

HERRAMIENTAS ESTADÍSTICAS PARA RESOLVER CONTRASTES DE HIPÓTESIS CON CONTENIDO BIOLÓGICO: SU USO EN ECOLOGÍA DEL SIGLO XXI

Alejandro MARTÍNEZ ABRAÍN¹, David CONESA² & Daniel ORO¹

¹ IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Mallorca, ESPAÑA
abraín@uib.es (+34961610847)

² Grup d'Estadística Espacial i Temporal en Epidemiologia i Medi Ambient.
Departament d'Estadística i Investigació Operativa, Universitat de Valencia.
C/ Dr. Moliner 50, 46100 Burjassot (Valencia), ESPAÑA

RESUMEN

A menudo la formación que han recibido durante la carrera los jóvenes investigadores tiene notables carencias en los aspectos prácticos de diseño experimental, análisis de datos e interpretación de resultados, lo cual limita de manera decisiva el provecho científico futuro de sus actividades. Eso es especialmente cierto en nuestros días, ya que vivimos una revolución importante en el campo de la metodología estadística e incluso en el procedimiento de hacer inferencia (el salto matemático desde las propiedades de nuestra muestra de datos a las de los parámetros desconocidos de la población, nuestro objeto de estudio), que afecta no sólo a los ecólogos sino a muchas otras disciplinas científicas como la medicina, la psicología o la sociología. Recordamos aquí, a los investigadores noveles, que la vieja idea de la existencia de una sola herramienta estadística (la comprobación de hipótesis nulas sin contenido biológico) ha de cambiarse hoy por el paradigma ejemplificado mediante la “caja de herramientas” en la que existen diversos tipos de herramientas a nuestra disposición, cuya elección depende de la complejidad de nuestros problemas. Dado que la información está dispersa en numerosos artículos y tratados específicos de estadística y diseño experimental, nuestra idea aquí es tratar de presentar un guión actualizado de las principales herramientas disponibles para resolver contrastes de hipótesis con contenido biológico en ecología en el siglo XXI, identificando la utilidad y limitaciones de cada una de ellas, con el fin de mejorar la cultura estadística de los jóvenes investigadores. Finalmente ofrecemos, con fines prácticos, una selección de referencias bibliográficas para el buen uso de cada herramienta. **Palabras clave:** herramientas estadísticas, hipótesis con contenido biológico, estadística Bayesiana, criterios de información, cultura estadística, relevancia biológica.

ABSTRACT

It is common place to find young researchers with serious problems regarding experimental design, data analysis and correct interpretation of results. These limitations are a serious handicap for their future scientific endeavour. This is specially the case nowadays because we are living a substantial revolution within the field of techniques for data analysis and even statistical inference procedures (the mathematical jump from sample data to the unknown population parameters), which affects not only ecology but other scientific disciplines such as medicine, psychology or sociology. We recall here that

the old idea of having just one tool available to answer ecological questions (the testing of biologically non-informed null hypothesis) has now been substituted by the paradigm of the “tool box”, where we can find several tools available depending on the complexity of the problem we are dealing with. Since gathering information on each tool requires an important searching effort, our purpose here is to offer an updated guideline of the main tools available to solve biologically-informed contrasts in ecology in the 21st century, and to correctly interpret our results, identifying the weaknesses and strengths of each tool. We also present, with practical purposes, a selection of relevant publications which can be consulted to master the tools in the box.

Key words: statistical tool box, biologically-informed hypotheses, Bayesian statistics, information criteria, statistical culture, biological relevance.

EL MÉTODO QUE USAN LOS CIENTÍFICOS

Todo maestrillo tiene su librillo. Los pintores y los escultores siguen un método para llenar los lienzos en blanco o para vaciar la piedra informe. Los científicos responden preguntas sobre el mundo siguiendo un método también. El método es lo que distingue nuestra actividad de otras. Tanto los artistas como los científicos perseguimos en el fondo dar respuesta a las grandes preguntas, pero lo hacemos siguiendo métodos distintos. Éste que tienes frente a tus ojos es el librillo de los científicos (muy resumido, bastante digerido y puesto al día), especialmente el librillo de los ecólogos del siglo XXI.

Observar la naturaleza y hacerse preguntas. Podemos empezar con un ejemplo. Pongamos que el resultado de una traslocación de un grupo de tortugas hembra, realizado con fines conservacionistas, es una larga etapa inicial de vagabundeo y una etapa final de asentamiento en una zona concreta (lo cual sabemos porque las tortugas fueron marcadas con radioemisores). La pregunta que surge de manera natural es: ¿qué ha ocurrido para que las tortugas hayan decidido estabilizarse? Lo importante de este punto es que para llegar a formularse preguntas interesantes hay que estar en contacto con lo que ocurre en la naturaleza. Desde el despacho rara vez surgen preguntas interesantes.

Formular hipótesis concretas. Las hipótesis son las posibles explicaciones que se nos ocurren para responder a la pregunta, basándonos en nuestra información previa de la especie y del medio en el que vive. La información previa (la experiencia acumulada) es por tanto muy importante, así que el “subjetivismo” es parte ineludible del proceso de responder preguntas. Es importante recalcar que las hipótesis no son preguntas sino posibles explicaciones. Por ejemplo en el caso de la pregunta anterior se nos pueden ocurrir varias hipótesis: a) en la zona de estabilización hay mayor densidad de machos, b) en la zona de estabilización hay mayor disponibilidad de recursos alimenticios o c) como sabemos de antemano que las tortugas emplean el olfato para orientarse puede que la zona de estabilización sea la que huele más parecido a la zona

de donde proceden las tortugas. Por supuesto hay muchas más hipótesis posibles y depende de la experiencia del investigador no dejarse en el tintero la hipótesis más cercana a la realidad, aunque de esto nunca podemos estar seguros. Seguramente no hay hipótesis en ecología que contemplen todas las variables en juego o de una manera que responda a la realidad. Siempre son simplificaciones que empleamos para separar el ruido de la información. De nuestras hipótesis se pueden derivar predicciones específicas, que con el paso del tiempo pueden ayudar a confirmar nuestras hipótesis o a sospechar de su falsedad. Para nosotros los ecólogos, las hipótesis quedan planteadas en forma de palabras, sin embargo, si queremos obtener conclusiones en términos estadísticos, es decir, aportar evidencia estadística, debemos reformular las hipótesis en términos de parámetros de una distribución de probabilidad. En el caso de la hipótesis anterior en la que la zona de estabilización tiene una mayor densidad de machos, nuestro parámetro de interés sería la proporción de machos y la hipótesis a contrastar si dicho parámetro supera 0.5.

Diseñar un experimento para la obtención de datos. En ecología llamamos experimento a cualquier protocolo de toma de datos dirigido a responder una pregunta, aunque sólo incluya datos “observacionales” y no puramente experimentales. La diferencia estriba en que en el primer procedimiento el investigador no tiene un control específico sobre los tratamientos aplicados, sino que estos se producen de manera natural (se les suele llamar estudios de campo), mientras que en los estudios experimentales el investigador controla el estudio eligiendo cómo, a quién o a qué aplicar los tratamientos, pudiendo así llegar a interpretaciones más relevantes (p. ej., establecer relaciones de causalidad). En ecología hay que saber estar atento e identificar los “experimentos naturales” que nos brinda la naturaleza. Medir los efectos sobre la biodiversidad de un parche de selva tropical tras un huracán puede ser una buena manera de aprovechar un experimento natural, especialmente si teníamos estimaciones de diversidad del sitio antes del huracán. La palabra natural no implica que los experimentos hayan de ser causados por la naturaleza. También son experimentos “naturales” los causados por el hombre, como los efectos de una moratoria de pesca sobre las aves marinas, que se alimentan parcialmente del descarte de los barcos, o un vertido petrolífero que las deja sin sus presas naturales. El resultado final del experimento es la obtención de muestras. Lo ideal sería no necesitar muestras, es decir, trabajar con la población entera, pero eso es raramente posible. Además, hemos de tener presente los recursos disponibles antes de abordar el diseño del experimento pues no siempre es posible acceder a un número elevado de muestras. Por eso es tan importante que la aleatorización esté presente en el proceso de selección de las muestras: el muestreo ha de ser lo más aleatorio posible. Tenemos que hacer todo lo que esté en nuestra mano por tomar muestras al azar, de modo que obtengamos el máximo de representatividad de la población de estudio. En realidad la única manera de muestrear estrictamente al azar es la que

realizamos cuando metemos todos los elementos de una población (p. ej., los números de un sorteo) en una chistera y escogemos uno a ciegas. Eso es imposible en ecología así que nuestros muestreos son sólo semi-azarosos. En gran medida por ello los estudios en ecología recogen sólo una parte pequeña de la enorme variabilidad de los ecosistemas y sus organismos, y es por ello que ésta es una disciplina que no suele responder simultáneamente a grandes preguntas (como la física) sino que sus practicantes deben conformarse, en la mayoría de los casos, con responder a pequeñas piezas del infinito rompecabezas que representa el planeta y su funcionamiento. Este punto es importante porque los jóvenes que se incorporan a la investigación en ecología normalmente creen (creencia loable, por supuesto, y hasta indispensable) que se disponen a encontrar respuestas inequívocas a las preguntas formuladas, y la sorpresa es grande cuando descubren que lo que se consigue suele ser responder a una pequeña parte de la gran variabilidad que se descubre allí afuera. La representatividad de nuestra muestra aumenta al aumentar el tamaño de la muestra, disminuyendo así la variabilidad de las estimaciones de los parámetros y aumentando por tanto la precisión de dichas estimaciones. De ahí la importancia de contar con muestras grandes. Una muestra de tamaño adecuado es, sobre todo, aquella que consigue acotar bien la variabilidad de las estimaciones, aunque también la que nos permite detectar efectos de la magnitud deseada. Los muestreos que se basan en muestras pequeñas pueden llevar a una gran variación entre las estimaciones individuales de los parámetros. De todos modos es conveniente tener información previa sobre el sistema que vamos a muestrear ya que si está compuesto de diferentes “estratos”, claramente identificables, el muestreo se deberá realizar de manera aleatoria dentro de cada uno de los estratos (es lo que se denomina un muestreo aleatorio estratificado). No hay que confundir el error de muestreo con los errores de bulto (p. ej., identificar mal una especie, etiquetar mal un tubo, medir dos veces el mismo huevo, etc.). Es en este punto cuando debemos seleccionar el modelo estadístico (GLM, estructura espacial, temporal, etc.) que va a describir la propia experimentación. En otras palabras, tendremos que concretar qué diseño vamos a utilizar. Así, por ejemplo, uno de los aspectos más importantes de la experimentación es que a menudo es necesario contar con un control. El control nos permite comparar y extraer conclusiones en presencia y ausencia de una intervención. Si queremos saber si las tortugas traslocadas se han movido mucho o poco, más de lo habitual o menos, la única manera de saberlo es seguir en la misma zona a un grupo de hembras no traslocadas, de similar edad y condición física, además de a las traslocadas, y mediante los mismos métodos. En ocasiones, no basta con tener un control y un tratamiento si existe algún factor ambiental que pueda estar influenciando a la variable de interés, además del propio tratamiento. Supongamos que se quiere saber si una rapaz social, que cría sobre los tejados de las casas, se reproduce mejor cuando las colonias son grandes (por los múltiples beneficios de la socialidad) que cuando

son pequeñas. Pero además sabemos que hay dos tipos de casas: unas son altas y nuevas (lo que haría más difícil la entrada de depredadores) y otras son bajas y viejas (lo que facilita la entrada de depredadores). Por tanto para saber si criar en colonias grandes o pequeñas afecta al éxito reproductor necesitamos un diseño en bloques aleatorizados. Tendremos que seguir el éxito reproductor de colonias grandes ubicadas sobre casas malas y buenas y también seguir colonias pequeñas ubicadas sobre casas malas y buenas, para poder separar el efecto de la calidad de las casas y del tamaño de colonia sobre el éxito de cría. En el caso anterior se dice que hemos controlado por un factor, teniéndolo en cuenta en el modelo. Para concluir este apartado conviene resaltar que hay que tener siempre en mente que todo el proceso, además de aleatorio, debe ser lo más transparente y reproducible que sea posible, ya que la ciencia avanza subiéndose a los hombros de los que nos han precedido y comprobando mil y una veces que los resultados son consistentes. Los resultados de un único trabajo rara vez pueden ser concluyentes.

Recuadro 1: Falta de independencia de los datos.

Replicar (repetir) es necesario para poder captar toda la variabilidad posible de la población de estudio. Para ello las observaciones han de ser independientes. Si queremos saber de qué depende la longitud del brazo de una persona, no nos sirve de nada medir 10 veces un mismo brazo, pero sería de gran ayuda medir la longitud de 10 brazos distintos seleccionados de manera aleatoria. El primer ejercicio es interesante sólo si lo que queremos es asegurarnos de la precisión de la medida o calcular la tasa de crecimiento si tomamos las 10 medidas en momentos distintos. En el segundo caso medimos realmente variabilidad en la población, porque las unidades experimentales son independientes. Al escogerlos al azar evitamos sesgos. Realmente en ecología es más difícil de lo que parece conseguir datos que sean completamente independientes en el espacio y en el tiempo, es decir, que no muestren autocorrelación espacial o temporal (colinearidad), porque una de las características principales de los sistemas naturales es que todo está relacionado dentro y entre ellos. Hay dos tipos de falta de independencia de datos. En el primer caso dos datos no son independientes entre sí por tener en común un factor externo (p.ej., no es un muestreo independiente muestrear dos plantas hija que provienen de la misma madre). En el segundo caso un dato está influido directamente por otro (p. ej., si queremos medir la tasa de crecimiento de una especie vegetal no deberíamos elegir individuos que estén muy próximos porque el crecimiento de una planta afecta al crecimiento de otra cercana por competición de recursos). Estrictamente hablando se reserva el término pseudoreplicación al segundo caso. El primer caso la falta de independencia de los datos es una característica estructural y a menudo es posible analizar los datos utilizando modelos que recojan dicha estructura, como, por ejemplo, las series temporales, la geoestadística, contrastes filogenéticos independientes, etc., pero el segundo caso no tiene más arreglo que muestrear plantas bien separadas entre sí, mediante un buen diseño experimental *a priori*.

Analizar los datos e interpretar los resultados

Tras realizar el experimento y obtener los datos el paso final consiste en analizarlos: calcular los estadísticos, elegir la herramienta estadística adecuada (entre las disponibles en nuestra caja de herramientas) e interpretar los resultados. Si tenemos la suerte de poder trabajar con todos los individuos de una población la inferencia estadística (es decir, la extrapolación desde la muestra de datos a los deseados parámetros poblaciones) no es necesaria. Basta con que empleemos la estadística descriptiva, en forma de histogramas (gráficos de frecuencias) por ejemplo. Hasta ahora existía una manera predominante de realizar este análisis de los datos: contrastar hipótesis por pares. Pero no un par cualquiera, sino una hipótesis sin contenido biológico, la llamada hipótesis nula (H_0), frente a una hipótesis alternativa, que simplemente siempre apunta en dirección contraria. La hipótesis nula se emplea porque es una hipótesis “falsificable” (como diría Karl Popper) es decir una hipótesis para la cual es posible demostrar que es falsa. La idea es que si demostramos que la nula es falsa nos podemos quedar con la otra, con la hipótesis alternativa (H_1). Un ejemplo de hipótesis nula falsificable es “en esta laguna no hay carpas”. Basta con que encontremos una carpa para rechazar la hipótesis nula. La hipótesis “en esta laguna hay carpas” no es falsificable ya que si salimos a la laguna y muestreamos y no encontramos ninguna no podemos asegurar que no haya carpas; puede ser que simplemente no las hayamos encontrado. Este procedimiento (llamado NHST del inglés “Null Hypothesis Statistical Testing”) tiene sin embargo numerosos problemas ya que las preguntas que nos podemos plantear están totalmente mediatizadas por el contraste posterior que vamos a realizar. Sólo podemos plantearnos preguntas dicótomas, de blanco o negro, cuando en realidad hay un gran número de preguntas que requieren de una extensa gama de tonos grises entre ambos extremos. Además por regla general solemos saber de antemano que la nula es falsa. Por ejemplo, si yo me pregunto si la longitud media de los gavilanes macho es menor que la longitud media de los gavilanes hembra me estoy haciendo una pregunta estúpida (lo que se llama una “nula tonta” en inglés) ya que yo sé antes de empezar que van a ser diferentes y lo que me interesa realmente es saber en cuanto difieren, sobre todo si sé de antemano que cierto grado de diferencia tiene relevancia biológica en algún sentido (lo que constituye una pregunta mucho más interesante que las que permite plantearse el NHST). Además el hecho de que se pueda rechazar la hipótesis nula no hace que la alternativa sea cierta necesariamente (puede haber muchas otras alternativas posibles que no hemos considerado). Pero no desesperen. Hay maneras de plantear contrastes de hipótesis nulas que sean mucho más informativos y, sobre todo, hay maneras de contrastar a la vez todas las hipótesis que nos parezcan relevantes para responder a una pregunta. En el resto de este trabajo vamos a presentar las herramientas con las que contamos en nuestra caja de herramientas para tal fin.

HERRAMIENTAS TRADICIONALES PARA REALIZAR CONTRASTES CON CONTENIDO BIOLÓGICO

Contraste de hipótesis nulas con contenido biológico desde el punto de vista frecuentista.

Una de las principales cuestiones que nos planteamos los ecólogos incluye el análisis de las diferencias entre dos poblaciones. Éste es un claro ejemplo en el que contestar si hay diferencias entre ambas poblaciones es irrelevante, ya que no existen dos seres vivientes que sean iguales. Por el contrario, la hipótesis más apropiada sería valorar cómo de grande es la diferencia entre ambas poblaciones. Supongamos que estamos interesados realmente en el contraste de una de estas hipótesis con contenido biológico. El primer paso es realizar una prueba de potencia *a priori* (o prospectiva). Esto lo puedes hacer fácilmente consiguiendo en la red software muy sencillo, gratuito, como Gpower: (<http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>). La potencia de una prueba es la capacidad que tiene para detectar un efecto si éste existe. Un valor habitual de la potencia suele ser 80%, por lo que el error de tipo II (al que se suele llamar β y que representa la probabilidad a largo plazo que tenemos de equivocarnos cuando no podemos rechazar la hipótesis nula y concluimos erróneamente que el tratamiento no tiene efecto) toma el valor del 20% dado que la potencia es $1-\beta$.

Recuadro 2. Compromiso entre alfa y beta.

No obstante, si en nuestro caso es más arriesgado cometer un error de Tipo II que un error de Tipo I podemos encontrar el equilibrio adecuado entre los dos tipos de error al hacer nuestra prueba de potencia. En materia de toma de decisiones, como en biología de la conservación, por ejemplo, es más peligroso cometer un error de Tipo II (decir que no pasa nada porque se haga una actuación cuando sí que pasa) que un error de Tipo I (decir que sí tiene efecto una actuación cuando no la tiene) ya que en el segundo caso la actuación no se realiza. En cada caso habría que estudiar el compromiso más adecuado entre alfa y beta, en lugar de tomarlo como algo predeterminado.

El objetivo de la prueba *a priori* es obtener el tamaño de muestra necesario para determinar si dos poblaciones difieren en más de una magnitud dada. Esa diferencia de magnitud entre poblaciones (sopesada por la varianza de ambas poblaciones) se denomina “tamaño del efecto” (“effect size” en inglés). Corresponde al investigador, tirando de su experiencia previa de nuevo (de su subjetivismo informado), establecer qué diferencia considera relevante biológicamente en cada caso. El error de tipo I (llamado α) se suele fijar por convenio en el 5%. Eso significa que tenemos una probabilidad del 5% de equivocarnos a largo plazo (es decir tras infinitas repeticiones

que en realidad nunca hacemos) al rechazar la nula. Es importante aclarar aquí que la coletilla “a largo plazo” no está de adorno. De hecho es fundamental entender que los métodos clásicos se llaman “métodos frecuentistas” porque la validez de los resultados depende de la frecuencia a largo plazo de los mismos. Sigamos. Una vez introducidos los datos básicos (potencia deseada, alfa y tamaño del efecto) obtenemos el tamaño de muestra necesario para nuestro experimento. Con esta muestra (por supuesto obtenida lo más al azar posible) podemos asegurar que la significación estadística y la relevancia biológica van emparejadas. Es decir, si rechazamos la hipótesis nula es porque las poblaciones difieren al menos en la magnitud de interés y, en caso contrario, podemos aceptar la nula y concluir que difieren en menos, porque disponemos de potencia para afirmarlo. Cuando no realizamos una prueba de potencia previa no podemos concluir nada si obtenemos resultados estadísticamente no significativos. Todo lo que se puede hacer es no rechazar la nula, pero no aceptarla. De ahí que todos los investigadores anhelan conseguir p-valores menores de 0.05, porque si salen mayores lo único que podemos decir es que no sabemos nada. Eso provoca que las revistas estén sesgadas hacia la publicación de resultados estadísticamente significativos (los llamados resultados “positivos”). Los resultados negativos no los quiere nadie, porque no saben si se deben a que no hay diferencias realmente o a que no se han detectado por bajo tamaño de muestra. Tal y como hemos dicho, ese problema se solventa al hacer la prueba de potencia *a priori* y escoger bien el tamaño de muestra. De ese modo los resultados negativos pasan a ser tan valiosos como los positivos. Además, en la versión típica de comprobación estadística de “nulas tontas” un p-valor estadísticamente significativo no indica nada sobre la magnitud de las diferencias (puede que sean muy pequeñas las diferencias pero haberlas detectado al ser la muestra muy grande) y tampoco es una medida de evidencia contra la nula. Es decir, un p-valor más pequeño no indica una evidencia mayor contra la nula. Todo lo que podemos decir es si el p-valor está a un lado u otro del valor de corte escogido, si el alumno está aprobado o reprobado; da igual si se reprueba con un 4.9 o con un 1.3. En esto difieren las dos visiones clásicas de las que emergió, como un híbrido, el contraste de hipótesis nulas. Para Ronald Fisher, el primer padre de la criatura, el valor exacto de p (p. ej., $p=0.032$) es empleado como medida de la evidencia de la nula (que no se acepta ni se rechaza). Sin embargo J. Neyman y E. Pearson introducen el concepto de alfa (fijado típicamente en el 5%) y el de hipótesis alternativa y con este último el de potencia y magnitud del “effect size”, rechazando la hipótesis nula cuando el p-valor es menor que alfa. Un p-valor de 0.03 significa que la probabilidad de haber obtenido mis datos (u otros más extremos), es muy baja, de tan sólo el 3%, asumiendo que la nula es cierta claro y que repitiéramos el experimento infinitas veces. Conviene tener presente esta diferencia ya que en realidad no es correcto mezclar ambos paradigmas, como por ejemplo realizando una

prueba de potencia y ofreciendo después el valor exacto del p-valor como si con ello aportásemos más información que diciendo simplemente si es mayor o menor que el punto de corte previamente acordado (alfa). Si queremos un punto de corte más arriesgado hay que fijarlo *a priori*, antes de empezar el experimento. Hacerlo *a posteriori* es hacer trampa.

También es importante tener en mente que si se opta por el procedimiento de contraste de hipótesis con contenido biológico, mediante el empleo de pruebas de potencia prospectivas, dicho procedimiento debería repetirse independientemente para cada una de las variables a analizar, ya que para cada variable se debe especificar el “effect size” que se considera biológicamente relevante. Estrictamente hablando, no es correcto aquello de calcular el tamaño de muestra para comprobar un efecto y, aprovechando que tenemos al animal o planta en mano, medir diez variables más.

Intervalos de confianza desde el punto de vista frecuentista.

Los intervalos de confianza (IC) son una herramienta de utilización muy extendida; sin embargo su uso en ecología como alternativa a los contrastes de hipótesis es reciente. Lo primero que hay que tener claro es qué es un intervalo de confianza. Sorprendentemente, la gran mayoría de los usuarios de los IC los interpretan erróneamente al pensar que dentro del intervalo que obtenemos (pongamos uno del 95%) se encuentra el verdadero valor del parámetro con una probabilidad del 95%. Como veremos más adelante, dicha interpretación sólo es adecuada en el caso de los intervalos de credibilidad de la estadística Bayesiana. Los intervalos de confianza en realidad tienen una interpretación frecuentista, es decir, de significado a largo plazo, tras infinitas repeticiones (un significado “asintótico” en términos matemáticos). Si repetimos nuestro experimento muchas veces, en cada una de las repeticiones obtendremos un intervalo de confianza distinto. Un intervalo del 95% significa que, si pudiéramos repetir de manera indefinida el proceso de repetición, en proporción, 95 de cada 100 de esos intervalos incluiría el verdadero valor del parámetro, el cual se puede encontrar en cualquier punto entre los límites superior e inferior de cada intervalo individual (el punto central que proporcionamos en cada caso es sólo una propiedad de los datos, de nuestra muestra). Es decir, a largo plazo, con infinitas repeticiones, tendríamos una probabilidad de 0.95 de escoger un intervalo que incluyese el valor real del parámetro. El problema es que nosotros hacemos un único experimento y, por tanto, esperamos (confiamos) que nuestro intervalo sea uno de los que realmente contiene al valor real. No siempre es fácil obtener intervalos de confianza. Para ello es necesario disponer de la distribución en el muestreo (“sampling distribution”) del estadístico que estamos utilizando para aproximar el parámetro de interés. Cuando no se dispone de dicha distribución, una buena manera de obtener intervalos de confianza es mediante la técnica de “bootstrap”, es decir, mediante muestreo con reemplazo de nuestra muestra de datos.

Este método se basa en la obtención de una distribución muestral aproximada. En concreto, si obtenemos, por ejemplo, 1000 submuestras al azar, a partir de nuestra muestra inicial, podemos generar 1000 estimaciones distintas del parámetro, ordenarlas después de menor a mayor y escoger los dos valores que limitan el $\alpha/2\%$ de valores extremos, por arriba y por abajo. Este método se ajusta bastante a la filosofía frecuentista, ya que hace múltiples repeticiones, si bien la validez de la estimación está muy sujeta a que nuestra muestra inicial fuese realmente representativa de la población.

Los intervalos de confianza contienen mucha información. Tanta como los contrastes de hipótesis con o sin contenido y mucho más. Vamos a tratar de resumirlo mediante un ejemplo. Digamos que estamos interesados en saber si una población local de sapos comunes pone más huevos que otra población cercana de la misma especie. A tal efecto contamos los huevos puestos en 20 charcas de ambos sitios y obtenemos una media aritmética y el intervalo de confianza de la diferencia de las medias (un parámetro que suele obtenerse por defecto con los programas estadísticos clásicos, aunque no le prestemos mayor atención normalmente). Digamos que ese intervalo del 95% de la diferencia de las medias tiene como límites inferior y superior 0.1 y 0.3. Esto nos informa de varias cosas, a) dado que el intervalo de la diferencia no incluye el 0 las diferencias entre poblaciones son estadísticamente significativas b) si nuestra hipótesis nula informada era, por ejemplo, que la diferencia entre poblaciones es de un 40% podemos afirmar que las diferencias no son tan grandes entre poblaciones ya que el límite superior es una diferencia del 30% y finalmente c) podemos juzgar sobre el grado de precisión de la estimación, sobre su grado de incertidumbre; en este caso, ya que las diferencias oscilan entre un 10 y un 30%, podemos decir que la estimación está regularmente acotada, probablemente porque nuestra muestra es pequeña y presenta una gran variabilidad. Es importante tener en cuenta que el intervalo de confianza (al igual que el error típico o estándar) proporciona información sobre la distribución al muestreo, es decir, de la distribución que obtendríamos tras infinitas repeticiones de nuestro muestreo dentro del paradigma frecuentista, al contrario la desviación típica (o estándar) que sólo proporciona información sobre variabilidad en nuestras muestras de datos. La amplitud del intervalo de confianza del 95% es básicamente 2 veces el error típico (estrictamente hablando es 1.96 veces el SE). Y el error típico es en realidad la desviación típica de la distribución al muestreo. El intervalo de confianza es una propiedad de los datos pero permite inferir sobre el posible valor del parámetro porque es una propiedad de nuestros infinitos muestreos, que se acercan supuestamente al valor real (y fijo) del parámetro, si hemos muestreado sin sesgos de ningún tipo.

HERRAMIENTAS ALTERNATIVAS PARA RESOLVER CONTRASTES BIOLÓGICAMENTE INFORMADOS.

Criterios de información teórica.

Con todos los problemas que acumula la técnica del contraste de hipótesis nulas, aún es salvable, especialmente en su formato “duro” (con prueba de potencia *a priori* especificando el “effect size” de interés) si realmente podemos solventar nuestro problema respondiendo una pregunta de manera dicótoma (sí/no). Pero esto sólo es aplicable a problemas sencillos. En cuanto aumentamos un poco la complejidad y tenemos problemas de múltiple causalidad, el blanco o negro queda apartado y se hace necesario tener en cuenta toda la escala de grises. Es decir, se impone el uso de múltiples hipótesis. Así, en nuestra pregunta original de por qué acaban estabilizándose las tortugas en una zona en particular, podríamos formular diversas hipótesis complejas:

1. Se estabilizan ahí porque es donde la densidad de machos alcanza su mayor valor y además porque es donde más humedad encuentran.
2. Se estabilizan ahí porque es donde más densidad hay de alimento apropiado y además porque es donde huele más parecido al lugar de origen de las tortugas.
3. Se estabilizan ahí porque están agotadas y su condición física está por debajo de un determinado valor crítico.
4. Se estabilizan ahí porque es el lugar en el que el sustrato es más adecuado para la excavación de los nidos, con densidades altas de machos.

En fin, en función de nuestro conocimiento previo de los requerimientos de la especie (de nuevo el subjetivismo informado en acción) podríamos continuar planteando hipótesis distintas. El objetivo del “juego” consiste en medir de alguna manera el grado de evidencia de cada una de ellas, con la única ayuda de los datos observados (sin que importen los datos extremos no observados, es decir, sin requerir las infinitas repeticiones frecuentistas). Para ello, tal y como hemos comentado antes, hace falta dominar el difícil arte de poner nuestras hipótesis en forma de modelos matemáticos. Desde esta perspectiva, nuestra sugerencia para un ecólogo medio, es recurrir a la ayuda de un estadístico experto en modelaje. El estadístico nos ayudará a seleccionar entre todos los modelos estadísticos disponibles en el mercado (Modelos Lineales Generales, Modelos Lineales Generalizados, Modelos Generalizados Aditivos, Modelos Lineales Mixtos, ZIP, Loess, Tree, AdaBoost, etc.) para expresar las hipótesis en el formato adecuado. Una vez tengamos las hipótesis en forma de modelos científicos tendremos que tratar de medir de alguna manera el grado de “proximidad a la realidad”. Desde el punto de vista de la estadística frecuentista (en el siguiente apartado veremos como aborda este problema la

estadística Bayesiana), una de las estrategias disponibles es el empleo de criterios de información teórica, un método que combina la pérdida de información teórica con la verosimilitud clásica, la “likelihood” inglesa (L), que es una medida de la probabilidad de obtener nuestros datos dado el valor del parámetro (lo opuesto a lo que buscamos en realidad). En concreto se suele calcular el denominado $AIC = -2\ln(L) + 2K$ que es el Criterio de Información de Akaike, un estadístico que sopesa el grado de ajuste del modelo [medido por su deviance = $-2 \ln(L)$] y su complejidad (medida por el número de parámetros empleados, K , para reducir la falta de ajuste del modelo). El AIC relaciona la verosimilitud con la cantidad relativa de información de Kullback-Leibler perdida al emplear cada modelo para aproximarse a la verdad. Es importante resaltar que si el número de parámetros es grande en relación al tamaño de muestra ($n/K < \text{aproximadamente } 40$) es más apropiado utilizar el llamado AIC corregido. El valor del AIC es relativo, no absoluto, es decir, sólo es útil si se calcula para varios modelos (hipótesis), permitiendo contrastar la pérdida de información de cada uno. Cuando el AIC entre modelos es muy distinto no hay problema (escogemos siempre el que tenga menor AIC, porque representa menor pérdida de información, o el mejor compromiso entre ajuste y complejidad) pero si los AIC de varios modelos son parecidos surge la duda de con cuál quedarnos. Para ello podemos calcular el denominado peso de Akaike “ w_i ” que sopesa la verosimilitud del modelo individual respecto al conjunto de modelos contemplados y re-escala a 1, de modo que los valores de w_i pueden considerarse absolutos, moviéndose entre 0 y 1 como una probabilidad. La verosimilitud (L) se obtiene a partir de la diferencia en AIC de un modelo concreto respecto al modelo con el menor AIC (el más parsimonioso). Si se quiere contrastar entre sí solamente la fuerza relativa de la evidencia de dos hipótesis es posible hacerlo mediante el denominado “Evidence Ratio” (E), que es básicamente el ratio entre los pesos de Akaike (w_i) de ambos modelos. Así pues mediante el uso de los criterios teóricos de información se pueden realizar contrastes entre pares de hipótesis (entre pares muy distintos al formado por una nula y su alternativa) pero también entre un número mayor de ellas. Es importante recalcar que el objetivo final es comparar diversos modelos, en función del grado de apoyo de los datos al modelo. Y que los distintos modelos no tienen porqué estar anidados (es decir, ser “subconjuntos” del modelo general). En la aproximación clásica, por el contrario, pretendíamos obtener un único modelo mediante simplificación por pasos a partir de un modelo general que contuviese todos los factores, covariables e interacciones de interés. Ahora retenemos el modelo con menor AIC (el más parsimonioso, es decir, el que ha conseguido un mejor ajuste del modelo con el menor número de parámetros, o sea, el que se espera que pierda la menor cantidad de información teórica) como referente, aunque recordando que no por más parsimonioso es más cierto. De este modo nos evitamos el problemático proceso de toma de decisiones dicótomas en cadena, de la estadística clásica, para la

selección del “mejor” modelo, que en ningún momento valora la evidencia disponible a favor de las hipótesis (modelos) alternativas. Otra gran ventaja de los criterios de información es la posibilidad de realizar el denominado “model averaging” o “multi-model inference”. Con esta técnica sopesamos la magnitud del efecto estimado para cada modelo con su peso de Akaike (w_i) y de este modo podemos obtener una estimación ponderada de la magnitud del efecto de una variable y de su grado de incertidumbre.

Estadística Bayesiana

Los criterios de información teórica son hasta ahora lo más parecido que tenemos a trabajar con la probabilidad de la hipótesis nula o del parámetro, dados los datos, nuestro verdadero y último objetivo. En el caso de los contrastes de hipótesis nulas tradicionales nos teníamos que conformar con una estimación de la probabilidad de obtener nuestros datos o datos más extremos en el caso de que la nula fuese cierta (el p-valor). Con los criterios de información la situación mejoró ostensiblemente ya que la probabilidad de nuestros datos, y sólo de nuestros datos, sin necesidad de pensar en datos más extremos (dado el parámetro o la hipótesis nula) se contrasta con la verosimilitud de todos los demás modelos en el cálculo del AIC y los pesos de Akaike. La forma en la que la estadística Bayesiana aborda la inferencia nos permite dar un paso adelante muy importante en relación con nuestra idea final de evaluar la relevancia biológica en ecología. La forma de hacerlo es sopesando la verosimilitud de cada modelo con la información previa disponible sobre el parámetro o la hipótesis. Mediante la regla de Bayes (que escribimos abajo) se combinan la probabilidad *a priori* del parámetro $P(\theta)$ y la verosimilitud del modelo (obtenida a partir de nuestra única muestra de datos, $L(\theta)=P(\text{Datos}|\theta)$) para obtener la distribución de probabilidad posterior del parámetro dados los datos $P(\theta|\text{Datos})$. Es decir, actualizamos la información previa de que muchas veces disponemos con la información aportada por los datos específicos de nuestra muestra para obtener información sobre el parámetro.

La regla de Bayes

$$P(\theta | \text{Datos}) = \frac{P(\text{Datos} | \theta)P(\theta)}{P(\text{Datos})}$$

La estadística Bayesiana no es un invento moderno; es en realidad un modo de hacer inferencia más antiguo que el de la máxima verosimilitud desarrollado por Fisher, pero quedó históricamente apartada debido a la imposibilidad de resolver las complejas integrales que requiere el cálculo de las probabilidades posteriores.

Recientemente, con el desarrollo de la computación moderna, ha renacido de sus cenizas y está viendo un creciente desarrollo que sin duda continuará en años venideros para acabar probablemente imponiéndose a la aproximación frecuentista cuando se desarrolle software más amigable. La gran ventaja de la estadística Bayesiana es que nos permite abordar de manera mucho más sencilla el problema de la inferencia. Esto se debe a que realizar inferencia se reduce a explicar con detalle la distribución *a posteriori*. Esto antes podía llegar a ser un problema ya que en la mayoría de las situaciones la distribución *a posteriori* tiene una forma analítica intratable. Sin embargo ahora, gracias a las técnicas de simulación por cadenas de Markov (MCMC) es relativamente sencillo poder obtener al menos una aproximación simulada de la distribución *a posteriori*. Hay varios procedimientos para obtener las cadenas de Markov, aunque el más habitual es Gibbs sampling que da nombre al software más comúnmente empleado hoy en día en estadística Bayesiana: WinBUGS (Bayesian inference Using Gibbs Sampling), disponible en <http://www.mrc-bsu.cam.ac.uk/bugs/>. Los modelos jerárquicos bayesianos nos permiten abordar de manera muy sencilla modelos complicados (diseños anidados, diseños con componente espacial, etc.) debido a la propia estructura jerárquica con la que relacionan los datos, los parámetros y los hiper-parámetros. En el universo frecuentista hablar de probabilidades del parámetro carece de sentido ya que los parámetros son considerados desconocidos pero fijos, cosa que no ocurre en la estadística Bayesiana, donde son considerados como variables aleatorias. La forma en la que aparecen las distribuciones de probabilidad en el contexto frecuentista es a través de las distribuciones que siguen los estimadores con los que aproximamos los parámetros (conocidas como distribuciones en el muestreo o “sampling distribution”). Cuando no se dispone de información previa la estadística Bayesiana nos permite utilizar las denominadas “prior planas o no informativas”. En ese caso las estimaciones clásicas del parámetro, obtenidas haciendo inferencia por máxima verosimilitud (es decir dando valores al parámetro hasta encontrar el valor que maximiza la probabilidad de haber obtenido nuestros datos), y las medias de la probabilidad posterior son muy parecidas, porque basamos la inferencia sólo en la información aportada por los datos. Cuando se cuenta con información previa buena (que puede ser simplemente el resultado de nuestro trabajo anterior sobre el tema) los resultados de los análisis basados en criterios de información pueden diferir de manera importante en relación a los bayesianos. En cualquier caso, aún sin *priors* informativas, la aproximación Bayesiana es siempre mejor que la clásica porque se basa sólo en nuestro conjunto de datos y no en “nuestros datos más todos los valores más extremos no observados”, como requiere el paradigma clásico de comprobación de hipótesis y además nos permite interrogar a la distribución *a posteriori* y contestar muchas preguntas acerca del parámetro que son impensables en el marco frecuentista. Por tanto, mediante la aproximación Bayesiana extraemos el mayor

provecho posible de nuestros datos ya que no partimos siempre de una situación, muchas veces ficticia, de desconocimiento inicial del problema sino que aprovechamos el conocimiento que se ha ido acumulando antes de que nosotros nos interesásemos en esa cuestión, introduciéndolo como probabilidad *a priori*.

Otra de las grandes ventajas de la Bayesiana es la obtención de “intervalos de credibilidad” en los que se encuentra con seguridad el valor del parámetro con la probabilidad deseada. Por ejemplo, si tenemos un intervalo de credibilidad del 95% con límites (0.3-0.5) podemos afirmar que el verdadero valor del parámetro está contenido entre dichos límites con una probabilidad del 95%. La estadística Bayesiana es especialmente apropiada para problemas relacionados con la toma de decisiones ya que contamos con una medida natural de incertidumbre cual es la probabilidad del parámetro (o de la hipótesis nula). Además, en lugar de dar una respuesta automática, en cada caso hay que pensar lo que esa probabilidad significa para nosotros, los expertos en el problema en cuestión. No es lo mismo tener una probabilidad del 90% de ganar en las carreras de caballos que tener un 90% de estrellarse en un avión. En el primer caso apostaríamos ese día y en el segundo caso no tomaríamos ningún avión. Al igual que en el caso de los criterios de información es posible la comparación múltiple de modelos y su selección mediante el denominado DIC (Deviance Information Criterion) o la comparación entre pares de modelos mediante el BIC (Bayesian Information Criterion), un equivalente del “Likelihood ratio test” frecuentista y del “Evidence Ratio” de los criterios de información. En la práctica se pueden abordar mediante estadística Bayesiana tanto los análisis más sencillos (medias, regresiones, correlaciones, análisis de varianza, análisis de proporciones) como los muy complejos (análisis de captura-recaptura, análisis de viabilidad, series temporales, etc.), ya que los modelos bayesianos son extremadamente flexibles. También se puede corregir por sobredispersión (la varianza observada de los datos es mayor que la predicha) y pseudoreplicación. El software disponible no es todavía tan amigable como el disponible para los análisis frecuentistas, ya que requiere escribir los códigos para el modelo pero cada vez hay más obras disponibles en las que se facilitan dichos códigos de modo que tan sólo hay que adaptarlos a nuestro caso particular. Aunque quizás sea más práctico reservar la estadística Bayesiana para tratar con problemas complejos, si se dispone de información previa vale la pena incluso abordar los problemas ecológicos sencillos “a la Bayesiana”, ya que los resultados pueden diferir en tal caso de manera sustancial, sobretodo si se dispone de información previa fiable. (véase recuadro 3)

En definitiva las herramientas disponibles en nuestra caja de trabajo para contrastar hipótesis ecológicas con contenido biológico han aumentado notablemente en los últimos años. Que haya herramientas más complejas a nuestro alcance no significa que tengamos necesariamente que usarlas, ya que lo más importante es la pregunta y no el medio para responderla. Hay muchas preguntas de planteamiento

simple (aunque de gran peso) que se pueden solucionar haciendo buen uso de las herramientas más tradicionales, incluso tan sólo mediante estadística descriptiva. No obstante, la ecología es una ciencia que trata directamente con la complejidad y a menudo es necesario recurrir al planteamiento de las múltiples hipótesis para ver luz en un problema. En este sentido se podría decir que los criterios teóricos de información y la estadística Bayesiana son las herramientas más adecuadas para analizar datos en ecología, mientras que los contrastes clásicos de hipótesis quedarían para aquellos casos en los que los planteamientos dicótomos son de interés. Cualquier desafío como el de dominar (aunque sea a nivel de usuario) las herramientas aquí descritas supone un reto que impresiona especialmente cuando nos encontramos en el “campamento base”, usando una metáfora alpinista: la simple mirada hacia la cumbre nos puede invadir con ciertas dosis de angustia y de dudas. Pero aún siendo angustias y dudas fundamentadas (pues hay que hacer un gran esfuerzo para ello) el mejor consejo es proceder despacio, de campamento en campamento, y a ser posible con la inestimable ayuda de alguien que domine la materia y nos arroje la luz que necesitamos cuando esas dudas se repitan en la ascensión. También es importante recordar que las técnicas modernas no están exentas de una mala utilización. Su uso no es suficiente. Si por algo se distinguen de las más tradicionales es por la necesidad de pensar cada paso del proceso de análisis. Pensar es la esencia de la ciencia y ese trabajo no se lo puede uno ahorrar en ninguna de las fases, incluida la delicada etapa del análisis de datos. Disponer de una buena cultura estadística es fundamental no sólo para poder analizar tus datos sino, especialmente, para interpretar tus resultados correctamente. La literatura científica está plagada de interpretaciones erróneas de lo que significan los resultados obtenidos y éste es un serio obstáculo que dificulta la acumulación de conocimiento y el avance la ecología como ciencia. Un buen consejo para concluir es tener claro el problema ecológico que uno tiene y escoger luego la herramienta que se lo resuelva de la manera más sencilla y no proceder nunca al contrario.

Recuadro 3: Análisis de proporciones: comparación entre las distintas herramientas.

Supongamos que estoy interesado en saber si hay paridad en el sex-ratio de los pollos de una población de aves (los pollos se pueden sexar mediante el color del iris). Es decir, quiero averiguar si la proporción de machos de la población es igual a la de hembras. Para ello tomo una muestra grande de pollos ($n=810$) y los sexo, encontrando que 404 de ellos son machos y 406 hembras. Para testar mi hipótesis paridad en el sex ratio de la población puedo seguir varias aproximaciones alternativas. Aquí las presentamos y discutimos:

1) Testado clásico de hipótesis nulas sin contenido biológico.

Mi $H_0 = \% \text{ de machos} = \% \text{ de hembras} = 0.5$ en la población y mi hipótesis alternativa $H_1 =$ número de machos es distinto al número de hembras.

Con los resultados de mi muestreo planteo una prueba de bondad de ajuste en la que contrasto las frecuencias observadas con las esperadas en el caso de que las proporciones entre sexos fuesen iguales. El resultado de tal prueba es que $\chi^2_c = 0.00$, g.l. = 1, p-valor = 1.00. Es decir no puedo rechazar la hipótesis nula de igualdad. Pero no puedo concluir que sea cierta. Simplemente podemos decir que no tenemos evidencia estadística suficiente para decir que no sea cierta, lo cual es muy poco concluyente.

2) Testado clásico de hipótesis nulas con contenido biológico.

En este caso, reconozco que una proporción de sexos exactamente de 1:1 es muy improbable que se de y establezco una magnitud del efecto (effect size) que para mi tenga sentido biológico. Un effect size de 0.1 se considera bajo por convenio, así que decido hacer una prueba de potencia prospectiva con esa magnitud del efecto, para saber que tamaño de muestra necesito para tener una potencia del 80% (es decir un error de Tipo II del 20%), estableciendo un error de Tipo I *a priori* del 5%. Un sencillo cálculo en GPower me indica que necesito una muestra total de 785 pollos. Como trabajo con 810 pollos voy por buen camino. Ahora llevo a cabo mi prueba de bondad de ajuste y obtengo los resultados expuestos en el punto anterior, pero puedo aceptar mi hipótesis nula informada: que las proporciones de machos y hembras difieren en una magnitud pequeña.

3) Criterios de información teórica (Modificado de McCarthy 2007)

La verosimilitud de la primera hipótesis es igual al número de machos partido por el número total de pollos. La verosimilitud de la segunda hipótesis es su complementario. Dado que empleamos los datos para estimar el parámetro $K=1$.

Hipótesis	Verosimilitud	AIC	w_i
% machos = 0.5	0.4987	0.6085	0.4987
% machos = distinto de 0.5	0.5013	0.6189	0.5013

El Evidence Ratio ($E=1.00$) sugiere que ambas hipótesis son igualmente probables, lo que sugiere paridad entre sexos.

4) Aproximación Bayesiana (modificado de McCarthy 2007)

Construimos un sencillo modelo en el que nuestros datos sobre sex-ratio proceden de muestrear al azar en una distribución binomial; definimos un prior plano por medio de una distribución uniforme que oscila entre 0 y 1 y especificamos que nuestros datos son los 404 machos encontrados.

```
model
{x~dbin(r,810) #los datos proceden de una distribución binomial con n=810
r~dunif(0,1) #prior para el sex ratio
}
list(x=404) # se encuentran 404 machos
```

Corremos 100.000 iteraciones de la simulación (con un “burn-in” de 10,000) y obtenemos que la media de la distribución de la proporción de machos es de 0.4994, con un credible interval del 95% que oscila entre 0.465 y 0.5339. Concluimos que no es descartable que la proporción de sexos sea de 0.5 pero, ya que hemos obtenido la distribución de probabilidad del parámetro (de la proporción de machos), podemos preguntarnos cual es la probabilidad de que la proporción sea exactamente 0.5 o de que oscile entre 0.49 y 0.51. Esas probabilidades son respectivamente de 0 y 0.43, con lo cual podemos concluir que la población de pollos presenta un grado de paridad de sexos importante, pero también se constata el absurdo de plantear una hipótesis nula de igualdad a 0.5 como se hace en el contexto frecuentista. En definitiva obtenemos una conclusión mucho más sólida que las anteriores ya que podemos asignar a nuestras hipótesis probabilidades concretas, empleando una medida natural de su incertidumbre.

ANEXO I

Ofrecemos un listado de bibliografía donde se pueden encontrar las “instrucciones” para manejar más adecuadamente la caja de herramientas, lo que nos ayudará a escoger la herramienta más adecuada para cada problema.

Sobre estadística clásica y ecología cuantitativa básica

Piñol, J. & Martínez-Vilalta, J. 2006. *Ecología con números: una introducción a la ecología con problemas y ejercicios de simulación*. Lynx Edicions, Barcelona.

Zar, J.H. 1996. *Biostatistical Analysis*. 3rd edn. Upper Saddle River, New Jersey.

Sobre diseño experimental

Quinn, G.P. & Keough, M.J. 2004. *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge.

Scheiner, S.M. & Gurevitch, J. 2001. *Design and analysis of ecological experiments*. Oxford University Press, Oxford.

Steidl, R.J., Hayes, J.P. & Schaubert, E. 1997. Statistical power analysis in wildlife research. *J. Wildlife Manage.* 61: 270-279.

Sobre críticas al contraste clásico de hipótesis

Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., & Thomason, N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* 20: 1539-1544.

Germano, J. D. 1999. Ecology, statistics, and the art of misdiagnosis: The need for a paradigm shift. *Environ. Rev.* 7: 167-190.

Gigerenzer, G. 2004. Mindless statistics. *The Journal of Socio-Economics* 33: 587-606.

Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). 1997. *What if there were no significance tests?* Lawrence Erlbaum Associates, Publishers, London.

Hobbs, N.T. & Hilborn, R. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. *Ecol. Appl.* 16: 5-19.

James, F.C. & McCulloch, C.E. 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? *Annu. Rev. Ecol. Sys.* 21: 129-166.

Johnson, D.H. 1999. The insignificance of statistical significance testing. *J. Wildlife Manage.* 63: 763-772.

Lukacs, P.M., Thompson, W.L., Kendall, W.L., Gould, W.R., Doherty, P.F., Burnham, K.P. & Anderson, D.R. 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. *J. Appl. Ecol.* 44: 456-460.

Martínez-Abraín, A. & Oro, D. 2005. Can ornithology advance as a science relying on significance testing? A literature review in search of a consensus. *Ardeola* 52: 377-387.

Martínez-Abraín, A. 2007. Are there any differences? A non-sensical question in ecology. *Acta Oecol.* 32: 203-206.

Martínez-Abraín, A. 2008. Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica* (In press).

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martínez del Río, C. 2005. Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* 42: 4-12.

Stephens, P.A., Buskirk, S.W. & Martínez del Río, C. 2006. Inference in ecology and evolution. *Trends Ecol. Evol.* 22: 192-197.

Williams, B.K., Nichols, J.D. & Conroy, M.J. 2001. *Analysis and management of animal populations.* Academic Press, London

Sobre intervalos de confianza

Belia, S., Fidler, F., Williams, J. & Cumming, G. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Meth.* 10: 389-396.

Cumming, G. & Finch, S. 2005. Confidence intervals and how to read pictures of data. *Am. psychol.* 60: 170-180.

Cumming, G., Fidler, F. & Vaux, D.L. 2007. Error bars in experimental biology. *J. Cell Biol.* 177: 7-11.

Sobre criterios de información teórica.

- Anderson, D.R. & Burnham, K.P. 2002. Avoiding pitfalls when using information-theoretic methods. *J. Wildlife Manage.* 66: 912-918.
- Anderson, D.R., Burnham, K.P. & Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildlife Manage.* 64: 912-923.
- Anderson, D.R., Burnham, K.P. & White, G.C. 2001. Kullback-Leibler information in resolving natural resource conflicts when definitive data exist. *Wildlife Soc B* 29: 1260-1270.
- Anderson, D.R., Link, W.A., Johnson, D.H. & Burnham, K.P. 2001. Suggestions for presenting the result of data analyses. *J. Wildlife Manage.* 65: 373-378.
- Burnham, K.P. & Anderson, D.R. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Res.* 28: 111-119.
- Burnham, K.P. & Anderson, D.R. 2002. *Model selection and multi-model inference: a practical information theoretic approach.* Springer-Verlag, New York.
- Burnham, K.P. & Anderson, D.R. 2004. Multimodel Inference: understanding AIC and BIC in model selection. *Sociol. Meth. & Res.* 33: 261-304.
- Hilborn, R. & Mangel, M. 1997. *The ecological detective: Confronting models with data.* Princeton University Press, Princeton.
- Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martínez del Río, C. 2005. Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* 42: 4-12.

Sobre estadística Bayesiana

- Brooks, S.P., King, R. & Morgan, B.J.Y. 2004. A Bayesian approach to combining animal abundance and demographic data. *Anim. Biodivers. Conserv.* 27: 515-529.
- Clark, J.S. 2005. Why environmental scientists are becoming Bayesians? *Ecol. Lett.* 8: 2-14.
- Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol. Appl* 6: 1036-1046.
- Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (Eds.). 2004. *Bayesian data analysis.* Chapman & Hall, London.
- King, R. & Brooks, S.P. 2004. Bayesian analysis of the Hector's Dolphin data. *Anim. Biodivers. Conserv.* 27: 343-354.
- Link, W.A., Cam, E., Nichols, J.D. & Cooch, E.G. 2002. On bugs and birds: Markov Chain Monte Carlo for hierarchical modelling in wildlife research. *J. Wildlife Manage.* 66: 277-291.
- McCarthy, M.A. 2007. *Bayesian methods for ecology.* Cambridge University Press, New York.
- McCarthy, M.A. & Masters, P. 2005. Profiting from prior information in Bayesian analyses of ecological data. *J. Appl. Ecol.* 42: 1012-1019.
- Wade, P.R. 2000. Bayesian methods in conservation biology. *Conserv. Biol.* 14: 1308-1316.

Recibido: 13 de septiembre de 2007

Aceptado: 27 de febrero de 2008