



Falacias y limitaciones de la significación estadística: ¿qué significa realmente $p < 0,05$?

Autora: Ana Belén Salamanca Castro 

* **Dirección de contacto:** nureinvestigacion@fuden.es

Diplomado y Grado en Enfermería. Máster en Cuidados Perinatales y la Infancia. Máster en Salud y Género. Experto en Metodología de la Investigación en Ciencias de la Salud. Directora de la revista NURE Investigación.

Si hay alguna expresión que impregna los artículos originales de los estudios enmarcados en el paradigma positivista es la de “estadísticamente significativo”. La significación estadística se mide mediante el valor de p (p -valor) y, en Ciencias de la Salud se ha consensuado, de forma arbitraria, que la probabilidad de cometer un error al afirmar que existe una diferencia estadísticamente significativa debe ser menor del 5% (el consabido $p < 0,05$), pero: ¿obtener una $p < 0,05$ es una garantía de calidad de los resultados?, ¿qué implica, para la práctica clínica, que un resultado sea estadísticamente significativo?, o, en esa misma línea nos podríamos preguntar: ¿un menor valor de p significa una mayor asociación o relación causal? En este y el próximo editorial intentaré explicar estos aspectos y expondré algunas reflexiones sobre las garantías y falacias, o las bondades y limitaciones del p -valor y por qué lleva décadas siendo cuestionado.

Comencemos dando unos pasos hacia atrás para poder responder estas preguntas:

Cuando seleccionamos una muestra de una población, obtenemos unos resultados que describen lo que ha ocurrido en esa muestra pero, generalmente, nuestro objetivo es generar conocimiento y, por tanto, ir más allá de la muestra estudiada. Nuestro objetivo, en este tipo de diseños, es generalizar, inferir, los hallazgos de la muestra a la población a la que pertenece y representa (esto último requiere que la muestra se haya seleccionado mediante técnicas probabilísticas y tenga un tamaño adecuado, también calculado mediante métodos estadísticos).

Es en este salto de la muestra a la población, en este paso de los datos hallados en la muestra a los que se estiman en la población, cuando pasamos de la estadística descriptiva (con sus medidas de tendencia central y de

desviación) a la estadística inferencial (con sus intervalos de confianza y sus contrastes de hipótesis) y es en este último caso cuando entra en juego el p -valor, puesto que es el fruto del análisis que permite poner a prueba a la hipótesis nula (que postula que no existen diferencias) y, en base al resultado de este cálculo, dirimir si existen o no diferencias desde un punto de vista estadístico.

Esta toma de decisiones, como Macía et al. (1) afirman se asemeja mucho al desarrollo de un juicio (en este caso, a quien se juzga es a la hipótesis nula). Siguiendo a Macía et al., de la misma forma que en un juicio se recogen pruebas para probar que el acusado es culpable (de partida, en un estado de derecho se presupone la inocencia), en el juicio que supone el contraste de hipótesis, las pruebas recogidas (los datos hallados en la muestra) se destinan a probar que la hipótesis es nula (ya que se presupone que lo es). Tras el desarrollo del juicio, la decisión que se toma al finalizar el juicio lo que permite es decir que el acusado no es culpable (porque es la culpabilidad lo que se ha intentado comprobar, no la inocencia). Si el acusado no es culpable, se concluye que entonces es inocente y por tanto, queda en libertad ya que no hay pruebas que prueben su culpabilidad. En la analogía, que Macía et al. brillantemente establecen, la decisión que se toma tras el contraste de hipótesis permite decidir que la hipótesis nula no es cierta (porque es esto lo que se ha contrastado en el análisis estadístico, no la veracidad de la hipótesis alternativa). Si la hipótesis nula no es cierta, se concluye entonces que la cierta es la hipótesis alternativa y se acepta que existen diferencias estadísticamente significativas, ya que los datos hallados en la muestra nos sitúa en la zona de rechazo de la hipótesis nula (ver Figura 1).

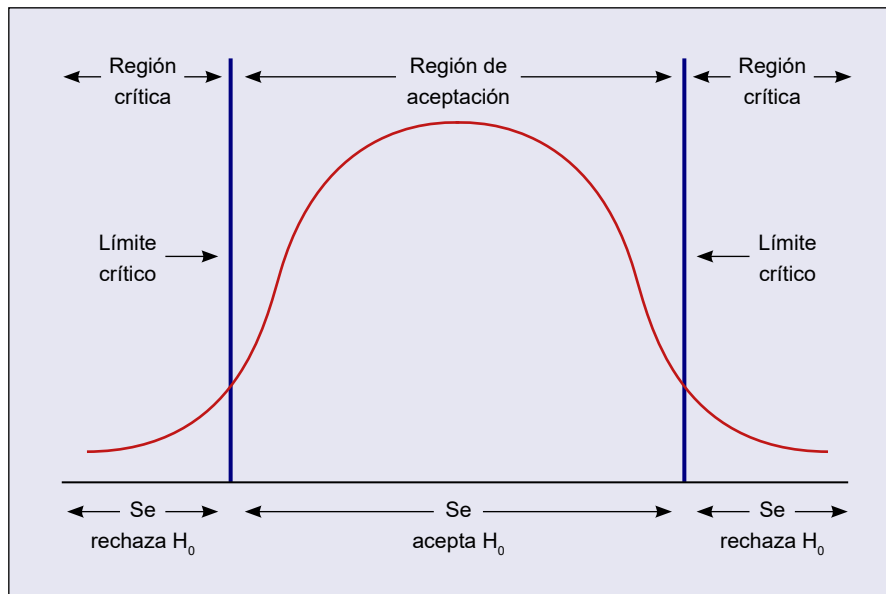


Figura 1. Zonas de aceptación y rechazo de la hipótesis nula.
Fuente: Salamanca Castro AB. Fase empírica de la investigación. p. 404.

Rechazar la hipótesis nula supone, como ya se ha explicado, que podemos asegurar que la probabilidad de que los resultados obtenidos en la muestra fueran diferentes en otra muestra seleccionada en la población de la que procede es menor del 0,05 (o, en términos de porcentajes, del 5%). Esto significa que, al menos en el 95% de los casos, obtendríamos el mismo resultado y sólo en un 5% de las posibles muestras seleccionadas podríamos obtener un resultado diferente.

¿Por qué nos movemos en términos de probabilidades? Porque no estamos decidiendo directamente (declarar inocente al acusado o probar que la hipótesis alternativa es cierta) sino a partir de negar lo contrario (el acusado no es culpable o la hipótesis nula no se acepta) y ello hace que podemos equivocarnos en esa toma de decisiones. De igual manera que, por error, podríamos condenar a un inocente o dejar libre a un culpable, en el símil del juicio; podríamos cometer el error tipo I o el tipo II, en el análisis que supone el contraste de hipótesis.

		Diferencias en la población (N)	
		SI	NO
Diferencias en la muestra (n)	SI	Existen diferencias (no hay error)	Error tipo I
	NO	Error tipo II	No existen diferencias (no hay error)

Tabla 1. Tipos de errores en la inferencia estadística.
Fuente: Salamanca Castro AB. Fase empírica de la investigación. p. 410.

Por tanto, el p-valor, ese $p < 0,05$, nos informa de la probabilidad de que nuestras inferencias a la población sean fruto del azar, y por ello, pudieran ser erróneas; o dicho de otro modo, nos informa de la probabilidad de que no hayamos aceptado (hayamos rechazado) la hipótesis nula erróneamente (puesto que deberíamos haberla aceptado), que es el denominado error tipo I (tabla 1). Como Macía et al. Indica, en el símil del juicio, equivaldría a condenar a un inocente (1).

En Ciencias de la Salud las consecuencias de este error (tipo I) pueden ser graves; puesto que al cometer este error inferimos que sí existen diferencias significativas (que las diferencias observadas no son debidas al azar), cuando realmente, sólo han existido en nuestra muestra, pero no existen en la población (y, por tanto, no se observarían esas diferencias si hubiésemos seleccionado otra muestra). Si comentemos el error tipo I nos estamos equivocando al afirmar que existe una asociación o relación de causalidad entre las variables porque, en base a unos resultados que realmente son fruto de la casualidad, estamos. Por eso se estableció que, en Ciencias de la Salud, el p-valor, la probabilidad de cometer ese error no podía ser mayor del 5%, en términos de porcentaje (o, en términos probabilísticos, del 0,05).

Como se puede deducir fácilmente, el error tipo I puede reducirse al aumentar el tamaño muestral (ya que cuanto mayor sea la muestra, menor será la influencia que pueda tener el azar en los datos obtenidos), y por tanto, la probabilidad de cometer el error tipo I se reduce (el p-valor). Llegamos entonces a una de las principales limitaciones o falacias del p-valor: si la muestra es muy grande (mayor de lo que se haya estimado a través de cálculo del tamaño muestral), encontraremos diferencias estadísticamente significativas, puesto que si la muestra es excesivamente grande, la probabilidad de que el azar distorsione los resultados obtenidos será muy reducida. Por este motivo es importante calcular un tamaño muestral conforme a las asunciones de las que debemos partir en función del tipo de diseño que se desarrolle y ajustarse a este tamaño estimado, puesto que si la muestra es mucho menor, no vamos a poder garantizar la representatividad de la muestra o podríamos no encontrar diferencias aunque realmente existan (cometeríamos entonces, el error tipo II) y, si es muy superior, vamos a obtener, sí o sí, diferencias estadísticamente significativas.

Por tanto, en respuesta a la primera pregunta formulada: ¿obtener una $p < 0,05$ es una garantía de calidad de los resultados?, podemos responder que no. Un valor de $p < 0,05$, como espero que haya quedado claro, nos informa que la probabilidad de cometer el error tipo I (no aceptar o rechazar la hipótesis nula siendo esta cierta) se encuentra dentro del mínimo establecido por consenso para las Ciencias de la Salud y que por tanto, obtener ese valor de p nos permite afirmar que existen diferencias estadísticamente significativas entre las variables que se hayan analizado. No obstante, si realizamos una buena valoración crítica de

los resultados, debemos considerar también cuál ha sido el tamaño muestral (para asegurarnos de que el valor obtenido no haya sido fruto de una estratagema estadística).

Por otro lado, los estudios denominados negativos (que son aquellos en los que se ha aceptado la hipótesis nula, porque el p-valor obtenido nos sitúa en la zona de aceptación de la hipótesis nula) también pueden ser considerados de calidad, puesto que conocer qué variables no se encuentran relacionadas también tiene un importante valor informativo. Es cierto que, en ocasiones, estos estudios no son publicados; bien porque los propios autores no lo envían a ninguna revista al sentirse frustrados por no haber podido aceptar su hipótesis de partida (la hipótesis alternativa) o bien porque las revistas no suelen publicar este tipo de artículos (no es el caso de *Nure Investigación*). Unos y otros debemos considerar que si el estudio tiene validez interna y la metodología y el análisis de los datos ha sido adecuada, los resultados pueden ser considerados de calidad y relevantes, independientemente del p-valor obtenido.

REFERENCIAS BIBLIOGRÁFICAS

1. Macía Antón MA, Moreno González E, Reales Avilés JM, Rodríguez MIñón Cifuentes P, Villarejo Rivas A. Estimación de parámetros y contraste de hipótesis. En: Macía Antón MA, Moreno González E, Reales Avilés JM, Rodríguez MIñón Cifuentes P, Villarejo Rivas A. Diseños de investigación y análisis de datos. Sanz y Torres; 2014.
2. Salamanca Castro AB. Fase empírica de la investigación. En: Salamanca Castro AB. El aeiou de la investigación en Enfermería. 2a ed. Fuden; Madrid: 2018.