COMUNICACIÓN BIOMÉDICA

INTERVALOS DE CONFIANZA Y NO VALORES *P*: ESTIMACIÓN EN VEZ DE PRUEBAS DE HIPÓTESIS[1]

Martin J. Gardner[2] y Douglas G. Altman[3]

El excesivo énfasis en las pruebas de evaluación de hipótesis —y en el uso de valores P *para diferenciar los resultados significativos de los no significativos*[4]*— ha relegado el uso de la estimación y los intervalos de confianza, procedimientos mucho más útiles para interpretar los resultados de un estudio. En la investigación médica los autores suelen estar interesados en la magnitud de la diferencia de un resultado medido en dos grupos diferentes, no en una simple indicación de si esa diferencia es o no estadísticamente significativa. El intervalo de confianza presenta un recorrido de valores, basado en el resultado muestral, en el que puede encontrarse la diferencia poblacional. Se indican algunos métodos de cálculo de intervalos de confianza para medias y diferencias entre medias, y para proporciones y diferencias entre proporciones. También se dan sugerencias para la representación gráfica.*

Los resultados principales de una investigación, tanto en el cuerpo de un artículo como en su resumen, deben expresarse mediante intervalos de confianza apropiados para el estudio en concreto.

Introducción

En los últimos 20 ó 30 años el uso de la estadística en las revistas médicas se ha incrementado enormemente. Una consecuencia desgraciada de ese fenómeno ha sido que el énfasis en los resultados básicos se ha desplazado hacia una con-

[1] Esta traducción del artículo "Confidence intervals rather than P values: estimation rather than hypothesis testing" (*British Medical Journal*, 1986;292:746–750) se publica con autorización de la revista *British Medical Journal* y de los autores.

[2] BSC, PHD, profesor de estadística médica. Consejo de Investigaciones Médicas (CIM), Unidad de Epidemiología Ambiental (Universidad de Southampton), Hospital General de Southampton. Dirección postal: MRC Environmental Epidemiology Unit (University of Southampton), Southampton General Hospital, Southampton SO9 4XY, Reino Unido.

[3] BSC, bioestadístico. División de Estadística Médica, CIM, Centro de Investigación Clínica, Harrow, Middlesex.

[4] En este artículo hemos preferido la notación de Mainland (*1*) y hemos usado *P* para la probabilidad asociada al resultado de una prueba de evaluación de una hipótesis nula, no *p* que indica proporción (véase apéndice 2). Esto es contrario a la convención de Vancouver, pero estadísticamente es más aceptado y también se ajusta más a las normas estadísticas publicadas en *British Medical Journal*.

centración indebida en las pruebas de hipótesis. En estas pruebas, los datos se examinan en relación a una hipótesis estadística "nula", práctica que ha llevado a la creencia errónea de que el objetivo de los estudios debe ser obtener "significación estadística", cuando, en realidad, el objetivo de la mayor parte de las investigaciones médicas es determinar la magnitud de algún factor de interés.

Por ejemplo, en una investigación de laboratorio puede estudiarse la diferencia de concentraciones medias de un componente hemático entre pacientes con cierta enfermedad y personas que no la tienen, mientras que en un ensayo clínico puede evaluarse la diferencia de pronóstico —en porcentajes de curación, remisión, recidiva o supervivencia— de pacientes con cierta enfermedad tratados de distintas formas. La diferencia obtenida en ese estudio solo será una estimación de lo que realmente nos interesa: el resultado que se habría obtenido si se hubieran investigado todos los sujetos elegibles (la "población") y no solo una muestra de ellos. Lo que los autores y lectores querrán saber es en qué medida la enfermedad modificó la concentración hemática media o hasta qué punto el nuevo tratamiento modificó el pronóstico, y no solo el nivel de significación estadística.

El uso excesivo de pruebas de hipótesis a expensas de otros métodos de interpretación de los resultados ha llegado a tal punto que, a menudo, en el cuerpo de los artículos y en los resúmenes se citan los niveles de significación sin mencionar siquiera las verdaderas concentraciones, proporciones, etc., o sus diferencias. Muchos deducen de las pruebas de hipótesis que siempre puede darse un simple "sí" o "no" como resultado fundamental de una investigación médica, lo cual es obviamente falso. La utilización de las pruebas de hipótesis con esa finalidad tiene un valor muy limitado (2).

Aquí comentamos la lógica en la que se basa un enfoque estadístico diferente, el uso de intervalos de confianza; estos intervalos son más informativos que los valores *P*, y nosotros recomendamos su uso en los artículos publicados en esta revista (*British Medical Journal*) u otras. No queremos decir con esto que tenga que haber intervalos de confianza en todos los artículos; en algunos casos, por ejemplo cuando los datos son puramente descriptivos, los intervalos de confianza no son apropiados y otras veces, obtenerlos exige una técnica muy compleja, o es imposible.

Presentación de los resultados de un estudio: limitaciones de los valores *P*

Las típicas aseveraciones "$P < 0{,}05$", "$P > 0{,}05$" o "P = NS" dan poca información sobre los resultados de un estudio y se basan en el consenso arbitrario de utilizar el nivel de significación estadística del 5% para definir dos posibles resultados: significativo o no significativo. Esto no sirve para casi nada y, además, favorece la vagancia intelectual. Incluso cuando se indica el valor *P* en concreto, no se proporciona información alguna sobre las diferencias entre los grupos de estudio. Rothman señaló esto en 1978, y a la vez defendió el uso de los intervalos de confianza (3); recientemente, dicho autor y sus colaboradores han repetido la propuesta (4).

Presentar valores *P* aislados puede hacer que se dé a estos valores más importancia de la que tienen. En concreto, se tiende a convertir la significación estadística en sinónimo de importancia médica o interés biológico. Pero pequeñas dife-

rencias sin interés real pueden ser estadísticamente significativas cuando el tamaño muestral es grande, mientras que efectos clínicamente importantes pueden no ser estadísticamente significativos solo porque el número de sujetos estudiados fue escaso.

Presentación de los resultados de un estudio: intervalos de confianza

Mucho más útil es presentar los valores muestrales (o "estadísticos") como estimaciones de los resultados que se obtendrían si se estudiara toda la población. La falta de precisión de un valor muestral —por ejemplo, la media— que depende tanto de la variabilidad del factor que se investiga como del tamaño limitado del estudio, puede ilustrarse muy bien mediante un intervalo de confianza.

Un intervalo de confianza nos lleva de un solo valor estimado —la media muestral, la diferencia entre las medias muestrales, etc.— a un recorrido de valores que se consideran plausibles para la población. La amplitud del intervalo de confianza basado en el valor muestral depende en parte del error estándar de ese valor, y de ahí que sea función tanto de la desviación estándar como del tamaño muestral (véase el apéndice 1 para una breve descripción de la importante diferencia conceptual —a menudo no comprendida— entre desviación estándar y error estándar). También depende del grado de "confianza" que queremos asociar con el intervalo resultante.

Supongamos que en un estudio en el que se comparan muestras de 100 varones diabéticos y 100 varones no diabéticos de cierta edad se encuentra una diferencia de 6,0 mm Hg de tensiones diastólicas medias y que el error estándar de esta diferencia de medias muestrales es de 2,5 mm Hg (datos similares a las diferencias entre medias detectadas en el estudio de Framingham (5)). El intervalo de confianza de 95% para la diferencia de medias poblacionales es de 1,1 a 10,9 mm Hg y se muestra en la figura 1 junto con los datos originales. La técnica de cálculo del intervalo de confianza consta en el apéndice 2.

En pocas palabras, esto significa que hay 95% de posibilidades de que el rango indicado incluya la diferencia de tensiones arteriales medias "de la población", es decir, el valor que se hubiera obtenido de haber incluido toda la población de diabéticos y no diabéticos a la que pretende referirse el estudio. Dicho más exactamente, en un sentido estadístico, el intervalo de confianza significa que, si se llevara a cabo una serie de estudios idénticos con diferentes muestras de la misma población, y se calculara un intervalo de confianza del 95% para la diferencia entre las medias muestrales calculadas en cada estudio, entonces, a la larga, 95% de esos intervalos de confianza incluirían la diferencia poblacional entre medias.

El tamaño muestral influye en la magnitud del error estándar y este, a su vez, influye en la amplitud del intervalo de confianza. Esto se ilustra en la figura 2, que muestra el intervalo de confianza de 95% para muestras con medias y desviaciones estándar idénticas a las anteriores, pero de tamaño igual a la mitad (50 diabéticos y 50 no diabéticos). Al disminuir el tamaño muestral se reduce la precisión y aumenta la amplitud del intervalo de confianza, en este caso cerca de 40%.

El investigador puede elegir el grado de confianza asociado con un intervalo de confianza, aunque la elección más habitual es 95%, igual que 5% es el nivel de significación estadística más usado. Cuando se requiere mayor o menor confianza pueden construirse otros intervalos; la figura 3 muestra intervalos de confianza del 99%, del 95% y del 90% para los datos de la figura 1. Como es lógico, para tener mayor confianza en que la diferencia poblacional se halla dentro de un intervalo se requieren intervalos mayores. En la práctica, casi nunca se usan intervalos con otros niveles de confianza que los de 99, 95 o 90%.

FIGURA 1. Tensiones arteriales sistólicas en 100 diabéticos y en 100 no diabéticos. Las medias son respectivamente 146,4 y 140,4 mm Hg. La diferencia de 6,0 mm Hg entre las medias muestrales se indica a la derecha junto con el intervalo de confianza de 95% que va de 1,1 a 10,9 mm Hg.

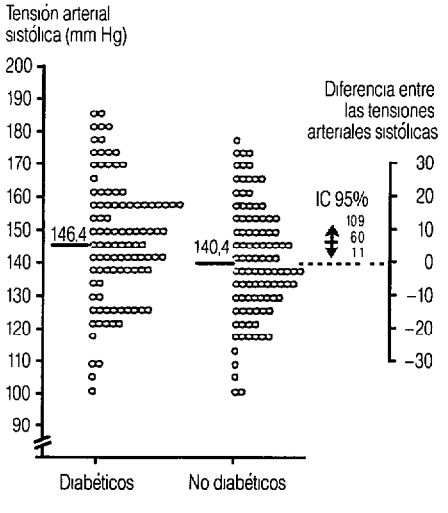


FIGURA 2. Esquema similar al de la figura 1, en este caso con los resultados de dos muestras de tamaño igual a la mitad de las muestras anteriores, es decir, 50 sujetos cada una. Las medias y desviaciones estándar son iguales a las de la figura 1, pero el intervalo de confianza de 95% es más amplio, de −1,0 a 13,0 mm Hg, debido al menor tamaño de las muestras

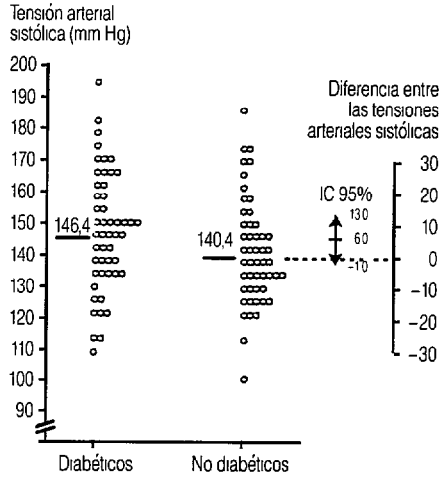
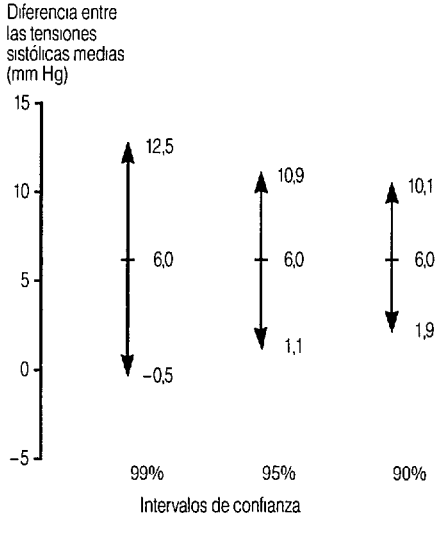


FIGURA 3. Intervalos de confianza asociados con diferentes grados de "confianza", usando los mismos datos de la figura 1



En el apéndice 2 se exponen algunos métodos para calcular intervalos de confianza para medias, proporciones y sus diferencias. También pueden calcularse intervalos de confianza para estadísticos tales como pendientes de rectas de regresión o riesgos relativos (*6*). Cuando los datos observados no pueden considerarse procedentes de una distribución normal [gausiana] el cálculo del intervalo de confianza puede no ser fácil (véase el apéndice 2).

Los intervalos de confianza solo dan cuenta de los efectos de la variación muestral sobre la precisión de valor estimado y no sirven para corregir sesgos de diseño, realización, análisis u otros errores ajenos al proceso de muestreo.

Intervalos de confianza y significación estadística

La determinación de un intervalo de confianza y la realización de una prueba bilateral de hipótesis son dos procedimientos estadísticos estrechamente relacionados. Cuando se determina el intervalo de confianza es posible deducir el resultado de la prueba de hipótesis al nivel correspondiente de significación estadística. La escala del lado derecho de la figura 1 incluye el punto que representa una diferencia igual a cero entre las tensiones arteriales medias de diabéticos y no diabéticos. Esta diferencia nula entre medias corresponde a la "hipótesis nula" y, tal como ilustra la figura 1, está fuera del intervalo de confianza de 95%. Esto indica que si se aplica una prueba *t* para datos no apareados, la diferencia entre las medias muestrales será estadísticamente significativa al nivel de 5%. Sin embargo, la figura 3 muestra que el valor *P* es mayor que 1% porque cero está dentro del intervalo de confianza de 99%, de manera que $0{,}01 < P < 0{,}05$. Por el contrario, si cero hubiera estado dentro del intervalo de confianza de 95%, el resultado habría sido no significativo al nivel de 5%. La figura 2 muestra un ejemplo similar para muestras de menor tamaño.

El intervalo de confianza de 95% cubre un amplio recorrido de posibles diferencias de las medias poblacionales, a pesar de que la diferencia muestral entre medias es distinta de cero a un nivel de 5% de significación estadística. El intervalo de confianza de 95% muestra en concreto que el estudio es compatible tanto con una pequeña diferencia de alrededor de 1 mm Hg como con una diferencia de hasta 10 mm Hg en las tensiones arteriales medias. Sin embargo, es mucho más probable que la diferencia entre las medias poblacionales se halle hacia la mitad del intervalo de confianza, no hacia los extremos. El intervalo de confianza es amplio, pero la mejor estimación de la diferencia poblacional es 6,0 mm Hg, la diferencia entre las medias muestrales.

Este ejemplo muestra la falta de precisión de la diferencia entre medias muestrales observadas como estimador del valor poblacional, lo cual queda claro en cada uno de los tres intervalos de confianza de la figura 3. También queda de manifiesto lo poco apropiado que es considerar la significación estadística aislada de las estimaciones numéricas.

El intervalo de confianza proporciona un recorrido de posibilidades para el valor poblacional, no una dicotomía arbitraria basada tan solo en la significación estadística. El intervalo proporciona más información, a expensas de la precisión del valor *P*. Sin embargo, el valor *P* real es útil cuando acompaña al intervalo de confianza, y lo mejor es dar ambos. No obstante, si hay que excluir alguno, es mejor eliminar el valor *P*.

Propuesta de estilo de presentación

Dicho con pocas palabras, lo único que proponemos es que se den intervalos de confianza en vez de errores estándar. Esto ha de animar a dejar a un lado el énfasis actual en la significación estadística. Para los resultados principales de una investigación, nuestra propuesta es que se dé una información estadística completa que incluya estimados muestrales, intervalos de confianza, estadísticos de prueba y valores *P*, suponiendo que detalles básicos tales como los tamaños muestrales y las desviaciones estándar ya se han dado previamente en el texto. Al exponer los resultados deben incluirse al menos todos los relativos a la o las hipótesis originales del estudio y todos los que consten en el resumen.

Para el ejemplo que comentamos, la presentación de los resultados en el texto podría ser como sigue:

> La diferencia entre las medias de las tensiones sistólicas de las muestras de diabéticos y no diabéticos fue de 6 mm Hg, con un intervalo de confianza de 95% de 1,1 a 10,9 mm Hg; el estadístico *t* de prueba fue 2,4, con 198 grados de libertad y un valor *P* asociado de 0,02.

Dicho brevemente:

> Media 6,0 mm Hg, IC 95% 1,1 a 10,9; $t = 2,4$, g.l. = 198, $P = 0,02$.

El valor *P* exacto obtenido de la tabla de valores *t* es 0,01732, pero una o dos cifras significativas son suficientes (2); esta *P* está comprendida entre 0,01 y 0,05, como podía deducirse de los intervalos de confianza previamente determinados. A menudo es necesario dar un recorrido para el valor *P* (por ejemplo, $0,3 < P < 0,4$), porque las tablas publicadas solo tienen unas pocas cifras significativas.

Los dos extremos de un intervalo de confianza son presentados a veces como límites de confianza. Sin embargo, el término "límite" sugiere que no hay nada más allá y, en ese sentido, puede desorientar, ya que, por supuesto, el valor poblacional no siempre está dentro del intervalo de confianza. Además esa práctica encierra el peligro de que uno u otro de los "límites" sea citado por sí solo, sin añadir los demás resultados, con las consecuencias nocivas subsiguientes. Por ejemplo, centrarse solamente en el extremo superior y no tener en cuenta el resto del intervalo de confianza daría una visión falsa de los resultados al exagerar las diferencias halladas. Inversamente, citar solo el extremo inferior podría subestimar incorrectamente la diferencia. El intervalo de confianza es, pues, preferible porque se enfoca en todo el recorrido de valores.

Puede usarse esta misma notación para presentar intervalos de confianza en los cuadros. Una columna encabezada con "intervalo de confianza del 95%" o "IC 95%" contendría en las filas los intervalos "1,1 a 10,9", "48 a 85", etc. Los intervalos de confianza también pueden representarse en las figuras, donde son preferibles al muy usado error estándar, que a menudo se muestra solo en una dirección a partir de la estimación muestral. Si pueden mostrarse también los datos individuales, lo cual suele ser posible en caso de pequeñas muestras, se proporciona aun más información. Así, en la figura 1, a pesar de la superposición considerable de los dos conjuntos de datos muestrales, el desplazamiento de la media queda ilustrado por el intervalo de confianza de 95%, que excluye cero. Para muestras apareadas, las diferencias individuales pueden representarse muy bien en un gráfico.

El ejemplo que aquí se ha dado de la diferencia de dos medias es bastante común. Las medias muestrales tienen cierto interés en sí mismas, pero las in-

ferencias realizadas a partir de un estudio tienen relación principalmente con la diferencia. Por lo tanto, dar intervalos de confianza para cada media por separado no es muy útil, porque esos intervalos no suelen indicar la precisión de la diferencia o su significación estadística (*7, 8*). La diferencia principal revelada por un estudio debe ser mostrada directamente y no vagamente a partir de medias (o proporciones) aisladas.

En los trabajos en los que las comparaciones estadísticas relacionadas con las hipótesis iniciales son pocas, es recomendable usar intervalos de confianza para todos los resultados. Cuando nos encontramos ante comparaciones múltiples, siempre surge el problema de la interpretación, ya que algunos intervalos de confianza excluirán el valor "nulo" —por ejemplo, una diferencia cero— simplemente por la propia variación muestral. Esto es lo mismo que ocurre cuando se calculan múltiples valores *P*, situación en la que probablemente no todas las diferencias estadísticamente significativas representan efectos reales (9). Es necesario tener buen juicio a la hora de analizar comparaciones múltiples; hay que tener en cuenta el número de comparaciones realizadas al calcular los intervalos de confianza y los valores *P*, para evitar confusiones, tanto de los autores como de los lectores (2).

Conclusiones

En nuestra opinión, el uso excesivo de las pruebas de hipótesis a expensas de métodos más informativos de interpretación de los datos es una forma insatisfactoria de evaluar y presentar los datos estadísticos de las investigaciones científicas. Preferimos el uso de intervalos de confianza, que presentan los resultados directamente en la escala de la magnitud medida. Proponemos una notación para intervalos de confianza que pretende que su significado resulte lo más claro posible.

Los intervalos de confianza, que también tienen su ligazón con el resultado de las pruebas de hipótesis, deben llegar a ser el método estándar de presentación de los datos estadísticos de los resultados principales.

Agradecimiento

Agradecemos la colaboración de la redacción del *British Medical Journal* en la elaboración de este trabajo y de las propuestas que contiene. También agradecemos su labor a las personas que leyeron amablemente y criticaron constructivamente el borrador del artículo; la Srta. Brigid Grimes realizó una encomiable labor de mecanografía.

Apéndice 1: Desviación estándar y error estándar

Cuando se presentan datos numéricos, independientemente de que se indique o no su significación estadística, a menudo se proporcionan datos estadísticos adicionales. Dos valores estadísticos muy citados son la desviación estándar y el error estándar, que a menudo se confunden (*10–14*).

La desviación estándar es una medida de la variabilidad individual del factor investigado, por ejemplo, las concentraciones de alcohol en sangre [alcoholemias] en una muestra de conductores; es un índice descriptivo. Por el contrario, el error

estándar es una medida de la incertidumbre de un valor muestral. Por ejemplo, el error estándar de la media indica la incertidumbre de la alcoholemia media de la muestra de conductores como estimador de la media de toda la población de conductores. La desviación estándar es de interés para indicar la variabilidad individual; el error estándar tiene interés para los valores estadísticos descriptivos tales como medias, proporciones, diferencias, pendientes de rectas de regresión, etc. (2).

El error estándar de un estadístico depende tanto de la desviación estándar como del tamaño muestral, y supone un reconocimiento implícito de que una muestra no sirve para determinar con exactitud el valor poblacional. De hecho, si se toma otra muestra en circunstancias idénticas, casi con toda seguridad obtendremos una estimación diferente del mismo valor poblacional. El estadístico muestral es impreciso y el error estándar es una medida de tal imprecisión. Por sí mismo, el error estándar tiene un valor limitado, pero es posible usarlo para generar un intervalo de confianza, que tiene una interpretación muy útil.

Apéndice 2: Métodos para determinar intervalos de confianza

A continuación se dan fórmulas para calcular intervalos de confianza (IC) para medias, proporciones y sus diferencias. En general, siempre se trata de sumar y restar al estadístico muestral un múltiplo de su error estándar (*EE*). El principio es válido para otros estadísticos, por ejemplo, coeficientes de regresión, pero no es un principio universal.

INTERVALOS DE CONFIANZA PARA MEDIAS Y SUS DIFERENCIAS

Los intervalos de confianza para medias se construyen usando la distribución *t*, si los datos tienen una distribución normal (gausiana). Para diferencias entre dos medias, la desviación estándar (*DE*) de ambos grupos de estudio debe ser similar. Esto está implícito en el ejemplo dado en el texto y en el ejemplo explicado a continuación.

Una sola muestra

El intervalo de confianza para la media poblacional se obtiene usando la media ($\bar{x}$) y el error estándar (*EE*) de una muestra de tamaño *n*. Para este caso, $EE = DE/\sqrt{n}$. De manera que el intervalo de confianza va de:

$$\bar{x} - (t_{1-\alpha/2} \times EE) \quad \text{a} \quad \bar{x} + (t_{1-\alpha/2} \times EE),$$

siendo $t_{1-\alpha/2}$ el valor de la tabla de valores *t* para $n - 1$ grados de libertad y una "confianza" de $100(1 - \alpha)\%$. Para un IC de 95%, α es 0,05, para un IC de 99%, α es 0,01, y así sucesivamente. Los valores *t* se hallan en las tablas de los textos de estadística o, también, en *Documenta Geigy* (*15*). Para un IC de 95%, el valor *t* es cercano a 2,0 cuando las muestras son de más de 20 individuos; para muestras más pequeñas, *t* es bastante mayor que 2,0.

Dos muestras

Muestras de elementos no apareados. El intervalo de confianza para la diferencia entre dos medias poblacionales se obtiene de manera similar. Supongamos que $\bar{x}_1$ y $\bar{x}_2$ son las dos medias muestrales, s_1 y s_2 las desviaciones estándar correspon-

dientes y n_1 y n_2 los tamaños muestrales. En primer lugar, necesitamos una estimación de la desviación estándar "combinada", que viene dada por la fórmula

$$s = \sqrt{\frac{(n - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Una vez obtenida esta estimación de la desviación estándar "combinada", el error estándar de la diferencia entre las dos medias muestrales será:

$$EE_{\text{dif}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

El intervalo de confianza es entonces de

$$\overline{x_1} - \overline{x_2} - (t_{1-\alpha/2} \times EE_{\text{dif}}) \quad \text{a} \quad \overline{x_1} - \overline{x_2} + (t_{1-\alpha/2} \times EE_{\text{dif}}),$$

siendo $t_{1-\alpha/2}$ el valor de la distribución t con $n_1 + n_2 - 2$ grados de libertad.

Si las desviaciones estándar difieren considerablemente no es adecuado un estimador combinado, a menos que pueda hallarse una transformación de escala adecuada. Por lo demás, obtener un intervalo de confianza es más complejo (*6*).

Muestras de elementos apareados. Este caso incluye las mediciones repetidas —por ejemplo, en momentos diferentes, o en circunstancias diferentes sobre los mismos sujetos— o las comparaciones de casos y controles emparejados. Para ese tipo de datos los extremos de los intervalos de confianza se determinan con las mismas fórmulas que se mencionaron en el caso de una sola muestra; en este caso $\bar{x}$ y *DE* son la media y la desviación estándar de las diferencias individuales entre sujetos o entre paciente y control.

Ejemplo resuelto: dos muestras de elementos no apareados

Se midió la tensión arterial en 100 varones diabéticos y en 100 no diabéticos de edades comprendidas entre 40 y 49 años. Las tensiones sistólicas medias fueron 146,4 mm Hg (*DE* 18,5) en los diabéticos y 140,4 mm Hg (*DE* 16,8) en los no diabéticos, lo cual representa una diferencia de 6,0 mm Hg entre ambas medias muestrales.

Usando las fórmulas antes mencionadas, calculamos el estimador combinado de la desviación estándar:

$$s = \sqrt{\frac{(99 \times 18{,}5^2) + (99 \times 16{,}8^2)}{198}} = 17{,}7 \text{ mm Hg},$$

y el error estándar de la diferencia entre las dos medias muestrales es:

$$EE_{\text{dif}} = 17{,}7\sqrt{\frac{1}{100} + \frac{1}{100}} = 2{,}50 \text{ mm Hg}.$$

Para calcular el IC 95%, el valor $t_{0,975}$ con 198 grados de libertad es 1,97. De manera que el IC 95% es de:

$$6,0 - (1,97 \times 2,50) \quad \text{a} \quad 6,0 + (1,97 \times 2,50),$$

es decir, de 1,1 a 10,9 mm Hg, como se ilustra en la figura 1.

Supongamos ahora que las muestras hubieran sido de tan solo 50 varones cada una y que las desviaciones estándar hubieran sido las mismas. La desviación estándar combinada seguiría siendo 17,7 mm Hg, pero el error estándar de la diferencia sería ahora:

$$EE_{\text{dif}} = 17,7 \sqrt{\frac{1}{50} + \frac{1}{50}} = 3,54 \text{ mm Hg.}$$

El valor $t_{0,975}$ con 98 grados de libertad es 1,98, y el IC 95% es el siguiente:

$$6,0 - (1,98 \times 3,54) \quad \text{a} \quad 6,0 + (1,98 \times 3,54),$$

es decir, de $-1,0$ a 13,0 mm Hg, como consta en la figura 2.

Para las muestras originales de 100 diabéticos y 100 no diabéticos los valores apropiados de $t_{0,995}$ y $t_{0,95}$ con 198 grados de libertad para calcular los IC de 99% y 90% son 2,60 y 1,65, respectivamente. Así, el IC 99% es:

$$6,0 - (2,60 \times 2,50) \quad \text{a} \quad 6,0 + (2,60 \times 2,50),$$

es decir, de $-0,5$ a 12,5 mm Hg (figura 3), y el IC 90% es:

$$6,0 - (1,65 \times 2,50) \quad \text{a} \quad 6,0 + (1,65 \times 2,50),$$

es decir, de 1,9 a 10,1 mm Hg (figura 3).

Tamaños muestrales e intervalos de confianza

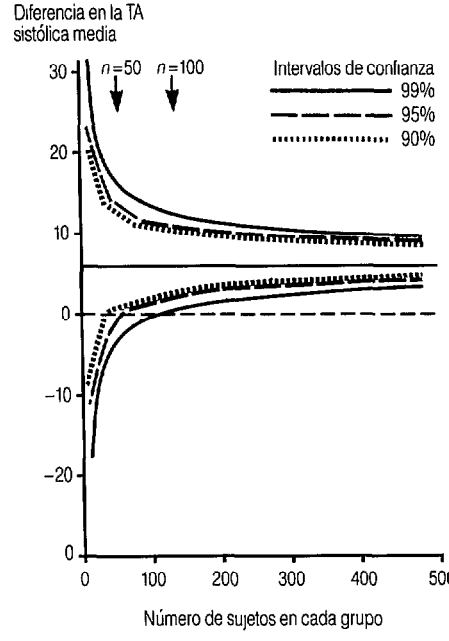
En general, el incremento del tamaño muestral reducirá la amplitud del intervalo de confianza. Si suponemos las mismas medias y desviaciones estándar que en el ejemplo, la figura 4 muestra los intervalos de confianza de 99%, 95% y 90% resultantes para la diferencia entre las tensiones arteriales medias para tamaños muestrales de hasta 500 en cada grupo. La reducción de la amplitud del intervalo de confianza que se obtiene cuando se incrementa aun más el tamaño muestral disminuye rápidamente a medida que aumenta dicho tamaño muestral.

Datos de distribuciones no gausianas

A veces hay que cambiar la escala de los datos muestrales para conseguir una distribución aproximadamente normal (gausiana). La razón más habitual es que la distribución de las observaciones está sesgada, con una "cola" muy larga hacia los valores altos. La transformación logarítmica es la más usada.

Para una sola muestra pueden calcularse la media y el intervalo de confianza con los datos transformados y luego transformar dichos resultados a la escala original de medida. Esto es preferible a presentar los resultados en unidades de, digamos, logaritmo de mm Hg. Cuando la distribución de las observaciones tiene mucho sesgo o los datos son de alguna manera problemáticos, la mediana a menudo es preferible a la media como medida de tendencia central. La mediana es compatible con métodos de análisis no paramétrico y pueden calcularse intervalos de confianza para ella (*15*).

FIGURA 4. Intervalos de confianza resultantes de medias y desviaciones estándar iguales a las de la figura 1, indicadas en el ejemplo resuelto explicado en el texto. La gráfica muestra el efecto del aumento del tamaño muestral hasta 500 sujetos en cada grupo. Las dos líneas horizontales indican una diferencia nula entre las dos medias (línea discontinua – – –) o una diferencia de 6,0 mm Hg entre las medias (línea continua ———) como la hallada en el estudio. Las flechas indican los intervalos de confianza representados en las figuras 1–3 correspondientes a tamaños muestrales de 100 y 50 en cada grupo.



TA = tensión arterial.

Cuando se trata de dos muestras, la única transformación apropiada es la logarítmica. Para muestras apareadas o no apareadas, el intervalo de confianza para la diferencia de las medias de los datos transformados tiene que transformarse de nuevo a la escala originaria. Cuando se usa una transformación logarítmica, el antilogaritmo de la diferencia de las medias muestrales en la escala transformada es un estimador de la razón de las dos medias (geométricas) poblacionales; el intervalo de confianza para dicha razón se obtiene tomando los antilogaritmos de los extremos del intervalo de confianza en la escala transformada.

INTERVALOS DE CONFIANZA PARA PROPORCIONES Y SUS DIFERENCIAS

Los intervalos de confianza para proporciones o diferencias entre dos proporciones se construyen de manera similar. Las fórmulas que se dan a continuación no deben usarse para pequeñas muestras, por ejemplo, menos de 50 en cada grupo, y tampoco para proporciones fuera del intervalo 0,1 a 0,9. Puede practicarse una corrección de continuidad (*16*), como se hace a veces para la prueba de χ^2 de la diferencia entre proporciones en una tabla 2 × 2.

Una sola muestra

Si p es la proporción observada de sujetos con cierta característica del total de una muestra de tamaño n, el error estándar de p es $EE = \sqrt{p(1 - p)/n}$. El intervalo de confianza del $100(1 - \alpha)\%$ va entonces de:

$$p - (N_{1-\alpha/2} \times EE) \quad \text{a} \quad p + (N_{1-\alpha/2} \times EE),$$

donde $N_{1-\alpha/2}$ es el valor correspondiente al percentil $100(1 - \alpha/2)$ tomado de la tabla de la distribución normal estandarizada, que consta en casi todas las tablas estadísticas corrientes. Para un intervalo de confianza de 95%, $N_{1-\alpha/2} = 1{,}96$; este valor no depende del tamaño muestral, como ocurre cuando se trata de intervalos de confianza para medias.

Dos muestras

Muestras de elementos no apareados. El intervalo de confianza para la diferencia entre dos proporciones poblacionales se construye en torno a la diferencia $p_1 - p_2$ entre las proporciones observadas en las dos muestras. El error estándar de $p_1 - p_2$ viene dado por la fórmula:

$$EE_{\text{dif}} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}.$$

El intervalo de confianza es entonces:

$$p_1 - p_2 - (N_{1-\alpha/2} \times EE_{\text{dif}}) \quad \text{a} \quad p_1 - p_2 + (N_{1-\alpha/2} \times EE_{\text{dif}}),$$

donde $N_{1-\alpha/2}$ se determina de la misma manera que en el caso de una sola muestra.

Muestras de elementos apareados. Supongamos que cada uno de los n sujetos de una muestra ha sido examinado dos veces para detectar la presencia o ausencia de cierta característica. Los datos pueden tabularse así:

Característica en el momento		
1	2	Número de sujetos
Presente	Presente	a
Presente	Ausente	b
Ausente	Presente	c
Ausente	Ausente	d
Total		n

Las proporciones de sujetos con la característica en cada una de las dos ocasiones son $p_1 = (a + b)/n$ y $p_2 = (a + c)/n$. La diferencia entre ambas será entonces $p_1 - p_2 = (b - c)/n$. El error estándar de esta diferencia es:

$$EE_{\text{dif}} = \frac{1}{n}\sqrt{b + c - \frac{(b - c)^2}{n}}.$$

El intervalo de confianza para $p_1 - p_2$ es entonces:

$$p_1 - p_2 - (N_{1-\alpha/2} \times EE_{\text{dif}}) \quad \text{a} \quad p_1 - p_2 + (N_{1-\alpha/2} \times EE_{\text{dif}}),$$

donde $N_{1-\alpha/2}$ se halla tal como se explicó en el caso de una sola muestra.

Ejemplo resuelto: dos muestras de elementos no apareados

Se estudió la respuesta al tratamiento entre 160 pacientes asignados al azar a uno de dos tratamientos, A y B. Los resultados fueron los siguientes:

	Tratamiento	
Respuesta	A	B
Mejoría	61	45
Sin mejoría	19	35
Total	80	80

Las proporciones que tuvieron mejoría fueron $p_A = 0{,}76$ y $p_B = 0{,}56$ (o sea, 61/80 y 45/80) para los tratamientos A y B, respectivamente, lo cual indica una proporción excedente de mejoría de 0,20 para el tratamiento A. En porcentaje, 76% de los pacientes sometidos al tratamiento A mejoraron, frente a 56% que mejoraron con el tratamiento B. Esto sugiere que un 20% adicional puede mejorar si se usa A en vez de B.

El error estándar de la diferencia $p_A - p_B = 0{,}20$ se calcula a partir de la fórmula para el caso de muestras no apareadas, es decir:

$$\sqrt{\frac{0{,}76 \times 0{,}24}{80} + \frac{0{,}56 \times 0{,}44}{80}} = 0{,}073.$$

De manera que el IC 95% es:

$$0{,}20 - (1{,}96 \times 0{,}073) \quad \text{a} \quad 0{,}20 + (1{,}96 \times 0{,}073),$$

es decir, de 0,06 a 0,34. Así, aunque 20% es el mejor estimador de la diferencia de porcentajes de pacientes que mejoran, el IC 95% va del 6% al 34%, lo cual muestra la imprecisión debida al pequeño tamaño de la muestra.

La habitual prueba de χ^2 para estos datos da un $\chi^2 = 7{,}16$, con un grado de libertad y un valor $P = 0{,}007$, de forma que el nivel de significación estadística es coherente con el IC 99% (siendo $N_{0,995} = 2{,}58$), que va de 0,01 a 0,39.

Observación técnica

Cuando se trata de datos cuantitativos y medias hay una correspondencia directa entre la metodología del intervalo de confianza y la de la prueba t de la hipótesis nula al nivel de significación estadística correspondiente. Sin embargo, no se puede decir exactamente lo mismo cuando se trata de datos cualitativos y proporciones. La razón es que para las pruebas habituales de hipótesis se usan estimadores del error estándar diferentes de los explicados aquí para la determinación de intervalos de confianza. La falta de correspondencia directa es pequeña y no implica cambios de interpretación. Además, a veces es posible obtener intervalos de confianza más exactos usando estimadores del error estándar del estadístico muestral en los mismos límites de confianza, tal como ha propuesto Cornfield para riesgos relativos (*17*).

Referencias

1. Mainland D. Statistical ritual in clinical journals: is there a cure? —I. *Br Med J.* 1984;288:841–843.
2. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J.* 1983;286:1489–1493.
3. Rothman K. A show of confidence. *N Engl J Med.* 1978;299:1362–1363.
4. Poole C, Lanes S, Rothman KJ. Analysing data from ordered categories. *N Engl J Med.* 1984;311:1382.
5. Kannel WS, McGee DL. Diabetes and cardiovascular risk factors: the Framingham study. *Circulation.* 1979;59:8–13.
6. Armitage P. *Statistical methods in medical research.* Oxford: Blackwell; 1971.
7. Browne RH. On visual assessment of the significance of a mean difference. *Biometrics.* 1979;35: 657–665.
8. Altman DG. Statistics and ethics in medical research: VI—presentation of results. *Br Med J.* 1980;281:1542–1544.
9. Jones DR, Rushton L. Simultaneous inference in epidemiological studies. *Int J Epidemiol.* 1982;11: 276–282.
10. Gardner MJ. Understanding and presenting variation. *Lancet.* 1975;i:230–231.
11. Feinstein AR. Clinical biostatistics XXXVII: demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clin Pharmacol Ther.* 1976;20:617–631.
12. Bunce H, Hokanson JA, Weiss GB. Avoiding ambiguity when reporting variability in biomedical data. *Am J Med.* 1980;69:8–9.
13. Altman DG. Statistics in medical journals. *Statistics in Medicine.* 1982;1:59–71.
14. Brown GW. Standard deviation, standard error: which "standard" should we use? *Am J Dis Child.* 1982;136:937–941.
15. Diem K, Lentner C, eds. *Documenta Geigy. Scientific tables.* 7ª ed. Basic: Geigy; 1970.
16. Fleiss JL. *Statistical methods for rates and proportions.* 2ª ed. Chichester: Wiley; 1981:29–30.
17. Breslow NE. Day NE. *Statistical methods in cancer research: volume 1 — the analysis of case-control studies.* Lyon: International Agency for Research on Cancer; 1980:133–134.

Nota a los lectores:

En la próxima sección de "Comunicación biomédica" del *Boletín* se presentarán las opiniones de otros autores sobre la polémica planteada en torno a la utilidad relativa de los valores *P* y de los intervalos de confianza.