INFERENCIA EVIDENCIAL… NO HAY DOS SIN TRES

Casellas[1], J.
[1]Dep. Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona. joaquim.casellas@uab.cat

INTRODUCCIÓN

En la actualidad existen dos paradigmas inferenciales mayoritarios (frecuentista y Bayesiano), los cuales se ven ampliamente representados en la mayoría de áreas de conocimiento, sin que la mejora genética animal sea una excepción. No obstante, existen alternativas y alguna de ellas se ha sugerido ya como una posibilidad real dentro del campo de la genética (Strug, 2018). Esta no es otra que la inferencia evidencial, centrada en caracterizar el nivel de evidencia a favor o en contra de una determinada hipótesis (o modelo estadístico), y sustentado sus divergencias en relación a los paradigmas mayoritarios en base a la interpretabilidad y objetividad de sus resultados.

EVIDENCIA ESTADÍSTICA

Los trabajos originales de Barnard (1949), Hacking (1965) o Edwards (1972), entre otros, establecen las bases para replantear la inferencia estadística como la medida del nivel de evidencia aportado por un conjunto de datos en favor de una hipótesis (en relación a otra). Aunque puede parecer trivial, el concepto y definición de "evidencia estadística" representa el punto de partida de la inferencia evidencial, en contraposición a las corrientes mayoritarias de inferencia estadística, tanto frecuentista como Bayesiana. La inferencia evidencial basa sus procedimientos en el cociente de verosimilitudes como medida de la evidencia estadística (Hacking, 1965), tomando como punto de partida el Principio de Verosimilitud que establece que la función de verosimilitud captura toda la evidencia inherente a la muestra analizada en relación a los parámetros del modelo; aunque este principio fue deducido formalmente por Birnbaum (1962), su uso y aplicaciones puede ya encontrarse en los trabajos de Sir RA Fisher, cuarenta años antes (Fisher, 1922). El concepto de "evidencia estadística" establece dos propiedades básicas para su validez:

a) Interpretabilidad (el nivel de evidencia tiene la misma interpretación práctica independientemente del tamaño muestral).
b) Objetividad (el nivel de evidencia no varia en función del investigador).

Aunque pueda resultar sorprendente, ni la inferencia frecuentista ni la Bayesiana cumplen de forma estricta ambos criterios.

INTERPRETABILIDAD

La inferencia frecuentista fundamenta buena parte de su base teórica en los test de hipótesis y *p*-valores, un híbrido ciertamente peculiar si consideramos que cada uno de ellos fue desarrollado casi al mismo tiempo por Neyman y Pearson (1933) y Fisher (1925), con posiciones abiertamente enfrentadas. Tal como fue descrito por Fisher en su momento, el *p*-valor únicamente determina la probabilidad empírica del error de tipo I (rechazar la hipótesis nula cuando esta es realmente cierta), y la misma lógica se aplicaría en el caso del test de hipótesis de Neyman-Pearson. Estas aproximaciones en ningún caso pretende caracterizar el nivel de evidencia a favor de una u otra hipótesis, aunque a menudo se malinterpreten en este sentido (*p.ej*., a menor *p*-valor, mayor evidencia a favor de la hipótesis alternativa).

Desde un punto de vista evidencial, la probabilidad de observar evidencias erróneas (errores de tipo I y II en la inferencia frecuentista) debería converger a 0 a medida que aumenta el tamaño de la muestra (Bickel, 2012), tal como se observa en la Figura 1. No obstante, la propia definición de los test de hipótesis frecuentistas (Neyman y Pearson, 1993) imposibilitan este patrón al fijar un error de tipo I constante (típicamente 0,05), independientemente del tamaño de la muestra. Dentro de este contexto, el resultado de cualquier test de hipótesis, tanto expresado en términos de *p*-valor como de aceptación/rechazo de la hipótesis nula, no se puede interpretar en términos de evidencia estadística sin tener en cuenta el tamaño de la muestra analizada; de hecho, se tiende a asumir que un determinado *p*-valor aportará mayor evidencia en contra de la hipótesis nula en tamaños de muestra pequeños que en grandes (Royall, 1997). Tal como describió Goodman y Royall (1988), en muchos casos un *p*-valor de

0,05 se corresponde a un cociente de verosimilitudes que indica evidencias abrumadoras a favor de la hipótesis nula siempre que la muestra sea suficientemente grande.

OBJETIVIDAD

Desde un punto de vista teórico, los procedimientos de inferencia Bayesiana toman información de dos fuentes claramente diferenciadas, inherentes a la expresión

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Por un lado, los propios datos a analizar (**y**) bajo el modelo en cuestión, que se combinan e integran en el término $p(\mathbf{y}|\boldsymbol{\theta})$, dando lugar a una expresión en nada distinta a las funciones de verosimilitud frecuentistas o evidenciales (al menos, en términos algebraicos). En segundo lugar, la distribución *a priori* de los parámetros del modelo ($p(\boldsymbol{\theta})$), que caracterizaría el conocimiento previo del investigador sobre los mismos.

Tomando los factores de Bayes (Kass y Raftery, 1995) como la metodología estándar dentro la inferencia Bayesiana para la comparación de modelos (*i.e.,* comparación de hipótesis), resultan evidentes sus limitaciones en cuanto a objetividad. El cociente entre la probabilidad posterior de dos modelos distintos incluye el término $p(\boldsymbol{\theta})$ de ambos y, en consecuencia, condiciona el resultado a la opinión misma (*a priori*) del investigador. Dos investigadores distintos deberían obtener distintos resultados, dado que resulta muy poco probable que ambos tengan exactamente el mismo grado de conocimiento previo en relación a los parámetros del modelo, a menos que open por *a prioris* genéricos que, en si, desvirtuarían la esencia misma de la inferencia Bayesiana (no obstante, en la práctica, se usan por comodidad). No sería impensable que las diferencias *a priori* de los investigadores pudieran, incluso, invertir el sentido del mismo factor de Bayes en casos extremos, condicionando de manera evidente el resultado del análisis a fuentes de información, ajenas a los datos y, en buena medida, subjetivas.

En el caso de la inferencia evidencial, se evitaría esta injerencia por parte del investigador, dado que el cociente de verosimilitudes no incluye información ajena a los datos a analizar, más allá del modelo mismo de análisis.

UMBRALES DE EVIDENCIA

De la misma forma que la inferencia frecuentista establece un umbral como criterio para discernir entre ambas hipótesis (típicamente, α=0,05), y la inferencia Bayesiana dispone también de umbrales para el factor de Bayes con el objetivo de categorizar el resultado a favor de un u otro modelo (*e.g.*, Jeffreys, 1961), la inferencia evidencial toma como referencia los umbrales 8 y 32 para distinguir entre evidencia débil (1 a 8), moderada (8 a 32) y fuerte (>32) a favor del modelo situado en el denominador del cociente de verosimilitudes (Royall, 1997, 2000). No obstante, se han sugerido otros umbrales posibles, algunos compartidos con los de inferencia Bayesiana (Jeffreys, 1961), y también pueden adaptarse fácilmente a situaciones de *multiple testing*.

Más allá de la controversia que pueda generar la entrada en escena de un tercer paradigma inferencial, este debería verse como una aproximación alternativa a las inferencias frecuentista y Bayesiana, todas ellas centradas en responder preguntas distintas aunque, quizá demasiado a menudo, se confundan en un mismo sentido y objetivo.

REFERENCIAS BIBLIOGRÁFICAS

• Barnard, G.A. 1949. Statistical Inference. J. Royal Stat. Soc. B 11: 115-149. • Bickel, D.R. 2012. The strength of statistical evidence for composite hypotheses: inference to the best explanation. Stat Sinica 22: 1147-1198. • Birnbaum, A. 1962. On the foundations of statistical inference. J. Am. Stat. Assoc. 57: 269-306. • Edwards, A.F. 1972. Likelihood. An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference. Cambridge University Press, Cambridge, Reino Unido. • Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. Phil. Trans. R. Soc. A 222: 594-604. • Fisher, R.A. 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edimburgo, Reino Unido. • Goodman, S.N. & Royall, R. 1988. Evidence and scientific research. Am. J. Public Health 78: 1568-1574. • Hacking, I. 1965. Logic of Statistical Inference. Cambridge University

Press, Cambridge, Reino Unido. • Jeffreys, H. 1961. Theory of Probability. Oxford University Press, Oxford, Reino Unido. • Kass, R.E. & Raftery, A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90: 773-795. • Neyman, J. & Pearson, E. 1933. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. R. Soc. Lond. A 231: 289-337. • Royall, R. 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York, NY, EUA. • Royall, R. 2000. On the probability of observing misleading statistical evidence (with discussion). J. Am. Stat. Assoc. 95: 760-767. • Strug, L.J. 2018. The evidential statistical paradigm in genetics. Genet. Epidemiol. 42: 590-607.

Agradecimientos: Este trabajo se enmarca en el proyecto CGL2016-80155-R.

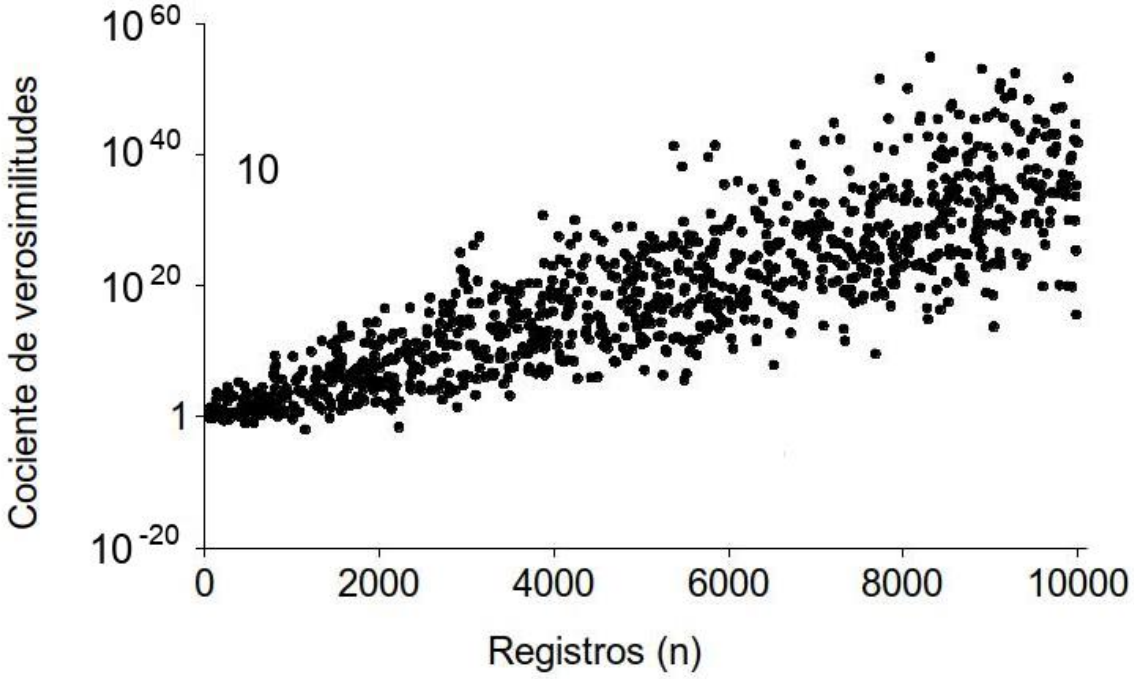

Figura 1*. Cociente de verosimilitudes (hipótesis nula vs hipótesis alternativa) en datos simulados de asociación genética tipo GWAs simulados bajo hipótesis nula verdadera.*

EVIDENTIAL STATISTICS… THINGS ALWAYS COME IN THREES

ABSTRACT: Frequentist and Bayesian methodologies provide key tools for animal breeding, and they have attracted almost all statistical attention during the last decades. Nevertheless, a third inferential paradigm was suggested long time ago and could be of special interest in the near future. This is known as evidential inference and relies on the ratio of likelihood functions as the reference statistic to compare hypothesis. The main objective of evidential inference focuses on calculating the strength of statistical evidence favouring (or disfavouring) a given hypothesis in comparison with an alternative one, this guaranteeing both objectivity (*i.e.,* the strength of evidence does not vary from one researcher to another) and interpretability (the strength of evidence has the same practical interpretation for any sample size). These two criteria become essential when comparing with frequentist and Bayesian inferences.

Keywords*:* Bayes factor, evidential statistics, hypotesis testing, statistical evidence