece que la comparación de instituciones y

> el intento de entender por qué las instituciones difieren es una actividad extremadamente importante y se lleva a cabo mejor en un espíritu de colaboración que en uno de confrontación. Es quizá el único método seguro para obtener información objetiva que pueda llevarnos, a la larga, a un mejor entendimiento. *El problema real con los procedimientos simplistas que hemos pretendido criticar es que distraen tanto la atención como los recursos de este objetivo más digno.*»
>
> Goldstein y Spiegelhalter, [19, p. 406]

Sería difícil encontrar una declaración mejor que exprese los objetivos que deberían ser compartidos por cualquiera involucrado en la evaluación de la investigación.

Referencias

[1] R. Adler, The impact of impact factors, *IMS Bulletin*, **36**:5 (2007), 4. Disponible en `http://bulletin.imstat.org/pdf/36/5`

[2] M. Amin y M. Mabe, Impact factor: use and abuse, *Perspectives in Publishing*, **1** (octubre de 2000), 1–6. Disponible en `http://www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf`

[3] P. D. Batista, M. G. Campiteli, O. Kinouchi y A. S. Martinez, Universal behavior of a research productivity index. arXiv: physics, v1 (2005), pp. 1–5. Disponible en `arXiv:physics/0510142v1`

[4] P. D. Batista, M. G. Campiteli y O. Kinouchi, Is it possible to compare researchers with different scientific interests?, *Scientometrics*, **68**:1 (2006), 179–189. Disponible en `http://dx.doi.org/10.1007/s11192-006-0090-4`

[5] C. Bergstrom, Eigenfactor: measuring the value and presitige of scholarly journals, *College & Research Libraries News*, **68**:5 (mayo de 2007). Disponible en `http://www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may07/eigenfactor.cfm` (véase también `http://www.eigenfactor.org/methods.pdf`).

[6] J. Best, *Damned lies and statistics: untangling the numbers from the media, politicians, and activists*, University of California Press, Berkeley, 2001.

[7] S. Bird y otros, Performance indicators: good, bad, and ugly (informe de un grupo de trabajo sobre monitorización del rendimiento de los servicios públicos), *J. R. Statist. Soc. A*, **168** (2005), parte 1, 1–27. Disponible en `http://dx.doi.org/10.1111/j.1467-985X.2004.00333.x`

[8] T. Brooks, Evidence of complex citer motivations, *Journal of the American Society for Information Science*, **37**:1 (1986), 34–36. Disponible en `http://dx.doi.org/10.1002/asi.4630370106`

[9] A. L. Carey, M. G. Cowling, P. G. Taylor, Assessing research in the mathematical sciences. *Gazette of the Australian Math Society*, A. L. Carey, **34**:2 (mayo de 2007), 84–89. Disponible en `http://www.austms.org.au/Publ/Gazette/2007/May07/084CommsCarey.pdf`

[10] S. E. COZZENS, What do citations count? The rhetoric-first model, *Scientometrics*, **15**:5–6 (1989), 437–447. Disponible en http://dx.doi.org/10.1007/BF02017064

[11] L. EGGHE, Theory and practice of the g-index, *Scientometrics*, **69**:1 (2006), 131–152. Disponible en http://dx.doi.org/10.1007/s11192-006-0144-7

[12] EVIDENCE REPORT 2007, *The use of bibliometrics to measure research quality in the UK higher education system* (un informe elaborado para el Research Policy Committee of Universities, UK, por Evidence Ltd., una compañía especializada en análisis en interpretación del rendimiento de la investigación. Evidence Ltd. tiene una «alianza estratégica» con Thomson Scientific.) Disponible en http://bookshop.universitiesuk.ac.uk/downloads/bibliometrics.pdf

[13] J. EWING, Measuring journals, *Notices of the AMS*, **53**:9 (2006), 1049–1053. Disponible en http://www.ams.org/notices/200609/comm-ewing.pdf

[14] E. GARFIELD Citation indexes for science: A new dimension in documentation through association of ideas, *Science*, **122** (julio de 1955), 108–111. Disponible en http://garfield.library.upenn.edu/papers/science1955.pdf

[15] E. GARFIELD, Citation analysis as a tool in journal evaluation, *Science*, **178** (1972), 471–479. Disponible en http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf

[16] E. GARFIELD, Why are the impacts of the leading medical journals so similar and yet so different? Current Comments, **2** (12 de enero de 1987), 3. Disponible en http://www.garfield.library.upenn.edu/essays/v10p007y1987.pdf

[17] E. GARFIELD, Long-term vs. short-term journal impact (part II), *The Scientist*, **12**:14 (6 de julio de 1998), 12–3. Disponible en http://garfield.library.upenn.edu/commentaries/tsv12(14)p12y19980706.pdf

[18] E. GARFIELD, Agony and the ecstasy—the history and meaning of the journal impact factor. Presentado en el *International Congress on Peer Review and Bibliomedical Publication*, Chicago, 16 de septiembre de 2005. Disponible en http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

[19] H. GOLDSTEIN Y D. J. SPIEGELHALTER, League tables and their limitations: Statistical issues in comparisons of institutional performance, *J. R. Statist. Soc. A*, **159**:3 (1996), 385–443. Disponible en http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A3%3C385%3ALTATLS%3E2.0.CO%3B2?5 y http://dx.doi.org/10.2307/2983325

[20] P. HALL, Measuring research performance in the mathematical sciences in Australian universities, *The Australian Mathematical Society Gazette*, **34**:1 (2007), 26–30. Disponible en http://www.austms.org.au/Publ/Gazette/2007/Mar07/26HallMeasuring.pdf

[21] J. E HIRSCH, An index to quantify an individual's scientific research output, *Proc. Natl. Acad. Sci. USA*, **102**:46 (2005), 16569–16573. Disponible en http://dx.doi.org/10.1073/pnas.0507655102

[22] A. L. Kinney, National scientific facilities and their science impact on non-biomedical research, *Proc. Natl. Acad. Sci. USA*, **104**:46 (2007), 17943–17947. Disponible en http://dx.doi.org/10.1073/pnas.0704416104

[23] S. Lehmann, A. D. Jackson y B. E. Lautrup, Measures for measures, *Nature*, **444**:21 (2006), 1003–1004. Disponible en http://www.nature.com/nature/journal/v444/n7122/full/4441003a.html

[24] S. Macdonald y J. Kam, Aardvark et al.: quality journals and gamesmanship in management studies, *Journal of Information Science*, **33** (2007), 702–717. Disponible en http://dx.doi.org/10.1177/0165551507077419

[25] B. R. Martin, The use of multiple indicators in the assessment of basic research, *Scientometrics*, **36**:3 (1996), 343–362. Disponible en http://dx.doi.org/10.1007/BF02129599

[26] B. R. Martin y J. Irvine, Assessing basic research, *Research Policy*, **12** (1983), 61–90. Disponible en http://dx.doi.org/10.1016/0048-7333(83)90005-7

[27] L. Meho y Y. Kiduk, Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar, *Journal of the American Society for Information Science and Technology*, **58**:13 (2007), 2105–2125. Disponible en http://dx.doi.org/10.1002/asi.20677

[28] J. F. Molinari y A. Molinari, A new methodology for ranking scientific institutions, aceptado en *Scientometrics*. Disponible en http://imechanica.org/files/paper.pdf

[29] R. Monastersky, The number that's devouring science, *Chronicle Higher Ed.*, **52**:8. Disponible en http://chronicle.com/free/v52/i08/08a01201.htm

[30] M. Rossner, H. Van Epps y E. Hill, Show me the data, *Journal of Cell Biology*, **179**:6 (17 de diciembre de 2007), 1091–1092. Disponible en http://dx.doi.org/10.1083/jcb.200711140

[31] P. O. Seglen, Why the impact factor for journals should not be used for evaluating research; *BMJ*, 314:497 (15 de febrero de 1997). Disponible en http://www.bmj.com/cgi/content/full/314/7079/497

[32] A Sidiropoulos, D. Katsaros y Y. Manolopoulos, Generalized h-index for disclosing latent facts in citation networks, V1, *arXiv:cs*. Disponible en arXiv:cs/0607066v1[cs.DL]

[33] M. J. Stringer, M. Sales-Pardo y L. A. Nunes Amaral, Effectiveness of journal ranking schemes as a tool for locating information, *PLoS ONE* 3:2 (2008), e1683. Disponiible en http://dx.doi.org/10.1371/journal.pone.0001683

[34] Thomson: Journal Citation Reports, 2007 (página web de Thomson Scientific). Disponible en http://scientific.thomson.com/products/jcr/

[35] Thomson: Selection, 2007 (página web de Thomson Scientific). Disponible en http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/

[36] THOMSON: IMPACT FACTOR (página web de Thomson Scientific). Disponible en http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/

[37] THOMSON: HISTORY (página web de Thomson Scientific). Disponible en http://scientific.thomson.com/free/essays/citationindexing/history/

[38] THOMSON: FIFTY YEARS (página web de Thomson Scientific). Disponible en http://scientific.thomson.com/free/essays/citationindexing/50y-citationindexing/

NOTAS

[1]Esta cita fue atribuida a Einstein en el *Reader's Digest*, octubre de 1977. Parece que procede de esta cita real: «No se puede negar que la meta suprema de toda teoría es hacer los elementos básicos irreducibles tan simples y tan pocos como sea posible sin tener que rendirse a la adecuada representación de un simple dato de experiencia.». De «Sobre el método de la Física teórica», la lección Herbert Spencer, dada en Oxford (10 de junio de 1933); también publicada en *Philosophy of Science*, vol. 1, núm. 2 (abril de 1934), páginas 163–169.

[2]A pesar de que en esta sección nos concentramos en el factor de impacto de *Thomson Scientific*, debemos mencionar que Thomson promueve el uso de otras dos estadísticas. También se pueden derivar estadísticas similares basadas en promedios de números de citas para revistas a partir de otras bases de datos, incluyendo Scopus, Spires, Google Scholar y (para matemáticas) la base de datos de citas de Math Reviews. La última consiste en citas de más de 400 revistas matemáticas desde el período entre 2000 y el presente, identificadas como artículos que fueron listados en Math Reviews desde 1940; incluye más de 3 millones de citas.

[3] *Thomson Scientific* señala (marzo de 2008) que indexa revistas en las siguientes categorías:

- Matemáticas (217)
- Matemáticas aplicadas (177)
- Matemáticas interdisciplinarias (76)
- Física matemática (44)
- Probabilidad y estadísticas (96)

Las categorías se solapan, y el número total de revistas es aproximadamente 400. Como contraste, *Mathematical Reviews* incluye artículos de bastante más de 1 200 revistas cada año, y considera más de 800 artículos como «troncales» (en el sentido de que cada artículo de la revista está incluido en Math Reviews). Zentralblatt cubre un número semejante de revistas matemáticas.

[4]La base de datos de citas de *Mathematical Reviews* incluye (marzo de 2008) más de 3 millones de referencias en aproximadamente 400 artículos publicados desde 2000 hasta el presente. Las referencias se corresponden a artículos en la base de datos MR y se extienden durante varias décadas. Al contrario de lo que pasa con el *Science*

Citation Index, se incluyen las citas a libros y a artículos. Es un hecho curioso que, a grandes rasgos, el 50 % de las citas sean a artículos publicados en la pasada década; el 25 % a artículos que aparecieron en la década anterior; el 12,5 % a artículos en la década anterior, y así sucesivamente. Este tipo de comportamiento es especial para cada disciplina, por supuesto.

[5]La distribución sesgada combinada con una ventana tan estrecha (se usan solo las revistas de un año como fuente de citas y cinco años como objetivo) significa que un gran número de artículos tienen ninguna o muy pocas citas. Esto hace que sea intuitivamente obvio que dos artículos escogidos al azar sean a menudo equivalentes.

El hecho de que muchos artículos no tengan citas (o solo unas pocas) es también una consecuencia del largo tiempo de citas para las matemáticas: los artículos necesitan a veces muchos años para acumular citas. Si escogiéramos períodos más largos de tiempo tanto para la revista de procedencia como para los años objetivo, entonces los números de citas se incrementarían sustancialmente y sería más fácil distinguir revistas por el comportamiento respecto de las citas. Esta es la aproximación que se usa en [33] para analizar citas. Muestran que para períodos de tiempo suficientemente largos, la distribución de números de citas para artículos normales parece ser log-normal. Esto proporciona un mecanismo para comparar dos revistas comparando las distribuciones, y es ciertamente más sofisticado que usar factores de impacto. De nuevo, sin embargo, esto considera solo citas y nada más.

[6]Para ilustrar cuánta información se pierde cuando se usa solo el índice h, aquí tenemos un ejemplo de la vida real de un matemático distinguido en mitad de su carrera que ha publicado 84 artículos de investigación. La distribución de citas tiene el aspecto de la figura 6.

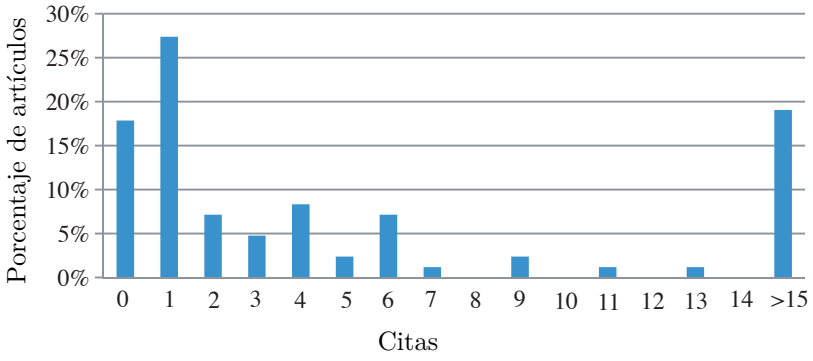

Figura 6: Un ejemplo real: distribución de citas de un matemático distinguido en mitad de su carrera.

Notemos que algo menos del 20 % de las publicaciones tienen 15 o más citas. La distribución de los números reales de citas para estos 15 artículos es la que aparece en la figura 7.

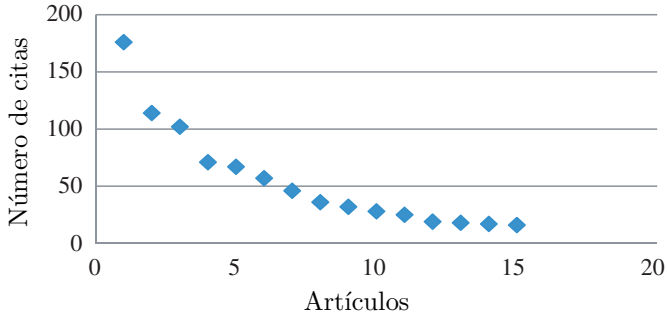

Figura 7: Un ejemplo real: distribución del número de citas para los 15 mejores artículos de un matemático distinguido en mitad de su carrera.

En el análisis de Hirsch, sin embargo, toda esta información se desperdicia. Solo se recuerda que el h-índice es 15, lo que significa que los 15 mejores artículos tienen 15 o más citas.

[7]El algoritmo en [5] usa un método de ordenación de páginas para dar a cada cita un peso, y entonces calcula un «factor de impacto» usando las medias ponderadas para citas. Los algoritmos de ordenación de páginas tienen mérito porque tienen en cuenta el «valor» de las citas. Por otro lado, su complejidad puede ser peligrosa porque los resultados finales son más difíciles de entender. En este caso, todas las «autocitas» se descartan (esto es, todas las citas de artículos en una revista dada J a artículos publicados en J durante los cinco años precedentes). Estas no son «autocitas» en ningún sentido normal de la palabra, y una mirada a algunos datos de la base de datos de citas de Math Reviews sugiere que esto descarta aproximadamente un tercio de todas las citas.

El algoritmo en [33] es interesante, en parte porque intenta tratar las diferentes escalas temporales para las citas así como el problema de comparar artículos elegidos aleatoriamente de una revista con los de otra. De nuevo, la complejidad de los algoritmos hace difícil para mucha gente evaluar sus resultados. Una hipótesis notable se desliza en la página 2 del artículo: «Nuestra primera suposición es que los artículos publicados en la revista J tienen una distribución normal de "calidad"...». Esto parece contradecir la experiencia común.

ROBERT ADLER, TECHNION-ISRAEL INSTITUTE OF TECHNOLOGY
Correo electrónico: robert@ieadler.technion.ac.il

JOHN EWING (PRESIDENTE), AMERICAN MATHEMATICAL SOCIETY
Correo electrónico: jhe@ams.org

PETER TAYLOR, UNIVERSITY OF MELBOURNE
Correo electrónico: pgt@ms.unimelb.edu.au

TRADUCIDO POR RAMÓN ESTEBAN ROMERO, INSTITUT DE MATEMÀTICA PURA I APLICADA, UNIVERSITAT POLITÈCNICA DE VALÈNCIA, CAMÍ DE VERA, S/N, 46022 VALÈNCIA
Correo electrónico: resteban@mat.upv.es
Página web: http://personales.upv.es/~resteban