■ Rev Méd Chile 2005; 133: 1111-1115

MEDICINA BASADA EN EVIDENCIA

Intervalos de confianza

Roberto Candia B, Gianella Caiozzi A.

Confidence intervals

En artículos previos hemos abordado cómo analizar en forma crítica la validez de un estudio de terapia[1-3] y cómo expresar los resultados con distintas medidas de efecto (riesgo absoluto, riesgo relativo, número necesario para tratar)[4]. Así, al momento de aplicar los resultados de un estudio, lo hacemos utilizando el número que se nos entrega, lo que conocemos como estimador puntual. Si el estudio se volviera a realizar en condiciones idénticas, pero con una nueva muestra, es probable que el resultado no sea exactamente igual, ya que el valor que se nos entrega es una aproximación del **valor real**. El **valor real** es el que se obtendría al aplicar la intervención a la **población completa**[5], entendiendo población como el total de pacientes idénticos a los del estudio dentro de la población general. Este es el valor que realmente nos interesa aplicar en la práctica clínica.

Utilizando los datos de un estudio podemos estimar un rango en el que se encuentra con alta probabilidad el valor real, y es precisamente este rango lo que conocemos como intervalo de confianza.

Este artículo pretende ayudar a los clínicos a comprender e interpretar un intervalo de confianza, su relación con el tamaño muestral y advertir las diferencias comparativas con el valor P.

INTERVALO DE CONFIANZA (IC): DEFINICIÓN Y PROPIEDADES

El intervalo de confianza describe la variabilidad entre la medida obtenida en un estudio y la medida real de la población (el **valor real**). Corresponde a un rango de valores, cuya distribución es **normal** y en el cual se encuentra, con alta probabilidad, el **valor real** de una determinada variable. Esta «alta probabilidad» se ha establecido por consenso en 95%. Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el **valor real** de un parámetro con 95% de certeza[5-8].

Para comprender y hacer intuitivo el concepto de intervalo de confianza utilizaremos un ejemplo clásico[6]:

Supongamos que tenemos una moneda, la cual puede o no estar balanceada. Así, después de varios lanzamientos, la probabilidad que el resultado sea sello variará desde 0 (todas las veces cara, es decir, una moneda balanceada) hasta 1 (todas las veces sello, nuevamente balanceada), pasando por 0,5 (la mitad de las veces sello y las otras cara, lo que equivale a una moneda no balanceada). Como no conocemos la verdadera naturaleza de la moneda, vamos a experimentar con ella.

Recibido el 2 de junio, 2005. Aceptado el 5 de julio, 2005.
Unidad de Medicina Basada en Evidencia, Pontificia Universidad Católica de Chile. Santiago, Chile.

Iniciamos el experimento con 2 lanzamientos, uno es cara y el otro es sello. La probabilidad de que el resultado sea sello fue 0,5, con lo que podríamos concluir que la moneda no está balanceada, sin embargo, ¿con sólo 2 lanzamientos podemos concluir con total certeza que esa es la naturaleza de la moneda? La respuesta es no, por lo tanto ¿cuál es el rango de valores donde se encuentra el valor real? Dado que el azar pudo influir en este resultado, uno acepta que el rango de valores reales posibles es amplio, incluso desde uno tan bajo como 0 a uno tan alto como 1, por lo tanto aún no estamos seguros de la naturaleza de nuestra moneda.

Considerando lo anterior, ampliamos el experimento y realizamos 8 nuevos lanzamientos (10 en total), resultando 5 caras y 5 sellos. Nuevamente el resultado es 0,5, sin embargo, ahora intuitivamente nos percatamos que la verdadera naturaleza de la moneda se encuentra en un rango menos amplio. Por ejemplo, es poco probable que después de 10 lanzamientos 9 sean sello, menos aún que todos lo sean, sin embargo, aún es factible que 8 ó 7 ó 6 sí lo sean. Así, nuestro nuevo rango puede variar entre 0,2 y 0,8, pero con un alcance: todos advertimos que si bien 0,8 y 0,2 son posibles, los valores centrales (0,4 y 0,6) lo son más aún, siendo 0,5 el más probable.

Decidimos seguir experimentando, realizando 90 nuevos lanzamientos (100 en total), resultando 50 caras y 50 sellos. Nuevamente el resultado es 0,5, advirtiendo que cada vez es más probable que la verdadera naturaleza de nuestra moneda es el de una no balanceada, pero aún con un rango de variabilidad que podríamos estimar entre 0,4 y 0,6 (es decir, que después de 100 lanzamientos, el resultado real varíe entre 40 y 60 sellos).

Realizamos 1.000 lanzamientos, resultando 500 sellos y 500 caras, con lo que estamos aún más seguros que nuestra moneda no está balanceada (nuestro rango puede ser 0,45 a 0,55 o menor).

El ejemplo anterior nos permite aclarar varios conceptos:

- La «verdadera naturaleza» de nuestra moneda (si está balanceada o no) corresponde al **valor real.**
- El rango de valores reales posibles, es decir, el rango donde se encuentra la verdadera naturaleza de nuestra moneda, corresponde al IC.
- El valor real más probable corresponde al estimador puntual del estudio, en este caso 0,5.
- Finalmente, advertimos la relación inversa entre la amplitud del IC y el tamaño muestral: si consideramos que el número de lanzamientos representa el n de la muestra, observamos que mientras más pequeño es el n más amplio es el IC. A mayor número de lanzamientos (mayor n) más certeza tenemos que el resultado del experimento se acerca al valor real, por lo tanto el IC es más estrecho[5-8].

Para llevar a la práctica el concepto vamos a recurrir al ejemplo utilizado en el artículo anterior: la comparación de una nueva droga A versus una droga B en la prevención de AVE en pacientes con antecedente de accidente isquémico transitorio (AIT) (Tabla 1)[4].

Tabla 1. Resultados de un ensayo clínico hipotético[4]

Tratamiento	AVE		Total
	Sí	No	
Droga A	10	250	260
Droga B	42	480	522
Total	52	730	782

Al analizar estos datos se obtiene una reducción absoluta del riesgo (RRA) de 4,2% con 95% de intervalo de confianza de 0,9% a 7,5%. Esto quiere decir que el **valor real**, es decir, el resultante al aplicar la intervención a la población total de pacientes con AIT, está con 95% de probabilidad entre un RRA de 0,9% a 7,5%, siendo el valor más probable 4,2%. Si aumentamos el n de la muestra a 20.000 obtendríamos nuevamente un RRA de 4,2%, pero con un intervalo de confianza más estrecho, de 3,5% a 4,9% (Fórmula en apéndice 1).

Apéndice 1. Fórmula de intervalo de confianza:

$$\text{Estimador puntual} \pm 1{,}96 \times \sqrt{\frac{p1\,(1-p1)}{n1} + \frac{p1\,(1-p2)}{n2}}$$

Donde:
p1 Tasa de eventos grupo 1
p2 Tasa de eventos grupo 2
n1 n grupo 1
n2 n grupo 2

Interpretación de un IC

El intervalo de confianza es una medida de precisión que permite al clínico evaluar 2 aspectos de un resultado (estimador puntual):

1. Si existe diferencia estadística significativa.
2. Si tal diferencia es relevante para recomendarla a mis pacientes (relevancia clínica).

Para analizar si existe o no diferencia estadística significativa debemos observar los extremos del IC. Independiente si el estimador puntual muestra beneficio o daño, debemos verificar si alguno de los extremos del IC pasa sobre la línea del no efecto. Si es así, existe la posibilidad de que el valor real corresponda al no efecto o incluso tenga un efecto opuesto al esperado. En este caso no existiría diferencia estadísticamente significativa entre aplicar o no la intervención (Figura 1)[7,8].

Cuando un estudio demuestra un efecto con significación estadística (es decir el extremo del IC no cruza ni toca la línea del no efecto), el clínico debe definir cuál es el beneficio mínimo necesario para recomendar la terapia, lo que llamaremos **umbral**. Así, nuestro estudio hipotético demuestra beneficio estadístico significativo, siendo el beneficio mínimo probable un RRA de 0,9%. El que este beneficio tenga relevancia clínica depende del tipo de evento prevenido o favorecido, los efectos adversos de la droga A v/s la droga B, el costo, las circunstancias clínicas, etc. Si el evento a prevenir es banal, o si la droga A tiene muchos efectos adversos y es más cara que B, nuestro umbral va a ser alto, por lo tanto el beneficio demostrado en nuestro estudio no sería relevante[7,8] (Figura 2).

Al contrario, si el evento a prevenir es relevante en sí mismo (por ej: mortalidad o invalidez), o si la nueva droga es más barata y sin efectos adversos, tal vez con demostrar un RRA de sólo 0,5% nos basta para recomendarla (umbral), por lo tanto nuestro estudio no sólo demuestra diferencia estadísticamente significativa, sino que también beneficio relevante para el paciente (Figura 3).

Así, para evaluar beneficio clínico, primero debemos establecer un umbral mínimo de beneficio, el que depende del tipo de evento a prevenir o favorecer los efectos adversos, costos, etc. de la nueva droga, y luego observar el beneficio mínimo probable que muestra el estudio, que corresponde

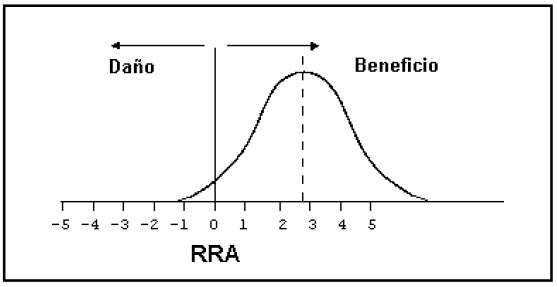

Figura 1. Estudio hipotético cuyo estimador puntual informa un RRA 2,8%, pero cuyo IC sobrepasa la línea del no efecto, por lo tanto es posible que el valor real sea daño. No existe diferencia estadística significativa en este estudio.

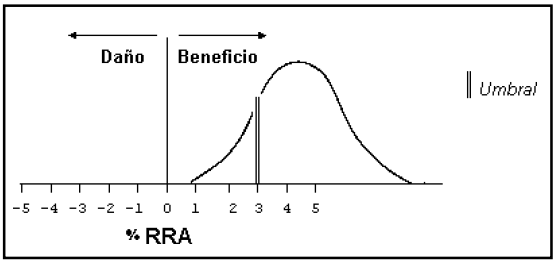

Figura 2. Estudio hipotético que informa beneficio estadístico significativo, sin embargo, el IC pasa sobre el beneficio mínimo necesario para recomendar la terapia (umbral, RRA 3%). El beneficio mínimo demostrado (RRA 0,9%) no es suficiente para recomendar la terapia.

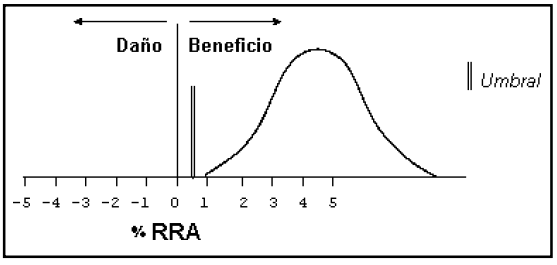

Figura 3. Estudio hipotético que informa beneficio estadístico significativo. El IC no sobrepasa el beneficio mínimo necesario para recomendar la terapia (umbral, RRA 0,5%). El beneficio mínimo demostrado (RRA 0,9%) es suficiente para recomendar la terapia.

al extremo del IC más cercano a la línea del no efecto. Si el extremo del IC no sobrepasa el umbral se asume que el beneficio mínimo probable es suficiente para recomendar la nueva terapia.

Existe la posibilidad que la nueva droga hiciese daño (RRA negativo). El proceso es similar al anterior, estableciendo un umbral máximo de daño tolerable, y observando el extremo del IC que más se acerca a la línea del no efecto. Si la nueva droga genera más daño con una diferencia estadísticamente significativa, debemos observar si el extremo del IC sobrepasa ese umbral. Si no lo hace se asume que el daño mínimo probable es más alto que lo tolerable, por lo tanto se está en condiciones de rechazar la nueva terapia[7,8] (Figura 4).

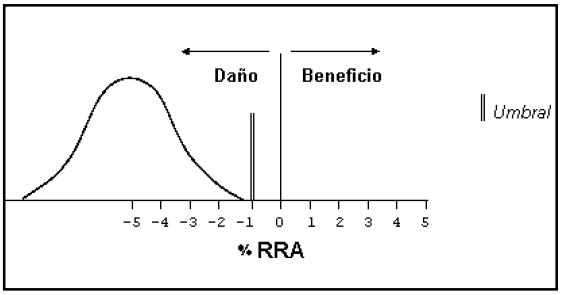

Figura 4. Estudio hipotético que informa daño estadístico significativo. El IC no sobrepasa el daño mínimo establecido como umbral. El daño mínimo demostrado es suficientemente importante para rechazar la terapia.

Valor P

Al comparar dos grupos en un estudio podemos demostrar que no existe diferencia entre ambos (**hipótesis nula**) o que sí la hay (**hipótesis alternativa**)[9,10]. El valor P es un test de hipótesis que nos ayuda a afirmar con cierto nivel de seguridad (por consenso se usa 95%, que se expresa como P <0,05) que una de las hipótesis es la correcta. Para nuestro ejemplo, la hipótesis nula corresponde a la igualdad de resultados al usar la droga A o B, mientras que la hipótesis alternativa supone que una de ellas es mejor que la otra en prevenir la enfermedad.

El valor P representa la probabilidad que una diferencia observada entre 2 grupos sea sólo debida al azar, es decir, la probabilidad que la hipótesis nula sea verdadera a pesar de observar diferencia en un estudio[7-9]. Como toda probabilidad, puede tener valores desde 0 a 1. Valores más cercanos a 1 indican que existe una alta probabilidad que las diferencias observadas sean sólo por azar, es decir, apoya la hipótesis nula. En cambio, valores más cercanos a 0 apoyan la hipótesis alternativa.

Apliquemos este concepto a nuestro ejemplo, en que se obtiene un RRA de 4,2% con un valor P <0,05 (p=0,039). Si asumimos como valor real que la droga A es igual a B (hipótesis nula) y pudiéramos repetir el estudio muchas veces, el P <0,05 nos dice que en menos de 5% de las ocasiones se observaría tal diferencia entre ambas, sólo por azar. Dicho de otra forma, en la mayor parte de las ocasiones la diferencia observada no se debe al azar, por lo tanto rechazamos la hipótesis nula y establecemos que existe diferencia estadística significativa[9,10].

El valor P se correlaciona en forma muy estrecha con el intervalo de confianza, ya que si uno muestra diferencia estadística significativa el otro también lo hace, y viceversa. Sin embargo, el valor P, a diferencia del IC, no nos entrega información respecto al rango en el que se encuentra la magnitud del efecto de un determinado tratamiento (valor real), por lo que sólo nos habla de diferencias estadísticas significativas, sin permitirnos evaluar si esta diferencia es relevante para mi paciente. Por ejemplo, un resultado significativo (P <0,05) podría incluir diferencias clínicamente irrelevantes, y resultados no significativos (P >0,05) podrían esconder una diferencia clínicamente importante entre 2 tratamientos si el estudio no incluye un tamaño muestral adecuado (un estudio con bajo poder puede no mostrar una diferencia que realmente sí existe)[8].

De esta forma, aunque el valor P mide la fuerza de una asociación, siempre es útil el intervalo de confianza para complementar la evaluación de la magnitud del efecto de una intervención y poder realizar una interpretación adecuada de los resultados de un estudio.

Conclusiones

Al leer un estudio es muy importante interpretar los resultados en forma correcta. Esto supone

comprender el significado del estimador puntual y de sus medidas de precisión, lo que permite extrapolar los datos a la población de interés. Tanto el análisis de un intervalo de confianza como el de un valor P nos permiten determinar diferencias estadísticas significativas, sin embargo sólo el IC nos permite evaluar el rango de valores donde posiblemente se encuentra el valor real, y por lo tanto, permite realizar una mejor interpretación y aplicación clínica de los resultados.

Referencias

1. Peñaloza B, Candia R. ¿Por qué vale la pena randomizar un estudio de terapia? *Rev Méd Chile* 2004; 132: 1007-14.
2. Letelier LM, Manríquez JJ, Claro JC. El «ciego» en los ensayos clínicos ¿importa? *Rev Méd Chile* 2004; 132: 1137-9.
3. Capurro D, Gabrielli L, Letelier LM. Importancia de la intención de tratar y el seguimiento en la validez interna de un estudio clínico randomizado. *Rev Méd Chile* 2004; 132: 1557-60.
4. Rivera S, Larrondo F, Ortega J. Evaluación de resultados de un artículo de tratamiento. *Rev Méd Chile* 2005; 133: 593-6.
5. Whitley E, Ball J. Statistics review 2: Samples and populations. *Critical Care* 2002; 6: 143-8.
6. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians 2. Interpreting study results: confidence intervals. *CMAJ* 1995; 152: 169-73.
7. Montori V, Kleinbart J, Newman T, Keitz S, Wyer P, Moyer V, Guyatt G. Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *CMAJ* 2004; 171: 611-5.
8. Montori V, Kleinbart J, Newman T, Keitz S, Wyer P, Moyer V et al. Tips for teachers of evidence-based medicine 2: Confidence intervals and p values. *CMAJ* 2004; 171, Online 1-12 (http://www.cmaj.ca/cgi/data/171/6/611/DC1/1).
9. Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. *Critical Care* 2002; 6: 222-5.
10. Shakespeare T, Gebski V, Veness M, Simes J. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *Lancet* 2001; 357: 1349-53.