Ecosistemas y Recursos Agropecuarios

Prueba de hipótesis en la investigación forestal, agropecuaria y en la ecología: retos y malentendidos sobre el uso de los niveles de significancia de 0.05 y 0.01

Hypothesis testing in forestry, agriculture and ecology: Use and overuse of the 0.05 and 0.01 significance levels

Pablo Antúnez[1], **Ernesto Alonso Rubio-Camacho**[2,3], **Christoph Kleinn**[4*]

[1]Inventory and Remote Sensing. Faculty of Forest Sciences and Forest Ecology. University of Goettingen. Büsgenweg 5, D-37077 Göttingen, Germany-División de Estudios de Posgrado, Universidad de la Sierra Juárez, Av. Universidad S/N, Ixtlán de Juárez, 68725 Oaxaca, México.
[2]Department Ecoinformatics, Biometrics & Forest Growth. Georg-August University of Goettingen. Büsgenweg 4. 37077 Goettingen, Germany.
[3]INIFAP-CIRPAC C.E. Centro Altos: Av. Biodiversidad núm. 2470, Col. Las Cruces, CP. 47600. Tepatitlán, Jalisco, México.
[4]Inventory and Remote Sensing. Faculty of Forest Sciences and Forest Ecology. University of Goettingen. Büsgenweg 5, D-37077 Göttingen, Germany.

*Autor de correspondencia: pantunez4@gmail.com

Carta al editor
Recibida: 08 de junio 2020
Aceptada: 13 de marzo 2021

Como citar: Antúnez P, Rubio-Camacho EA, Kleinn C (2021) Prueba de hipótesis en la investigación forestal, agropecuaria y en la ecología: retos y malentendidos sobre el uso de los niveles de significancia de 0.05 y 0.01. Ecosistemas y Recursos Agropecuarios 8(1): e2616. DOI: 10.19136/era.a8n1.2616

Generalidades de la significancia estadística

La estadística es una herramienta metodológica importante para estudiar e interpretar fenómenos y eventos en diferentes disciplinas, a partir de datos empíricos, en un marco de protocolos y normas estadísticas. Siendo la estadística inferencial, y en particular, la prueba de hipótesis uno de los instrumentos fundamentales en la toma de decisiones al examinar el atributo de una población a partir de una muestra. La importancia de la prueba de hipótesis radica en la necesidad de medir el grado de fiabilidad o significación de los resultados de un estudio.

En este documento se enlistan algunos atributos de la significancia estadística dignos de tomarse en cuenta para verificar la pertinencia de su uso y aplicación. Se hace una breve recapitulación histórica de la prueba de hipótesis y se discute la probabilidad (P) como un componente importante de esta. Con ello, los autores esperan clarificar algunos puntos en torno a este tema y puedan servir de apoyo para realizar interpretaciones más idóneas. Es importante destacar desde un inicio que la significancia estadística es un concepto metodológico "*artificial*" basado en una serie de suposiciones dignas de ser comprendidas para evitar interpretaciones erróneas.

1. La significancia estadística informa sobre las características de una población a partir de una muestra limitada, al desconocerse con certeza el parámetro de interés de una población determinada.

2. Al aplicar la inferencia estadística clásica, se asume que la muestra se obtiene aleatoriamente con estimadores insesgados de la media y de la varianza. Por lo tanto, al obtener los datos de la muestra sin un diseño estadístico o sin estimadores insesgados, se prohíbe aplicar el concepto de la significancia estadística. Por ejemplo; en inventarios forestales, comúnmente se opta usar el muestreo sistemático y, al no haber estimadores insesgados de la variancia, debe observarse con atención los niveles de significancia y los márgenes del error al aplicar las pruebas estadísticas. De manera similar en ecología, al usar diseños de parcelas "*k-tree*" (también llamados "parcelas de recuentos fijos"; por decir, los 6 árboles más cercanos al punto de muestro), habría que evitar estimadores insesgados (véase Kleinn y Vilčko (2006) y trabajos citados por estos autores). Si las muestras se toman mediante el diseño de las parcelas "*k-tree*", se prohíbe aplicar pruebas estadísticas; o al menos, una interpretación idónea de los resultados no sería posible.

E. ISSN: 2007-901X

3. La significancia estadística informa exclusivamente acerca de lo que puede decirse de una población a partir de una muestra. Nada más. En principio, dista de la relevancia práctica cuyo concepto difiere del primero y, por lo tanto, debe interpretarse de manera separada, en virtud de que, por una pequeña diferencia podría haber "significancia" pero sin tener ninguna relevancia en el contexto técnico. De forma similar, existe la posibilidad de observar una diferencia grande - potencialmente relevante en términos prácticos - pero sin *significancia estadística*, debido a una muestra muy pequeña.

4. El punto tres adquiere mayor relevancia en ciertos casos; por ejemplo, al examinar las diferencias de las medias poblacionales, una diferencia significativa podría indicar una diferencia efectiva y real entre los promedios estimados, pero dicha significancia también podría deberse al tamaño de las muestras comparadas. Entonces, teóricamente podría lograrse una "diferencia pequeña" pero significativa con aumentar el tamaño de la muestra.

5. Lo dicho anteriormente aplica al hacer inferencia estadística en el contexto de las pruebas de significancia, aunque también en los análisis de correlación y de regresión. Por ejemplo, en este último caso, la probabilidad de encontrar "significancia de la regresión" se incrementa con el tamaño de la muestra; no obstante, no siempre hay una correspondencia con la dispersión de los datos (alta dispersión de la nube de puntos) ni con el coeficiente asociado a la pendiente de regresión (un coeficiente muy pequeño).

¿Por qué se usan los niveles de significancia de 0.05 y 0.01?, probablemente por convención, convirtiéndose muchas veces en un mecanismo de toma de decisiones dicotómicas, al observar únicamente los valores de P, a pesar de que no existe un respaldo teórico para hacerlo (Clark 2004, Dahiru 2008). Para entender por qué esta forma de proceder no es del todo adecuada, conviene recordar qué es la prueba de hipótesis y qué indica el valor de P.

La definición de la prueba de hipótesis varía según los autores, pero en esencia, se refiere a un conjunto de procedimientos cuyo propósito es validar si la información de una muestra sostiene o refuta un hecho o conjetura sobre la naturaleza de una población (Pagano 1999, Hines y Montgomery 2002, Lind *et al*. 2001). El término estadístico que refiere a la conjetura se conoce como hipótesis estadística, denotada comúnmente como H_0, que puede ser verdadera o falsa. La hipótesis estadística se plantea a manera de enunciado sobre la distribución de probabilidad de una variable aleatoria (Hines y Montgomery 2002) que requiere de variables aleatorias independientes e idénticamente distribuidas (Lehman y Romano 2008). El enunciado planteado, representa la información que contiene los datos de interés. Por consiguiente, una prueba de hipótesis se inicia con una afirmación o suposición, sobre uno o más parámetros de una población (Lind *et al*. 2001).

Origen de la prueba de hipótesis: Dos teorías antagónicas

Los primeros ensayos lógicos relacionados con la prueba de hipótesis se remontan hacia 1857-1936 con la prueba de ji al cuadrado (χ^2) de Karl Pearson (Lehman 1993), aunque los planteamientos formales de la prueba de hipótesis se presentaron a principios del siglo XVIII (Stigler 1986, Lehman 1993) y la propuesta metodológica como tal, se desarrolló entre 1915 y 1933 como un resultado del análisis de dos escuelas de pensamiento; por un lado, de Ronald Fisher (1890-1962) y por el otro, de Jerzy Neyman (1894-1981) junto con Egon Pearson (1895-1980) (Lehman 1993, Batanero 2000). La principal diferencia entre estas dos teorías, no radica en los cálculos sino en las concepciones y su razonamiento subyacente (Batanero 2000). El enfoque de Fisher se caracteriza por definir únicamente una hipótesis (la nula = H_0) y a partir de ella, con base en la distribución muestral del estadístico de prueba, se estima la probabilidad de una muestra de datos para decidir su rechazo o no rechazo. El enfoque de Neyman-Pearson, en cambio, se caracteriza por la adición de una hipótesis alternativa (H_A) en contraposición con la hipótesis nula, lo que conduce a la definición de dos regiones: región de rechazo y región de no rechazo,

E. ISSN: 2007-901X

además de los errores asociados a la decisión sobre la hipótesis nula denominados errores Tipo I y Tipo II (Levin 1998, Batanero 2000, Inzunsa-Cazares y Jiménez-Ramírez 2013). De forma simplificada, estos errores ocurren al rechazar la H_0 siendo verdadera (error Tipo I); o bien, al no rechazar H0 siendo falsa (error Tipo II) (Zar 2010). Estos dos tipos de errores, son los conocidos universalmente, aunque se sugieren otros más (Daniel y Onwuegbuzie 2000).

Entonces, el valor de P podría ser definido como la probabilidad, bajo el supuesto de no diferencia (H0), de obtener un resultado igual o más extremo al observado; por ejemplo, la diferencia de la media muestral entre dos grupos (Dahiru 2008, McLean y Ernest 1998, Wasserstein y Lazar 2016). Además, representa la probabilidad de obtener un resultado similar en un experimento repetido en igualdad de condiciones (Molina-Arias 2017). El valor de P, es pues, una medida de evidencia a *posteriori* mientras que el nivel de significancia, asociado al error tipo I, indica una tasa de error a *priori* (Lehmann 1993, Hubbard y Ballari 2003, Sterne y Smith 2001).

Los niveles de significancia de 0.05 y 0.01

El uso de los valores de 0.05 y 0.01 se ha hecho habitual y esto se debe, en parte, a la divulgación de su autor principal: Fisher (Sterne y Smith 2001, Moran 2004), pero tanto Fisher como Neyman-Pearson, advirtieron que no hay una regla rígida establecida para los umbrales ni para los valores de las tasas de error (Senn 2001, Rebasa 2003), sino dependerá del tipo de problema y de la magnitud del error que el investigador está dispuesto a tolerar. Por ello, Monterrey y Gómez-Restrepo (2007) enfatizan que en ningún caso es válido y correcto considerar como valores universales, únicos e inflexibles los límites 0.01 y 0.05, sino deben cambiarse y adaptarse a la naturaleza de cada investigación.

¿Qué sugiere el valor de P?

En este sentido, es pertinente recordar que el valor de P sugiere la evidencia en contra de la hipótesis nula. Un valor de P más grande que un nivel de significancia indica que es compatible con la hipótesis nula dada la reducida discrepancia (Rebasa 2003). El rechazo de la hipótesis nula, no sugiere total certeza, dado que el parámetro evaluado, aún puede caer en la zona especificada para la hipótesis alternativa (Batanero 2000). En este último caso, el error radica en que no se descarta que los resultados puedan ocurrir debido al azar.

Según los principios estadísticos, la obtención y uso apropiado de los valores de P, podrían resumirse en tres pasos: (1) establecer la probabilidad del error Tipo I (alfa) antes de observar los datos, (2) calcular el valor de P usando un muestreo aleatorio y (3) emplear la regla de decisión pudiendo rechazarse o no la hipótesis nula (Kuffner y Walker, 2019). El uso de los valores de P, según los principios estadísticos, evita principalmente los sesgos en la literatura científica al reducir los reportes selectivos o la práctica conocida en inglés cómo "*P-hacking*" donde el investigador se esmera por superar el umbral de $alfa$ (α) y obtener resultados confirmatorios y estadísticamente significativos cuyo propósito principal es incrementar la probabilidad de aceptación del manuscrito en una revista científica en lugar de un reporte basado en la evidencia científica (Grabowski 2016). Por esta razón, algunas revistas como "*Basic and Applied Social Psychology*" han prohibido el uso del valor de P en sus publicaciones (Trafimow y Marks 2015).

Todo lo anterior nos conduce a que es importante incluir en los resultados los valores de P de forma explícita, aun cuando estos no hayan superado el nivel de significancia establecido; además, es pertinente incluir los intervalos de confianza cuando sea posible, los cuales aportan un rango plausible de valores en el cual, el parámetro poblacional de interés se localiza. El intervalo de confianza se complementa con los valores de P, al expresar el grado de incertidumbre del estadístico además de aportar una información descriptiva de la variable de interés (Candia y Caiozzi 2005, Newcombe y Merino-Soto 2006). La inclusión del intervalo de confianza adquiere relevancia mayor en estudios de sucesos o eventos "escasos"; por ejemplo, al estudiar las tasas de deforestación donde el valor porcentual de interés (tasa de deforestación) a menudo es pequeño (por

ejemplo, 1%). En estos casos, el riesgo del error de estimación se incrementa y es de suponerse que, al contrastar estimadores provenientes de estos datos, las diferencias serían mínimas.

En general, es deseable que el analista de datos comprenda los fundamentos y la lógica que subyace a las pruebas de hipótesis, de esta forma se hará una correcta aplicación de esta herramienta al reportar los resultados de investigación. Así mismo, conocer su perspectiva histórica ayuda a entender mejor sus bases teóricas y el aporte de cada corriente que le dieron origen y sustento. Estas bases permiten comprender con mayor claridad las bondades y limitaciones de la prueba de hipótesis y el uso de los valores de probabilidad (P).

Los autores desean agradecer a dos revisores anónimos y al editor de la revista ERA por las sugerencias para mejorar el texto.

LITERATURA CITADA

Batanero C (2000) Controversies around the role of statistical tests in experimental research. Mathematical thinking and learning 2: 75-97.

Candia R, Caiozzi A (2005). Intervalos de confianza. Revista médica de Chile 133: 1111-1115.

Clark ML (2004) Los valores P y los intervalos de confianza: ¿en qué confiar? Revista Panamericana de Salud Publica/Pan American Journal of Public Health 15: 293?296.

Dahiru T (2008) P-value, a true test of statistical significance? A cautionary note. Annals of Ibadan postgraduate medicine 6: 21-26.

Daniel LG, Onwuegbuzie AJ (2000) Towards an Extended Typology of Research Errors. https://files.eric.ed.gov/fulltext/ED449166.pdf.

Evwcombe RG, Merino-Soto C (2006) Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas. Interdisciplinaria, 23(2): 141-154.

Grabowski B (2016) “P < 0.05” Might Not Mean What You Think: American Statistical Association Clarifies P Values. JNCI: Journal of the National Cancer Institute 108: 4-5.

Hines WW, Monthomery DC (2002) Probabilidad y Estadística para Ingeniería. Compañía editorial continental, 11va edición. México. 834 p.

Hubbard R, Bayarri MJ (2003) P values are not error probabilities. Institute of Statistics and Decision Sciences, Working Paper 03-26: 27708-0251.

Inzunsa-Cazares S, Jiménez-Ramírez JV (2013) Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipótesis. Revista latinoamericana de investigación en matemática educativa 16: 179-211.

Kleinn C, Vilčko F (2006) A new empirical approach for estimation in k-tree sampling. Forest Ecology and Management 237: 522-533.

Kuffner TA, Walker SG (2019) Why are p-Values Controversial? American Statistician 73: 1-3. Doi: 10.1080/00031305.2016.1277161.

Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association 88: 1242-1249.

Levin JR (1998). What If There Were No More Bickering About Statistical Significance Tests? Research in Schools 5: 43-53.

E. ISSN: 2007-901X

Lind DA, Mason RD, Marchal WG (2001) Estadística para administración y economía. McGraw-Hill/Interamericana editores. 3^{a} edición. México. 573p.

McLean JE, Ernest JM (1998) The role of statistical significance testing in educational research. Research in the Schools 5: 15-22.

Molina-Arias M (2017) ¿Qué significa realmente el valor de p? Rev Pediatr Aten Primaria 19: 377-381.

MonterreyP, Gómez-Restrepo C (2007) Aplicación de las pruebas de hipótesis en la investigación en salud. ¿Estamos en lo correcto? Universitas Médica 48: 193-206.

Moran JL (2004). Point of view: A farewell to P values? Critical Care and Resuscitation 6: 130-7.

Pagano RR (1999) Estadística para las ciencias del comportamiento. International Thomson. 5^{a} ed. México. 548p.

Rebasa P (2003) Entendiendo la "p < 0,001". Cirugía Española 73: 361-365.

Senn S (2001). Two cheers for P-values? Journal of Epidemiology and Biostatistics 6: 193-204.

Sterne JA, Smith GD (2001) Sifting the evidence? what's wrong with significance tests? Physical Therapy 81: 1464-1469.

Stigler SM (1986) The history of statistics: The measurement of the uncertainty before 1900. Harvard University Press. Massachussetts and London, England. 432p.

Trafimow D, Marks M (2015) Editorial. Basic and Applied Social Psychology 37: 1-2. Doi: 10.1080/01973533.2015.1012991.

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: Context, process, and purpose. American Statistician 70: 129-133.

Zar JH (2010) Biostatistical analysis. 5ta ed. Prentice Hall Inc. New Jersey. United States. 944p.