J. de Mata Donado Campos,
G. Orejas Rodríguez-Arango[1]

*An Esp Pediatr 1999;50:81-90.*

# Epidemiología y metodología científica aplicada a la Pediatría (III): Introducción al análisis estadístico en epidemiología

## Introducción

En investigación en Ciencias de la Salud, cuando se describe la existencia de una relación entre una variable exposición (variable independiente) y una variable respuesta (variable dependiente), es preciso analizar si esa relación existe realmente o está motivada por el azar, por sesgos diversos o por un efecto de confusión. Este artículo se centra sobre el papel de la suerte en este contexto.

La influencia del azar en la investigación epidemiológica es relevante por dos razones fundamentales. En primer lugar, prácticamente siempre se trabaja con muestras de población, no con poblaciones completas. Las muestras varían, son diferentes, cada vez que se seleccionan por el "error aleatorio del muestreo". La inferencia estadística se encarga de analizar cómo los resultados observados en esas muestras se acercan a los verdaderos o, lo que es lo mismo, cuan próximos serían esos resultados respecto a los obtenidos con la población de origen. Obviamente, cuanto mayores sean las muestras, más se aproximarán los resultados a los verdaderos. En segundo lugar, cuando se repite un experimento varias veces, el azar puede modificar los resultados. Es fácil de comprender. Si tiramos al aire una moneda perfectamente equilibrada un millón de veces, el 50% de las veces obtendremos cara y el 50% cruz. Si arrojamos la misma moneda 10 veces y repetimos varias series de la misma forma, el porcentaje de caras o cruces oscilará para cada serie dentro de unos límites amplios sólo por causa del azar. En investigación biomédica sucede exactamente lo mismo. Si repetimos un estudio varias veces los resultados diferirán, en mayor o menor cuantía, de una vez a otra, de forma aleatoria y dependiendo del tamaño de la muestra.

Por otra parte, la metodología científica en general, y la epidemiología como disciplina estrechamente relacionada con la elección de opciones de salud pública en particular, exigen continuamente al investigador la toma de decisiones basadas en criterios claramente explícitos.

En este tercer capítulo de la serie "Epidemiología y metodología científica aplicada a la Pediatría" se introducen una serie de conceptos estadísticos básicos para el estudio epidemiológico dirigidos a afrontar esta problemática. Inicialmente, se abordan las nociones de contraste de hipótesis y pruebas de significación, haciendo énfasis en el concepto y los determinantes del valor de la conocida "*P*". Posteriormente, se define el concepto de "intervalo de confianza", se expone cómo calcular los intervalos de confianza más utilizados en la práctica epidemiológica y se analiza la relación entre el intervalo de confianza y el valor "*P*". Por último, se discuten los conceptos de potencia estadística y tamaño de la muestra.

Departamento de Epidemiología y Bioestadística. Escuela Nacional de Sanidad - Instituto de Salud Carlos III. Madrid. [1]Clinamat-Medycsa. Madrid.
*Correspondencia:* Gonzalo Orejas. Camino de los Tilos 127. 33429 La Fresneda-Siero (Asturias).

## Contraste de hipótesis y pruebas de significación

El contraste de hipótesis trata de discernir, mediante la aplicación de una prueba estadística, si la asociación encontrada en una investigación es debida a que en realidad existe esa asociación o más bien a que la asociación ha sido ocasionada por el azar. Gráficamente, si se estudia, por ejemplo, el efecto hipotensor de un fármaco A en niños hipertensos y se encuentra que con su uso se reduce en promedio la tensión arterial diastólica (Tad) 20 mmHg respecto al no tratamiento o al tratamiento con placebo, tendríamos que decidir si existe una verdadera asociación entre el fármaco A y el efecto hipotensor observado o si este efecto no fue causado por el fármaco, sino que fue debido al azar. En realidad, debemos elegir entre la hipótesis nula ($H_0$), que dice que el efecto hipotensor del fármaco A es igual a 0 ($H_0$: Tad tras fármaco A – Tad sin tratamiento = 0), o la hipótesis alternativa ($H_1$), que implica que el fármaco A modifica la tensión arterial diastólica de los niños hipertensos ($H_1$: Tad tras fármaco A – Tad sin tratamiento ≠ 0) (como se mencionó en el primer artículo de esta serie[(1)], si se utilizan medidas de asociación basadas en cocientes, como, por ejemplo, el riesgo relativo, la $H_0$ sería igual a 1 al ser numerador y denominador iguales; la $H_1$, al contrario, diferiría de 1). En consecuencia, la hipótesis nula ($H_0$) representa la alternativa de que no hay relación entre la exposición y la respuesta o que los grupos que se comparan son semejantes, mientras que la hipótesis alternativa ($H_1$) presenta la eventualidad de que sí existe una verdadera asociación o que los grupos difieren realmente en el factor a estudio.

La probabilidad de obtener un efecto determinado en una investigación asumiendo que la hipótesis nula es cierta es lo que valora la prueba de significación y se cuantifica en estadística como valor de "*P*". Es decir, el test de significación estima la probabilidad de obtener un determinado resultado por la influencia del azar al admitir como cierta la hipótesis nula. Así, si en el ejemplo anterior hemos detectado una reducción media de la tensión arterial diastólica de 20 mmHg con el fármaco A, siendo el valor de

*P* igual a 0.20, significa que, suponiendo que el fármaco A no tiene efecto sobre la tensión arterial, la probabilidad de obtener ese resultado simplemente por azar es del 20%. Por el contrario, si el valor de *P* fuera 0.01, diríamos que, suponiendo que el fármaco A no tiene efecto sobre la tensión arterial, la probabilidad de obtener ese resultado por azar es del 1%. En el primer caso, ya que la probabilidad del 20% es relativamente alta, no podemos descartar que la hipótesis nula sea falsa y la damos por cierta. Sin embargo, en el segundo, puesto que la probabilidad del 1% es muy baja, podemos rechazar la hipótesis nula como falsa y admitir que el fármaco A tiene un efecto hipotensor "estadísticamente significativo". Conviene subrayar que un resultado estadísticamente no significativo señala que los datos de la investigación son compatibles con la hipótesis nula, pero no que ésta sea cierta. Por otro lado, un resultado estadísticamente significativo indica que la población origen de la muestra presenta una variación respecto a la hipótesis nula[2]. En cualquier caso, es preciso recordar que en biología la suerte se refiere a las múltiples e impredecibles fuentes de variación que afectan una respuesta concreta en una investigación determinada. Es la llamada "variabilidad biológica", que en teoría podría justificar cualquier tipo de resultado[3]. Ello implica que, antes de la consideración de la prueba de significación en sí misma, procede el análisis riguroso de los resultados a la luz de los conocimientos disponibles.

El valor de *P* es un valor continuo que oscila entre 0 y 1. En general, se admite arbitrariamente que el valor de *P* es estadísticamente significativo cuando es menor o igual a 0.05, es decir, cuando la probabilidad de que ocurra ese suceso si la hipótesis nula fuese cierta sea igual o inferior al 5%. En este caso, por lo tanto, se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa sin gran riesgo de error. Sin embargo, deben ser los responsables de cada investigación los que fijen "a priori" cual es el nivel de significación exigido según las consecuencias de los resultados de su estudio concreto. Por ejemplo, en caso de que se esté evaluando un fármaco barato, sin ningún efecto secundario, podría ser conveniente llevar el umbral de la significación estadística a la probabilidad del 10% (*P* = 0.1). A la inversa, si se tratase de un fármaco con importantes efectos secundarios, con el que nos interesa ser altamente selectivos, seríamos más exigentes en el nivel de significación, pudiendo reducirlo hasta una probabilidad del 1% (*P* = 0.01). No obstante, casi siempre se adopta como nivel de significación el límite arbitrario ya citado del 5% (*P* = 0.05), aceptando que un suceso que ocurre cada 20 veces es demasiado infrecuente como para que sea causado solamente por el azar. De hecho, sólo una de cada 20 veces será ocasionado por la suerte.

Las hipótesis que se contrastan pueden ser unilaterales o bilaterales. La hipótesis unilateral (*P* con una cola) se refiere a que la asociación entre variables tiene un sentido positivo o negativo específico. En cambio, la bilateral (*P* con dos colas) admite la posibilidad de que la asociación pueda existir en ambos sentidos, positivo y negativo. En el ejemplo anterior la hipótesis unilateral sólo consideraría la posibilidad de que la acción del fármaco A disminuyese la tensión arterial diastólica ($H_0$: Tad tras fármaco A = Tad sin fármaco A; $H_1$: Tad tras fármaco A < Tad sin fármaco A). La hipótesis bilateral contemplaría ambas situaciones, que el fármaco A pudiese reducir o incrementar la tensión arterial diastólica ($H_0$: Tad tras fármaco A = Tad sin fármaco A; $H_1$: Tad tras fármaco A > o < Tad sin fármaco A). En Ciencias de la Salud, aunque son frecuentes las situaciones en que sólo es probable el efecto en un único sentido, lo habitual es adoptar la posición más conservadora de evaluar la hipótesis alternativa bilateralmente. Sería extraño comprobar que el fármaco hipotensor A aumenta la tensión arterial diastólica en vez de reducirla. No obstante, normalmente se considera la hipótesis alternativa bilateral y se calcula el valor de *P* con dos colas, correspondiente en nuestro ejemplo a la hipótesis de que el fármaco A pueda aumentar o reducir la tensión arterial diastólica de los sujetos estudiados. Lógicamente, al ampliar el rango de posibilidades, el valor de *P* será mayor cuando se considera la hipótesis bilateral que cuando se evalúa la unilateral o, lo que es lo mismo, será más difícil obtener un resultado estadísticamente significativo cuando se estima una hipótesis bilateral que cuando sólo se considera la hipótesis unilateral respecto al mismo caso. En general, las hipótesis unilaterales sólo se aplican cuando existe una hipótesis bien fundamentada "a priori" y el objetivo del estudio es incrementar la precisión de una estimación en la que la dirección del efecto es conocida o cuando el estudio está diseñado específicamente para rebatir un hallazgo previo[4]. Al igual que sucedía con el nivel de significación, los investigadores deben decidir al planificar el estudio si les interesa efectuar pruebas de significación bilaterales o unilaterales y la determinación adoptada debe quedar bien especificada en la metodología del trabajo.

Cuando hablamos de pruebas de significación sólo estamos considerando probabilidades y nunca certezas. Por ello, es importante tener en cuenta las dos posibilidades de error que siempre hay que contemplar en estas situaciones. Por un lado, cuando aceptamos la hipótesis alternativa, porque el valor de la *P* es inferior a 0.05, todavía existe una probabilidad, ciertamente muy pequeña, que coincide con el valor de la *P*, de que el resultado obtenido haya sido debido al azar y no porque exista una verdadera asociación entre las variables estudiadas. Sería equivalente al resultado falso positivo de una prueba diagnóstica. Es lo que se conoce como error tipo I. La probabilidad de cometer un error tipo I se denomina probabilidad $\alpha$ (Fig. 1). En nuestro ejemplo en el que el fármaco A reducía la tensión arterial diastólica 20 mm Hg, si la *P* obtenida era 0.20, la probabilidad $\alpha$ de cometer un error tipo I sería del 20%. Si la *P* fuese de 0.01, la probabilidad $\alpha$ de cometer un error tipo I sería del 1%.

Al mismo tiempo, cuando se obtiene una *P* no significativa estadísticamente, es decir, cuando no se puede rechazar la hipótesis nula $H_0$, es posible cometer el error llamado tipo II. Este error consiste en que, aunque la *P* sea mayor de 0.05 y no se pueda rechazar la hipótesis nula, siempre existe una probabilidad, denominada $\beta$, por la que la hipótesis alternativa $H_1$ es la realmente cierta (Fig. 1). Sería, pues, equivalente a un resultado falso negativo en una prueba diagnóstica, en el que existe un efecto importante, pero que no es detectado. La probabilidad $\beta$ habitualmente es desconocida por el investigador. Se considera, en ge-

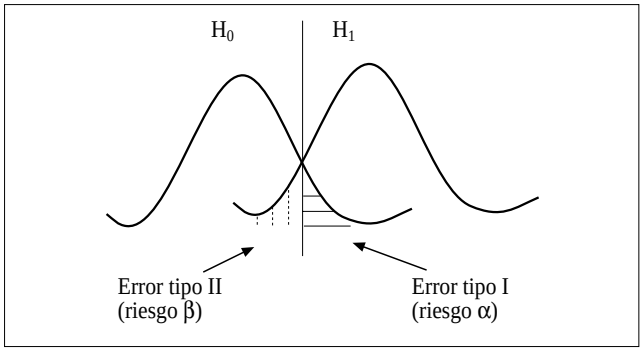



**Figura 1.** Representación gráfica de las probabilidades α y β.

neral, aceptable cuando es inferior al 20%. Los investigadores deberían preguntarse por esta probabilidad β siempre que obtengan un resultado no significativo. Un efecto poco acusado, datos de la investigación no suficientemente informativos y el error aleatorio del muestreo, contribuyen conjuntamente a incrementar la probabilidad β de cometer un error tipo II[5]. Al final de artículo se verá como la probabilidad β está íntimamente relacionada con el concepto de potencia estadística y debe ser conocida para el calcular el tamaño necesario de la muestra. En la tabla I se resume el significado de los errores tipo I y II.

Las pruebas estadísticas (Tabla II) son aplicadas a partir de los datos de la investigación para obtener un "estadístico", valor que se contrasta con el de una población estándar, dándonos a conocer la probabilidad de que el valor observado esté justificado sólo por la suerte, siempre presuponiendo que la hipótesis nula sea la correcta. El investigador concluirá rechazando (prueba significativa) o aceptando (prueba no significativa) la hipótesis nula. Aunque las pruebas son específicas dependiendo de la propia hipótesis evaluada y de la distribución de los datos analizados, todas comparten básicamente la misma estructura en cuanto que son función de un numerador, la diferencia entre los valores observados en el estudio y los que hubieran aparecido de ser la hipótesis nula cierta, y de un denominador, la variabilidad de la muestra. La prueba

Tabla I Significado del error tipo I y tipo II

| *Prueba* | ***$H_0$ verdadera*** | ***$H_0$ falsa*** |
|---|---|---|
| **No significativa** (no se rechaza $H_0$) | Correcto | Error tipo I →Riesgo β |
| **Significativa** (se rechaza $H_0$) | Error tipo I →Riesgo α | Correcto |

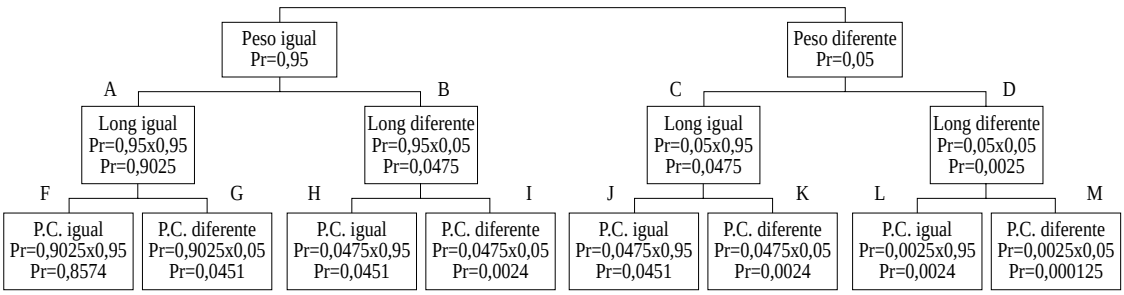
será significativa o el valor de *P* más bajo cuanto mayor sea la diferencia entre los valores observados y teóricos, y cuanto menor sea la variabilidad de la muestra o lo que es lo mismo, cuanto mayor sea el tamaño de ésta. Los libros de estadística explican detalladamente el desarrollo e interpretación de todas estas pruebas[6,7].

Son dos las circunstancias que condicionan de forma importante el valor de la *P* en una prueba estadística. En primer lugar, como es lógico, la magnitud del efecto que se está midiendo. Si la diferencia de tensión arterial que pretendemos detectar es sólo de 5 mmHg será mucho más difícil que la *P* sea significativa que si la diferencia fuese de 40 mmHg. Por tanto, cuanto mayor sea la magnitud del efecto de la intervención, más fácil será obtener un resultado estadísticamente significativo. En segundo lugar, influye de forma importante el número de sujetos que intervienen en la observación. Si sólo participan 5 niños en la investigación sobre el fármaco hipotensor, incluso aunque la diferencia del efecto sea grande, será difícil obtener una *P* inferior a 0.05. Por el contrario, si el número de niños es muy elevado, incluso aunque el efecto de la intervención haya sido muy pequeño, será posible obtener un resultado significativo. Esto sucede porque al aumentar el número de observaciones, igual que sucedía al tirar la moneda al aire, disminuye la variabilidad de los resultados. Esta circunstancia constituye una limitación importante a las pruebas de significación, ya que, en última instancia, el obtener un resultado significativo dependerá del tamaño de la muestra. Con un tamaño suficientemente amplio, cualquier efecto puede presentar significación estadística. Por ello, es fundamental distinguir entre un efecto "estadís-

Tabla II Pruebas estadísticas más usadas en epidemiología

| | |
|---|---|
| **Para investigar el significado estadístico de una diferencia** | |
| Entre dos o más proporciones | Chi cuadrado ($\chi^2$) |
| Entre dos proporciones, cuando el número de observaciones es pequeño | Prueba exacta de Fisher |
| Entre medianas | U de Mann-Whitney |
| Entre medias | T de Student |
| Entre 2 o más medias | F de Sneadecor |
| **Para describir el alcance de la asociación** | |
| Entre una variable independiente continua y una variable dependiente continua | Coeficiente de regresión de Pearson |
| Entre una variable independiente ordinal y una variable dependiente ordinal | Coeficiente de regresión de Spearman |
| **Para modelar el efecto de variables múltiples** | |
| En caso de variable respuesta dicotómica | Regresión logística |
| En caso de variable respuesta tiempo-dependiente | Riesgos proporcionales de Cox |

*Modificado de Fletcher RH y cols. Clinical Epidemiology. The Essentials. Baltimore: Williams & Wilkins; 1996*




Probabilidad que el peso sea diferente Pr=0,05
Probabilidad que el peso o la longitud sean diferentes Pr=B+C+D=0,0975
Probabilidad que el peso, la longitud o el perímetro cefálico sean diferentes Pr=G+H+I+J+K+L+M=0,1426

*Long: longitud; P.C.: perímetro cefálico.*

**Figura 2.** Comparaciones múltiples (explicación en el texto).

ticamente significativo" y "clínicamente importante". La significación estadística traduce una categorización arbitraria de los resultados de la investigación, que no tiene nada que ver con la importancia clínica o biológica del efecto estudiado. El hecho de obtener una *P* de 0.05 o de 0.0005 no debe traducirse, en absoluto, en una mayor o menor magnitud del efecto observado. El investigador o el lector estará mucho más interesado en conocer si el efecto es "clínica o biológicamente importante", que en determinar el nivel de "significación estadística". Pequeñas diferencias sin ningún interés clínico pueden ser estadísticamente significativas con muestras suficientemente amplias y viceversa, efectos de importancia clínica podrían ser estadísticamente no significativos. Los intervalos de confianza, como se verá a continuación, nos permiten tener una idea más apropiada de ambas circunstancias.

La problemática de cómo el valor de *P* se ve afectado por la repetición de las pruebas estadísticas es un importante aspecto que los investigadores siempre deben de tener presente. Imaginemos que tenemos dos cohortes similares de niños nacidos en dos hospitales A y B. Suponemos las medidas antropométricas de peso, longitud y perímetro cefálico iguales para ambas cohortes. Fijando el nivel de significación estadística en una $P \leq 0.05$, la probabilidad $\alpha$ de que, por ejemplo, el peso difiera sólo por el azar será del 5%. Si se consideran simultáneamente, la probabilidad $\alpha$ de que el peso o la longitud difieran sólo por azar asciende al 9.75%. Si ahora consideramos las 3 variables peso, longitud y perímetro cefálico a la vez, la probabilidad $\alpha$ de que alguna de las 3 varíe significativamente con respecto a las del otro grupo será del 14.26% (Fig. 2). La probabilidad de que ninguna de las 3 variables difiera significativamente entre los grupos habrá disminuido del 95% al 86%. Siguiendo la misma línea argumental, si los investigadores llegaran a realizar 30 comparaciones estadísticas entre las dos cohortes de sujetos, la probabilidad de que sólo por el azar apareciesen una o más pruebas significativas sería del 78.54%[8].

Para abordar este problema derivado de las comparaciones múltiples se puede recurrir a la desigualdad de Bonferroni[9], según la cual se puede calcular la probabilidad Pr de encontrar un resultado significativo en un proceso de comparaciones múltiples aplicando la fórmula siguiente:

$$Pr = 1 - (1- \alpha)^n$$

siendo n el número de comparaciones efectuadas. En el ejemplo anterior de las 30 comparaciones, la probabilidad de que alguna de estas comparaciones derive en un resultado significativo sólo por el efecto de la suerte será:

$$Pr = 1- (1- 0.05)^{30} = 0.7854 \text{ ó } 78.5\%.$$

La corrección de Bonferroni[9] propone conservar el nivel global de significación en 0.05, requiriendo para cada una de las pruebas el nivel de significación de 0.05/n, siendo n el número de comparaciones a efectuar. Esta técnica exige niveles de significación muy bajos para cada una de las comparaciones individuales y asume que las comparaciones son independientes entre sí, lo que no siempre se cumple. Por otra parte, el método de Bonferroni se fundamenta en una hipótesis nula general (todas las hipótesis nulas son simultáneamente verdaderas), que generalmente carece de interés alguno para el investigador y supone un incremento considerable del riesgo de cometer un error de tipo II[10]. La solución óptima será planear las comparaciones a realizar antes del análisis, reduciéndolas a las menos posibles[11]. Una aplicación juiciosa de las pruebas estadísticas y limitarse a contestar la hipótesis fundamental planteada en el diseño de la investigación, acotará el impacto de la inflación del error tipo I consecuencia de la repetición de pruebas estadísticas[10,12].

Otro problema habitual relacionado con la significación estadística se refiere a que, con frecuencia, por cuestiones éticas, económicas o la simple curiosidad del investigador se efectúan análisis estadísticos de la investigación antes de completar el tamaño de la muestra requerido. Por ejemplo, investigaciones en Tailandia y en Costa del Marfil encaminadas a estudiar la reducción de la transmisión perinatal del VIH mediante pautas cortas de tratamiento

con AZT fueron recientemente interrumpidas por razones éticas, al comprobar que las madres tratadas con la pauta corta presentaban una transmisión inferior a las tratadas con placebo[13]. Sin embargo, como las pruebas de significación requieren un tamaño de la muestra mínimo, esta interrupción conlleva un coste estadístico para cada análisis prematuro, siendo la probabilidad de alcanzar un resultado significativo mayor de la deseada. Se han descrito diversas manipulaciones estadísticas para obviar este problema, pero su aplicación es demasiado compleja para el no estadístico[14-16]. La mejor sugerencia en este caso es no precipitar el cálculo de la significación estadística hasta que la investigación no esté finalizada. Si por razones éticas no es posible aguardar, lo recomendable sería recurrir a un profesional de la estadística. Otra posibilidad es planear esta eventualidad con antelación, utilizando un diseño secuencial, que prevé los análisis intermedios de los datos[11].

## Estimación del efecto. Intervalos de confianza

La estimación puntual del efecto es el número individual que mejor informa sobre un conjunto de datos obtenidos en una investigación determinada. En el ejemplo que hemos seguido en este artículo, 20 mmHg es la reducción de tensión arterial diastólica originada por el fármaco a estudio. Pero, en realidad, este número no es más que un punto de una escala continua con infinitos valores posibles, por lo que matemáticamente la probabilidad de que la estimación puntual sea correcta será uno dividido entre infinito ($1/\infty$), es decir, 0. Esta información debe ser, entonces, complementada con una medida que oriente sobre el error aleatorio de la investigación, lo que se consigue por medio del intervalo de confianza (IC). El IC no es más que el rango de valores alrededor de la estimación puntual que, considerando el grado de variabilidad de los datos y contando con que no existan sesgos, en un porcentaje determinado de veces que se repita la investigación siempre incluirá esa estimación puntual. Es decir, si el IC del 95% del ejemplo es 15-25 mmHg indica que, en ausencia de sesgos, cuando repitamos la misma investigación, un 95% de las veces la diferencia de tensión arterial diastólica oscilará entre esos valores y sólo un 5% de las veces estará fuera de esos límites. Una interpretación más sencilla, pero también válida, es que el IC del 95% es el rango de valores sobre el que podemos estar un 95% seguros de que incluye al valor que correspondería a la población de origen[17].

El cálculo del IC se basa en la magnitud observada (d), que es lo que se intenta cuantificar (por ejemplo, la diferencia entre medias), y en el error estándar (EE) de esa estimación, según la fórmula:

$$IC\ 95\% = d \pm 1.96 \times EE$$

(esta fórmula variará según la naturaleza de la variable respuesta que se mide y según el nivel de confianza deseado, pero siempre mantendrá esta estructura general). Es posible construir IC para diseños de estudios clínicos (diferencia de medias o proporciones, riesgos relativos, razones de odds, número necesario a tratar (NNT)), para estudios diagnósticos (sensibilidad, especificidad, valores predictivos) y para estimaciones derivadas de metaanálisis y estudios caso-control. En la tabla III se proporcionan las fórmulas del cálculo de los IC más populares en investigación epidemiológica.

La amplitud del IC depende, pues, del error estándar y del nivel de confianza que deseemos asociar con el intervalo. El error estándar, que se calcula dividiendo la desviación estándar entre la raíz cuadrada del número de sujetos incluidos en la muestra, refleja la imprecisión derivada de trabajar con muestras y no con poblaciones. El error estándar de la media de una muestra concreta indica la proximidad de la media de esa muestra a la media verdadera de la población de donde procede la muestra. En consecuencia, con muestras pequeñas el error estándar será grande y el IC amplio. Con muestras grandes, el error estándar será pequeño y el IC estrecho. Por ejemplo, si suponemos que la diferencia de tensión arterial diastólica entre los niños tras tomar el fármaco A o el placebo fue de 20 mmHg con una desviación estándar de 32 mm Hg, podemos calcular el IC del 95% para una muestra de 400 pacientes

$$IC\ 95\% = 20 \pm 1.96 \times 32/\sqrt{400} = 16.9 - 23.1\ mmHg$$

Sin embargo, para una muestra de 10 pacientes,

$$IC\ 95\% = 20 \pm 1.96 \times 32/\sqrt{10} = 0.17 - 39.8\ mmHg$$

Estos resultados se exponen gráficamente en la figura 3.

Por otra parte, el nivel de confianza lo fija arbitrariamente el investigador. Cuanto mayor sea el nivel de confianza más amplio será el intervalo de confianza, puesto que menos posibilidades habrá de que el verdadero valor de la población esté fuera de ese rango. Veíamos como el IC del 95% de la media de la reducción de tensión arterial de nuestro ejemplo oscilaba entre 15 y 25 mmHg. El IC del 90% oscilaría entre 18 y 22 mmHg. Habitualmente se maneja como nivel de confianza adecuado el 95%, pero tampoco es infrecuente utilizar el 90 o el 99%. Como sucedía con el nivel de significación estadística, los investigadores deben especificar el nivel de confianza a priori. Cuando este nivel de confianza se aleja del más aceptado, el 95%, una justificación clara debe incluirse en el apartado de metodología[18]. No obstante, al comunicar los resultados es recomendable notificar la diferencia entre las medias y su error estándar, además del IC. Así, cualquier lector podrá calcular fácilmente el IC para cualquier nivel de confianza.

A partir del IC es posible inferir el resultado de la prueba de significación. Cuando el valor correspondiente a la hipótesis nula (0 para las diferencias, 1 para los cocientes) queda comprendido fuera del IC, deducimos la existencia de una diferencia estadísticamente significativa, pudiendo rechazar la hipótesis nula con un riesgo $\alpha$ asociado a ese intervalo 1-$\alpha$. Cuando el valor de la hipótesis nula se encuentra dentro del IC se puede deducir que es verosímil y la hipótesis nula no debe ser rechazada[2]. Una ventaja sobresaliente del IC es que permite evaluar la importancia clínica o biológica del efecto que estamos cuantificando, orientando, además, sobre su precisión. A diferencia de las pruebas de sig-

Tabla III Construcción de intervalos de confianza del 95%.

**1. Proporción (ej., sensibilidad, especificidad, tasas, etc.)**

$$p \pm (1.96 \times \sqrt{\frac{p \times (1-p)}{n}})$$

p = proporción; n = número de sujetos
Ej. La sensibilidad de una prueba diagnóstica en 40 niños fue del 35%.
p=0.35; n=40. IC 95%: 0.2 a 0.5

**2. Diferencia de medias**

$$(x_1 - x_2) \pm (1.96 \times \sqrt{s_p^2 \times (\frac{1}{n_1} + \frac{1}{n_2})})$$

$x_1, x_2, n_1, n_2$: medias aritméticas y tamaño de la muestra en los grupos 1 y 2. $s_p^2$: varianza acumulada.
Ej. La concentración media de triglicéridos plasmáticos en 2 grupos de 80 y 90 niños fue de 100 y 75 mg/dl.
$x_1$: 100 mg/dl; $x_2$: 75 mg/dl; $n_1$: 80; $n_2$: 90; $s_p^2$:1980

$$(100 - 75) \pm (1.96 \times \sqrt{1980 \times (\frac{1}{80} + \frac{1}{90})})$$

IC 95%: 11.6 a 38.4 mg/dl

**3. Diferencia de proporciones entre 2 grupos**

$$(p_2 - p_1) \pm (1.96 \times \sqrt{(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})})$$

$p_1, p_2, n_1, n_2$: proporciones y tamaño de la muestra en los grupos 1 y 2.
Ej. Diferencia entre la proporción de niños que mejoran tras la administración del fármaco X (40%) en 75 pacientes o de un placebo en 70 (25%).

$$(0.40 - 0.25) \pm (1.96 \times \sqrt{(\frac{0.25(1-0.25)}{70} + \frac{0.40(1-0.40)}{75})})$$

IC 95%: 0 a 30%. Este concepto corresponde a la **reducción absoluta de riesgos** (RAR), tras una intervención determinada. El **número necesario a tratar** (NNT) sería el inverso de la RAR (1/15) y su intervalo de confianza el inverso a los límites del IC del RAR (1/0 y 1/30)[31].

**4. Riesgo relativo** (RR)

$$RR\ exp \pm [1.96 \times \sqrt{\frac{1 - a/(a+c)}{a} + \frac{1 - b/(b+d)}{b}}]$$

Ej. De 194 recién nacidos con la fibronectina fetal positiva, 57 nacieron pretérmino. De 2734 con la fibronectina fetal negativa, 246 fueron pretérmino[32].

$$3.26\ exp \pm [1.96 \times \sqrt{\frac{1 - 57/(57+137)}{57} + \frac{1 - 246/(246+2488)}{246}}]$$

Límite superior IC 95%: $3.26e^{0.245} = 4.19$. Límite inferior IC 95%: $3.26e^{-0.245} = 2.55$.

**5. Odds ratio** (OR)*****

$$OR\ exp \pm [1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}]$$

Ej. De 11 niños asmáticos que sufrieron parada cardiorrespiratoria, en 10 el test cutáneo para aeroalergenos había sido positivo. En 99 niños asmáticos que no la habían padecido, el test cutáneo fue positivo en 31 niños[33].

$$21.9\ exp \pm [1.96 \times \sqrt{\frac{1}{10} + \frac{1}{1} + \frac{1}{31} + \frac{1}{68}}] = 21.9\ exp \pm (1.96 \times 1.36)$$

Límite superior IC 95%: $21.9e^{2.67} = 316.2$. Límite inferior IC 95%: $21.9e^{-2.67} = 1.5$.
*Esta fórmula es sólo una aproximación válida en caso de muestras relativamente grandes

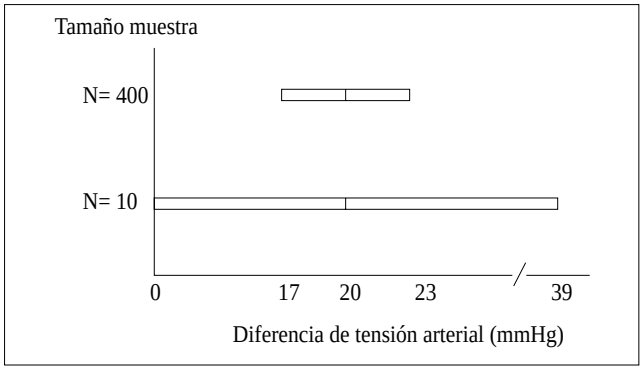



**Figura 3.** Variación del intervalo de confianza según el tamaño de la muestra (explicación en el texto)

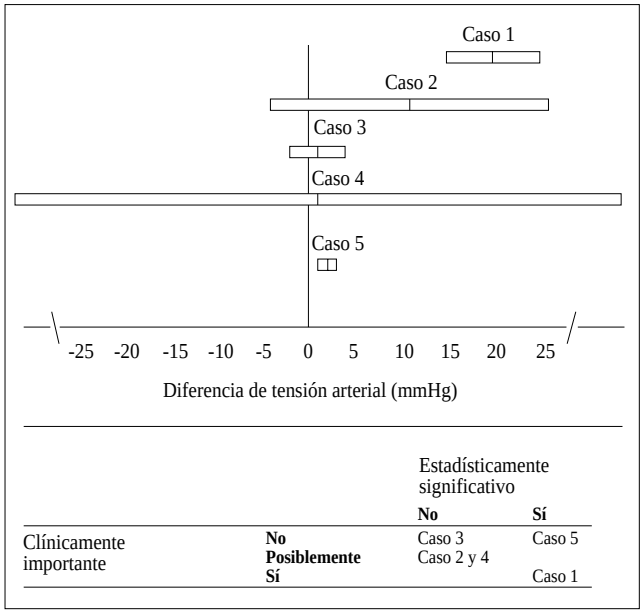



**Figura 4.** La interpretación del intervalo de confianza (ver texto).

nificación que se basan en una dicotomía arbitraria de significativo-no significativo y que se expresan mediante un número, la *P*, que por sí mismo tiene una interpretación difícil y que en absoluto orienta sobre la intensidad de la relación entre las variables ni sobre la magnitud del efecto, el IC facilita el conocer el conjunto de valores que pudiera previsiblemente adoptar la población estudiada en unas circunstancias determinadas y, consecuentemente, la magnitud del efecto estudiado.

Con variaciones sobre este ejemplo del fármaco hipotensor, trataremos de ilustrar todas las posibilidades que se desprenden del estudio del IC (Fig. 4). Si la tensión arterial diastólica de los niños estudiados se ha reducido al emplear el fármaco A en 20 mmHg (IC 95%: 15 a 25 mmHg) (caso 1), estamos viendo perfectamente cuál ha sido la magnitud del efecto de ese fármaco A. Además, ya que el valor correspondiente a la hipótesis nula, es decir, que la reducción de la tensión arterial fuera 0 mmHg, no está comprendido entre los límites del IC, podemos también asegurar que el resultado ha sido estadísticamente significativo. Pero el estudio del IC, como se ha dicho, aún nos puede proporcionar más información. Imaginemos que la reducción de tensión arterial tras ingerir el fármaco A ha sido de 10 mmHg (IC 95%: –5 a +25 mmHg) (caso 2). Podemos deducir que el resultado no ha sido estadísticamente significativo, ya que el IC incluye el valor 0 de la hipótesis nula. También vemos que es un intervalo ancho, con el límite inferior cercano a 0 y el límite superior indicando un efecto apreciable. Ello sugiere que nuestra muestra seguramente es pequeña y que hay un efecto potencial. Los investigadores deberían insistir en aumentar el tamaño de la muestra, ya que es muy probable que encontrasen un efecto importante. Al contrario, si la reducción de tensión arterial fuera de 1 mmHg (IC 95%: –2 a +4 mmHg) (caso 3), estaríamos ante un resultado preciso en el que no se aprecia la existencia de un efecto ni clínicamente importante, ni biológicamente significativo, por lo que no estaría justificado persistir en la justificación de la hipótesis original. En el caso que la reducción de tensión arterial fuese también de 1 mmHg, pero oscilase entre –30 mm y +30 mmHg (IC 95%: -29 a +31 mmHg) (caso 4), el resultado sería sumamente impreciso, pero ya que no se denota la existencia de ningún efecto, posiblemente tampoco merecería la pena ampliar la muestra del estudio en busca de una mayor precisión. También cabe la posibilidad de que la reducción de tensión arterial fuera de 2 mmHg (IC 95%: 1 a 3 mmHg) (caso 5). En esta ocasión, el resultado es preciso, ya que el rango de valores que abarca es estrecho, y estadísticamente significativo, puesto que no incluye el valor de 0, correspondiente a la hipótesis nula. Sin embargo, vemos que el efecto clínico es prácticamente despreciable y, por tanto, de escasa utilidad. En todos estos supuestos anteriores, el valor de la *P* correspondiente a la prueba de significación únicamente nos habría informado sobre si el resultado era o no estadísticamente significativo. Es evidente, entonces, que la información que los investigadores proporcionan mediante los IC es mucho más completa y valiosa que la que proporcionan las pruebas de significación. No obstante, autores tan prestigiosos como Feinstein[18] consideran adecuado aportar en las publicaciones científicas ambos tipos de información: los IC para mostrar los rangos potenciales de resultados y los valores de la *P*, para que investigadores y lectores tengan un punto de referencia y no resulten confundidos por un intervalo con un nivel de confianza elegido, de alguna manera, arbitrariamente.

## Introducción al cálculo del tamaño de la muestra. Diseño del muestreo

Una vez que hemos decidido el problema que queremos investigar, definida la hipótesis que deseamos probar y descritas las variables que vamos a utilizar, es preciso acometer la recolección de datos mediante la aplicación de los métodos elegidos. Para ello es necesario conocer cuántos datos debemos recoger, es decir, debemos calcular el tamaño muestral. Lo ideal en cualquier investigación sería estudiar a todos los individuos que compartan la características objeto de la investigación, por ejemplo, a todos los hipertensos de un país o a todos los niños obesos entre 8 y 12 años de edad en una ciudad determinada. A este conjunto de personas

que incluye a TODOS los individuos se le denomina "universo" o "población diana". En la realidad esto no es viable, ya que razones de tipo de económico, logístico, ético, etc., obligan a efectuar el cálculo del tamaño de la muestra antes de iniciar una investigación epidemiológica. Si nuestra muestra fuese insuficiente, la investigación supondría un desaprovechamiento de recursos y el sometimiento innecesario de los pacientes a una experimentación que no servirá para obtener ninguna conclusión científicamente válida. Por el contrario, si la muestra fuese excesiva, más individuos de los necesarios serían expuestos a la investigación, privándoles quizás de la mejor conducta terapéutica o diagnóstica posible, al menos a una parte de ellos. En esta situación se malgastarían parte de los recursos empleados, ya que las conclusiones podrían ser obtenidas con una muestra menor.

Así, lo que se hace en la práctica es estudiar una parte de ese universo, pero con la condición de que esa parte represente fielmente a todos los miembros del universo concreto que se pretende estudiar. O sea, el objetivo será estudiar una "muestra" representativa del universo. La muestra que finalmente estudiamos procede de lo que se denomina "población accesible". Existe una serie de pasos intermedios para llegar desde la población diana hasta la población accesible, pero los obviamos en este momento, ya que su estudio queda fuera de los objetivos de este artículo. Una vez que tenemos la población accesible se procede a seleccionar la muestra, lo que se conoce como diseño del muestreo.

Existen dos tipos de diseño de muestreo: el probabilístico y el no probabilístico. Un muestreo probabilístico utiliza un proceso aleatorio, el azar, para garantizar que cada miembro de la población accesible tenga una probabilidad específica (la misma para cada miembro) de ser seleccionado. Al contrario, un muestreo no probabilístico no se basa en el azar para la selección de los sujetos de la muestra, con lo que no todos los miembros de la población accesible disfrutarán de la misma probabilidad de ser elegidos.

Dentro de estos tipos de diseño de muestreo se describen a su vez varios subtipos. En este artículo, sin embargo, vamos a comentar sólo un subtipo de muestreo probabilístico, el más empleado en la práctica habitual, el "muestreo aleatorio simple".

En el muestreo aleatorio simple conocemos a todos y a cada uno de los integrantes de la población accesible. El procedimiento consiste en enumerarlos y elegir al azar el número de sujetos previamente calculado. Por ejemplo, si vamos a realizar un estudio en una comunidad sobre la obesidad infantil y hemos calculado que necesitamos estudiar a 400 niños entre 8 y 12 años, el proceso a seguir sería el siguiente. Primero, dispondríamos de un listado de todos los niños que viven en la comunidad y estén comprendidos entre esas edades (población accesible). Segundo, enumeraríamos a cada uno de los niños. Tercero, mediante un programa informático o una tabla de números aleatorios seleccionaríamos los números hasta completar la muestra necesaria de 400 niños.

## Error de muestreo o error muestral

Antes de proceder al estudio del cálculo del tamaño de la muestra, conviene revisar el concepto de error muestral.

Como ya hemos visto a lo largo de todo este artículo, la muestra es sólo una representación de todo el universo y, aunque seamos muy cuidadosos al seleccionar esta muestra, normalmente habrá una diferencia entre los valores estadísticos obtenidos de la muestra con los derivados del universo. Veíamos al inicio del artículo como si lanzamos al aire una moneda perfectamente equilibrada 20 veces podemos obtener un 45% de caras y un 55% de cruces. Sin embargo, en el universo (lanzar la moneda al aire infinitas veces), el porcentaje de caras y cruces es el mismo, e igual al 50%. A esta diferencia se denomina "error muestral real". La única manera de conocer el error muestral real sería lanzar al aire infinitas veces la moneda o, lo que es lo mismo, estudiar a todo el universo, pero esto es virtualmente imposible. Por lo tanto, cuando hablamos de error muestral, no nos referimos al error muestral real, que nos es desconocido, sino a un error muestral determinado estadísticamente, de tipo genérico, válido para todas las muestras posibles del mismo tamaño. Así, este error sirve para darnos los límites formados por la media de la muestra más/menos el error en cuestión, dentro de los cuales se debe encontrar la media del universo, con un grado de probabilidad determinado y especificado por el investigador (90%, 95%, 99%, etc.).

Por ejemplo, si realizamos un estudio (con un grado de probabilidad del 95%) entre 400 niños, de 8 y 12 años de edad, y encontramos que el 15% presenta obesidad con un error muestral del 3.5%, significa que la proporción media de obesidad en todos los niños de 8 y 12 años de edad (universo) estará, con una probabilidad del 95%, entre 15 ± 3.5%, es decir, entre el 11.5% y el 18.5%.

## Cálculo del tamaño muestral

Una vez revisados los conceptos de muestreo aleatorio simple y de error muestral expondremos como calcular el tamaño mínimo necesario de sujetos que se necesitan para realizar un estudio.

Existen varias formas de calcular tamaños muestrales. Nosotros expondremos sólo la más simple, que es la correspondiente al cálculo del tamaño muestral para encuestas poblacionales o estudios descriptivos. Se deben cumplir dos condiciones para utilizar este tipo de cálculo de tamaño muestral. Primera, el diseño del muestreo debe ser un muestreo aleatorio simple. Segunda, la variable estudiada debe ser dicotómica (obesidad sí, obesidad no) o cualquier otra respuesta con dos alternativas (tensión arterial diástólica menor de 110 mmHg versus tensión arterial diástólica mayor de 110 mmHg).

*Ejemplo.* En una comunidad con 150000 niños entre los 8 y 12 años de edad (población accesible) pretendemos realizar un estudio sobre obesidad infantil. Sabemos por estudios previos realizados o por la bibliografía revisada que la frecuencia o prevalencia de obesidad infantil en este tipo de población es del 15% (porcentaje de obesidad infantil en el universo) con una variación del 3.5% (error muestral). Deseamos calcular el tamaño mínimo de la muestra para probar, con una probabilidad del 95%, si ese porcentaje de obesidad en el universo es el mismo que en nuestra población accesible de 150000 niños.

La fórmula a utilizar sería:

$$T = \frac{Z \times Z \times [P \times (1-P)]}{D \times D}$$

siendo T, el tamaño muestral que queremos calcular; Z, el valor de la Z-score en una curva normal para una probabilidad del 95%, que es igual a 1.96[6,7]; P, la prevalencia o frecuencia de la variable expresada como proporción, igual a 0.15; D, el error muestral expresado como proporción, igual a 0.035.

Sustituyendo en la fórmula,

$$T = \frac{1.96 \times 1.96 \times [0.15 \times (1-0.15)]}{0.035 \times 0.035} = 399$$

Por lo tanto, necesitamos estudiar, como mínimo, a 399 niños entre 8 y 12 años de edad para probar si el porcentaje de obesidad en esa franja de edad en nuestra comunidad es del 15 ± 3.5%.

Si el número de pacientes calculado fuese imposible de reclutar por no disponer de tiempo o de medios necesarios, o por ser poco frecuente la enfermedad estudiada, la única alternativa será ampliar el error muestral. Así, si en lugar de 3,5% se estima en 7%, el número de sujetos requerido sería de 100 niños.

Ahora bien, nos podemos preguntar para qué necesitamos conocer el tamaño de la población accesible, en nuestro caso de 150000, para calcular el tamaño muestral, si este valor, como vemos, no se utiliza en la fórmula.

Para responder a esta pregunta nos basta conocer que las poblaciones se clasifican como infinitas si tienen mas de 100000 miembros y como finitas si tienen menos de 100000. En nuestro caso estaríamos ante una población infinita. Ello significa que a partir de una población accesible de más de 100000 individuos, el cálculo del tamaño muestral permanece constante. Es decir, una mayor población no implica que haya que seleccionar mayor muestra. En nuestro ejemplo, se demuestra que si en vez de tener una población de 150000 niños tuviéramos otra de 1000000 de niños, en vez de 399 niños en la muestra, necesitaríamos 400 niños.

En el caso de que la población accesible tuviera menos de 100000 sujetos, al tamaño muestral calculado habría que aplicarle el factor de corrección para poblaciones finitas.

*Ejemplo*. Supongamos que queremos hacer el estudio anterior, no en una comunidad, sino en un hospital con una serie de 1000 historias clínicas de niños entre 8 y 12 años de edad. ¿Qué número mínimo de historias necesitaríamos estudiar para probar que la prevalencia de obesidad infantil es del 15 ± 3.5%?

Al tamaño muestral calculado anteriormente se le debe aplicar la corrección de poblaciones finitas.

$$n = \frac{T}{1 + (T/P)} = \frac{399}{1 + (399/1000)} = 286$$

Así, si nuestra población accesible fuera de 1000 individuos precisaríamos estudiar sólo a 286 niños en lugar de los 399 obtenidos anteriormente.

Es preciso señalar que cuando se calcula el tamaño muestral mínimo necesario, no se puede obviar el porcentaje de no respuestas. Este porcentaje es el número de personas que no van a querer participar en el estudio. Así, si pensamos que el 10% de la población rechazará participar en el estudio hay que calcular un tamaño muestral un 10% superior, para que podamos seguir manteniendo el tamaño muestral mínimo estimado.

Podemos deducir que, para calcular el tamaño de la muestra en la estimación de una proporción se debe decidir en primer lugar la precisión o error muestral, lo que se hace en cierta forma arbitrariamente. Posteriormente, se debe conocer la prevalencia de la enfermedad estudiada en la población de origen. Si la prevalencia es desconocida, la solución óptima será otorgar al producto P x (1-P) el valor máximo posible, que ocurre cuando P = 0.5[19].

Las fórmulas para el cálculo del tamaño de la muestra en la estimación de una media o en los problemas de contraste de hipótesis con dos muestras, aunque se fundamentan en las mismas bases teóricas, son progresivamente más complejas. Su exposición y análisis superan claramente los propósitos de este artículo. Mayor información está disponible en la literatura especializada[11,20-22].

## Potencia estadística

Una vez realizado un estudio, especialmente cuando el resultado no permite rechazar la hipótesis nula, es decir, cuando no es estadísticamente significativo, el investigador está obligado a calcular y comunicar la probabilidad de que el resultado hubiese sido estadísticamente significativo en caso de que realmente existiese un efecto[23]. Esta probabilidad (1-β), complementaria a la probabilidad β definida anteriormente, se denomina potencia estadística. Mide, por tanto, la capacidad para detectar una determinada asociación si ésta realmente existe.

Cuanto mayor sea la magnitud del efecto considerado, el tamaño de la muestra estudiada y el riesgo α establecido y cuanto menor sea la variabilidad de la variable respuesta, mayor será la potencia estadística. Sin embargo, en la práctica es la magnitud del efecto investigado lo que condiciona de forma más fundamental la potencia estadística[24]. Por ello, en la investigación clínica es necesario tratar de detectar los menores efectos clínicamente importantes, para posibilitar que la prueba posea suficiente potencia. Una práctica común totalmente rechazable consiste en concretar la magnitud del efecto que se desea detectar en función del número de sujetos disponibles para el estudio. Como ya se ha comentado, lo primero debiera ser siempre establecer la magnitud del efecto según su transcendencia clínica. Una vez decidido cual es el mínimo efecto clínicamente importante, si se precisa incrementar la potencia estadística, se aumentará el tamaño de la muestra. Si es necesario, el investigador dispone de diversas estrategias para tratar de incrementar la potencia estadística de su estudio, tanto en la fase de diseño, como en la de análisis del mismo (Tabla IV).

Aunque en puridad el nivel de potencia estadística requerido en el estudio debería ser establecido en función de las consecuencias de cometer un error tipo II, se acepta como límite aceptable el valor del 80%. Sin embargo, si se trata de un ensayo clínico, siempre que no existan sesgos, podría ser conveniente acometer un estudio concreto con una potencia inferior al 80% cuando no sea posible reclutar suficientes sujetos, ya que las revisiones sistemáticas y las técnicas de metaanálisis posibilitarán el aprovechamiento

Tabla IV Estrategias para incrementar la potencia estadística de un estudio de investigación epidemiológico

| *Fase de diseño* | *Fase de análisis* |
|---|---|
| Elección de diseños que apoyan demostración de relaciones de causalidad (de más a menos, ensayos clínicos, cohortes, casos y controles, estudios transversales). | Analizar las variables medidas utilizando, si es posible, una escala cuantitativa. |
| Selección de una población homogénea, sensible a la exposición, con una distribución próxima al 50% en cuanto a sujetos expuestos o enfermos y no expuestos o controles. | Utilizar, siempre que se pueda, pruebas paramétricas, diseños de medidas repetidas e hipótesis unilaterales. |
| Medición exacta de la exposición y de la enfermedad. | Acotar periodos de exposición de interés. |
| Aumentar el tamaño de la muestra, distribuyendo el número de sujetos entre los grupos a estudio uniformemente. | Reducir la variabilidad de los parámetros estudiados controlando los factores de confusión. |

del trabajo[25].

En definitiva, si el resultado del análisis de un estudio no es estadísticamente significativo y la potencia estadística es elevada, el investigador deberá admitir con una probabilidad β de cometer un error tipo II que no existe asociación entre exposición y respuesta o, mejor, que esa posible asociación no alcanza la dimensión catalogada de clínicamente importante. Al contrario, si la potencia estadística es baja no será posible establecer conclusiones del análisis efectuado. En cualquier caso, ante un estudio negativo, el cálculo de la potencia estadística es obligado[11,20].

El cálculo del tamaño de la muestra y de la potencia estadística de una prueba se puede encontrar en los textos apropiados[11,20,26]. Afortunadamente, existen programas informáticos específicos para efectuar dicho cálculo[27-29] e, incluso, páginas web con listados de software apropiado para este menester[30].

## Bibliografía


1 Orejas Rodríguez-Arango G, Martínez Navarro JF. Epidemiología y metodología científica aplicada a la pediatría (I). Introducción. Medidas de frecuencia, asociación e impacto. *An Esp Pediatr* 1998;**49**:313-320.

2 Cava F, Fernández GC, Cava C, Doménech JM. Utilización de los intervalos de confianza para presentar los resultados en las revistas biomédicas. *Med Clin (Barc)* 1993; **100**: 597.

3 Rothman KJ. Significance questing. *Ann Int Med* 1986; **105**:445-447.

4 Hennekens CH, Buring JE. Epidemiology in medicine. Boston/Toronto: Little, Brown and Company; 1987. p. 243-271.

5 Rothman KJ. The role of statistics in epidemiologic analalysis. En: Rothman KJ, ed. Modern epidemiology. Boston. Little, Brown and Company; 1986. p. 115-129.

6 Colton T: Estadística en medicina. Barcelona: Salvat; 1988.

7 Rosner B: Fundamentals of biostatistics, 4ª edición. Belmont: Duxbury Press; 1995.

8 Brown GW. P values. *AJDC* 1990; **144**: 493-495.

9 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995; **310**: 170.

10 Perneger TV. What´s wrong with Bonferroni adjustments. *BMJ* 1998; **316:** 1236-1238.

11 Argimón Pallás JM, Jiménez Villa J: Métodos de investigación. Aplicados a la atención primaria de salud. Barcelona: Mosby/Doyma Libros; 1991. p. 151-166.

12 O´Brien PC. The appropriateness of analysis of variance and multiple comparison procedures. *Biometrics* 1983; **39**: 787-788.

13 *MMWR* 1998; **47**: 151-154.

14 Pocock SJ. Size of cancer trials and stopping rules. *Br J Cancer* 1978; **38**: 757-766.

15 Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1982; **38**: 153-162.

16 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Eng J Med* 1987; **317**: 426-432.

17 Sackett DL, Richardson WS, Rosemberg W, Haynes RB. Evidence-based medicine. How to practice and teach EBM. New York: Churchill Livingstone; 1997. p. 227-234.

18 Feinstein AR. P-values and confidence intervals: Two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998; **51**: 355-360.

19 Sentís J. Reflexiones sobre el tamaño de la muestra en los trabajos de investigación. *Med Clin (Barc)* 1997; **108**: 512-516.

20 Hulley SB, Cummings SR: Diseño de la investigación clínica. Un enfoque epidemiológico. Barcelona: Ed. Doyma; 1993.

21 Luna del Castillo J, Martín Andrés A. Y ahora ¿cuántos individuos tomo?. Algunas ideas básicas sobre el tamaño de la muestra: I. Tamaño de la muestra en un problema de estimación. *Atención Primaria* 1990; **7**: 64-67.

22 Luna del Castillo J, Martín Andrés A. Y ahora ¿cuántos individuos tomo?. Algunas ideas básicas sobre el tamaño de la muestra: II. Tamaño de la muestra en un problema de contraste de hipótesis con dos muestras. *Atención Primaria* 1990; **7**: 90-93.

23 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994; **272**: 122-124.

24 Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Ann Allergy Asthma Immunol* 1997; **78**: 5-16.

25 Peipert JF, Metheny WP, Schulz K. Sample size and statistical power in reproductive research. *Obstet Gynecol* 1995; **86**: 302-305.

26 Selvin S. Statistical analysis of epidemiologic data. New York: Oxford University Press; 1991. p. 71-89.

27 Systat, Inc., Evanston, IL, USA.

28 Epi-Info 6.04b. Centers for Disease Control and Prevention, Atlanta, GA, USA.

29 Epistat, Epistat Services, Richardson, TX, USA.

30 http://www.interchg.ubc.ca/cacb/power/

31 Solís Sánchez G, Orejas Rodríguez-Arango G. Epidemiología y Metodología Científica aplicada a la Pediatría (II): Diseños en investigación epidemiológica. *An Esp Pediatr* 1998; **49**: 527-538.

32 Goldenberg RL, Iams JD, Mercer BM, Meis PJ, Moawad AH, Copper RL, Das A, Thom E, Johnson F, McNellis D, Miodovnik M, Van Dorsten JP, Caritis SN, Thurnau GR, Bottoms SF, and the NICHD MFMU