Revista Chilena de Historia Natural
68: 12 - 18, 1995

EDITORIAL

Uso e importancia de la inferencia estadística: algunos problemas frecuentes detectados en la Revista Chilena de Historia Natural

Use and importance of statistical inference: some common problems detected in Revista Chilena de Historia Natural

La estadística es una de las herramientas más usadas en biología, y particularmente en ecología, para refutar, dar sustento o asignar importancia relativa a proposiciones de interés. Sin embargo, con frecuencia no advertimos hasta qué punto la decisión biológica sobre los resultados obtenidos está condicionada por la inferencia estadística asociada, y a la inversa, hasta qué punto la información biológica y características de un estudio condicionan la inferencia estadística a usar. Una de las prácticas más usuales en biología es someter a prueba una hipótesis mediante la formulación de un diseño experimental y la evaluación posterior de sus resultados. No obstante, en este proceso los biólogos cometemos reiteradamente una gran cantidad de errores, de distinta importancia, principalmente debido a que desestimamos, subestimamos o sobreestimamos la importancia de los requerimientos y resultados de las pruebas estadísticas usadas. Recientemente, por ejemplo, en un comentario destinado a servir de consejo editorial para autores y revisores de trabajos científicos, Fowler (1990) llamó la atención sobre los diez errores estadísticos de caracter general más comunes (existen por supuesto otros varios análisis generales sobre el empleo de la estadística por los biólogos; e.g., Seaman & Jaeger 1990, y referencias incluidas en Camus & Lima 1995). Por otra parte, Yoccoz (1991), en un comentario similar, abordó problemas más específicos y de mayor importancia relativos a una práctica ampliamente difundida: las pruebas de significancia estadística. Yoccoz advirtió que "la literatura está infiltrada por una formación de conceptos erróneos sobre el uso e interpretación de las pruebas de significancia", y pese a que las limitaciones de estas pruebas son ampliamente conocidas por los estadísticos, se han transformado en una verdadera "religión de la estadística" (Salsburg 1985 *fide* Yoccoz 1991). Buena parte de los problemas proviene del uso descuidado de estas pruebas o de no apreciar claramente las operaciones estadísticas subyacentes a toda prueba de hipótesis, lo que ha generado una serie de creencias no justificadas. Las más comunes son: a) considerar que una prueba de significancia es de rigor en una publicación, por lo cual muy pocos escribirían que la población A difiere de la población B "sin agegar la palabra mágica 'significativo' o alguna fórmula como prueba de t, $P < 0.05$", y b) considerar al valor de 0.05 como "el límite absoluto entre dos mundos: diferencia por un lado, igualdad por el otro" (Yoccoz 1991). Este último caso provoca la tendencia muy difundida de asignar mayor importancia a un valor de probabilidad "no significativo" a medida que se acerca a 0.05, para lo cual hay varias expresiones de uso común como "marginalmente significativo", "cerca de la significancia", "bordear o aproximarse a la significancia" y otras. También es común que estas probabilidades sean consignadas como $P > 0.05$ o simplemente N.S., sin importar que el valor calculado sea 0.055 o 0.99. Es importante recordar que 0.05 es una convención que por su extendido uso se ha llegado a confundir con un criterio objetivo de decisión. El peligro es que, tomado como dogma, el límite 0.05 puede operar en la práctica como elemento modulador de teoría, ya que de acuerdo a él descartamos o aceptamos resultados que posteriormente pueden proyectarse a nuevos análisis usando el mismo criterio. Esto es, según Yoccoz (1991), parte del problema principal aso-

ciado al mito del 0.05: confundir significancia biológica con significancia estadística. No obstante, en biología generalmente no existe ninguna razón para escoger *a priori* un nivel de significancia dado, mientras que en la industria, por ejemplo, sí es posible escoger niveles de significancia de acuerdo al error tipo I (rechazar Ho verdadera) usando teoría de decisión, lo que se logra aplicando funciones de costo/beneficio bien definidas. En biología, en contraste, muy rara vez podría determinarse con claridad una función de este tipo, y el nivel se asocia más bien a las circunstancias de la prueba, por lo que es importante analizar en cada caso la potencia de la prueba (práctica que casi nunca aplicamos), es decir la probabilidad de que una prueba particular rechaze Ho a un nivel de signifcancia (alfa) particular cuando Ho es realmente falsa. La potencia es función del alfa, del tamaño muestral, y del "tamaño del efecto" o magnitud del fenómeno biológico observado. Muy pocas veces se efectúa este análisis, aunque hay algunos intentos en la literatura para extender su uso (e.g., Toft & Shea 1983, Rotenberry & Wiens 1985). Sin embargo, es necesario estimar el tamaño esperado del efecto, lo cual es ciertamente difícil y poco probable de establecer objetivamente.

Una alternativa extrema al uso de las pruebas de significancia tradicionales es renunciar a la ilusión de la objetividad y operar directamente con la subjetividad involucrada en las decisiones, por ejemplo usando inferencia Bayesiana (e.g., Berger & Berry 1988). Pero también es saludable simplemente prestar atención a los errores que se cometen con más frecuencia. Para esto es importante tener presente que el alfa no es un criterio de decisión fijado en 0.05, sino una tasa de error que estima la proporción de muestras raras donde la inferencia es falsa a partir de una serie grande de muestras. Además, la prueba de significancia es un procedimiento para medir la consistencia de los datos con una hipótesis nula (buscando la contradicción o inconsistencia con Ho) usando un estadígrafo o criterio de prueba. En términos generales, el estadígrafo es función de los datos observados, y se compara con una variable aleatoria que describe la distribución de ese estadígrafo bajo las condiciones establecidas en la Ho. El procedimiento permite obtener el conocido valor observado de P, que en términos gruesos corresponde a la probabilidad obtenida desde el grado de diferencia entre el estadígrafo asociado a la distribución bajo Ho y aquel observado. Debe notarse que el efecto biológico no está incorporado directamente en la prueba, y por ello el valor de P debiera interpretarse sólo como el nivel de significancia alcanzado bajo las condiciones específicas de la prueba. Es claro que un cambio en las condiciones de la prueba (e.g., aumento del tamaño de muestra) puede producir un resultado diferente y, eventualmente, hacer cambiar nuestra conclusión biológica si sólo se toma en cuenta el límite de 0.05. Por ello, usar el criterio "significativo" cuando P es menor que 0.05 y "no significativo" cuando P es mayor o igual que 0.05 implica dicotomizar arbitrariamente el resultado de la prueba abandonando la información real en favor de una decisión. Por otro lado, al resultado de la prueba muchas veces se sobreimponen los efectos derivados del no cumplimiento de supuestos estadísticos generales (e.g., independencia de las observaciones) o específicos (e.g., normalidad), lo que puede distorsionar el procedimiento estadístico usado en mayor o menor grado dependiendo de su robustez.

Algunos problemas frecuentes asociados o derivados de errores en el uso e interpretación de pruebas de significancia son: 1) confundir el valor observado de P con la probabilidad de que Ho sea verdadera; 2) confundir el valor observado de P con el efecto biológico asociado a los datos observados o con la fuerza de ese efecto; 3) considerar que rechazar Ho es comprobar el efecto postulado; 4) validar un efecto en función de los criterios de prueba (e.g., sumas de cuadrados, F y P en un ANDEVA) sin considerar los atributos asociados al efecto (e.g., media y error estándar); 5) no considerar la potencia de la prueba; 6) validar una relación de regresión usando la ecuación de predicción y el valor de P sin analizar residuos, error estándar de los parámetros, etc.; 7) validar la linealidad de una relación entre dos variables cuando el

coeficiente de correlación es "significativo" y considerar que si no es "significativo" las variables son independientes; 8) comparar valores de P obtenidos en diferentes estudios bajo diferentes condiciones; 9) validar una diferencia significativa entre dos medias sin considerar posibles deficiencias en el muestreo, diferencias en tamaño de las muestras o efecto aislado del tamaño de muestra (muchas veces una diferencia estadística se alcanzará como resultado de aumentar el n); 10) el problema de mayor importancia: confundir significancia biológica con significancia estadística, ya que se debiera revisar la robustez del modelo teórico para decidir qué tan grande debe ser una diferencia para considerarla biológicamente significativa, procedimiento que debiera efectuarse al comenzar un estudio y que además depende de los objetivos de éste (véase Yoccoz 1991 y Underwood 1990 para mayores detalle sobre parte de los problemas indicados).

Estos problemas son suficientemente comunes para que tanto autores, revisores y editores deban tenerlos presentes al elaborar o analizar un manuscrito. Para hacer un diagnóstico rápido de esta situación en la ciencia nacional, realicé un análisis de los trabajos de investigación publicados durante 1992 y 1993 en la Revista Chilena de Historia Natural, sin considerar los trabajos que no contenían información cuantitativa o que sólo la empleaban en forma ilustrativa, sin efectuar comparaciones o inferencias. Seleccioné un total de 46 artículos, todos ellos conteniendo decisiones o conclusiones biológicas elaboradas a partir de datos obtenidos para ese estudio, o de datos de otros estudios. De ellos 32 incluyeron uno o más análisis estadísticos, y 14 no utilizaron estadística. Para cada trabajo examiné la presencia y frecuencia de ocho tipos de error, siete relativos al uso e interpretación de pruebas de significancia y uno definido como no usar inferencia estadística cuando debió haberse usado. Los errores fueron: 1) descartar un resultado indicando sólo "no significativo" (N.S.) sin indicar el nivel de significancia alcanzado; 2) dar relevancia a un resultado por estar "cerca de la significancia"; 3) validar un resultado de inferencia usando sólo los criterios de prueba, sin presentar una estimación de la diferencia real entre los datos; 4) presentar tablas de ANDEVA con información insuficiente (e.g., falta de media y error estándar de tratamientos o parámetros de regresión); 5) asimilar la probabilidad observada al tamaño del efecto; 6) notación insuficiente o deficiente (e.g., X^2, $P < 0.05$; diferencia no significativa, $P < 0.05$; dife-

TABLA 1

Frecuencia de ocurrencia de 8 tipos de errores estadísticos, asociados a la inferencia basada en pruebas de significancia, encontrados en 46 artículos de investigación publicados en Revista Chilena de Historia Natural durante 1992 y 1993. Los tipos de error se citan en la Tabla en forma resumida (ver detalles en el texto) y no están ordenados de acuerdo a su importancia.

Occurrence frequency of eight types of statistical errors, related with inferences based on significance tests, found in 46 research articles published in Revista Chilena de Historia Natural during 1992 and 1993. Errors are listed below in summarized way (see details in the text) and are not ranked according their importance.

TIPO DE ERROR	NUMERO DE ARTICULOS CONTENIENDO EL ERROR	FRECUENCIA (%)
Indicar sólo N.S. sin probabilidad	16	17.2
Indicar cercanía a la significancia	0	0
No dar estimación de diferencia real	8	8.6
ANDEVA con información insuficiente	10	10.8
Probabilidad = tamaño del efecto	1	1.1
Notación deficiente o insuficiente	27	29.0
Interpretación confusa o contradictoria	17	18.3
Sin análisis estadístico	14	15.1

rencia significativa, P=0.75); 7) confusión en la interpretación de resultados de la prueba o contradicción entre resultados de la prueba y conclusiones; 8) error por omisión, o falta de inferencia estadística siendo importante o necesaria.

Escogí los tipos de error entre varios posibles, aunque considerando su frecuencia o relevancia, pero no los evalué de acuerdo a su grado de importancia ya que tal calificación es dependiente del contexto y de la apreciación del que analiza. Tampoco consideré el número de veces que un tipo de error aparecía en un mismo artículo porque se asociaba a la cantidad de información analizada, entre otros factores. Como punto de partida, la revisión de los trabajos indicó que los autores en general escogen las pruebas y tienen en cuenta sus supuestos de manera adecuada, y la mayoría se preocupa de fundamentar estadísticamente sus conclusiones. Esto revela en parte la importancia que se concede al uso de análisis estadístico en la mayoría de los trabajos, a pesar de su muy diferente naturaleza. Sin embargo, esto no reflejó que se concediera igual importancia a un conocimiento más detallado de la estadística. La Tabla 1 muestra la frecuencia con que un tipo de error fue encontrado en los 46 artículos, destacando dos aspectos interesantes. Por una parte, los problemas de notación son los más frecuentes, lo que podría estar relacionado simplemente con descuido o errores de tipeo no corregidos. Sin embargo, las otras frecuencias más altas fueron para la interpretación de las pruebas y el uso del valor de probabilidad, sugiriendo que la notación podría estar reflejando un manejo o conocimiento poco adecuado del significado de los elementos de la prueba (e.g., estadígrafos, dócima, distribución bajo la hipótesis nula, nivel de significancia, etc.) o de sus relaciones. Por otra parte, las más bajas frecuencias fueron para la equivalencia entre probabilidad y tamaño del efecto y para destacar cercanía de la significancia. Esto pareciera indicar que los autores asignan importancia a no cometer esos errores, pero una revisión del contexto de los trabajos sugiere más bien una tendencia a creer que el límite de 0.05 no puede ser relativizado, y que los análisis tienden a situarse en el mundo dicotómico de no significativo versus significativo, o igual versus distinto.

Por otro lado, la distribución de frecuencia de tipos de error para los 32 artículos que efectuaron análisis estadísticos fue bastante esperable (Tabla 2), sugiriendo una distribución normal con algún sesgo hacia la derecha, aunque habría sido deseable obtener un sesgo a la derecha mucho mayor. La distribución actual muestra que los trabajos conteniendo dos y tres tipos de error representan el 62.5 % del total, en 6.25 % no se encontraron errores y sólo en 6.25 % se encontraron cinco o seis de ellos. En ningún trabajo se detectaron más de seis tipos de error simultáneamente, y la distri-

TABLA 2

Distribución de frecuencia del número de tipos de error contenidos en cada artículo. Ningún artículo incluyó más de 6 tipos de error simultáneamente.

Frequency distribution of the number of error types contained in each article. No article included more than 6 error types simultaneously.

NUMERO DE TIPOS DE ERROR POR ARTICULO	NUMERO DE ARTICULOS	FRECUENCIA RELATIVA	FRECUENCIA ACUMULATIVA
0	2	6.25	6.25
1	4	12.50	18.75
2	10	31.25	50.00
3	10	31.25	81.25
4	4	12.50	93.75
5	1	3.125	96.875
6	1	3.125	100.00

bución acumulada muestra que exactamente la mitad de los trabajos mostraron sólo dos tipos de error o menos. Sin embargo, debe recordarse que la importancia del error no está considerada, y por tanto dos trabajos que contengan por ejemplo tres tipos de error cada uno no necesariamente son comparables en términos de la calidad de su análisis estadístico.

Aparte de los datos anteriores, y parcialmente debido a la forma de la distribución de frecuencia de tipos de error, exploré la posible asociación (medida como correlación) entre el número de tipos de error en un artículo y su extensión en número de páginas. En promedio, el número de tipos de error por artículo fue de 2.53 ± 0.23 ($E.E_{.}$), y la proporción de tipos de error por página y por artículo fue de 0.257 ± 0.022 (E.E.). Para analizar la correlación introduje algunos criterios arbitrarios que entregaron resultados que si bien podrían considerarse antojadizos fueron bastante sugerentes. La correlación entre las variables involucró tres situaciones: a) excluyendo sólo los trabajos sin errores, b) excluyendo además los trabajos de autores no sudamericanos, y c) excluyendo además trabajos de autores sudamericanos con trayectoria reconocida en el conocimiento y manejo de pruebas estadísticas. La Tabla 3 muestra los resultados para los tres casos descritos. Este análisis incurrió en el error de no estimar la potencia de la prueba por la dificultad de establecer el tamaño esperado del efecto, pero el problema se amortigua al considerar que los trabajos se eliminaron en función de criterios *a priori* sin seleccionarlos por su desviación de la nube de puntos, y además la disminución del tamaño de muestra desde a) a c) incidiría más bien en una disminución de la potencia. La asociación encontrada fue positiva (su linealidad no importa en este análisis) y lo interesante fue que tanto los coeficientes de correlación como los niveles de significancia respectivos aumentaron desde el caso a) al c). Esto muestra que efectivamente los trabajos removidos oscurecían la tendencia encontrada en c), lo que se debió a que ellos presentaban las más bajas proporciones de errores por página. Esto inmediatamente sugiere que los autores sudamericanos, excluyendo a aquellos con mayor formación estadística, en general pondrían menor énfasis en el planteamiento, presentación e interpretación de resultados estadísticos. Por otra parte es sugerente (y poco reconfortante) que la relación encontrada en el caso c) sea positiva, indicando simplemente que mientras más páginas escribimos más errores cometemos.

Sin duda es posible modificar los resultados anteriores con un examen más detenido de los trabajos antes de enviarlos a publicación. Es claramente deseable que en un análisis de este tipo no se encuentre ninguna relación entre la extensión de un artículo y el número de errores que contiene, y por supuesto que el número de errores por artículo sea cercano a cero. La responsabilidad de mejorar la calidad del análisis de datos recae principalmente en los autores de un trabajo, pero también sobre los revisores y editores, que debieran involucrarse íntimamente en el proceso.

TABLA 3

Análisis de correlación (coeficiente de Pearson) entre el número de tipos de error por artículo y el número de páginas por artículo usando tres criterios arbitrarios definidos *a priori*: A: excluyendo artículos con cero error, B: excluyendo artículos de autores no Sudamericanos además de A, C: excluyendo artículos de autores Sudamericanos con formación estadística reconocida además de A y B.

Correlation analysis (Pearson coefficient) between the number of error types per article and the number of pages per article using three arbitrary criteria selected *a priori:* A: excluding zero error articles, B: excluding articles by non Southamerican authors besides A, C: excluding articles by Southamerican authors with qualified statistical expertise besides A and B.

CRITERIO	TAMAÑO de MUESTRA	r PEARSON	PROBABILIDAD
A	30	0.21	0.279
B	28	0.37	0.051
C	24	0.61	0.012

PATRICIO A CAMUS

Departamento de Ecología,
Facultad de Ciencias Biológicas,
P. Universidad Católica de Chile, Casilla 114-D,
Santiago, Chile

LITERATURA CITADA

BERGER JO & DA BERRY (1988) Statistical analysis and the illusion of objectivity. American Scientist 76: 159-165.

CAMUS PA & M LIMA (1995) El uso de la experimentación en ecología: supuestos, limitaciones, fuentes de error, y su estatus como herramienta explicativa. Revista Chilena de Historia Natural 68: .

FOWLER N (1990) The 10 most common statistical errors. Bulletin of the Ecological Society of America 71: 161-164.

ROTENBERRY JT & JA WIENS (1985) Statistical power analysis and community-wide patterns. American Naturalist 125: 164-168.

SEAMAN JW & RG JAEGER (1990) Statisticae dogmaticae: a critical essay on statistical practice in ecology. Herpetologica 46: 337-346.

TOFT CA & PJ SHEA (1983) Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist 122: 618-625.

UNDERWOOD AJ (1990) Experiments in ecology and management: their logics, functions and interpretations. Australian Journal of Ecology 15: 365-389.

YOCCOZ NG (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72: 106-111.