Potencia de una prueba estadística: aplicación e interpretación en ecología del comportamiento

Javier Quesada[1, 2] y Jordi Figuerola[3]
[1] *Institut Català d'Ornitologia, Barcelona*
Correo electrónico: analisi@ornitologia.org
[2] *Unitat associada d'ecologia evolutiva i comportamental, Museu Ciències Naturals, Barcelona*
[3] *Departamento de Ecología de Humedales, Estación Biológica de Doñana (CSIC)*

Resumen

La verificación de hipótesis requiere un planteamiento previo sobre diversas cuestiones, entre las que el tamaño de muestra a analizar y la fiabilidad de los resultados obtenidos son aspectos fundamentales. La potencia de una prueba estadística permite determinar la fiabilidad de dichas pruebas así como el tamaño muestral necesario para abordar el estudio deseado. El objetivo de este artículo es definir qué es la potencia de una prueba estadística, explicar su cálculo y describir sus aplicaciones, que son particularmente importantes en estudios comportamentales.

Introducción

Cuando un investigador se formula una hipótesis que desea verificar surgen inmediatamente una serie de problemas o cuestiones a resolver. ¿Cuál es el sistema ideal para verificar la hipótesis?, ¿podré abordar experimentalmente esa cuestión?, ¿cuáles son las variables importantes que debo controlar? Para garantizar una mayor probabilidad de éxito en la empresa es recomendable plantearse en esta fase cuales son las herramientas estadísticas con que vamos a analizar los datos, qué factores serán fijos, cuales aleatorios, etc. Sería muy extraño que en distintas partes de este proceso de preparación no sugiera la pregunta: ¿qué tamaño muestral necesito para comprobar mi hipótesis? Esta cuestión es un elemento clave

cuando se plantea un experimento. Otra pregunta también muy recurrente es aquella que se refiere a la fiabilidad de una prueba estadística que ya se ha realizado: ¿me puedo fiar de la "*p*" que he obtenido? Esta pregunta se hace frecuentemente cuando los resultados de una prueba estadística no son significativos: ¿no lo son porque verdaderamente no hay un efecto, o el problema es que tenemos un tamaño muestral tan reducido que no somos capaces de detectar estadísticamente un efecto que por otro lado podría ser biológicamente importante? (ver [11]).

A la primera pregunta (el tamaño muestral adecuado) se puede contestar mirando por ejemplo un libro de estadística que, según el análisis estadístico que estemos realizando y la distribución de nuestros datos, nos hará una serie de recomendaciones generales para escoger el tamaño muestral más adecuado. Una recomendación que se utiliza (que se considera a veces como norma) entre la comunidad científica es que podemos empezar a realizar casi cualquier análisis estadístico a partir de un tamaño muestral superior a 30. Esto es debido a que la mayoría de las variables siguen el teorema central del límite (ver [18], pág. 62, para una discusión más detallada del tema). Asumiendo que tenemos un tamaño muestral adecuado y que los datos cumplen las asunciones del método estadístico utilizado podremos suponer que los resultados son fiables. Sin embargo este tipo de aproximaciones no dejan de tener un carácter demasiado generalizador ya que, por ejemplo, podríamos preguntarnos sobre la naturaleza del efecto que estamos midiendo y su relación con el tamaño muestral. Por ejemplo, para determinar si hay diferencias en masa corporal entre machos y hembras de Gorrión Común (*Passer domesticus*) posiblemente sí sería adecuado utilizar un tamaño muestral de 30 pero, si quisiéramos ver diferencias en la masa corporal de los machos de esta especie y los machos de águila real (*Aquila chrysaetos*) la propia intuición nos dice que con tan solo cinco individuos de cada grupo podríamos demostrar que hay diferencias significativas puesto que nunca un Gorrión Común llegará a pesar como un águila real. Sin embargo, si estuviéramos analizando la relación entre la concentración de un metal pesado en un animal y su posible efecto en el crecimiento de éste, la relación podría resultar muy sutil pero biológicamente importante, de manera que un tamaño muestral de 30 no sería suficiente para detectar tal relación. Por ello, no siempre el tamaño "recomendado" en los libros de estadística aseguraría que estuviéramos utilizando un tamaño muestral adecuado a las preguntas que pretendemos responder y así, la fiabilidad de nuestros resultados podría quedar comprometida.

Actualmente existen herramientas estadísticas de fácil manejo que nos permiten determinar la fiabilidad de nuestras pruebas estadísticas y el tamaño muestral necesario. El cálculo de la potencia de una prueba estadística y su estrecha relación con el tamaño muestral nos permite responder a las dos preguntas iniciales que nos habíamos planteado. A pesar de que el análisis de la potencia de una prueba estadística lleva empleándose muchos años en el campo de la medicina y la biología celular no ha sido hasta la década pasada que su uso se ha empezado a extender en ecología del comportamiento ([31,33]). Definir qué es la potencia de una prueba estadística, sus aplicaciones y como calcularla es el objetivo de este artículo.

Una introducción al uso de la estadística en la ciencia

La ciencia tiene como prioridad explicar los fenómenos que ocurren en nuestro entorno mediante el conocimiento sistematizado. La ciencia utiliza distintos métodos para la adquisición de conocimiento sobre un conjunto de hechos objetivos. La aplicación del método científico conduce a la generación de predicciones concretas, cuantitativas y comprobables referidas a hechos observables. Para obtener una interpretación objetiva de nuestra cuestión, necesitamos tomar los datos de forma que minimicemos el error de medida e intentemos reducir la subjetividad que confieren nuestros juicios cuando observamos un hecho ([12]).

La estadística es una herramienta más en el desarrollo del método científico la cual nos permite analizar los datos recabados ([18]). Sin embargo, ésta se basa en el estudio de la probabilidad de que un suceso ocurra. Por ello, se ha de tener presente que los resultados estadísticos obtenidos y aquellas conclusiones que se derivan de estos, se deben interpretar con cautela ya que siempre tendremos una probabilidad de que lo que estemos afirmando no sea cierto, algo que a menudo los científicos olvidamos.

Cuando un científico utiliza la estadística lo hace normalmente con cuatro propósitos fundamentales ([7]):

- **Estadística descriptiva:** resumir las variables para ver tendencias (Ejemplo: media, varianza, desviación típica, mediana, etc.)
- **Estadística para comparar dos o más grupos:** por ejemplo cuando queremos ver la diferencia en la masa corporal de los ciervos de dos cotos con hábitat distintos (*t* de Student, ANOVA, etc.)

- **Estadística para analizar relaciones entre dos variables determinadas:** por ejemplo, determinar si las distintas concentraciones de contaminantes en el ambiente y la coloración del plumaje producido por un ave se encuentran relacionados entre sí (regresión tipo II, correlación, etc.)
- **Estadística para determinar patrones y ordenar conjuntos de datos**. Es lo que denominamos Análisis multivariante (Análisis factorial, Análisis cluster, etc.)

En los tres últimos tipos de estadística mencionados se utiliza el contraste de hipótesis. En estos tipos de análisis se propone una asunción que consideramos verdadera (hipótesis) y utilizamos la estadística para calcular la probabilidad de que esta afirmación sea cierta o no a través de una fórmula o estadístico (7). Existen diferentes tipos de análisis que nos permiten contestar a las preguntas que nos hemos formulado. Según el tipo de hipótesis que tengamos (si analizamos diferencias entre poblaciones o relaciones entre variables), de cómo sean las variables que hemos registrado (continuas, categóricas) y la distribución de nuestra población (normal, Poisson, binomial, etc.) escogeremos un análisis estadístico u otro (24).

Por ejemplo, imaginemos que tenemos dos cotos de caza prácticamente iguales (cotos A y B) en los que se utilizan diferentes manejos, en uno se elimina toda la cobertura arbustiva y en el otro no. Queremos saber si esto puede influir en la condición física de los ciervos que se crían en las dos zonas. Para valorar la condición física elegimos la variable “Masa corporal del animal”. Como en la masa corporal influye la edad, el sexo y la estación del año solo utilizaremos para nuestro estudio los machos de 3 años cazados en otoño para estandarizar nuestros datos y evitar que nuestras conclusiones puedan estar afectadas por estas variables colaterales.

La **Hipótesis Nula** (H_0) es normalmente aquella que nos dice que no hay diferencias entre los grupos o que la relación entre dos variables es 0 (no existe tal relación), ya que lo que realmente cuantifica una prueba estadística es la probabilidad de que el suceso que estamos analizando se deba al azar y no al efecto que estamos midiendo (24). Sin embargo, son posibles otros tipos de hipótesis nula, por ejemplo que el parámetro de correlación r sea igual a 1, 0,5 o a cualquier otro valor entre 0 y 1.

En nuestro caso, la hipótesis nula que comprueba el análisis estadístico (*t* de Student) es que no existen diferencias en la masa corporal de los animales en la zona A y B, es decir: *Media de la Masa corporal en A*

= *Media de la Masa corporal en B*. Cuando realizamos nuestro análisis estadístico, obtenemos la probabilidad de que H_0 sea cierta (la tan famosa *p*), pero cuando esta probabilidad es muy baja no tenemos más remedio que descartar que ocurra H_0 (que no existan diferencias, o que no haya relación entre dos variables). De esta manera, lo más probable es que esté ocurriendo lo contrario a lo que dice la H_0, y que sí existan diferencias entre ambos grupos (o que existe una relación entre las dos variables si lo que estamos haciendo es correlacionar dos variables). Esto es lo que se denomina **Hipótesis alternativa (H_1)**.

¿Cuándo hemos de aceptar que la H_0 no es cierta? ¿Dónde se establece el límite que dice que la p es muy pequeña?

No existe un criterio matemático pero sí uno científico. Normalmente cuando la p que nos da el estadístico es menor que 0,05 se decide que la probabilidad de que ocurra H_0 es tan baja que lo que debe estar ocurriendo es la H_1 (las masas corporales de los ciervos en los dos cotos diferirán). En las distintas disciplinas sin embargo se utilizan diferentes valores de *p* (Geología = 0,1, Ciencias Naturales = 0,05, Toxicología = 0,01).

Si nos fijamos bien siempre estamos hablando de probabilidades. Por ejemplo, si trabajamos a un nivel de significación $p = 0,05$ y aceptamos H_1 hemos de observar que al menos queda un 5% de probabilidad de que la H_0 sea cierta, por lo que estamos asumiendo un posible error de equivocarnos, es decir de aceptar que ocurre H_1 cuando en realidad lo que ocurre es H_0. Así pues, cuando hacemos un solo análisis y obtenemos una $p < 0,05$ y aceptamos la H_1 estamos asumiendo un porcentaje de error de tal manera que pueden ocurrir dos cosas:

- Que aceptemos H_1 y la verdad sea que H_1 es cierta.
- Que aceptemos H_1 pero que verdaderamente ocurra H_0.

Este pequeño error que asumimos se denomina **Error de tipo I** y se denota por la letra griega α.

Imaginemos ahora que obtenemos una $p = 0,40$. En este caso aceptaríamos H_0 asumiendo que no existen diferencias entre los pesos de los ciervos de los dos cotos. Sin embargo, también existe una probabilidad de que en realidad esté ocurriendo H_1 y nuestro estudio sea incapaz de detectar las diferencias entre los dos grupos, debido a un deficiente diseño experimental, por ejemplo porque tengamos un tamaño muestral muy pequeño. Este posible error se denomina **Error de tipo II** y se denota β.

A cualquier científico le interesaría disminuir, o como mínimo acotar, los dos tipos de errores.

La potencia de una prueba estadística se calcula como 1-β y se entiende como la probabilidad de afirmar que la H_1 es verdadera ([22]). También se puede entender como la probabilidad de rechazar H_0 cuando de hecho es falsa ([20]). El interés de la potencia de una prueba estadística estriba en que los análisis estadísticos más comúnmente usados están diseñados para "controlar" el error de tipo I obviándose tradicionalmente el error de tipo II, por lo que el cálculo de la potencia de una prueba estadística también se puede entender como una medida de confianza del análisis que hemos realizado, principalmente cuando no hemos obtenido un resultado significativo. Así, si disminuimos el error de tipo II estaremos aumentando la potencia de la prueba estadística.

Al igual que el convenio de la *p*, se entiende que una potencia es adecuada para una prueba estadística cuando es superior a 0,80 (80%), o lo que es lo mismo, fijamos el error β = 0,20. Este convenio (llamado "five-eighty convention") es el más utilizado debido a Jacob Cohen ([3]) quien escribió un libro sobre el análisis de la potencia de pruebas estadísticas frecuentemente citado en ciencias naturales. Sin embargo, en los últimos años el uso de esa convención se ha cuestionado ya que, por una parte, el libro de Cohen estaba enfocado a estudios de psicología, y por otra, debido a una mala interpretación del texto. Este convenio está pensado para estudios en los que no haya ninguna base lógica para elegir otra combinación de α:β (ver [6, 28] para una discusión más profunda e intuitiva de esta cuestión).

Factores que influyen en la potencia de una prueba estadística

El cálculo de la potencia de una prueba estadística se basa en una fórmula general que relaciona cuatro parámetros: la potencia estadística (1- β), el nivel de significación (α), el tamaño muestral y el tamaño de efecto ([3,6,31]; ver Cuadro 1). Sin embargo, según el estadístico que estemos utilizando deberemos emplear una fórmula diferente ([3]).

Conocidos tres parámetros de esta fórmula se puede determinar el cuarto, si bien los más útiles son la potencia de una prueba estadística y el tamaño muestral.

Cuadro 1. ¿Qué es y cómo se calcula el tamaño de efecto?

De los tres parámetros implicados en el cálculo de la potencia de un análisis estadístico, el tamaño de efecto es quizás el que nos pueda resultar menos familiar y por este motivo merece la pena profundizar en la definición y cálculo del tamaño de efecto.

El tamaño de efecto puede interpretarse como una medida estandarizada de la distancia entre H_0 y H_1. El análisis de la correlación lineal entre dos variables es quizá el caso más fácil de comprender e ilustrar. Para la correlación lineal, el tamaño de efecto se corresponde con el parámetro de correlación r, un parámetro ampliamente utilizado en ecología. Para detectar como significativas correlaciones muy importantes (por ejemplo, r = 0,70) necesitaremos muchos menos individuos que para detectar correlaciones mucho menos fuertes (r = 0,20, n = 9 y 150 respectivamente, ver Figura 1).

Para otros análisis, no existe ningún parámetro normalmente utilizado que refleje el tamaño de efecto, y se requieren distintas fórmulas en función de la prueba estadística. Por ejemplo, para comparar las medias de dos grupos independientes el tamaño de efecto se calcula como la diferencia de las medias dividida por la desviación típica. Al realizar un análisis de potencia usando G-Power el programa nos sugerirá distintos valores para tamaños de efecto altos, medios o bajos, basados en las sugerencias de Cohen ([3]).

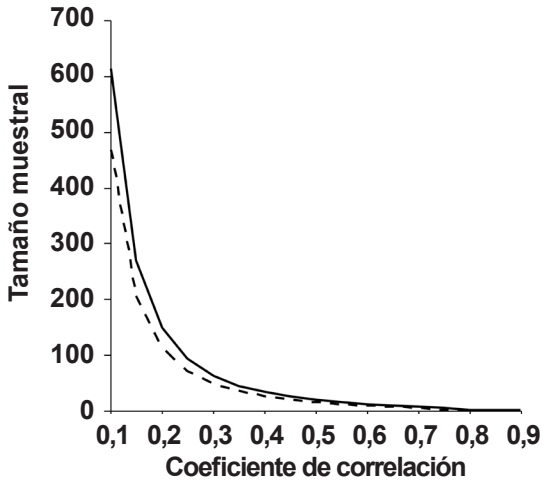

Figura 1. Tamaño muestral necesario para detectar como estadísticamente significativas correlaciones de distinto tamaño de efecto (r). Los tamaños muestrales necesarios son superiores para obtener un poder de 0,80 (línea continua) que para 0,70 (línea discontinua).

¿Cómo podemos aumentar la potencia de una prueba estadística?

La potencia de una prueba estadística depende básicamente del:

- Tamaño muestral
- Tamaño de efecto
- Nivel de significación

Tamaño muestral

Cuando realizamos una prueba estadística lo que estamos haciendo es realmente coger una muestra e inferir que lo que obtenemos con ella es lo que en realidad está ocurriendo en toda la población, ya que la totalidad de la población frecuentemente es imposible de medir (18). Para que podamos obtener conclusiones y determinar que lo mismo ocurre en esa población, la muestra indudablemente ha de ser representativa de la población. Así, cuanto mayor tamaño muestral tengamos mayor representatividad tendrá la población que medimos, más fiable será nuestro análisis, y por lo tanto la potencia de la prueba estadística aumentará. Por ejemplo, podríamos fiarnos más de una correlación positiva entre la abundancia de endoparásitos en los ciervos de una población y la masa de la cuerna si tuviéramos una muestra de 60 individuos que si hubiéramos hecho la correlación tan sólo con cinco.

Si lo planteamos al revés, para una potencia dada (80%), podremos definir el tamaño muestral necesario para que la potencia de nuestra prueba estadística sea alta de manera que nuestro análisis sea fiable y por tanto representativo de la población. En un diseño óptimo el número de individuos es el mismo para cada uno de los grupos considerados (diseño balanceado). Sin embargo, si consideraciones prácticas limitan el tamaño muestral alcanzable, es más práctico aumentar el número de individuos dentro del grupo más fácil o económico de capturar-manipular (2).

Tamaño de efecto

El tamaño de efecto es la diferencia entre la hipótesis nula y la hipótesis alternativa (22) o también el grado en el que el fenómeno se da en la población (3). Intuitivamente, el tamaño de efecto vendría a ser como una medida del “grado de diferencia” entre los dos grupos (sexo: macho y hembra; localidad: Lleida, Girona) que queremos detectar (si hablamos de pruebas que analizan diferencias), o bien el “grado de relación” entre

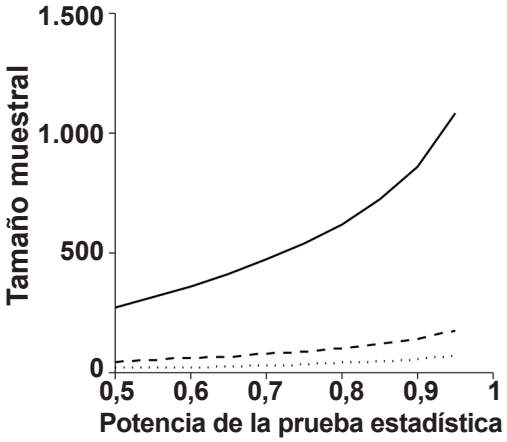

Figura 2. El tamaño muestral necesario para obtener un mismo valor de potencia de una prueba estadística está negativamente relacionado con el tamaño de efecto que queramos detectar. De este modo, para detectar diferencias significativas entre las medidas de dos grupos usando la prueba de la *t* de Student, el tamaño muestral necesario será muy elevado para un tamaño de efecto pequeño (0,20, línea continua), intermedio para tamaño de efecto intermedios (0,50, línea discontinua), o bajo para detectar tamaños de efecto grandes (0,80, línea de puntos). Del mismo modo para obtener mayores potencias es necesario aumentar el tamaño muestral de nuestros estudios.

dos variables que estamos midiendo. Por ejemplo, el tamaño de efecto es mayor cuando comparamos la masa corporal de machos de Gorrión Común y de Águila Calzada que cuando comparamos la masa corporal entre machos y hembras de gorrión.

Para un tamaño muestral dado y un nivel de significación fijado previamente por nosotros la potencia aumenta para grandes tamaños de efecto (gorrión *vs.* águila). Para pequeños tamaños de efecto (gorrión macho *vs*. gorrión hembra), la potencia será más baja porque será más difícil detectar las diferencias entre estos dos últimos grupos (ver Cuadro 1 y Figura 2). El tamaño de efecto no se puede variar ya que es una propiedad intrínseca de las poblaciones que estamos midiendo,

pero sí que se puede cambiar en diseños experimentales, por ejemplo aumentando la intensidad del tratamiento experimental para obtener mayores diferencias. Existen otras formas de incrementar el tamaño de efecto (revisadas en [2]). Por ejemplo, podemos incrementar el tamaño de efecto incluyendo covariables o factores aleatorios que reduzcan la variación dentro de la muestra. El uso de diseños de medidas repetidas en que los mismos individuos son medidos antes y después de un determinado tratamiento experimental también redundará en un aumento del tamaño de efecto.

Nivel de significación (α):

Es el porcentaje de error que asumimos para decir que rechazamos la hipótesis nula, es decir, el error de tipo I. Como comentamos al principio este error lo definimos nosotros en base a un convenio. Para niveles de significación más estrictos (α = 0,001) y con un tamaño muestral fijado, la potencia de la prueba estadística disminuye, ya que es más difícil demostrar que la hipótesis nula es cierta (recordemos que la potencia de una prueba estadística es la probabilidad de aceptar H_1 por lo que al ser más estrictos al poner el límite para rechazar H_0 también necesariamente seremos más estrictos para aceptar la hipótesis alternativa). Un ejemplo de incrementos en el valor de α para aumentar el poder estadístico lo tenemos en el uso de métodos de captura-recaptura para la estimación de parámetros poblacionales. Lebreton et al. ([16]) recomendaron el uso de α = 0,10 como criterio para excluir o incluir variables predictoras en un modelo debido al bajo poder estadístico de este tipo de modelos.

Aplicaciones del análisis de la potencia de una prueba estadística

Hemos comentado anteriormente que la potencia de una prueba estadística es una medida de la fidelidad de la p que hemos obtenido y, por otra parte, que tiene una estrecha relación con el tamaño muestral de manera que mayores tamaños muestrales determinarán una mayor potencia del análisis. Sin embargo, el tamaño de efecto es una propiedad intrínseca de la población de datos que estamos midiendo, y el nivel de significación lo fijamos nosotros previamente al análisis.

La relación directamente proporcional entre tamaño muestral y potencia de test nos permite trabajar en el diseño experimental a dos niveles:

- Antes de realizar un experimento podemos preguntarnos qué tamaño muestral necesitamos para alcanzar una potencia dada (**Análisis a priori**)
- Realizada ya la prueba estadística preguntarnos si el resultado de la p obtenido con el tamaño muestral utilizado tiene una potencia aceptable (> 80%) (**Test *Post-Hoc* o retrospectivo**) o bien, con los datos ya tomados cual es el tamaño de efecto que nuestros datos son capaces de detectar (**Análisis de sensibilidad**).

Análisis a priori

Cuando se diseña un experimento se ha de tener en cuenta el estadístico que se empleará, de manera que tendremos que preguntarnos antes de ir al campo: "¿qué tamaño muestral necesito para demostrar de forma fiable lo que está ocurriendo?".

La potencia del test nos permite contestar a esta pregunta. Por ejemplo, imaginemos que queremos determinar si existen diferencias en masa corporal entre las dos poblaciones de ciervos del ejemplo anterior y que utilizaremos una *t* de Student para testar si las poblaciones son diferentes. Podemos contestar a nuestra pregunta fijando la potencia a un 80% (según el criterio cinco-ochenta), a un nivel de significación determinado (normalmente $\alpha = 0,05$) y lo único que nos faltaría sería calcular el tamaño de efecto.

El cálculo del tamaño de efecto dependerá del análisis estadístico que estemos realizando (para calcular los tamaños muestrales de cada análisis en concreto ver [3,8,9]). En el caso de una prueba de la *t* de Student (de muestras independientes y varianzas iguales) nos basamos en la desviación estándar y la media de las dos poblaciones.

La pregunta es evidente: ¿cómo se averigua la desviación estándar de las poblaciones si todavía no he empezado el experimento?

Existen diferentes fuentes que nos pueden proveer de esta información ([22]):

- **Experimentos Pilotos:** realizar unos ensayos preliminares con tamaños muestrales muy pequeños (N = 3-10 cada grupo), de esta manera tendremos una idea de lo que está ocurriendo antes de hacer el experimento real. Evidentemente es una estimación pobre, pero será suficiente para determinar el tamaño muestral necesario. La realización de estudios pilotos es particularmente recomendable, porque además de permitir

calcular el tamaño muestral, nos permite ver aspectos metodológicos de nuestro experimento que no habíamos tenido en cuenta en el diseño previo y así evitar sorpresas desagradables al analizar los datos ([27]).

- **Bibliografía:** a través de las publicaciones podemos obtener los valores de las medias y con ello estimar la desviación estándar mediante los coeficientes de confianza, ya que éstos se calculan mediante la fórmula $X \pm 2\alpha$. En el caso de regresiones y correlaciones, el tamaño de efecto es el valor de la pendiente (R). Si no estuvieran publicados estos datos en el artículo siempre se puede escribir al autor y solicitarle estos datos. Esta estima es aún más pobre que la de antes pero siempre algo es mejor que nada. Otra posibilidad es prever si el tamaño de efecto será pequeño, mediano o grande y, teniendo en cuenta el tipo de análisis que haremos, tomar como referencia las sugerencias hechas por Cohen ([3]).

El cálculo del tamaño muestral tiene una serie de aplicaciones muy interesantes. A través del cálculo del tamaño muestral a una potencia dada podremos realizar experimentos con una fiabilidad adecuada, lo que aumentará notablemente la calidad de nuestros resultados (ver Cuadro 2).

También podremos predecir si nos interesa escoger o no una variable determinada. Si utilizarla requiere un tamaño muestral excesivo que nosotros no podemos asumir por costos de tiempo o dinero podremos desestimarla y cambiar por otra más barata o que requiera menos tiempo ([1,13,26]).

Otra interesante aplicación del análisis a priori es que al escoger el tamaño muestral necesario evitaremos utilizar más animales o plantas de los necesarios. Esto es particularmente interesante cuando trabajamos con especies delicadas de manipular o con problemas de conservación, o bien cuando el protocolo de experimentación requiere hacer sufrir a los individuos sujetos de estudio o su sacrificio.

El análisis a priori pone de manifiesto que en algunos casos se pueden utilizar tamaños muestrales pequeños para un análisis, lo que abre un camino muy importante para realizar buena estadística con especies amenazadas con las que por desgracia frecuentemente contamos con pocos efectivos. Para ver ejemplos consultar [23,26,29].

Cuadro 2. Un ejemplo práctico de la utilidad del cálculo de la potencia de un análisis.

En un reciente trabajo en *Animal Behaviour*, Ross MacLeod ([17]) presenta un bonito ejemplo de cómo el conocer la potencia de las pruebas estadísticas con que trabajamos puede mejorar los resultados de nuestros trabajos. A continuación se resumen las bases y principales aspectos de este trabajo. Los lectores pueden encontrar en el artículo original los detalles y referencias de este estudio (ver [17]).

La acumulación de reservas energéticas en forma de grasa en las aves reducirá el riesgo de morir de hambre pero de acuerdo con la física Newtoniana reducirá también la capacidad de vuelo. De acuerdo con esta predicción varios estudios han documentado cómo el incremento de la masa corporal durante la migración (que puede representar en muchos Paseriformes un incremento del 27-67% respecto la masa corporal "normal") reduce la velocidad y ángulo de ascenso de aves que escapan frente a un 'predador' humano. La masa corporal de las aves no solo cambia durante la migración, sino que presenta una variación diaria bastante acusada. Las aves acumulan reservas durante el día para sobrevivir durante la noche, con oscilaciones del 5-10% en la masa corporal de los individuos. Sin embargo todos los estudios realizados hasta la fecha han sido incapaces de detectar ningún cambio en la capacidad de vuelo de los individuos en respuesta a las variaciones diarias en masa corporal. Esto ha llevado a pensar que estos resultados podrían deberse a fallos en el diseño experimental, que se podrían dar fenómenos de compensación fisiológica de manera que pequeños cambios en la masa corporal no afectarían la capacidad de vuelo, o que otros factores no controlados podrían estar afectando los resultados.

Para analizar los efectos de la variación diaria en masa corporal sobre la capacidad de vuelo MacLeod ([17]) utiliza un nuevo protocolo experimental que resolvió algunos de los problemas metodológicos de los estudios preliminares. A pesar de esto, el autor sigue sin encontrar una relación significativa entre cambios en masa corporal y capacidad de vuelo. Sin embargo, la mayor novedad de este estudio es que utiliza la física newtoniana para predecir los cambios en capacidad de vuelo asociados a la variación observada en masa corporal. De este análisis se desprende que aunque la capacidad de vuelo de las aves esté cambiando en función de la variación en masa corporal, estos cambios son demasiado pequeños como para ser detectados como estadísticamente significativos con los tamaños muestrales utilizados en este tipo de estudio. Así por ejemplo, dado el aumento en masa corporal de la mañana a la tarde, esperaríamos un incremento del 1,3% en la velocidad de escape (estimado a partir de modelos de física newtoniana), pero el tamaño muestral utilizado en el experimento (14 individuos) solo permitiría detectar cambios superiores al 18% como estadísticamente significativos. Este es un buen ejemplo de cómo el calculo de la potencia de una prueba estadística no es solo necesario para el diseño del trabajo (ya sea observacional o experimental) sino también para la correcta interpretación de los resultados obtenidos.

Test *Post-Hoc* o retrospectivo

Ocurre a veces que ya hemos realizado el análisis sin haber tenido en cuenta el tamaño muestral, y lo que queremos es saber si esta *p* que hemos obtenido es fiable (independientemente de si aceptamos o no H_0).

Mientras que antes queríamos saber el tamaño muestral necesario para tener una potencia determinada (0,80 o superior) ahora lo que hemos hecho es un análisis estadístico concreto (*t* de Student, correlación de Pearson, ANOVA) con un tamaño muestral que hemos utilizado y a un nivel de significación de 0,05 (fijado de antemano por nosotros). En este caso, lo que queremos saber es la potencia de la prueba estadística que realicemos con estos datos, es decir cuan fiable es nuestro análisis.

Calculamos nuestro tamaño de efecto (esta vez con los datos del análisis) y obtenemos la potencia de test. Este tipo de análisis en el que calculamos la potencia de una prueba estadística ya realizada se denominan tests retrospectivos. A pesar de que son muy intuitivos y de uso tentador, han recibido muchas críticas y su uso no esta recomendado. Los motivos principales son que se asume que el tamaño de efecto observado en nuestra población es el tamaño de efecto real, cuando realmente es solo una muestra del tamaño de efecto real (14). Por ello, nosotros estaríamos calculado únicamente la potencia de test observada pero no la real. Por otro lado, la potencia de un análisis que ha dado un resultado significativo siempre será alta, mientras que la de una prueba no significativa será baja debido a la relación negativa entre ambos parámetros (14), reduciendo la utilidad real de este tipo de análisis. Una clara excepción es su uso en meta-análisis que permiten determinar los tamaños de efecto medio detectados en distintos estudios.

Análisis de sensibilidad

Esta es una interesante aplicación del cálculo de poder estadístico. Muchas veces ocurre que no encontramos diferencias significativas y obtenemos una *p* cercana o no a la significación. Podemos preguntarnos entonces, si esa falta de significación es debida a que efectivamente no existen diferencias o a que existe ruido en nuestros datos y no obtenemos la significación porque tenemos un tamaño muestral muy pequeño. Para ello lo único que es necesario es calcular el tamaño de efecto que podríamos detectar dado nuestro tamaño muestral, con un nivel de significación α y un poder estadístico que consideremos adecuados. Esto nos permite

determinar si las diferencias detectables en función de nuestro tamaño muestral son lo suficientemente pequeñas como para descartar que la falta de significación sea debida a problemas con el tamaño muestral.

Existen otras aproximaciones para calcular la potencia de una prueba estadística en análisis post-hoc. Para más detalles de este tipo de análisis se puede consultar [29,14,20,30]. Algunos estudios recientes ([4,19,20]) han sugerido que en lugar de los valores de potencia de una prueba estadística deberían indicarse los tamaños de efectos y sus intervalos de confianza del 95%.

La potencia de una prueba estadística en ecología del comportamiento

Conocer la potencia de las pruebas estadísticas que usamos mejora ostensiblemente la calidad de los trabajos científicos porque nos permite obtener una estima del tamaño muestral necesario y de esta manera realizar un mejor diseño de los experimentos. También es una herramienta que nos orienta sobre la fiabilidad de nuestros resultados. Estos aspectos, son particularmente importantes en estudios comportamentales ya que en ecología evolutiva se suele trabajar con tamaños de efecto muy pequeños y tamaños muestrales escasos. Es más, un estudio reciente ([15]) ha puesto de relevancia que los tamaños muestrales utilizados en ecología evolutiva son realmente insuficientes, principalmente aquellos que se refieren a tamaños de efecto pequeños, por lo que se ha incidido sobre la necesidad de utilizar tamaños muestrales adecuados. De hecho, ya existen algunas revistas del ámbito del comportamiento animal que requieren la potencia de test estadístico en aquellos resultados no significativos o marginalmente significativas (*p. ej. Animal Behaviour* o *Biological Conservation*) y el uso de la potencia de test cada día es más frecuente en algunas disciplinas de la ecología (por ejemplo, [10]). Sin embargo, en algunos casos, la posibilidad de obtener un tamaño muestral adecuado está limitada por la disponibilidad existente en la naturaleza (pongamos el caso de una mutación muy escasa que determina un comportamiento anómalo en un especie determinada). En este caso, se ha sugerido que se debería llegar a un compromiso entre la ratio α/β ([6, 11]). Así, en función de la hipótesis que tengamos, deberíamos asumir un mayor error de tipo I o II. Para ver una discusión del tema y ejemplos consultar [3,5,6,8,9].

Otro aspecto también frecuente en ecología del comportamiento es el uso de análisis no-paramétrico o de análisis multivariante. En el primer caso existe la recomendación general de calcular el equivalente al análisis paramétrico y añadir un 10% del tamaño muestral requerido ([8]). Otras posibilidades más objetivas tanto para análisis no-paramétrico así como para tests más complejos son realizar simulaciones de Montecarlo ([21,27]) o utilizar técnicas de *bootstrapping* ([31]).

Software utilizado para el cálculo de la potencia de una prueba estadística

El cálculo de la potencia de una prueba estadística en general no requiere grandes complicaciones cuando se realiza a mano pero hoy día se cuenta con un vasto conjunto de programas informáticos que nos ahorran mucho trabajo y tiempo.

Thomas y Krebs ([32]) realizaron una amplia e interesante revisión de todos los programas utilizados hasta la fecha. En este estudio valoraban diferentes aspectos de estos como el rango de tests que cubrían, la facilidad de uso o la facilidad de aprendizaje y daban sugerencias para escoger el más adecuado (ver también http://qssi.psu.edu/files/Backgrounder_Power.pdf para una revisión de aplicaciones en SAS y en R).

Estos autores dividieron el software en programas dedicados al cálculo de la potencia de una prueba estadística y el cálculo del tamaño muestral, programas que calculan sólo algunos de estos dos aspectos y paquetes estadísticos que llevan implementada una aplicación para el cálculo estadístico (por ejemplo SPSS 15.0 y STATISTICA 7.0). Algunos programas son gratuitos y pueden ser descargados por Internet o son aplicaciones on-line (GPOWER, PC-SIZE, POWSIM, UnifyPow, statpages.org; http://www.danielsoper.com) mientras que otros son comerciales (p.ej. STUDY-SIZE©, Power and Precision©, nQuery Advisor©). La diferencia radica en que los primeros suelen ser menos potentes y con menos aplicaciones que los comerciales, aunque contienen muchos de los estadísticos que utilizan los ecólogos del comportamiento. Algunos del primer grupo que pueden ser interesantes son el POWSIM, UnifyPow o GPOWER 3. Este último particularmente, en su última versión, además de ser muy intuitivo y con una guía de fácil comprensión, cubre un vasto conjunto de análisis utilizados con frecuencia en ecología ([9], disponible de http://www.psycho.

uni-duesseldorf.de/abteilungen/aap/gpower3/). También, una consulta en Internet nos proveerá de una rápida localización de muchos de los programas revisados por [32] y de otros recientemente publicados.

Podemos afirmar, por tanto, que tenemos al alcance una aplicación a usar relativamente nueva en el campo de la ecología del comportamiento, que nos permite dar mayor calidad a nuestros análisis y al diseño experimental. El software disponible en la actualidad cubre un amplio abanico de análisis frecuentemente utilizados por ecólogos del comportamiento, y que hoy se encuentra al alcance de cualquier científico o estudiante interesado en mejorar la calidad de sus estudios científicos.

Agradecimientos

Los autores agradecen a Joan Carles Senar la sugerencia de escribir este artículo en la revista, a Laura Gangoso y Esther del Val por sus comentarios sobre el manuscrito inicial y a Francisco Valera por su paciencia frente a los repetidos incumplimientos de plazos. Este artículo fue inicialmente escrito mientras JQ disfrutaba de una beca FPI (FP2000-6439) del Ministerio de Ciencia y Tecnología Español.

Referencias

1. Aaron, D.K. & Hays, V.W. 2004. How many pigs? Statistical power considerations in swine nutrition experiments. *Journal of Animal Science*, 82: E245-E254.
2. Bausell, R.B. & Li, Y.-F. 2002. *Power Analysis for Experimental Research*. Cambridge University Press.
3. Cohen, J. 1988. *Statistical Power Analysis for Behavioural Sciences*. Hisdle, New Jersey: Erlbaum.
4. Colegrave, N. & Ruxton, G.D. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behaviorual Ecology*, 14: 446-447
5. Di Stefano, J. 2001. Power analysis and sustainable forest management. *Forest Ecology and Management*, 154: 141-153.
6. Di Stefano, J. 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, 17: 707-709.
7. Dytham, C. 2001. *Choosing and using statistics: A biologist's guide*. Oxford: Blackwell Science Ltd.

8. Erdfelder, E., Faul, F. & Buchner, A. 1996. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28: 1-11.
9. Faul, F., Erdfelder, E., Lang, A.L. & Buchner, A. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39: 175-191.
10. Fidler, F., Burgman, M., Cumming, G., Buttrose, R. & Tomason, N. 2006. Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology*, 20: 1539-1544.
11. Forbes, L.S. 1990. A note on statistical power. *Auk*, 107: 438-439.
12. Gould, S.J. 2007. *La Falsa Medida del Hombre*. Drakontos Bolsillo, Barcelona.
13. Greenwood, J. 1993. Statistical power. *Animal Behaviour*, 46: 1011.
14. Hoenig, J. M. & Heisey, D.M. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statician*, 55: 1-6.
15. Jennions, M.D. & Moller, A.P. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioural Ecology*, 14: 438-445.
16. Lebreton, J.-D., Burnham, K.P., Clobert, J. & Anderson, D.R. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, 62: 67-118.
17. MacLeod, R. 2006. Why does diurnal mass change not appear to affect the flight performance of alarmed birds. *Animal Behaviour*, 71: 523-530.
18. Martín Andrés, A. & Luna del Castillo, J.D. 1995. *50 ± horas de Bioestadística*. Granada, España: Ed. Norma.
19. Martínez-Abraín, A. 2007. Are there any differences? A non-sensical question in ecology. *Acta Oecologica*, 32: 203-206.
20. Nakagawa, S. & Foster, G.F. 2004. The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, 7: 103-108.
21. Peres-Neto, P.R. & Olden, J.D. 2001. Assessing the robustness of randomization tests: examples from behavioural studies. *Animal Behaviour*, 61: 79-86.
22. Piedrafita, J. & Puig, P. 2001. Análisis estadístico, diseño experimental e interpretación de resultados. En: *Ciencia y tecnología en protección y experimentación animal*: 605-628 (J. Zúñiga, J.A. Tur Marí, S. Milocco y R. Piñeiro, Eds.). Madrid: McGraw-Hill.
23. Reed, J.M. & Blaustein, A.R. 1997. Biologically significant population declines and statistical power. *Conservation Biology*, 11: 281-282.
24. Robson, L. S, Shannon, H. S., Goldenhar, L. M. & Hake, A.R. 2001. Statistical Issues: Are the results significant? En: *Guide to evaluating the effectiveness of strategies for preventing work injuries: how to show whether a safety intervention really works*. Cincinatti: NIOSH. Publication nº 2001-119.
25. Sandin, L. & Johnson, R.K. 2000. The statistical power of selected indicators -metrics using macroinvertebrates for assessing acidification and eutrophi-

cation of running waters. *Hydrobiologia*, 422/423: 233-243.
26. Schwagmeyer, P.L. & Mock, D.W. 1997. How to minimize sample sizes while preserving statistical power. *Animal Behaviour*, 54: 470-474.
27. Scott, A.F., Andrew, J.T., Niclas, J., Jonathan, R.R. & Hugh, P.P. 2004. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters,* 7: 669-675.
28. Taylor, L.B. & Gerrodete, T. 1993. The uses of statistical power in conservation biology: the vaquita and northern spotted owl. *Conservation Biology*, 7: 489-500.
29. Thomas, L. 1997. Retrospective power analysis. *Conservation Biology*, 11: 276-280.
30. Thomas, L. & Juanes, F. 1996. The importance of statistical power analysis: An example from Animal Behaviour. *Animal Behaviour*, 52: 856-859.
31. Thomas, L. & Krebs, C.J. 1997. A review of statistical power analysis software. *Bulletin of the Ecological Society of America*, 78: 126-139.
32. Thompson, C.E. & Neill, A.J. 1993. Statistical power and accepting the null hypothesis. *Animal Behaviour*, 46: 1012.