Diseño experimental y consideraciones sobre el tamaño de muestra

Mª Belén Pintado

Unidad de Transgénesis CNB-CBMSO. Consejo Superior de Investigaciones Científicas

Introducción

El diseño de un experimento es el conjunto de todos aquellos pasos previos que deben realizarse para asegurar que los datos que se obtengan finalmente, permitan realizar un análisis objetivo y obtener generalizaciones válidas sobre el problema planteado. Antes de comenzar el experimento se debe especificar claramente el problema, los objetivos y las preguntas que la investigación debe contestar, es decir lo que se denomina en Estadística "contrastes de interés". Mead (1988) considera que el diseño experimental es un trabajo conjunto entre investigador y estadístico y debe desembocar en:

i) Definir claramente los objetivos inherentes al problema experimental que se desea analizar, estableciendo la estructura de tratamientos a realizar para proporcionar respuestas al problema planteado (lo que se denomina fase de concepción);

ii) Determinar los recursos experimentales disponibles (lo que se denomina fase de recursos);

iii) Hacer compatibles ambas acciones, reconociendo y teniendo en cuenta las restricciones prácticas.

Sin embargo, durante la fase de recursos el investigador siempre se plantea las siguientes preguntas: ¿cuántos sujetos son necesarios para poder establecer conclusiones estadísticas de forma concluyente? y ¿qué ocurre si no se puede alcanzar el tamaño necesario?

La respuesta a la primera pregunta son otras muchas preguntas como por ejemplo: ¿se ha realizado una experiencia similar anteriormente?, ¿cuál es el objetivo que se persigue alcanzar?, ¿con que significación estadística se desea concluir?, ¿cuál es el efecto que se espera obtener?. Todas ellas aportan información relevante para tratar de determinar el tamaño de muestra más adecuado para el experimento planteado. Sin embargo, si una vez fijado en función de los criterios establecidos por el investigador, dicho tamaño no se puede alcanzar por falta de recursos o tiempo, no se debe desechar el experimento. En este caso se deberá establecer claramente el tipo de conclusiones a las que se podrá llegar dado el tamaño de muestra que se ha podido alcanzar.

El objetivo de este artículo es presentar de forma lo más practica posible el problema de la determinación del tamaño de muestra en diferentes situaciones que son habituales dentro del diseño de experimentos. La formulación genérica que vamos a considerar viene dada por una variable o característica Y, denominada variable respuesta o de interés, que se encuentra clasificada según una variable o característica T, denominada variable predictora, clasificadora o explicativa, y que representa los diferentes tratamientos o situaciones que se piensa pueden afectar al comportamiento de la respuesta. Más concretamente, nos centraremos en estudiar los problemas asociados con el cálculo del tamaño de la muestra para la comparación de diferentes datos: dos proporciones, dos medias independientes, dos medias emparejadas y varias (k) medias independientes.

Para obtener el tamaño de muestra adecuado en cualquier experimento hay que establecer: el objetivo u objetivos del estudio; el diseño del experimento a realizar; la hipótesis de trabajo, y la determinación de la diferencia de los datos que se considerará significativa en el experimento planteado. A continuación detallamos cada uno de estos aspectos.

Objetivos del estudio

El establecimiento de los objetivos del estudio corresponde al investigador y tienen que ser planteados desde el inicio del experimento. Para ello es necesario hacer una reflexión concienzuda basada en una serie de preguntas cuyas respuestas deben ayudar a diseñar correctamente el experimento. Estas deben ser: ¿qué variable(s) identifica(n) el problema?, ¿cómo ha(n) de cuantificarse la(s) variable(s)?, ¿qué factores influyen sobre la(s) variables(s)?, ¿cuáles de ellos son interesantes para la investigación?, ¿cuáles son importantes pero no interesan y

cuáles no son importantes?, ¿sobre qué población se generalizarán los resultados obtenidos?, ¿cuántas veces y de qué forma se repetirá el experimento?.

La respuesta a todas estas cuestiones permite determinar el problema estadístico al que nos enfrentamos y por lo tanto, plantear el procedimiento de cálculo de tamaño de muestra adecuado para cada situación.

Diseño del experimento

Hay que especificar cómo se va a llevar a cabo el experimento. Eso significa por ejemplo, que si estamos interesados en comparar dos tratamientos deberemos especificar si los tratamientos se van a aplicar a grupos de sujetos diferentes o si se aplicaran secuencialmente sobre la misma población. Una vez se ha especificado el experimento que se desea llevar a cabo, comienza la tarea de la recogida de datos. Aunque esto suele parecer una tarea rutinaria, no es así. Debe tenerse en cuenta que los análisis y las conclusiones se basan en los datos recogidos y por tanto, un proceso de adquisición erróneo puede invalidar el experimento. En este proceso de adquisición de datos hay que considerar fundamentalmente:

- **Calidad de los datos**, que afecta a las conclusiones finales del estudio. Debe quedar muy claro, antes de comenzar el experimento, cuál es el proceso de toma y almacenamiento de datos, para evitar errores durante el proceso experimental. Un aspecto muy importante a tener en cuenta es el error cometido en las mediciones experimentales. La aparición de datos anómalos (*outliers*) debe tratarse como un problema importante y el investigador debe preguntarse: ¿estos datos pueden ser posibles o se deben a una mala recogida de la información, transcripción o a un error experimental de medición?, ¿puede ocurrir que se haya diseñado de forma incorrecta el experimento para que aparezcan estos datos?

- **Estructura de los datos**, que identifica las unidades experimentales, los grupos o bloques de tratamiento. Una documentación exhaustiva del método experimental se considera crucial para el análisis de los datos. La dependencia o independencia de los datos recogidos resulta fundamental, ya que determina de forma unívoca el tipo de análisis estadístico que puede utilizarse.

- **Mecanización de los datos**, que reduce la manipulación de los mismos en el proceso de recogida de información. Este proceso debe ser lo más sencillo posible para evitar el trasiego de documentación y ficheros que puede generar grandes pérdidas de datos y, en consecuencia, deficiencias severas en las conclusiones del estudio.

- **Análisis de datos**, que viene marcado por los objetivos, los tipos de datos, y la recogida de información realizada en el proceso experimental.

Hipótesis de trabajo y procedimiento para el cálculo de tamaño de muestra

Asociado con el objetivo y el diseño experimental establecido se encuentra la hipótesis de partida y la determinación del tamaño de la muestra. En la práctica, el tamaño de la muestra puede obtenerse a través de un análisis de precisión o análisis de potencia. Para realizar cualquiera de los dos es necesario controlar los denominados errores de tipo I y tipo II que vienen implícitos en cualquier análisis estadístico interesado en el estudio de la efectividad de un tratamiento o en la comparación de la efectividad de dos o más tratamientos. Como ejemplo supongamos que tenemos dos grupos de individuos y que a uno de ellos le suministramos un nuevo tratamiento y al otro le damos un placebo. Si no poseemos ningún tipo de información sobre el nuevo tratamiento, el contraste de interés en esta situación vendría dado por:

$$\left.\begin{array}{l} H_0: \text{El tratamiento no es efectivo} \\ H_A: \text{El tratamiento es efectivo} \end{array}\right\}, (6)$$

donde en la hipótesis nula (H_0) concluimos que la efectividad del nuevo tratamiento es igual al efecto producido por el placebo, mientras que en la hipótesis alternativa (H_A) concluimos que el tratamiento tiene una efectividad distinta a la producida por el placebo.

Para resolver de forma correcta cualquier contraste estadístico se deben tener en cuenta lo que se conoce como errores de tipo I y tipo II, cuyas probabilidades viene dadas por:

$$\alpha = P(\text{error tipo I}) = P(\text{rechazar } H_0 \text{ cuando es cierta}) \quad (7)$$
$$\beta = P(\text{error tipo II}) = P(\text{no rechazar } H_0 \text{ cuando es falsa}) \quad (8)$$

La probabilidad de error de tipo I (también conocido como error tipo α o falso positivo) es el que utilizan normalmente los investigadores para resolver cualquier tipo de contraste. Cuando el valor p del contraste planteado para los datos experimentales

es inferior a dicha significación se concluye que se puede rechazar la hipótesis nula, y por tanto, podemos concluir que el tratamiento tiene una efectividad distinta a la que proporciona el placebo. Por el contrario, si el valor *p* es superior a la significación, se concluye que la efectividad del tratamiento es similar a la que produce el placebo, ya que no es posible rechazar la hipótesis nula planteada.

Sin embargo, a la hora de establecer esta conclusión muchos investigadores se olvidan de la probabilidad de error de tipo II (también conocido como error tipo β o falso negativo), que es tan o más importante que la de tipo I. Con el error de tipo I rechazaríamos la hipótesis de que el tratamiento no es efectivo cuando realmente no lo es, mientras que con el error de tipo II, no rechazaríamos la hipótesis de que el tratamiento no es efectivo cuando realmente sí lo es. Para controlar este error se define la potencia del contraste como:

$$\text{Potencia} = 1-\beta = P(\text{rechazar } H_0 \text{ cuando es falsa}) \quad (9)$$

Por tanto, cuando estamos resolviendo cualquier contraste no podemos fijarnos únicamente en el valor *p* obtenido, ya que es necesario además reportar la potencia del contraste realizado. Potencias muy pequeñas con valores *p* significativos indican que a pesar de concluir que podemos rechazar la hipótesis nula tendríamos una probabilidad muy baja de rechazar dicha hipótesis cuando en realidad es falsa, con lo que quedarían invalidadas nuestras conclusiones experimentales. A la hora de establecer el tamaño de muestra se suele llegar a un compromiso entre ambos errores, ya que no se puede reducir uno sin alterar el otro. De forma habitual se suelen tomar:

$$\alpha=0.05,\quad \beta=0.2,\quad 1-\beta=0.8 \quad (10)$$

que establece que la probabilidad de rechazar H_0 cuando es cierta es del 5%, mientras que rechazarla cuando es falsa se sitúa en el 80%. Valores más exigentes provocarán tamaños de muestra más exigentes, y por tanto, se debe proceder con cuidado a la hora de cambiarlos. Además, estos valores deben fijarse antes de comenzar el experimento y no pueden ser alterados a lo largo del mismo para conseguir resultados significativos. El efecto directo de introducir dichos cambios es que el tamaño muestral elegido al comienzo del experimento deja de tener sentido y quedarían invalidados todos los resultados experimentales obtenidos.

El siguiente paso es determinar la prueba de contraste a utilizar para las hipótesis de partida planteadas. Por ejemplo, en el caso de la comparación de dos tratamientos (T1 y T2) en el que se desea comparar la efectividad media de cada uno de ellos, las hipótesis de partida podrían ser:

- **Prueba de igualdad:** Este es la prueba habitual de comparación de dos medias que realizan la mayoría de paquetes estadísticos. El objetivo en este caso es saber si podemos considerar que la efectividad media de ambos tratamientos puede considerarse distinta.

$$\left.\begin{aligned} H_0 &: \mu_{T1} = \mu_{T2} \\ H_A &: \mu_{T1} \neq \mu_{T2} \end{aligned}\right\} (1)$$

- **Prueba de no inferioridad:** Esta prueba debe utilizarse cuando se quiere concluir que los dos tratamientos tienen el mismo efecto cuando la diferencia de medias entre ellos es inferior a δ, que se conoce con el nombre de diferencia clínica relevante. Se debe utilizar cuando se sospecha que un tratamiento mejorará el comportamiento medio respecto del otro tratamiento en una cantidad δ. La dificultad para el investigador estriba en fijar dicho valor.

$$\left.\begin{aligned} H_0 &: \mu_{T1} - \mu_{T2} \geq \delta \\ H_A &: \mu_{T1} - \mu_{T2} < \delta \end{aligned}\right\} (2)$$

- **Prueba de superioridad:** Funciona de forma similar a la prueba anterior pero en este caso se quiere concluir que los dos tratamientos tienen el mismo efecto cuando la diferencia de medias entre ellos es superior a δ.

$$\left.\begin{aligned} H_0 &: \mu_{T1} - \mu_{T2} \leq \delta \\ H_A &: \mu_{T1} - \mu_{T2} > \delta \end{aligned}\right\} (3)$$

- **Prueba de equivalencia:** Esta prueba debe utilizarse cuando se quiere concluir que los dos tratamientos tienen el mismo efecto cuando la diferencia de medias entre ellos es en valor absoluto inferior a δ. Lo que se evalúa en realidad es si los tratamientos pueden considerarse equivalentes, es decir, que tienen un efecto similar.

$$\left.\begin{aligned} H_0 &: |\mu_{T1} - \mu_{T2}| \geq \delta \\ H_A &: |\mu_{T1} - \mu_{T2}| < \delta \end{aligned}\right\} (4)$$

Diferencia Clínica significativa

La determinación de la diferencia clínica significativa para las pruebas de no inferioridad, superioridad o equivalencia resulta un aspecto crucial para el establecimiento del tamaño de la muestra. Si se dispone de información previa sobre otra experiencia o tratamientos similares puede establecerse a partir de dicha información. Si no es así, deberá ser el investigador el que establezca el valor o valores que considere adecuados para el experimento a realizar. Hay que tener en cuenta que este valor condicionará de forma notable el cálculo del tamaño de la muestra final.

Para simplificar los cálculos del tamaño de la muestra se suele sustituir la diferencia clínica significativa por el tamaño del efecto o "*effect size*" (Cohen, 1988) que tiene en cuenta a la vez tanto la diferencia clínica significativa como la variabilidad asociada con dicha diferencia. En el caso de la comparación de dos medias el tamaño del efecto viene dado por:

$$\delta = \frac{\mu_{T1} - \mu_{T2}}{\sigma}, \quad (5)$$

siendo σ la varianza común para ambas poblaciones, cuya formulación depende de si asumimos que ambas poblaciones tienen varianzas iguales o distintas. Cohen (1998) sugiere valores que representan tamaños de efecto pequeños, medios y grandes. Un tamaño de efecto pequeño representa que las medias de ambas poblaciones son similares pero distintas, es decir, se espera que un tratamiento mejore el efecto del otro pero de forma muy reducida. Un tamaño de efecto grande representa que las medias de ambas poblaciones son muy diferentes, lo que implica que un tratamiento mejora de forma extensible respecto del otro tratamiento. En la sección de ejemplos mostraremos el comportamiento del tamaño del efecto en el cálculo del tamaño de muestra.

En la sección siguiente se presentan las situaciones prácticas más habituales y se muestra cómo proceder en cada una de ellas para el cálculo del tamaño de la muestra. Para ejemplificar los resultados se utiliza una aplicación creada a tal efecto mediante el programa R (R Core Team, 2014) y sus librerías shiny (RStudio and Inc., 2014) y pwr (Stephane Champely, 2012). Dicha aplicación estará disponible en la web del Órgano Evaluador de Proyectos de la Universidad Miguel Hernández (http://oep.umh.es) en un futuro próximo.

Para un desarrollo más completo del cálculo del tamaño de la muestra en ensayos clínicos se puede consultar Chow, Shao, y Wang (2003).

Casos Prácticos

En este apartado se presentan diferentes experimentos en los que se muestra el cálculo del tamaño de la muestra asociado. Más concretamente, trataremos las situaciones en las que interesa comparar dos medias, dos proporciones, o k medias poblaciones (prueba ANOVA). Para cada experimento plantearemos el objetivo, la hipótesis de trabajo, la formulación del tamaño del efecto específico para esa situación, y los diferentes valores considerados de tamaño de efecto, significación y potencia necesarios para el cálculo del tamaño de la muestra.

Tamaño de muestra para la comparación de dos medias

El problema de comparación de dos medias se plantea cuando el objetivo que se persigue es el estudio de una variable de tipo continuo (respuesta) frente a una variable de tipo categórico que identifica dos tipos de tratamientos (predictora). Más concretamente, lo que se pretende investigar es el comportamiento de la media de la variable respuesta para ambos tratamientos. Los contrastes de interés que se pueden plantear son los presentados en la sección "Hipótesis de trabajo". En este caso consideramos la prueba de igualdad dada en (1) y el tamaño de efecto dado en (5). En la Figura 1 se representa el cálculo del tamaño de la muestra para la prueba de igualdad de medias en diferentes situaciones en función del tamaño del efecto considerado. Consideramos valores de potencia igual a 0.7, 0.8, y 0.9, y valores del nivel de significación iguales a 0.01, 0.05, y 0.1. En línea discontinua naranja se presentan los resultados para tamaños de efecto pequeño (0.2), mediano (0.5) y grande (0.8) dados por Cohen (1988). Hay que destacar que los valores de tamaño de la muestra más pequeños resultan de las combinaciones menos exigentes, es decir, con valores de significación más grandes y potencias más pequeñas. Este comportamiento es natural debido a que estamos siendo menos exigentes con las probabilidades de error de tipo I y II. Por el contrario, los valores de tamaño de muestra más grandes se obtienen para la potencia más alta y significación más pequeña, es decir, cuando queremos que dichas probabilidades de error sean lo más pequeñas posible. Para valores estándar de potencia (0.8) y nivel de significación (0.05), el tamaño de la muestra para un tamaño de efecto grande es de 26 sujetos. Habitualmente, en los ensayos clínicos se suele considerar un porcentaje de descarte que permite tener sujetos de reserva en caso de que alguno de los seleccionados para la muestra desaparezca del estudio. Considerando un porcentaje de descarte del 10%, el tamaño de muestra para cada grupo bajo estudio sería de 29 sujetos.

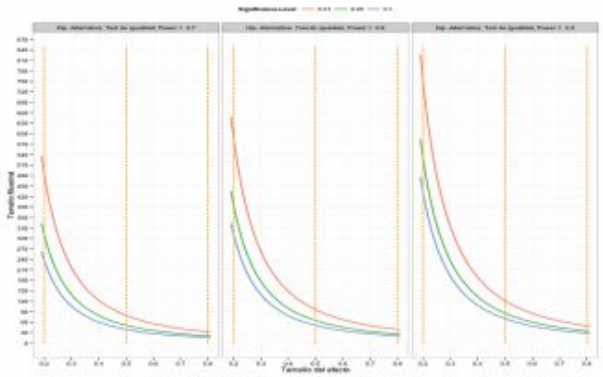

Figura 1.- Tamaño de la muestra para diferentes significaciones, potencias y tamaño de efecto en la comparación de dos medias mediante la prueba de igualdad.

Tamaño de muestra para la comparación de dos proporciones

Imaginemos un experimento en el que nos interesa registrar el porcentaje de sujetos que evolucionan de forma positiva frente a la aplicación de un tratamiento. Si en lugar de considerar un único grupo, tomamos dos grupos de sujetos de forma que en cada grupo aplicamos un tratamiento distinto, nos encontramos ante el problema de comparación de proporciones en dos poblaciones independientes. En este caso consideramos la prueba de igualdad

$$\left.\begin{array}{l} H_0 : \theta_{T1} = \theta_{T2} \\ H_A : \theta_{T1} \neq \theta_{T2} \end{array}\right\} \quad (11)$$

siendo θ_{T1} y θ_{T2} la proporción de sujetos que evolucionan de forma positiva en cada población respectivamente. En este caso el tamaño de efecto viene dado por la expresión

$$h = 2\arcsin(\theta_{T1}) - 2\arcsin(\theta_{T2}). \quad (12)$$

En la Figura 2 se representa el cálculo del tamaño de muestra para la prueba de igualdad de proporciones en diferentes situaciones en función del tamaño del efecto considerado. Consideramos valores de potencia igual a 0.7, 0.8, y 0.9, y valores del nivel de significación iguales a 0.01, 0.05, y 0.1. En línea discontinua naranja se presentan los resultados para tamaños de efecto pequeño (0.2), mediano (0.5) y grande (0.8) dados por Cohen (1988). Se puede observar que el comportamiento del tamaño de muestra es muy similar al observado para la prueba de dos medias. Para valores estándar de potencia (0.8) y nivel de significación (0.05), el tamaño de muestra para un tamaño de efecto grande es de 25 sujetos. Con un porcentaje de descarte del 10%, el tamaño muestral final para cada grupo sería de 28 sujetos.

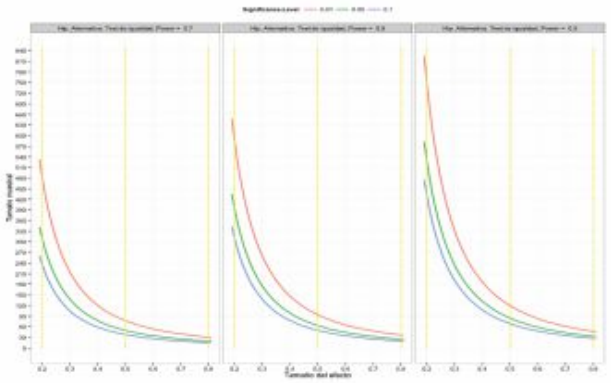

Figura 2.- Tamaño de muestra para diferentes significaciones, potencias y tamaño de efecto en la comparación de dos proporciones mediante la prueba de igualdad.

Tamaño de muestra para la comparación de k medias (ANOVA)

El problema de comparación de medias de más de dos grupos, o comúnmente llamada prueba ANOVA, es una generalización del problema de comparación de dos medias. En este caso el experimento considera más de dos grupos, y estamos interesados en comparar las medias de la variable de interés para diferentes grupos. El contraste de interés especifica que la hipótesis nula y la de igualdad de todas las medias, mientras que la hipótesis alternativa establece que hay al menos dos grupos cuyas medias pueden considerarse distintas. La expresión del tamaño del efecto es mucho más compleja y viene dada en Cohen (1988). En la Figura 3 se representa el cálculo del tamaño de muestra para la prueba ANOVA en diferentes situaciones en función del tamaño del efecto considerado. Consideramos únicamente el tamaño de potencia 0.8, valores del nivel de significación iguales a 0.01, 0.05, y 0.1, y número de grupos igual a 3, 4, 5 y 6. En línea discontinua naranja se presentan los resultados para tamaños de efecto pequeño (0.1), mediano (0.25) y grande (0.4) dados por Cohen (1988). El comportamiento del nivel de significación es similar a los anteriores casos estudiados, mientras que se puede ver que conforme se aumenta el número de grupos, el tamaño de muestra requerida es inferior. Para valores estándar de potencia (0.8), nivel de significación (0.05) y 5 grupos, el tamaño de la muestra para un tamaño de efecto grande es de 19 sujetos por grupo. Con un porcentaje de descarte del 10% el tamaño muestral final para cada grupo sería de 21 sujetos.

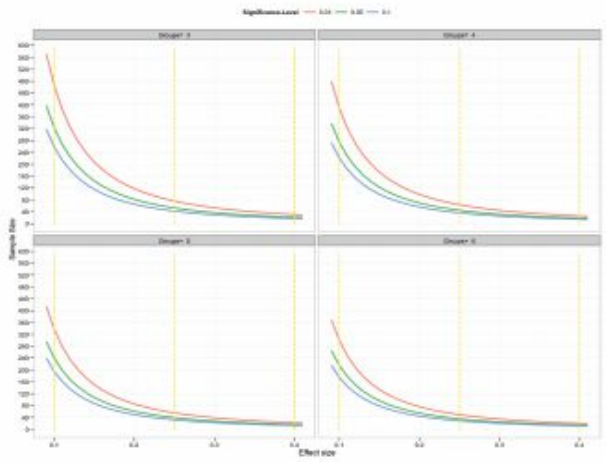

Figura 3.- Tamaño de muestra para diferentes significaciones, potencias y tamaño de efecto para la comparación de k medias (k=3, 4, 5 y 6) con potencia igual a 0.8.

Conclusiones

En este artículo se han pretendido mostrar las ideas principales involucradas en el cálculo del tamaño de la muestra en el diseño de experimentos. Se han presentado tres ejemplos típicos experimentales pero se recomienda la lectura de Cohen (1988) y Chow, Shao, y Wang (2003) para un estudio mucho más detallado del cálculo del tamaño de la muestra en muchas otras situaciones experimentales.

BIBLIOGRAFÍA

1. Chow S-C., Shao J., and Wang H. *Sample size calculations in clinical research*. CRC Press 2003.
2. Cohen J. *Statistical power analysis for the behavioral Sciences* 2nd ed. Lawrence Erlbaum Associates 1988.
3. Mead R. *The design of experiments*. Cambridge, New York: Cambridge University Press 1988.
4. R Core Team. R: A *language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.
5. Rstudio and Inc. *Shiny: Web Application Framework*. 2014, http://www.rstudio.com/shiny/.
6. Stephane Champely. *pwr: Basic functions for power analysis*. 2012, http://CRAN.R-project.org/package=pwr.

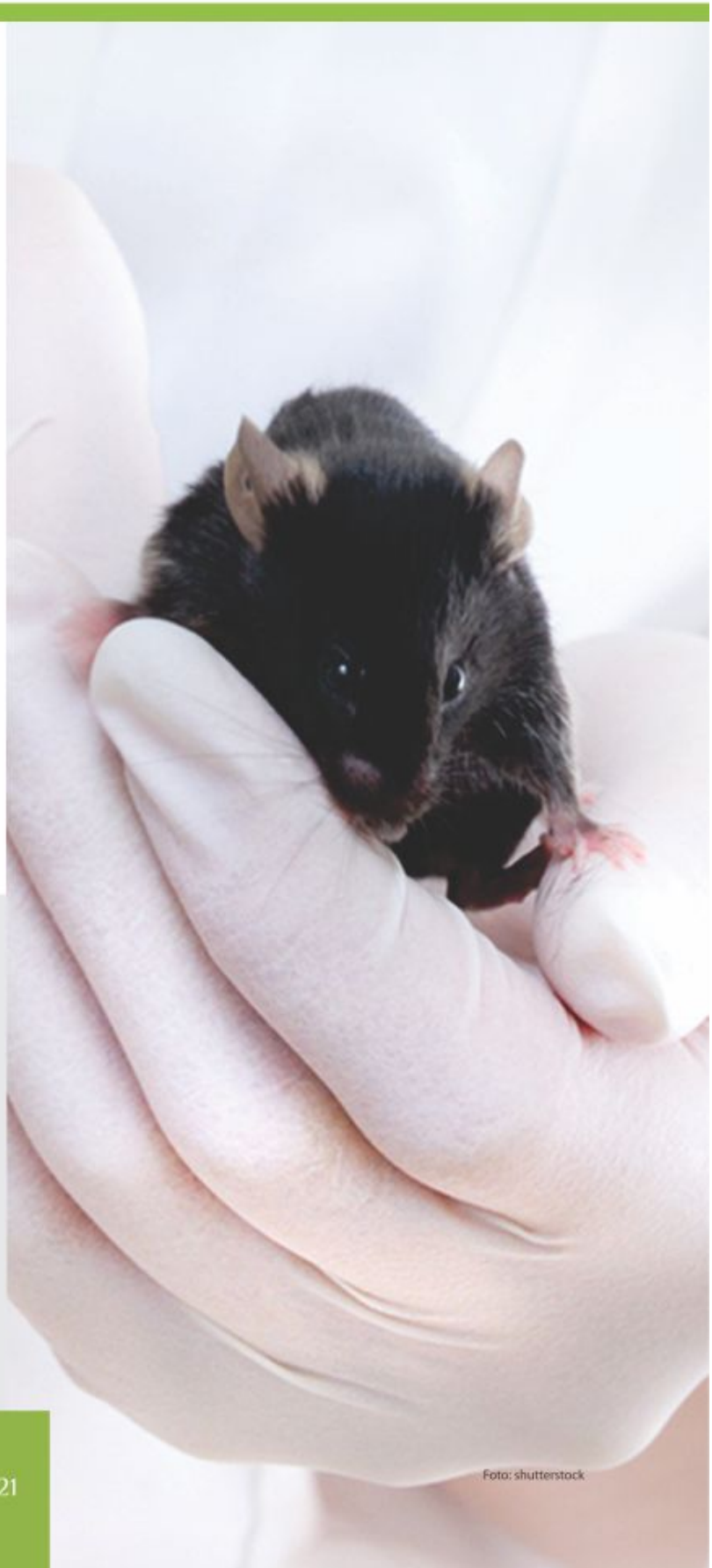

Foto: shutterstock