

## MEDICINA BASADA EN EVIDENCIA

### Comentario al artículo

RETIRE STATISTICAL SIGNIFICANCE  
Amrhein,V.,Greenland,S., McShane,B.  
NATURE,2019 567 305-307.

En 21 párrafos numerados se presenta el artículo esencialmente resumido y en partes con observación. Encabeza el texto una figura donde un aparente profesor es invitado a pasar a una habitación donde hay una serie de frascos de aspecto muy antiguo en medio de telarañas y un enorme esqueleto de cabeza y cuernos de aspecto similar a un mamut, ciertamente prehistórico. El aparente profesor lleva en sus manos un libro con el título **...significación estadística... Esto ya sugiere un prejuicio de gran tamaño.**

Y empieza el artículo...

I. Introduce el tema destacando que lo común es encontrar que donde se expresa que no hay diferencia entre dos grupos. Por ejemplo, el argumento reside en no encontrar una diferencia estadística significativa. Se señala que al menos alguien presente en esta declaración debe quedar perplejo si la tabla que informa datos numéricos muestra que efectivamente hay una diferencia. Esta última frase **si** debe dejar perplejo a muchos puesto que estaría planteando que la decisión tomada, **a ojo**, es más creíble que aquella en base a un análisis estadístico.

II. Los autores se preguntan como hacen los estadísticos para conducir a los científicos a negar diferencias que aquellos sin formación en estadística ven con claridad. Por varias generaciones los investigadores han sido advertidos que un resultado no estadísticamente significativo no **prueba** la hipótesis nula. Lo mismo, que un resultado estadísticamente significativo no **prueba** otra hipótesis. Aquí hay un asunto de probabilidad que es soslayado, supongo, voluntariamente. Cuando la probabilidad que una diferencia encontrada pueda ser explicada por mero azar sea muy baja se **decide** que no debe ser por azar, decisión arbitraria por cierto, aunque esa probabilidad se la reconoce como **nivel de error** en tal decisión.

III. Los autores indican que tienen unas propuestas para evitar que los científicos sean presa de estos errores conceptuales. Indican que **nunca** (esta afirmación es bastante concluyente y no concuerda con las menciones que más adelante advierten sobre la incertidumbre). Se debe concluir que no hay diferencia o asociación en base a un valor de p mayor que un umbral como 0,05, o equivalentemente porque un intervalo de confianza incluye cero. Tampoco que dos estudios están en conflicto porque uno tiene resultado estadísticamente significativo y el otro no. Cual es entonces la conducta a seguir. Esto no se expresa.

IV. Citan como ejemplo una serie de análisis acerca de los efectos de drogas antiinflamatorias. En vista que los resultados no fueron estadísticamente significativos, un grupo de los investigadores concluyó que la exposición a las drogas no estaba asociada con nueva fibrilación auricular y que los resultados contrastaban con los de un estudio anterior que mostraba significación estadística.

V. Después se dan ejemplos donde dos estudios, teniendo un valor central igual como resultado, difieren en cuanto uno se expresó como no-significativo en su conclusión aunque el otro indica generar valores significativos en su intervalo de confianza. El que los valores centrales sean iguales claramente no indica que los estudios debiesen generar iguales conclusiones. Pueden diferir no solo en método, en el tamaño de la muestra así como en el de la varianza para generar intervalos de confianza perfectamente diferentes y sugiriendo interpretaciones distintas.

VI. Se declara que es inútil concluir que los resultados no estadísticamente significativos muestran ..ausencia de asociación.. en circunstancias que el intervalo estimado incluye francos aumentos de riesgo. Consideran igualmente absurdo sostener que estos resultados han estado en franco contraste con información previa que muestra resultados idénticos. Como sea, cualquiera se preguntaría cuán improbable es que dos estudios

acerca de un mismo asunto generen resultados **idénticos**, como se ha señalado. Estas practicas, dicen, comunes muestran cuan poco confiables son los umbrales de significación estadística, los que nos pueden llevar a conclusiones falsas, según sostienen.

VII. Sobre cientos de artículos han encontrado que resultados estadísticamente no significativos son interpretados indicando ..sin diferencia.. o .. sin efecto.. en alrededor de la mitad de ellos. Señalan que el American Statistical Association Journal planteo una advertencia acerca del mal empleo de la significación estadística y los valores de  $p$ . Posteriormente, un numero especial de la revista presenta mas de 40 articulos acerca de la inferencia estadística en el siglo 21. Las menciones que anteriormente se han indicado, particularmente donde se comparan los resultados de artículos, **no corresponden estrictamente a las maniobras de inferencia estadística donde, a partir de una muestra se espera poder informar aspectos de la población de la que procede**. Introducen la colección con la advertencia ..No diga -estadísticamente significativo-. Los autores del articulo en comento señalan estar de acuerdo y lo que resulta muy llamativo es que transforman el asunto en un problema resolver o al menos a apoyar por medio de encuesta ya que dicen haber invitado a otros a pronunciarse a favor o en contra de estas ideas. 250 personas que respondieron estuvieron de acuerdo en un primer lapso. Tiempo después, ya tenían mas de 800 personas entre estadísticos, investigadores médicos y clínicos, biólogos y psicólogos de mas de 50 países. Esto es cosa seria, particularmente porque no se expresa el **como** debería interpretarse los resultados entonces.

VIII. Los autores aclaran que no están por proscribir los valores de  $p$ . De hecho sostienen que se puede usar como criterio de decisión en ciertas aplicaciones especializadas, como determinar si un proceso de manufactura esta de acuerdo con un estándar de control de calidad. Como puede ser!! El resultado seria dicotómico ..de acuerdo o en desacuerdo. Despues de haber planteado claramente que no se debería emplear el valor de  $p$  como elemento o criterio de decisión, creo que en este momento es exigible una fundamentación que justifique lo opuesto en el ejemplo dado y probablemente en otros no mencionados.

IX. Los autores consideran que los resultados estadísticamente significativos o estadísticamente no significativos hacen pensar a la gente que los ítems asignados de tal manera son **categorícamente** diferentes. Igualmente enfatizan que la falsa creencia que atravesar el umbral establecido de significación estadística es suficiente para mostrar que un resultado es **real, lo que** ha conducido a privilegiar tales resultados, distorsionando la literatura. Esto sugiere que algunos siguen esta conducta y otros no. Interesante seria que los autores señalaran como se conducen estos últimos.

X. Se insiste en el hecho que los estimados significativos estadísticamente así como los no significativos estadísticamente, están sesgados en magnitud, los primeros en mas y los segundos en menos. Encima de esto, el centrarse rígidamente en la significación estadística, estimula a los investigadores a elegir datos y métodos que conduzcan a obtener significación estadística o lo contrario. **Esto revelaria una manipulación no aceptable.**

Aun los resultados de estudios pre-registrados pueden estar sesgados por decisiones dejadas abiertas en el plan de análisis.

XI. Nuevamente los autores indican que no son de la idea de eliminar los valores de  $p$ , los intervalos de confianza, u otras medidas estadísticas. Solo enfatizan que no se les debe tratar categorícamente, lo que incluye evitar la dicotomizacion en estadísticamente significativo o no. Agregan que debemos acostumbrarnos a la incertidumbre y como ejemplo de conducta apropiada para ello cambiar el nombre de intervalo de confianza por intervalo de compatibilidad e interpretarlos de modo de evitar el exceso de confianza. Especificamente recomiendan centrarse en las implicancias practicas de los valores centrales y los limites. **Aqui debieran dar al menos un caso de lo propuesto.**

XII. Se señala que un intervalo que contiene el valor de la hipótesis nula, también contiene otros valores de mucha importancia practica. Dicen ..si usted escoge todos los valores dentro del intervalo como sin importancia practica cosa que no hacemos ni vamos a hacer.... usted podría entonces decir algo como ..nuestros resultados son mayormente compatibles con efectos sin importancia. CLARO, pero debería preguntarse cual es la probabilidad de obtener en el intervalo de confianza de un valor central **solo** valores sin interés practico.

XIII. Cuando se esta hablando de intervalo de compatibilidad los autores recomiendan considerar cuatro aspectos. En primer termino si bien el intervalo entrega los valores mas compatibles con los datos, aquellos que están fuera del intervalo no son propiamente incompatibles sino menos compatibles. En otras palabras son los menos probablemente compatibles. De acuerdo.

XIV. En segundo termino no todos los valores dentro del intervalo son igualmente compatibles. Hay que recordar la distribucion Gaussiana en que se distribuyen los valores alrededor del estimado puntual. Los autores señalarían que interpretando el estimado puntual, junto con reconocer su incerteza, le asegurara no efectuar declaraciones de ..no diferencia.. o de efectuar declaraciones exageradamente confiadas.

XV. Otro punto destacado es que el umbral de 0.05 permite un 95% utilizable para computar los intervalos, lo que es considerado una convención arbitraria. Dicen que se perpetua el problema de la significación estadística toda vez que que la dicotomizacion que se impone es manejada como un estándar científico.

XVI. Por ultimo, los autores recomiendan ser humilde, de suerte que el establecer compatibilidad como ha sido mencionado esta basada en la corrección de las presunciones estadísticas utilizadas para computar el intervalo. En la practica, tales presunciones están sujetas a una considerable incerteza.

XVII. Tales presunciones deben ser hechas lo mas claramente posibles y probar aquellas que son posibles, por ejemplo diagramando sus datos y ajustando modelos alternativos para finalmente publicar todos los resultados. Esto ya parece una misión muy cuesta arriba.

XVIII. Los autores destacan que lo que sea que muestren las estadísticas, es adecuado sugerir razones para los resultado obtenidos y discutir un margen de explicaciones potenciales evitando favorecer unos sobre otros. Las inferencias deben ser científicas, (acaso no lo han sido hasta ahora?) lo que va mas allá (hasta donde y como?) de lo meramente estadístico. Frecuentemente hay factores mas importantes que las mediciones de p o los intervalos. Aqui, nuevamente, se esperarí a un buen ejemplo que aclarase estos conceptos.

XIX. La objeción mas frecuentemente encontrada a la propuesta de retirar la significación estadística es que es necesario tomar decisiones de si o no. Pero para las elecciones que se requiere realizar en un conjunto de situaciones, no basta solamente con la información que proporciona la significacion estadística. UN ejemplo se requeriría en este momento, mas aun porque a continuación los autores hacen un nexo con la idea de seguir o no con una idea de investigación fundamentando esto en que no hay una coneccion sencilla entre un valor de p y los resultados probables de subsecuentes estudios.

XX. Que aspecto generaría la supresión de la significación estadística. Los autores esperan que las secciones de métodos y de tabulación de los datos sean mas detalladas y con mas matices. Al mismo tiempo se espera que enfaticen sus estimaciones y la falta de certeza de ellos -por ejemplo discutiendo acerca de los limites mas bajos y los mas altos de los intervalos. Ellos no confiaran en los tests de significación. Al presentar los valores de p debieran emplear valores exactos y no utilizar simplemente las expresiones de mayor o menor de 0.05. Muy bien, y entonces como leen los valores exactos y deciden algo con ellos. La decisión de interpretar y publicar los resultados no estará basada en umbrales estadísticos. Deben precisar en que estarán fundadas las interpretaciones. Dicen que las personas deben gastar menos tiempo con programas estadísticos y mas tiempo pensando. Bien, pensando en.....?

XXI. Los autores señalan que la propuesta de retirar la significación estadística, y emplear mas bien intervalos de compatibilidad no es una panacea. Aunque se eliminaran algunas malas practicas, bien pueden introducir nuevas. **CUALES?** Estiman que la comunidad científica debiera priorizar el examen de la literatura identificando abusos estadísticos. **Nuevamente un buen ejemplo vendría bien.** Un elemento positivo sería la erradicación de la categorización. El mal uso de la significación estadística ha hecho mucho daño a la comunidad científica y a aquellos que confían en el consejo científico. Se destaca que los valores de p, los intervalos y otras medidas estadísticas tienen su lugar, pero es momento de eliminar la significación estadística y reemplazarla por...mucho pensamiento.

P.S.

Hay varios aspectos que considerar.

1. La figura que encabeza el artículo debiera ser eliminada.
2. Los autores se centran notablemente en restarle valor a las interpretaciones que, de acuerdo a las cifras obtenidas de  $p$  en análisis estadístico se tomen decisiones dicotómicas, prefiriendo los valores absolutos de  $p$  **PERO** sin explicar como debe interpretarse.
3. Resulta muy sorprendente que apoyen sus planteamientos por medio de encuesta, método estadístico que tiene numerosas exigencias que no parecen haber sido contempladas.
4. No hay ejemplos en situaciones planteadas que los requerirían.
5. Llama la atención que se insista en que debemos acostumbrarnos a la incertidumbre. La impresión del suscrito es que la bioestadística consiste en el manejo racional de la incertidumbre.

**Dr. Gaston Duffau Toro**

Profesor de la Facultad de Medicina de la Universidad de Chile