Financial Post

Semana de la Ciencia Chatarra: Estadísticas no significativas

http://opinion.financialpost.com/2013/06/10/junk-science-week-unsignificant-statistics/

Stephen T. Ziliak, Especial para el Financial Post | 13/06/10 |

Última actualización: 13/06/12

Traducción: Alex Hill (www.et3m.net)

Stephen Ziliak sostiene que no cree en las afirmaciones estadísticas lanzadas acerca del descubrimiento del bosón de Higgs, o la "partícula de Dios."

Las demostraciones basadas en probabilidades generalmente no tienen significado alguno.

Quisiera creer, tanto como mi vecino, que los físicos especialistas en partículas han descubierto a un bosón de Higgs, la así llamada “partícula de Dios,” una con una masa de 125 gigaelectron voltios (GeV). Pero hasta ahora no estoy dispuesto a creer en las afirmaciones estadísticas que se están efectuando acerca de dicho descubrimiento. Dado que las afirmaciones acerca de la evidencia se basan en "significancia estadística" – es decir, en el número de desviaciones estándar que la señal observada se aleja de la hipótesis nula de “no hay diferencia” – las afirmaciones de los físicos no resultan creíbles. La significancia estadística es ciencia chatarra, y sus grandes montañas de tonterías están arruinando la investigación de muchos más que tan sólo los físicos de partículas.

Soy un economista. De manera que no confíen en mí acerca de bonos chatarra de última moda o del destino del sistema financiero mundial. Pero esto es algo que sí pueden creer, y querrán hacerlo: la significancia estadística huele mal. En ciencias estadísticas, desde la economía a la medicina, incluyendo algunas partes de la física y la química, la ubicua "prueba" de "significancia estadística" no puede, y no lo hará, demostrar la existencia de un bosón de Higgs, más allá de lo que puede demostrar la realidad de Dios, la existencia de una buena píldora contra el dolor, o la validez de una política monetaria poco controlada.

Un alejamiento estadísticamente significativo a partir de una hipótesis nula supuesta-como-verdadera es, por sí misma, demostración de nada en absoluto. De la misma manera, el fracaso en lograr una significancia estadística en un nivel de .05, u otro nivel estipulado, no constituye una demostración de que se haya descubierto algo de importancia.

Suena demasiado sencillo como para ser verdad, pero de hecho los dos problemas más fundamentales con la prueba de la significancia estadística nacen a partir de pequeños fragmentos de lógica errónea.

La prueba de significancia se inicia habitualmente con la estipulación de una "hipótesis nula". Expresada en términos algebraicos, se supone que, en promedio, un nuevo objeto A no es diferente de un objeto que nos resulta familiar B, donde los objetos podrían ser tipos de píldoras para bajar de peso, esquemas impositivos, o partículas físicas con diferentes nombres. Se reúnen los datos, ya sea en forma experimental o de otra manera, y luego se efectúa un cálculo para determinar la probabilidad de que datos mayores que aquellos que vemos en promedio pudieran haber ocurrido si, de hecho, no existe diferencia observable entre los objetos bajo estudio.

La significancia estadística es ciencia chatarra, y sus grandes montañas de sinsentidos están arruinando la investigación de muchos más que tan sólo los físicos de partículas.

El nombre formal para este extraño cálculo es "valor *p*". Si el valor *p* posee un número bajo – en ciencias sociales y en el mundo de los negocios, si *p* cae por debajo de .05, o una probabilidad de 1-en-20 – se afirma que el resultado del experimento es "estadísticamente significativo". La afirmación es que el nuevo objeto A, por ejemplo, es de una manera estadísticamente significativa diferente del objeto viejo B, porque las probabilidades de ver una diferencia mayor entre A y B – mayor que la diferencia que ustedes han visto – es pequeña. Se trata de una norma sumamente extraña.

De la misma manera, si *p* excede .05 (o cualquier línea que los científicos hayan trazado – más baja en física, más alta en negocios), se afirma que el resultado del experimento es no concluyente, y por ende puede ignorarse.

El procedimiento de prueba de hipótesis nula no es la única prueba de significancia pero es la más comúnmente utilizada y abusada de todas las pruebas. Para empezar, la prueba de significancia estadística formula la pregunta equivocada. La prueba emite la pregunta: "Suponiendo que la hipótesis nula es verdadera – que el bosón de Higgs (o cualquier otra cosa) no existe – ¿cuál es la probabilidad de encontrar un resultado al menos tan grande como el que hemos hallado en los datos?" Este cálculo de probabilidad es el valor de *p*.

Al formular la pregunta cuantitativa de la manera en la que lo hacen, los científicos que están evaluando la significancia han inadvertidamente invertido la ecuación fundamental de la estadística. Créase o no, han transpuesto su hipótesis y los datos, forzándolos a distorsionar de una manera grosera las magnitudes de los eventos probables – y aquí podemos ver por qué.

Se han dejado engañar por una lógica equivocada denominada, en estadística, la "falacia del condicional transpuesto". Si la señora Smith sufre un calambre esta semana, y fallece, uno no podría sencillamente concluir que la señora Smith probablemente haya fallecido por un calambre. Esto es porque la probabilidad de sufrir un calambre, dado que uno está muerto, no es igual a la probabilidad de que uno esté muerto, suponiendo que uno sufrió un calambre. Puede que la señora Smith haya fallecido por múltiples razones. Pero eso

constituye precisamente la inversa de la hipótesis y los datos que, no importa cuán ilógicos, continúan formulando los físicos de partículas en Ginebra – y la mayoría de los otros científicos en campos que van desde la economía hasta la medicina.

Demuestro, en un libro que escribí en colaboración con Deirdre N. McCloskey, The Cult of Statistical Significance - El Culto de la Significancia Estadística - (2008), que el procedimiento de pruebas de la hipótesis nula – otro nombre para la evaluación mediante significancia estadística – produce muchos errores de este tipo, con resultados trágicos para las economías, leyes, medicina e incluso vida humana en el mundo real.

En un escrutinio de publicaciones líderes que duró varias décadas, y que incluye publicaciones que van desde el *American Economic Review* al *New England Journal of Medicine*, hemos hallado que ocho o nueve de cada 10 artículos supone que la prueba de la significancia estadística demuestra importancia científica, económica, o de otra actividad humana, y que una ausencia de significancia estadística – "insignificancia" estadística o un valor de *p* mayor que .05 – indica una falta de importancia. La significancia estadística ni es necesaria ni es suficiente para demostrar un resultado físico, económico o médico. Pero las burocracias de la ciencia – desde aquellos que asignan fondos gubernamentales a los árbitros de publicaciones científicas – continúan insistiendo en demostraciones de significancia estadística, sin que importen los efectos reales económicos, médicos, físicos o de otro tipo que sean revelados por la evidencia total.

Consideremos nuevamente la falta de lógica del procedimiento de los físicos. La señal en los datos que se ha observado por encima del ruido de trasfondo (denotado como estando en 5 sigmas) es posiblemente un bosón de Higgs – eso es cierto. Pero en momentos de sobriedad – cuando se apagan los reflectores y desaparece la efervescencia en las bebidas – esos mismos físicos de partículas admiten que el jurado aún está encerrado y deliberando – que la protuberancia estadísticamente significativa podría tener "consistencia con" otras hipótesis plausibles no especificadas en sus modelos – de la misma manera en que la señora Smith pudiera haber padecido de algo que no fuera un calambre, y que probablemente haya sido el caso.

Esto resulta evidente por sí mismo. Un resultado estadísticamente significativo podría constituir evidencia de alguna otra partícula o campo – un Júpiter o Zeus o Prometeo tanto como el anticipado Higgs. Pero debido a que los modelos utilizados por los físicos no asignan un peso probabilístico a Higgs y a sus hipótesis competidoras, las creencias previas y posteriores acerca de todas las hipótesis permanecen estáticas, despreciadas y desconocidas.

Así, la reportada probabilidad de hallar un bosón de Higgs – medida, según afirman lógicamente los físicos, a través de su valor de *p* super-pequeño – es incorrecta. El valor de *p* sólo muestra la probabilidad de que no se obtuviesen datos que no fueron observados – es decir, partículas más pesadas que 125 GeV no fueron explicadas, con un alto valor de probabilidad, por la hipótesis nula de "no bosón de Higgs". Pero resulta imposible pasar de "Veo algo diferente de la hipótesis nula" a "Veo mi hipótesis favorita" sin agregar nuevas suposiciones, llevándonos desde la falacia del condicional transpuesto a afirmaciones claras acerca de la probabilidad de la hipótesis favorecida, tal como la de Higgs.

Los evaluadores de significancia no explican esas suposiciones adicionales, pero parecen satisfechos de poder efectuar inferencias adicionales basadas en esas suposiciones fallidas.

Existe un segundo e igualmente fundamental problema con la teoría y la práctica de la evaluación de significancia. La prueba no nos informa cuán grande o pequeño es el tamaño del efecto; no nos informa qué tan importante (o útil, o peligroso, o sorprendente) es el efecto, en una métrica de grande y pequeño – aquello que McCloskey y un servidor denominamos el "uuumff." Tal como observó el eminente estadístico Leonard "Jimmie" Savage en el libro *Foundations of Statistics* (1954), la significancia estadística nos informa qué decir, pero no qué hacer. Pensemos acerca de la calidad del debate entre géneros acerca de la "edad" como factor "significativo" en la evaluación de salud mediante mamografías. No podríamos culpar a las mujeres jóvenes por creer que la necesidad de la evaluación fuese un verdadero juego de cara o cruz.

Tomemos otro ejemplo: la caza de ballenas. En el mes de junio del año 2005 el gobierno japonés aumentó el límite del número de ballenas que se permite sacrificar en la Antártida – de un valor de 440 ballenas anuales a más de 1,000. En vista de la oposición internacional, el Sub-Comisionado Akira Nakamae afirmó ante los micrófonos del noticiero de la BBC: "Llevaremos a cabo JARPA-2 [el plan de sacrificios adicionales] según lo programado, porque el tamaño de la muestra se determina con el objeto de obtener resultados estadísticamente significativos." Utilizando los métodos de significancia estadística, el gobierno japonés incrementó el tamaño de la muestra para llegar a una conclusión de que el sacrificio de más ballenas era aceptable. El "sacrificar más ballenas" es "para ser más significativo" —mediante el incremento del tamaño de la muestra – como si fuese más preciso. Tristemente, esta lógica retrógrada es muy común en biología marina, tanto como en los experimentos de campo en economía en el Banco Mundial.

O consideremos la "significancia" de daño cometido en un caso que involucra miles de seres humanos, algunos de ellos muertos. A principios de la década del 2000, bastantes pacientes que consumían el medicamento Vioxx experimentaron la ira de la así llamada regla del 5% de significancia estadística.

La prueba clínica se llevó a cabo en el año 2000 y publicada en los *Annals of Internal Medicine* (2003). La compañía patrocinadora, Merck, informó que cinco pacientes que consumían Vioxx sufrieron problemas cardíacos – fatales o no – durante la fase clínica. Eso se comparaba con sólo un resultado negativo en el grupo de control, "una diferencia [en resultados negativos] que no alcanzó significancia estadística". La creencia errónea entre los científicos chatarra es que el fallar en alcanzar significancia estadística es equivalente a no encontrar diferencias importantes entre los dos resultados. Además de esto, los investigadores descubrieron que no habían reportado tres de un total de ocho resultados negativos – según parece para alcanzar una diferencia no significativa – el error opuesto de aquel cometido por los balleneros.

La prohibición de esta chatarra de estadística significativa parece posible. Incluso está de acuerdo con ello la Suprema Corte de Justicia de los Estados Unidos. El 22 de marzo del año 2011, en el caso Matrixx Initiatives, Inc. contra Siracusano, un caso importante en la ley de valores, la Suprema Corte rechazó en forma unánime el empleo de reglas de línea

brillante de significancia estadística como una manera de ocultar de los inversionistas información adversa.

El caso involucraba una medicina homeopática llamada Zicam, un remedio para el resfrío común a base de zinc, producido por Matrixx Initiatives, Inc. Cuando se aplica a través de las fosas nasales, la medicina provoca que algunos usuarios experimenten una sensación de quemado, mientras que otros sufren de anosmia, la pérdida permanente del sentido del olfato. Matrixx ignoró un número de reportes adversos que recibió de médicos y usuarios desde 1999. Un médico informó a la empresa que la toxicidad del zinc fue descubierta por los biólogos en la década de 1930. Cuando un médico se presentó en el popular programa televisivo *Good Morning America* en el año de 2004 e informó de la mala actuación de la empresa, el precio de las acciones de Matrixx se derrumbó. Nuevamente la empresa ocultó, ahora tras el argumento de que los efectos adversos de consumir Zicam por las fosas nasales eran – espérenlo – "estadísticamente no significativos."

La Corte Suprema rechazó unánimemente el argumento convencional. La Corte concluyó: "El argumento de Matrixx [acerca de los efectos adversos de Zicam] se basa en la premisa de que la significancia estadística es la única indicación confiable de causación. Esta premisa es errónea."

Durante la argumentación oral, el magistrado Breyer dijo sarcásticamente: "Esta significancia estadística siempre funciona y siempre no funciona". En otras palabras, la Suprema Corte está de acuerdo: reglas duras-y-rápidas de significancia son ciencia chatarra y necesitan ser eliminadas.

Stephen T. Ziliak es Profesor de Economía en la Universidad Roosevelt en Chicago.