REVISTA CUBANA DE CIENCIAS BIOLÓGICAS
http://www.rccb.uh.cu

CARTAS AL EDITOR

Sobre el verdadero valor de las pruebas de hipótesis.

¿Qué significa realmente la P de una prueba estadística y cuándo podemos decir que no?

On the true value of null hipothesis testing ¿What really mean the p-value in a statistical test and when can we said just No?

Dennis Denis Ávila

Dpto. Biología Animal y Humana
Facultad de Biología
Universidad de La Habana

* Autor para correspondencia:
ralonso@fbio.uh.cu

Recibido: 2008-10-07

Aceptado: 2008-11-20

La epistemología popperiana y el enfoque frecuentista de Fisher constituyen propiedades centrales del paradigma actual de la ciencia, caracterizado también por un fuerte mecanicismo cartesiano subyacente en los análisis y por la hipérbole del Método. Sin embargo, en muchas disciplinas se han comenzado a acumular razones contrarias al empleo de tales metodologías, lo cual hace avizorar un próximo cambio de paradigmas. Desafortunadamente, el empleo habitual y rutinario muchas veces se hace de forma automática y acrítica, sin tenerse en cuenta los inconvenientes de cada una.

Las pruebas de hipótesis constituyen una de las ramas fundamentales del análisis estadístico, ya que sirven de herramienta para establecer la probabilidad de que un conjunto de datos cumpla determinada condición. Independientemente de que su valor epistemológico está siendo severamente cuestionado, han sido históricamente los métodos de más amplio uso para precisar probabilidades asociadas a las inferencias. En este contexto, se establece una hipótesis nula (H), que por definición se establece siempre como la de «no diferencias» o «no efectos», y una hipótesis alternativa (H) que, frecuentemente, plantea el efecto que se quiere demostrar. La prueba se desarrolla con un conjunto de datos y a partir del valor de P obtenido se toma una decisión acerca de la hipótesis probada (nula). De tal resultado se concluirá si hay, o no, resultados estadísticamente significativos.

Ahora bien, la expresión «estadísticamente significativo» es hipnótica y seductora, pero demasiado fácil de malinterpretar. Generalmente, se asume que obtener, en una prueba de hipótesis, una P de 0,03 indica que hay una probabilidad muy baja de que el patrón observado en los datos sea resultado del azar y, por tanto, se decide que este debe ser verdadero y con algún significado biológico. Nada más lejos de la realidad. La interpretación de los resultados de las pruebas de hipótesis es una tarea muy compleja y depende más de lo que se

conoce previamente sobre el fenómeno estudiado, quede lo que se calcule a partir de los datos tomados. Resulta alarmante observar cómo el proceso de análisis de datos se ha mecanizado hasta el punto en que se ha perdido la libertad para oponerse aun resultado estadísticamente significativo y se ha hiperbolizado la importancia de las probabilidades resultantes de estas pruebas.

La vía para concluir que una hipótesis nula es falsa es la siguiente: si se obtiene una P muy pequeña –menor de 0,05 en la mayoría de los casos– ello indica que los datos son muy poco probables de obtener si la hipótesis nula es verdadera. Tal resultado puede significar, a su vez, dos cosas: una, que sucedió algo poco probable; dos, que la hipótesis nula realmente no era verdadera. En los datos no hay ningún otro elemento, pero ante estas opciones tomamos una decisión y rechazamos la igualdad; aunque nunca se sabe cuál de estos dos eventos ha realmente sucedido. Al fin y al cabo, las cosas poco probables suceden constantemente.

Ya de por sí, esta decisión ante el resultado de una prueba es arbitraria. Sin embargo, todo se complica aún más cuando vemos que el proceso de decisión depende del resultado de la P obtenida en la prueba realizada. Pero, ¿qué significa realmente este valor?

El valor de P de una prueba y su interpretación no es un proceso obvio. Sin embargo, resulta muy fácil «relajar» la definición formal y usar una más sencilla y aproximada. La tendencia natural al facilismo ha conllevado una distorsión total del valor de P: ¿a quién se le ocurre explicarle a un niño o aun joven la teoría sintético-moderna de la evolución para darle las razones del porqué las jirafas tienen el cuello largo, si la explicación de Lamarck es tan fácil de entender y aparenta parecerse a la realidad? ¿Qué es más fácil decir: «Se rechaza la hipótesis nula para un nivel de significación de 0,05» o «Hay un 95 % de probabilidades de que tal fenómeno ocurra»? ¡Viva el lamarckismo!

Como el ejemplo brindado, existen muchos mitos acerca de la P y su interpretación. Pero, en general, todos parecen coincidir en la creencia de que una P pequeña indica que la probabilidad de que se cumpla la hipótesis nula es muy baja y, por tanto, se rechaza, con lo que se presupone que los resultados demuestran la hipótesis alternativa. Por tanto, se etiqueta el resultado con la frase «estadísticamente significativo».

Sin embargo, ¿indica una P pequeña que hay altas posibilidades de que la diferencia entre las dos muestras comparadas sea real? Veamos, paso a paso, el siguiente ejemplo hipotético: si se quiere probar que una especie de pez (A) tiene un peso (biomasa) diferente de otra especie (B), se toma una muestra de cada especie y se comparan estadísticamente los pesos medios obtenidos. Para hacerlo correctamente, basándonos en la dispersión de las medidas del peso –determinada previamente por un muestreo piloto– y en una diferencia en peso que se consideraría biológicamente significativa, se elige un tamaño de muestra adecuado que permitiera tener un 80 % de potencia en la detección de esa diferencia, para un nivel de significación de P = 0,05.

Si se toman dos muestras con este tamaño apropiado, y se obtiene, como resultado de la comparación, una P = 0,03 –diferencias significativas ya que son menores de 0,05–, ¿indicaría ello que una especie pesa más que la otra? Generalmente se asume que sí; pero veamos cómo se interpretaría este mismo resultado si la realidad, que es desconocida para el investigador, tuviese tres escenarios diferentes:

1 Si realmente solo el 10 % de los individuos de la especie A es mayor que los de la especie B.

2 Si realmente el 80 % de los peces A son menores que los B.

3 Si realmente solo el 1 % de los peces A son diferentes a los B.

Recuérdese que está claro que en la mayoría de nuestros estudios no sabemos cuál es el contexto del resultado, por lo que no podemos predecir cuáles la realidad a la cual, precisamente, intentamos acercarnos realizando estas comparaciones. Ahora bien, procedemos a interpretar estadísticamente lo que se asumió al inicio para el análisis. Desde el punto de vista estadístico, como se asumió un nivel de significación de 0,05, ello indica que el riesgo máximo que se estaba dispuesto a correr de cometer un error del primer tipo –rechazar la hipótesis nula cuando deberíamos retenerla– es de 5 %. Por lo que, si las especies son de igual peso, como máximo, en un 5 % de los muestreos que se realicen van a aparecer valores de P significativos de la prueba –menores de 0,05– solo por azar. En estos casos, se detectará una diferencia que realmente no existe.

Por otra parte, como la potencia es del 80 %, significa que se asumió como aceptable tener unerror del segundo tipo –retener una hipótesis nula cuando debía haber sido rechazada– del 20 %. Por lo que, como máximo, se acepta que el 20 % de las muestras tomadas en dos especies de peces que sí difieren en peso podría dar valores de P mayores de 0,05 –no significativos– solo por azar. En estos casos, las pruebas fallarán en detectar la significación.

Lo anterior es lo que muestran esas probabilidades de error que se asumen desde un inicio. ¿Qué pasaría si, en lugar de tomar solo una muestra de individuos de cada especie y compararlas entre sí, se repite el muestreo y la comparación mil veces–obviando la posible tasa de error experimental si se siguiera ese procedimiento. Llevemos estas probabilidades de errores que se asumen a cantidades de comparaciones –de estas mil–, para cada escenario posible (tabla 1):

- Escenario I: si el 10 % de los peces dela especie A fueran más pesados que los peces B, un muestreo totalmente aleatorio debería rendir como resultado más probable que, de los 1000 pares demuestras, 100 reflejarán un aumento de peso y 900 no lo reflejarán. Ahora bien, de las 100 que sí lo reflejan, la prueba fallará en detectar el aumento en 20 de ellas –que es el 20 %: la tasa de error del segundo tipo– y sí lo detectarán correctamente en las 80 restantes. Luego, de las 900 comparaciones que no reflejaron diferencias de peso reales, en 45 la prueba, solo por azar, podría dar significativa, que es la tasa de error del segundo tipo asumida (5 %). En las 855restantes, la prueba sí detectará que no hubo diferencias.
- Escenario II: si el 80 % de los peces A son menos pesados que los B, probabilísticamente, 800 de las 1000 muestras reflejarán una verdadera diferencia del peso y 200, no. Pero, de estas 800 que sí lo reflejan, según el riesgo de cometer el error del segundo tipo que se asumió (20 %), como máximo se fallará en detectar la diferencia en 160 de ellas y sí se detectará en las 640 restantes. Pero, de los 200 pares de muestras que no reflejan un aumento real de peso, en 10 la prueba, solo por azar, podrá dar resultados significativos, en correspondencia al riesgo máximo de cometer error del segundo tipo asumido.
- Escenario III: si solo el 1 % de los individuos de la especie A son diferentes delos de la especie B, de las 1000 muestras solo 10 reflejarán esta diferencia y 990, no; pero de estas, la prueba fallará en detectarla en dos de ellas –el 20 %, error del segundo

Tabla 1. Probabilidades verdaderas de error asociadas al resultado de una prueba estadística con resultados significativos, en tres posibles escenarios.

Table 1. Real error probabilities asociated to statistical test result with significant outcome, in three possible scenaries.

ESCENARIOS	MUESTRAS QUE REALMENTE DIERON PESOS MAYORES	MUESTRAS QUE NO DIERON AUMENTO EN EL PESO	TOTAL
I (10 % DE A MAYOR QUE B)			
Cantidad de pruebas que detectarán diferencias significativas (P < 0,05)	80	45	125
Cantidad de pruebas que no detectarán diferencias significativas (P > 0,05)	20	855	875
TOTAL	**100**	**900**	**1000**
II (80 % DE A MENOR QUE B)			
Pruebas que detectarán diferencias significativas (P < 0,05)	640	10	650
Pruebas que no detectarán diferencias significativas (P > 0,05)	160	190	350
TOTAL	**800**	**200**	**1000**
III (1 % DE A MENOR QUE B)			
Pruebas que detectarán diferencias significativas (P < 0,05)	8	50	58
Pruebas que no detectarán diferencias significativas (P > 0,05)	2	940	942
TOTAL	**10**	**990**	**1000**

tipo– y sí la detectará correctamente en las 8 restantes. De los 990 pares de muestras que no reflejarán una diferencia en peso, en 50 la prueba, por azar, dará significativa.

Según los resultados tabulados, y una vez sumadas las columnas, en el contexto I, del total de 1000 pruebas, se obtendrá que hay significación en 125casos –correctos más los erróneos– y se dirá que no hay significación en 875. En el escenario II se detectará diferencia en 650 casos y se dirá que no hay significación en 350 casos. Y en el escenario III, la prueba estadística diría que hay significación en 58 casos y que no la hay en 942 casos.

Lo anterior sería lo esperable si se hicieran 1000 muestreos y 1000 pruebas en cada contexto. Pero solo se tomó una muestra simple, sin saber cuál contexto verdadero se cumple, y se hizo una única prueba estadística que dio un valor, estadísticamente significativo, de $P < 0,03$. Si se interpreta mecánicamente este resultado se llega a la conclusión de que los peces de la especie A son diferentes en peso a los de la especie B. ¿Cuál es la probabilidad real de que esto sea cierto? Calculémosla en cada contexto:

En el escenario I, de las pruebas estadísticas que se hicieran, 125 habrían dado significativas cuando solo en 80 muestras habría un aumento verdadero del peso. Por tanto: 80 / 165 = 0, 64 (64 %). Siesta fuera la realidad, aunque la prueba estadística reportara diferencias significativas hay un riesgo del 36 % de que este resultado fuera por azar.

En el escenario II, de las pruebas posibles a realizar, 650 habrían dado significativas y en 640 de estas las diferencias del peso habrían sido reales, o sea, 640/ 650 = 0,98 (98 %). El resultado de la P significativa tendría una seguridad del98 % de que fuese real (solo un 2 % de que fuese por azar).

En el escenario III, por último, ante la misma P significativa de la comparación, habría solo una probabilidad del 14 % de ser verdadera la diferencia. ¡La mayoría de los resultados significativos que se obtendrían serían por casualidad!

Por lo tanto, de lo anterior se deduce que el valor de la P nunca dice por sí solo la probabilidad real de que las diferencias o efectos detectados sean verdaderos o probables. Su interpretación siempre estará afectada por las probabilidades de error y, amenos que se tenga un conocimiento previo –y cierto– acerca de lo que existe en la realidad, no se puede interpretar adecuadamente un valor de P significativo. De ahí que las pruebas de hipótesis no son totalmente confiables, ni informativas. No pueden interpretarse los valores de una prueba a partir de «la nada». Sin embargo, en la mayoría delas ocasiones, no sabemos qué sucede en la realidad sobre lo que estamos estudiando. La interpretación de estos resultados requiere mucho sentido común, intuición y juicio, por lo que, bajo ningún concepto, puede ser mecánica o automática.

Este cuidado debe tenerse en cuenta no solo cuando obtenemos una prueba con un resultado «estadísticamente significativo» sino también cuando no lo tenemos. Al resultar una prueba en un valor de P superior a 0,05 no podemos, automáticamente, asumir que no existe el efecto que deseamos probar. Hay que analizar la potencia que se tuvo en la prueba para no caer en error del segundo tipo.

La potencia o poder de una prueba de hipótesis es un concepto que se aborda en la mayoría de los textos básicos de estadística y en los cursos de posgrado y pregrado de Biometría. Sin embargo, si bien es conocido, para la mayoría de las personas constituye una suerte de caja negra que está relacionada con los tipos de errores que se pueden cometer en las pruebas de hipótesis. Estos tipos de errores son dogmas tales que cualquier conocedor medio de aspectos estadísticos es capaz de recitarlos. A pesar de ello, la aplicación directa de este concepto a nuestras investigaciones es muy pobre o nula.

Independientemente del resultado de una prueba siempre existirá la probabilidad de equivocarnos. Si se encuentra un resultado significativo, se rechaza la hipótesis nula; pero cabe la posibilidad de que esta hipótesis sea realmente verdadera. En este caso, cometemos el error del primer tipo. Si no se alcanza la significación estadística, retenemos la hipótesis nula cuando, tal vez, debíamos rechazarla. En este caso, cometemos el error del segundo tipo. La probabilidad de cometer error del primer tipo viene dada por el nivel de significación que se asume –antes de tomar los datos– (α), y que es predeterminada en todos los casos: no depende del tipo de estudio ni de la calidad de los datos y, por tanto, es muy fácil de «definir». Por otra parte, la probabilidad de cometer error del segundo tipo –denotada con la letra β– depende directamente de los datos –su variabilidad, el tamaño de muestra, el tipo de prueba o el tamaño del efecto que se desea detectar– y generalmente requiere

cálculos más complejos. La potencia de una prueba estadística para detectar un efecto dado está determinada por 1-β.

Existe un error metodológico generalizado, que es darle más valor a los resultados que provengan de pruebas con significación estadística. Ello da lugar al llamado «sesgo de publicación», que se refiere al hecho de que en los artículos científicos existe una fuerte desproporción entre la cantidad de resultados significativos y los no significativos que se informan. Aun cuando se suele reconocer que no encontrar diferencia puede tener tanto valor científico como su caso contrario, nos decepcionamos cuando nuestras pruebas de hipótesis no dan significativas y nos alegramos cuando sí sucede. Sin embargo, cuando no se obtiene una diferencia estadísticamente avalada, la mayoría delas investigaciones se detienen allí. Algunos autores no se conforman, defienden sus convicciones y plantean posibles tendencias: o que el resultado se puede deber al pequeño tamaño de muestra, o a la alta variabilidad encontrada. ¿Es confiable este resultado negativo? Por supuesto que no, porque no se tiene seguridad de no estar cometiendo el otro error: no haber detectado una diferencia existente. No obstante, esta probabilidad nunca se expresa. ¿Por qué?

Esta situación se ha generado por varios factores. Primero, debido a la falta de conocimiento acerca de la gravedad que puede tener el cometer este error en particular; segundo, como resultado de la tendencia a aplicar esquemas rígidos y mecánicos para el análisis de los datos; tercero, porlas complejidades estadístico -matemáticas de su cálculo; y, en cuarto lugar, por la ausencia de programas estadísticos de uso general donde puedan realizarse.

En cada investigación específica, el cometer uno u otro de los errores puede tener diferentes implicaciones. En las investigaciones básicas puede dar origen, apoyar o generalizar una hipótesis falsa; pero en las investigaciones aplicadas, biomédicas o medioambientales la repercusión puede ser mucho más severa. Por ejemplo, si se quisiese comprobar el efecto del impacto humano sobre una comunidad, cometer un error del segundo tipo sería concluir que hay un impacto negativo –que en la realidad no existe– y esto podría conducir a proponer medidas de mitigación innecesarias, con lo cual se perderían recursos económicos o tiempo. Sin embargo, cometer un error del segundo tipo llevaría a negar un impacto existente, con lo cual las consecuencias serían no tomar medidas importantes para proteger el ecosistema. A corto plazo, los recursos económicos gastados pueden recuperarse; pero el daño al ambiente, no. Lo mismo sucedería, por ejemplo, si se intentara detectar la disminución de las poblaciones de una especie amenazada. Podemos errar y producir gasto de recursos o un alarmismo innecesario, o podemos errar y no hacer nada, permitiendo que la especie se extinga. En investigaciones biomédicas, las diferentes implicaciones de estos errores son mucho más claras. Por ejemplo, si se investiga la letalidad de una droga para utilizarla en personas, la hipótesis nula es que la droga no es letal. Si se detecta una significación inexistente, se le adjudica una acción perjudicial que no tiene y por tanto no se usa, con lo cual se impide su beneficio médico. Pero, si no se detecta una morbilidad que sí tiene, entonces se induce o sugiere su uso, con lo que se pone en riesgo la vida de las personas. Ninguno de los errores es bueno; pero, indudablemente, en estos casos el del segundo tipo tiene efectos más peligrosos a largo plazo. En la mayoría de las situaciones reales es mucho mayor la probabilidad de cometer error de segundo que de primer tipo, lo cual depende críticamente del tamaño de muestra. El cálculo de la potencia estadística permite discriminar entre la retención de una hipótesis nula por limitaciones del método o porque las probabilidades realmente la apoyan.

Muchas publicaciones carecen de análisis de potencias –declarados– ante resultados no significativos. Un ejemplo representativo se observa, por ejemplo, en la revista Biología, en el artículo de Denis *et al.* (2000) acerca de la morfometría del Aguaitacaimán (*Butorides virescens*) en las arroceras del Sur del Jíbaro, en Sancti Spiritus, Cuba. En la tabla II de ese trabajo (Denis *et al.*, 2000, p. 135) se muestra la comparación entre sexos de ocho variables morfométricas, de las cuales solo la longitud del pico resulta significativa. Los autores discuten la ausencia de dimorfismo en la especie; sin embargo, ninguna de las restantes comparaciones llegó a superar el 41 % de potencia, por lo cual estas conclusiones no son sostenibles. En varios artículos de esta revista se pueden detectar problemas de potencia estadística. Sin embargo, en la mayoría de los casos, ante resultados no significativos, no se reportan todos los elementos necesarios para comprobar la potencia del análisis realizado, como podrían ser el tamaño de efecto, el tipo de prueba, el tamaño de muestra el tipo de hipótesis –una o dos colas–, el nivel de significación y los estadísticos resultantes.

Actualmente, un gran número de programas estadísticos tienen incorporados módulos de análisis de potencia –Statistica lo incluye a partir de su versión 7.0, SPSS, SYSTAT, SAS, etcétera– y existe, además, una buena cantidad de pequeños programas específicos para estos análisis. Por lo tanto, en la actualidad la ausencia de programas ya no es una justificación. Autores, revisores y editores de nuestras revistas científicas deben tener muy claras las implicaciones de esta omisión en cada tipo de estudio y contribuir a desplazar el foco de atención del error del segundo tipo al del primer tipo.

¿Obtener una baja potencia estadística significa dejar de publicar un resultado obtenido? Para nada. No debe asumirse ciegamente que un análisis con poca potencia no tiene valor para ser publicado. Es necesario desligar completamente la significación estadística de la importancia biológica y tener en cuenta el uso específico y limitado que tienen las pruebas de hipótesis. De forma idónea, los investigadores siempre deseamos tener buenas potenciasen nuestros análisis, pero no siempre los recursos logísticos así lo permiten. Existen estudios o fenómenos que nunca tendrán recursos suficientes para garantizar, de una vez, toda la potencia ideal requerida; sin embargo, pueden servir de base para otro tipo de estudios integradores como los meta-análisis (Mann, 1990). Solamente se debe ser más cuidadoso a la hora de interpretar un resultado sin significación estadística y siempre, con la integridad que la investigación científica requiere, informar la potencia que se asocia a un resultado negativo o, como mínimo, asegurarse de dar todos los elementos para que los lectores puedan hacer el análisis, si así lo desean.

Aunque estos problemas asociados a las pruebas de hipótesis en la estadística frecuentista parecen reclamar a voces que no se empleen más tales métodos, hay que tener cuidado nuevamente con las apariencias y aprender de nuestros errores. La ciencia continúa su proceso de generalización, de inducción de procesos generales a partir demuestras; por lo tanto, es imprescindible el empleo de métodos estadísticos para evaluar su certidumbre. Las pruebas de hipótesis continuarán siendo parte del arsenal de métodos de los investigadores, solo hay que tener más cuidado en su uso indiscriminado. Rescatando la intención original de R. A. Fisher cuando propuso su valor de P, los resultados de una prueba pueden, de hecho, aportar una evidencia sobre un experimento; lo que no se debe es basar la conclusión en su resultado.

De esta misma forma quedó recogido en las Normas de Vancouver, desde su edición de 2001, para las publicaciones biomédicas: «Se evitará la dependencia exclusiva de las pruebas estadísticas de verificación de hipótesis, tal como el uso de los valores P, que no aportan ninguna información cuantitativa importante». La posición no debe ser rechazar el uso de las pruebas de hipótesis, sino darles un empleo más teórico y racional. La filosofía de la estadística debe contener de forma básica la comprensión de su limitado alcance y la aceptación de la probabilidad inevitable de estar cometiendo errores cuando se aplique.

En los momentos actuales, están tomando auge una serie de alternativas a las pruebas de hipótesis que posiblemente estén llamadas a sustituirlas (Steidl, 2006; Sleep *et al.*, 2007). Entre estas, se encuentra una serie de procedimientos que incluyen los meta-análisis y el desarrollo de inferencias basadas en la selección de modelos y estimación de parámetros a partir de técnicas bayesianas y por máxima verosimilitud (Hobbs y Hilborn, 2006; Anderson, 2008). Sin embargo, todavía las recetas frecuentistas dominarán el análisis de datos biológicos por algún tiempo, por lo cual la llamada de precaución es necesaria.

LITERATURA CITADA

Anderson, D.R. (2008): *Model based inference in the life science: a primer on evidence*, Springer Science, Nueva York.

Denis, D.; L. Mugica y M. Acosta (2000): «Morfometría y alimentación del Aguaitacaimán (*Butorides virescens*) en las arroceras del Sur del Jíbaro», *Biología,* vol. 14(2, La Habana, pp. 133-140.

Hobbs, N.T. y R. Hiborn (2006): «Alternatives to statistical hypothesis testing in ecology: a guide to self teaching», *Ecology Applied,* vol. 16(1, pp. 5-19.

Mann, C. (1990): «Meta-analysis in the breech», *Science,* vol. 249, pp. 479-480.

Sleep, D.J.H.; M.C. Drever y T.D. Nudds (2007): «Statistical versus biological hypothesis testing: response to Steidl», *Journal of Wildlife Management*, vol. 71, pp. 2120-2121.

Steidl, R.J. (2006): «Model selection, hypothesis testing and the risk of condemning analytical tools», *Journal of Wildlife Management*, vol. 70, pp. 1497-1498.

●●●