

## Cómo mejorar la comunicación de estadísticos inferenciales en ciencias de la salud

### How to improve inferential statistics reporting in health sciences

Ana María Ruiz-Ruano García<sup>a</sup>, Jorge López Puga<sup>b</sup>

<sup>a</sup> Departamento de Psicología Evolutiva y de la Educación, Facultad de Ciencias de la Educación, Universidad de Granada, España

<sup>b</sup> Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Facultad de Psicología, Universidad de Granada, España

#### Resumen

Las técnicas de inferencia estadística son esenciales para las ciencias de la salud. Gracias a estas herramientas estadísticas se pueden identificar, por ejemplo, factores de riesgo que afectan negativamente al estado de salud de las personas. Sin embargo, el proceso de comunicación científica y la relevancia de los hallazgos científicos pueden distorsionarse por el mal uso de estadísticos inferenciales como el p-valor o el factor de Bayes. En este trabajo proporcionamos siete directrices básicas que pretenden ayudar a interpretar y usar conjuntamente el p-valor clásico de un contraste de hipótesis en conjunción con los factores de Bayes. Aunque el factor de Bayes es menos conocido, y de introducción más reciente, que el p-valor es susceptible de ser utilizado superficial o erróneamente. Una comunicación más eficiente de los resultados de la investigación científica favorecería una mejor comprensión de estos y redundaría en mayores cotas de salud pública. Esperamos que estas directrices puedan ser de utilidad para personas de ciencia con poca experiencia, frente a la toma de decisiones políticas relacionadas con hallazgos científicos, en procesos editoriales y para el público en general.

Palabras clave: comunicación en salud; diseminación de la información; ciencias de la información; directrices; modelo estadístico; probabilidad; incertidumbre; teorema de Bayes.

#### Abstract

Statistic techniques for inference are essential for health sciences. Those techniques are useful to identify, for example, risk factors. However, the scientific communication process can be biased when inferential statistics are wrongly used. Here we provide seven guidelines to help readers to use the p-value and Bayes factor, two inferential statistics. Although the Bayes factor is less known than the p-value it is also prone to be misinterpreted and misused. A better scientific communication of research output would lead to a better understanding of scientific discoveries. As a result, this improvement in the information process would affect positively public health. We hope our guidelines to be helpful for researchers, reviewers, editors, policy decision makers and the general public.

Keywords: health communication; information dissemination; information sciences; guidelines; statistical model; probability; uncertainty; Bayes theorem.

## Introducción

Una de las piedras angulares del conocimiento científico es el análisis estadístico de los datos que son recogidos en estudios de campo o laboratorio (Box, 1976). Tras la recogida de datos y su posterior análisis, los equipos de investigación producen manuscritos que son enviados a revistas científicas especializadas o generalistas para difundir el conocimiento entre expertos y el público general. Este proceso de producción de textos científicos tiene como objetivo transferir conocimiento a la sociedad que pueda ser usado frente a la toma de decisiones. Por ejemplo, una investigación sobre los efectos perniciosos de una familia de sustancias (por ejemplo, las lipoproteínas de baja densidad) puede orientar políticas públicas destinadas a reducir el impacto negativo de estas sustancias en la población. Pese a que las técnicas de análisis estadístico, y los valores que éstas proporcionan, han sido concebidas para coadyuvar en esa toma de decisiones informada, en algunos casos estas mismas técnicas estadísticas suponen problemas tanto para expertos como para el público general.

Los problemas relacionados con el manejo de información estadística suponen una amenaza para la comunicación efectiva de los avances científicos. El proceso de comunicación científica en revistas especializadas, así como la divulgación periodística de este conocimiento, puede verse dañado cuando el rigor lingüístico no está en consonancia con el formalismo metodológico. En algunas ocasiones los problemas pueden ser debidos a cuestiones meramente informativas (por ejemplo, que en una comunicación científica no aparezca reflejada cierta información estadística), mientras que en otros casos las debilidades son más de índole conceptual (véase, por ejemplo, Prieto y Herranz, 2005). En cualquiera de los casos, disponer de guías que orienten sobre el uso de información estadística redundará en una mejor comunicación y transferencia de los resultados científicos. En este trabajo, precisamente, proporcionamos una serie de ideas breves y concisas que esperamos puedan ser de utilidad para mejorar el proceso comunicativo en el que el uso de estadísticos inferenciales goce de cierto protagonismo.

El marco de referencia estadístico en que se ha apoyado la comunidad científica en los últimos 75 años se encuadra bajo el concepto de Contraste de Significación de Hipótesis Nulas (o NHST, que alude a la expresión inglesa null hypothesis significance testing). En este contexto, el uso del p-valor de los tests estadísticos ha sido el punto clave en el que se han apoyado las

tomas de decisiones científicas. Sin embargo, el uso de este estadístico ha sido duramente criticado a lo largo de la historia y todavía sigue siendo objeto de duras discusiones (véase, por ejemplo, Altman & Krzywinski, 2017a, 2017b; Bakan, 1966; Orliczky, 2012; Stern, 2016; Wagenmakers, 2007). Los problemas relacionados con la interpretación de este estadístico no sólo afectan a los estudiantes universitarios o a los investigadores de un amplio rango de áreas de conocimiento; sino que, además, Haller y Krauss (2002) observaron que estos problemas también afectan al profesorado universitario que explica estos contenidos a su alumnado.

Una correcta interpretación y uso del p-valor redundará en una mejor difusión de la información científica. Dado que esa optimización de la transferencia del conocimiento científico favorecerá la toma de decisiones informada y mejorará la imagen pública de la empresa científica. Creemos que es esencial proporcionar una serie de ideas sencillas que permitan, tanto a legos como expertos, usar de mejor manera este estadístico. Por ello, nuestro objetivo con este trabajo es proporcionar un conjunto de directrices revisadas que ayuden a interpretar y usar tanto el p-valor estadístico de un test como el factor de Bayes. El factor de Bayes puede ser considerado como una alternativa reciente, o complemento bayesiano, al p-valor que puede ser utilizado en la toma de decisiones estadística (Jeffreys, 1948). Nuestro objetivo no es producir una guía exhaustiva sobre directrices de interpretación y uso del p-valor o del factor de Bayes (véase, por ejemplo, Wasserstein y Lazar, 2016); sino que, más bien, nos gustaría presentar una breve y concisa lista que, en nuestra opinión y experiencia, podría reducir la tasa de usos anómalos y malas interpretaciones cuando se llevan a cabo contrastes de hipótesis estadísticas. En la siguiente sección revisamos brevemente el significado y el origen del p-valor de un test estadístico, así como el del factor de Bayes. Seguidamente presentaremos una serie de directrices básicas que pueden ser de utilidad para interpretar y usar eficientemente estos estadísticos de manera conjunta.

### *Valoración de la incertidumbre en ciencia*

La misión de la ciencia es aproximarse a la comprensión de la naturaleza y el universo. Para abordar esta ardua tarea, las personas de ciencia han de enfrentarse y gestionar la incertidumbre que tienen sobre los sistemas naturales a los que se enfrentan. La reducción de esa incertidumbre se realiza utilizando modelos probabilísticos que permiten cuantificar nuestro estado de conocimiento sobre la naturaleza y el universo (Jeffreys, 1931).

Como indicábamos más arriba, el p-valor de un test

estadístico es, probablemente, el estadístico de apoyo a la inferencia científica más utilizado en nuestros días. Sin embargo, cada vez es más frecuente en revistas científicas especializadas el uso de factores de Bayes (Kass y Raftery, 1995) como complemento o alternativa al p-valor. En las siguientes subsecciones describiremos muy brevemente la lógica que subyace en estos dos estadísticos para, seguidamente, proporcionar una serie de directrices que faciliten su uso e interpretación y minimicen errores en la difusión de los resultados científicos.

### *El p-valor*

El p-valor de un test estadístico (como una prueba t o una prueba F en un análisis de varianza) puede entenderse como la probabilidad de haber observado los datos muestrales, o unos datos muestrales más extremos, si la hipótesis nula fuese cierta. Por lo general, aunque no siempre es así, esta hipótesis nula suele ser expresada por los investigadores como una hipótesis que nulifica el efecto de una manipulación experimental o el resultado esperado por alguna teoría científica. Por lo tanto, el p-valor de un test estadístico es una probabilidad condicional que podría simbolizarse como  $P(D|H_0)$ , es decir, la probabilidad de observar ciertos datos muestrales extremos dada una hipótesis nula.

En términos prácticos la interpretación que se hace de los p-valores es que cuanto más pequeños son, más improbables o inverosímiles se consideran los datos muestrales si, en realidad, la hipótesis nula fuese cierta. Por consiguiente, los investigadores típicamente rechazan la credibilidad de la hipótesis nula y tienden a confiar en la idea de que la alternativa (normalmente esa que alude a un efecto significativo de un factor de interés) es más verosímil.

### *El factor de Bayes*

El factor de Bayes es un estadístico desarrollado al auspicio de lo que ha venido a denominarse “estadística bayesiana” (véase, por ejemplo, Edwards, Lindman, y Savage, 1963). La estadística bayesiana, frente a la estadística “clásica” dentro de la cual surgió el p-valor, se preocupa también por probabilidades condicionadas pero el sentido y la interpretación es sensiblemente diferente. Mientras que, como hemos señalado, el p-valor se preocupa por la probabilidad de que unos datos sean verosímiles bajo la condición de que una hipótesis sea cierta; la estadística bayesiana pretende aproximarse a la probabilidad de que una hipótesis sea cierta considerando los datos muestrales observados, es decir,  $P(H_i|D)$ .

Los factores de Bayes están cobrando cada vez más importancia como complementos al p-valor en el contexto del contraste de hipótesis (véase, por ejemplo, Wagenmakers, 2007). El factor de Bayes ( $FB_{\alpha\beta}$ ) sirve para comparar la probabilidad de certeza de una hipótesis  $\alpha$  dados los datos muestrales, frente a la probabilidad de otra hipótesis competidora  $\beta$  considerando esos mismos datos muestrales. Más formalmente, podríamos expresar esta idea con la siguiente ecuación:

$$FB_{\alpha\beta} = \frac{P(H_\alpha | D)}{P(H_\beta | D)}$$

De esta manera, los factores de Bayes representan en qué cuantía una hipótesis es más probable que otra teniendo en cuenta los datos muestrales observados. Es decir, cuanto mayor es el resultado de esta fracción mayor será la verosimilitud de la primera hipótesis frente a la segunda. Por ejemplo, siguiendo el ejemplo que hemos presentado con letras griegas, si un factor de Bayes es igual a diez, podremos decir que la hipótesis alfa es diez veces más probable que la beta. Por el contrario, si el factor de Bayes es inferior a uno, la hipótesis más probable sería la beta. Y, por último, si el factor de Bayes fuese igual a uno, ambas hipótesis podrían considerarse igualmente probables teniendo en cuenta los datos muestrales.

### *Directrices*

La inferencia estadística, entendida como una forma de ir más allá de los datos muestrales para generalizarlos a una población objeto de estudio, es esencial para la ciencia. A su vez, los estadísticos como el p-valor y los factores de Bayes son capitales en los procesos de inferencia estadística. Creemos que el uso conjunto de los factores de Bayes y los p-valores es posible. Ambos se aproximan al proceso de inferencia de un modo diferente y, en cierto modo, pueden considerarse como complementarios.

Dado que los factores de Bayes están recibiendo poco a poco mayores cotas de atención por parte de la comunidad científica, creemos conveniente proporcionar una breve lista de directrices que favorezcan el buen uso conjunto de los p-valores y factores de Bayes. Estas directrices podrían ser de gran utilidad para autores, revisores, editores de revistas científicas, personas responsables implicadas en tomas de decisiones sobre aspectos de salud pública, periodistas que divulgan contenidos científicos, estudiantes universitarios, así como para el público en general. Como hemos indicado previamente, esta lista no es exhaustiva, sino que

pretende ser una guía básica o esencial sobre este tema. La persona interesada puede consultar algunas referencias de las utilizadas en este trabajo para poder acceder a listas más exhaustivas sobre este tema.

### 1. Evitar el uso estereotipado

En nuestra opinión, la primera y más importante directriz al utilizar p-valores y/o factores de Bayes es reflexionar críticamente sobre el significado de estos estadísticos en el contexto concreto en que van a ser aplicados. Es decir, utilizarlos evitando caer en un uso estereotipado y ritual de los mismos (Gigerenzer, 2004). Algunos autores han señalado que ha de producirse un cambio cultural para que esta tendencia estereotipada pueda ser superada (Orlitzky, 2012) y, por ello, nosotros invitamos a aquellas personas que van a usar estos estadísticos a reflexionar detenidamente sobre sus datos y lo que éstos sugieren en vez de confiar ciegamente en lo que puedan sugerir los p-valores o los factores de Bayes.

### 2. Utilizarlos en conjunción con otros estadísticos

Como sugerimos en el punto precedente, es conveniente no confiar ciegamente en los estadísticos inferenciales como el p-valor o el factor de Bayes. Más bien, es aconsejable valorar críticamente los estadísticos de tendencia central y de dispersión que hayan podido haberse obtenido al analizar los datos. Adicionalmente, conviene interpretar y utilizar otros estadísticos, también de naturaleza inferencial, como los intervalos de confianza (desde el punto de vista clásico) o los intervalos de credibilidad (desde el punto de vista bayesiano). Usar estadísticos que permitan valorar la potencia estadística, la magnitud o el tamaño del efecto es también recomendable. Este tipo de estadísticos permite valorar la relevancia clínica de los resultados de una investigación y el uso exclusivo del p-valor, por ejemplo, podría no ser tan útil como valorar la relevancia substantiva de los resultados. Como señalan Leek y Peng (2015), el uso del p-valor (o del factor de Bayes) es únicamente la “punta del iceberg”. Existen múltiples elementos del proceso de inferencia estadística que han de ser tenidos en cuenta para maximizar la validez de las conclusiones científicas.

### 3. Tener en cuenta la naturaleza condicional

No hay que perder de vista que tanto el p-valor como el factor de Bayes son estadísticos que valoran aspectos condicionados de la realidad (Marden, 2000). Como hemos visto, el p-valor es la probabilidad de obtener ciertos datos muestrales en el caso de que la hipótesis

nula fuese cierta. Es recomendable no perder de vista, por tanto, que este estadístico no es una probabilidad simple. Así, se evitaría caer en la conclusión errónea de que el p-valor valora la probabilidad de que una hipótesis sea cierta o incierta. Del mismo modo, el factor de Bayes también sopesa probabilidades condicionales. En su caso, como hemos indicado, valora la cantidad de evidencia que hay en favor de una hipótesis dados los datos observados empíricamente frente a la evidencia que hay en favor de otra hipótesis teniendo en cuenta esos mismos datos.

### 4. Precaución al usar/interpretar puntos de corte

Es muy común leer y escuchar expresiones como “las diferencias son estadísticamente significativas” en el contexto, por ejemplo, de los ensayos clínicos en los que se valora la eficiencia relativa de un nuevo fármaco frente al efecto de una sustancia placebo. En estos casos, esa expresión se acuña en los manuscritos científicos como consecuencia de haber comparado el p-valor obtenido con cierto punto de corte preestablecido. Creemos que hay varios argumentos que invitan a que expresiones como esa sean evitadas. En primer lugar, porque muchas veces se entremezclan diferentes ópticas interpretativas del marco estadístico-inferencial (Gigerenzer, 2004). Por otro lado, como señalaron Rosnow y Rosenthal (1989), ni Dios es capaz de discriminar a un valor de otro redondeando a la centésima. Es decir, valorar ciertos puntos de corte (uno de los más utilizados comúnmente es el famoso “0.05”) como fronteras inequívocas que disciernen entre lo “científicamente veraz” y lo “no certero” es una actividad que hace un flaco favor al progreso científico. Más bien, no debería perderse de vista que tanto el p-valor como el factor de Bayes son valores que se posicionan en un continuo que pueden ser interpretados como gradientes.

Aunque son menos populares, Jeffreys (1948) también introdujo puntos de corte orientativos para interpretar los factores de Bayes. Por ejemplo, sugirió que cuando el factor de Bayes se encuentra comprendido entre uno y tres podría considerarse que la evidencia en favor de una hipótesis frente a otra es anecdótica, mientras que cuando el factor de Bayes se encuentra entre 30 y 100 la evidencia es muy fuerte en favor a una de las hipótesis testadas. Aunque también se han proporcionado puntos de corte para los factores de Bayes con base logarítmica (véase, por ejemplo, Kass y Raftery, 1995), no podemos olvidar que todas estas categorías son arbitrarias y que sólo tienen un propósito orientativo. Es decir, no suponen, ni pretenden cosificar el pensamiento estadístico. Por ello, al igual que ocurre con los p-valores, sugerimos

que se utilicen estas categorías no olvidando que estos estadísticos pertenecen a un continuo, que pueden interpretarse como gradientes y que tienen sentido en un contexto aplicado concreto.

### 5. El p-valor o el factor de Bayes no son el fin

El objetivo de la ciencia no es encontrar verdades absolutas e inmutables. Más bien, la ciencia trabaja para aproximarse a la comprensión de la naturaleza. Las teorías y modelos científicos encuentran respuestas tentativas, parciales y momentáneas a las preguntas que van desafiando al intelecto humano. Algunas de esas respuestas son tan sugerentes, consistentes y consensuadas que desencadenan grandes progresos tecnológicos. En algunas ocasiones, sin embargo, algunos equipos de investigación se obcecán en equiparar los “efectos estadísticamente significativos” con el descubrimiento de la verdad. En ese afán por tratar de justificar que los datos obtenidos resuelven una pregunta científicamente relevante, algunos equipos de investigación abusan del p-valor y presumiblemente harían lo mismo con el factor de Bayes. Este abuso pasa por considerar que el p-valor, o el factor de Bayes, sirve para encontrar verdades científicas inequívocamente. Pero esta tendencia a “presionar los datos para que confiesen” únicamente conduce a prácticas indeseables (como el p-hacking) que atentan contra la empresa científica (dificultando, por ejemplo, los estudios de replicación) auspiciando el descrédito social de la misma (Munafò et al., 2017). Por consiguiente, nuestra recomendación es evitar que el núcleo de una investigación sea la interpretación de estadísticos como el p-valor o el factor de Bayes.

### 6. Clarificación de hipótesis bajo análisis

Es esencial que los estudios hagan explícitas las hipótesis que están evaluando. Y, lo que es más, que se especifique si los p-valores o factores de Bayes interpretados se están utilizando para testar hipótesis previamente establecidas o si los contrastes que se están realizando no habían sido planeados. Esta diferencia entre estudio confirmatorio (hipótesis o contrastes planteados ad hoc) y exploratorio (contrastos realizados post hoc) es esencial para evitar problemas de replicabilidad y, por consiguiente, para optimizar el desarrollo científico (Nosek, Ebersole, DeHaven, y Mellor, 2018). Por ello, es esencial que se especifique, cuando sea requerido, si los p-valores reportados son referidos a contrastes unilaterales (es decir, de regiones críticas de rechazo de una cola) o bilaterales (de regiones críticas de rechazo de dos colas). Del mismo modo, es esencial que se indique si los

factores de Bayes calculados son, por ejemplo, referidos a la superioridad explicativa de la hipótesis nula frente a la alternativa ( $FB_{01}$ ) o a la inversa ( $FB_{10}$ ).

### 7. Evitar lenguaje taxativo o juicios de valor sobre los resultados

Para terminar con esta breve lista de directrices, sugerimos que se modere el tono y la expresión verbal cuando se aluda a los resultados de un estudio. Por ejemplo, sugerimos que expresiones como “estadísticamente significativo” o “queda demostrado que” sean evitadas. Más bien, sugerimos que se utilicen expresiones como “los datos observados pueden considerarse como una evidencia en favor de” la hipótesis pertinente. También es aconsejable evitar juicios de valor sobre los resultados de un estudio científico. Por ejemplo, expresiones como “la significación estadística obtenida es buena” son deletéreas para favorecer una transmisión apropiada de los resultados científicos. Estas y otras expresiones teñidas de valor, ambiguas y rotundas contribuyen a dar la sensación de que la ciencia persigue verdades absolutas a la vez que destruyen el principio de incertidumbre que caracteriza la humilde empresa científica.

## Conclusiones

La comunicación eficiente de resultados de investigación (ya sea en revistas científicas especializadas, magazines científicos, en prensa escrita generalista, en medios de comunicación audiovisual o por medio de cualquier otro canal informativo) podría mejorarse si se tuviesen en cuenta algunas cuestiones básicas sobre el fundamento estadístico de la ciencia. Esto tendría consecuencias positivas tanto para la comunidad científica, últimamente desprestigiada por problemas relacionados con la “crisis de replicabilidad” o por temas de integridad científica, como para la sociedad en su conjunto. De hecho, algunos autores sugieren que debería favorecerse una especie de “alfabetización estadística” que pueda servir para que la comunicación científica (sobre todo aquella que tiene que ver con factores de riesgo para la salud y la integridad de las personas) sea eficiente a todos los niveles (Bond, 2009). Iniciativas como la que lidera el psicólogo alemán Gred Gigerenzer en el Harding Center for Risk Literacy del Max Planck Institute for Human Development van en esa línea de mejorar el proceso comunicativo de la ciencia, con el objetivo de ayudar a las personas a tomar decisiones óptimas respecto a su salud.

Hacer ciencia no es tarea fácil. No es fácil tampoco utilizar las herramientas estadísticas de inferencia. Comunicar óptimamente los resultados de la investigación

científica tampoco es sencillo. Quizá sean esas las razones que pueden explicar que, sistemáticamente, las personas de ciencia sucumban una y otra vez a los mismos errores a lo largo de la historia reciente en lo que toca a la interpretación de estadísticos inferenciales. En este trabajo hemos presentado siete directrices que pretenden favorecer un mejor uso de dos de los estadísticos inferenciales que más comúnmente se utilizan en la investigación en ciencias de la salud: el p-valor y el factor de Bayes.

Como hemos señalado, el p-valor tiene una historia más dilatada como herramienta de apoyo en la investigación científica, mientras que el factor de Bayes es más reciente. En cualquiera de los casos, al utilizarlos corremos el riesgo de hacer un uso inapropiado de los mismos lo que podría repercutir negativamente en el proceso de comunicación científica. En nuestra opinión no consiste en decantarse por uno o por otro como herramienta que guíe la investigación científica, aunque hay estudios de simulación que sugieren que teniendo en cuenta ciertos parámetros existen diferencias entre ellos (véase, por ejemplo, Ruiz-Ruano y Puga, 2018). Más bien, nuestra directriz general es que se aborde el análisis estadístico de los datos observados en una investigación desde un punto de vista lógico, global y contextualizado tratando de evitar rituales compulsivos o mecánicos culturalmente institucionalizados (Gigerenzer, 2004; Leek y Peng, 2015; Orlitzky, 2012).

Esperamos que estas directrices puedan servir para favorecer la discusión y la reflexión de la comunidad científica al respecto. Sería ideal que este trabajo pudiese estimular el desarrollo de directrices más sencillas y precisas que pudiesen ser de ayuda para el mejor uso de los estadísticos tratados. También esperamos que estas directrices puedan ser de utilidad para estudiantes universitarios que se inician en la investigación científica o que ayuden a los equipos de investigación consolidados con la interpretación de los resultados que generan. También esperamos que estas directrices puedan ser de utilidad en el proceso de revisión de manuscritos enviados en revistas, tanto a nivel de revisión por pares como en el proceso de decisión editorial. Por último, esperamos que estas notas puedan ser aprovechadas por periodistas que divulgan conocimientos científicos, por las personas que toman decisiones políticas relacionadas con hallazgos científicos y por el público en general.

## Referencias Bibliográficas

- Altman, N., y Krzywinski, M. (2017a). Points of significance: P values and the search for significance. *Nature Methods*, 14, 3-4. doi: 10.1038/nmeth.4120
- Altman, N., y Krzywinski, M. (2017b). Points of significance: Interpreting P values. *Nature Methods*, 14, 213-214. doi: 10.1038/nmeth.4210
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Bond, M. (2009, 28 de octubre). Decision-making: risk school. *Nature*, 461, 1189-1192. doi: 10.1038/4611189a
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791-799.
- Edwards, W., Lindman, H., y Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587-606. doi: 10.1016/j.socec.2004.09.033
- Haller, H., y Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, 7(1), 1-20. Retrieved from <https://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf>
- Jeffreys, H. (1931). *Scientific inference*. Londres, Inglaterra: Cambridge University Press.
- Jeffreys, H. (1948). *Theory of probability* (2nd Ed.). Oxford, Inglaterra: Oxford University Press.
- Kass, R. E., y Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Leek, J. T., y Peng, R. D. (2015, April 28). P values are just the tip of the iceberg. *Nature*, 520, 612. doi: 10.1038/520612a
- Marden, J. I. (2000). Hypothesis testing: From p values to Bayes factors. *Journal of the American Statistical Association*, 95, 1316-1320.
- Munafò, M., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior*, 1, 1-9. doi: 10.1038/s41562-016-0021
- Nosek, B., Ebersole, C., DeHaven, A., y Mellor, D. (2018).

- The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115, 2600-2606. doi: 10.1073/pnas.1708274114
- Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, 5, 199-228.
- Prieto, L., y Herranz, I. (2005). ¿Qué significa “estadísticamente significativo”? *La falacia del criterio del 5% en la investigación científica*. Madrid: Díaz de Santos.
- Rosnow, R. L., y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- Ruiz-Ruano, A. M., y Puga, J. L. (2018). Deciding on null hypothesis using p-values or Bayesian alternatives: a simulation study. *Psicothema*, 30, 110-115. doi: 10.7334/psicothema2017.308
- Stern, H. S. (2016). A test by any other name: P-values, Bayes Factors and statistical inference. *Multivariate Behaviour Research*, 51, 23-39. doi: 10.1080/00273171.2015.1099032
- Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779-804. doi: 10.3758/BF03194105
- Wasserstein, R. L., y Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70, 129-133. doi: 10.1080/00031305.2016.1154108