En búsqueda de una p < 0,05: el problema de las comparaciones múltiples

María E. Knorre, Mariano Bergier y Arturo M. Cagide
Servicio de Cardiología, Hospital Italiano. Buenos Aires, Argentina

RESUMEN

El artículo analiza el papel del azar y la estadística en la investigación médica a partir del concepto de significación estadística y el valor de *p*. Si bien la aleatorización en los ensayos clínicos permite controlar el azar, el riesgo de obtener resultados falsamente positivos persiste cuando se realizan múltiples comparaciones sin los ajustes estadísticos correspondientes.

Esta problemática se ilustra mediante una analogía con una apuesta ganadora en el casino, en la que se presenta un resultado exitoso sin considerar los intentos fallidos previos, situación comparable al análisis *post hoc* sin corrección. Se describen distintas estrategias para reducir este sesgo, como la definición *a priori* de hipótesis y puntos finales, la aplicación de correcciones estadísticas por comparaciones múltiples y la validación de los hallazgos en cohortes independientes. Finalmente, se destaca que la inteligencia artificial tampoco está exenta de este problema si no se aplican criterios estadísticos adecuados.

Palabras clave: análisis estadístico, significación estadística, análisis de varianza, investigación clínica, inteligencia artificial.

In Search of a p < 0.05: The Problem of Multiple Comparisons

ABSTRACT

This article examines the role of chance and statistics in medical research, focusing on statistical significance and the p-value. Although randomization in clinical trials helps control randomness, the risk of false positive findings persists when multiple comparisons are performed without appropriate statistical adjustment.

This issue is illustrated through an analogy involving a winning casino bet, in which a successful outcome is reported without disclosing previous failed attempts, a situation comparable to uncorrected post hoc analyses. Strategies to mitigate this bias are discussed, including the a priori definition of hypotheses and endpoints, the application of statistical corrections for multiple testing, and the validation of findings in independent cohorts. Finally, the article emphasizes that artificial intelligence is also susceptible to this problem when data are not properly managed.

Keywords: statistics as topic, statistical significance, analysis of variance, clinical trials as topic, artificial intelligence.

Autora para correspondencia: eugenia.knorre@hospitalitaliano.org.ar, Knorre ME.

Recibido: 1/12/2025 | Aceptado: 28/04/2025 | Publicado: 9/09/2026

DOI: http://doi.org/10.51987/rev.hosp.ital.b.aires.v46i3.1301

Cómo citar: Knorre ME, Bergier M, Cagide AM. En búsqueda de una p < 0,05: el problema de las comparaciones múltiples. Rev Hosp Ital B.Aires. 2026;46(3):e0001301

ISSN (en línea) 2314-3312

INTRODUCCIÓN

¿Existen puntos de contacto entre el azar y la medicina? En principio la respuesta es negativa.

En medicina, las decisiones tienen un fundamento lógico basado en evidencias y en la experiencia del profesional. El azar solo interviene cuando, en un ensayo de investigación experimental en forma aleatoria, se asigna el tratamiento o el control que recibirán los grupos en comparación. La base conceptual es que el azar asigna por igual aquellas variables que pudieran afectar la evolución, de modo que la única diferencia es el tratamiento, objetivo del estudio. Bajo esta premisa, una $p < 0,05$ indica que la posibilidad de que la conclusión resulte falsamente positiva es de solo *5* en 100.

Vamos ahora a contraponer las consideraciones anteriores con la siguiente experiencia de la vida real.

En un grupo de amigos de vacaciones, uno de ellos comenta que la noche anterior concurrió al casino donde obtuvo una ganancia suculenta apostando un pleno al número 18. Sus amigos sorprendidos preguntan detalles. El afortunado dice, "ocurre que..."

La anécdota, perfectamente creíble, queda condicionada por una sutileza inicialmente no comentada.

DESARROLLO

Sobre bolilleros y probabilidades (conceptos estadísticos)

Supongamos que en un bolillero se dispone de 1000 bolillas, cada una rotulada con una determinada diferencia, signo positivo o negativo; y cuya *distribución de frecuencias* corresponde a una curva de tipo campana, de modo que el valor medio de esas diferencias es igual a cero (Fig. 1). El sector sombreado representa el *5%* de las diferencias extremas (*50* bolillas) 2,*5%* con valores negativos (*25* bolillas) y 2,*5%* con positivos (*25* bolillas).

La posibilidad de obtener una bolilla con una diferencia, negativa o positiva de las existentes en el sector negro, es de solo el *5%*. Podemos incluir la bolilla en el bolillero y efectuar una nueva extracción repitiendo el experimento 100 veces, al cabo de los cuales habríamos obtenido *5* bolillas con la diferencia de las contenidas en el sector sombreado de la curva.

Del bolillero a la estadística inferencial aplicada

Dejemos el azar y vayamos a la investigación clínica[1-4].

Supongamos que un estudio observacional pretende estimar el valor de un indicador diagnóstico, pronóstico o el resultado de un tratamiento, o, un ensayo con diseño aleatorizado, pretende estimar el efecto de una determinada intervención. En ambos casos se están comparando dos muestras diferenciadas con la variable objeto del estudio (*punto final*) para demostrar si existe o no una "diferencia estadísticamente significativa".

Partiendo del criterio de que no hay diferencias (*hipótesis nula*), las probables diferencias halladas en cuanto al punto final entre ambas muestras tendrían una distribución (*distribución de frecuencias*) como la ilustrada en la figura 2 (similar a la Fig. 1) con un valor medio (*media*) de cero.

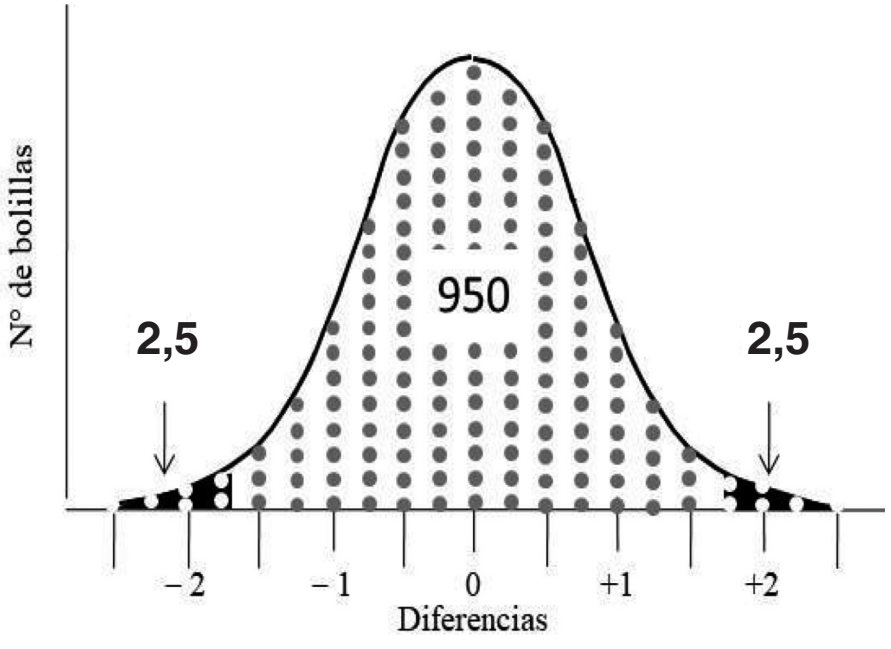

Figura 1. Esta figura describe la probabilidad de hallar en un bolillero una bolilla rotulada con una determinada diferencia (positiva o negativa): si en un bolillero se incluyen 1000 bolillas, 950 con diferencias en el rango de −1,5 a +1,5 y 50 con diferencias extremas, 25 con −1,5 y 25 con +1,5, la probabilidad de extraer una de estas bolillas es 5%. Dicha probabilidad se incrementa progresivamente si se repite la prueba resultado que luego de 100 intentos se habrán obtenido 5 bolillas con esas diferencias extremas (o si se prefiere, la relación puede expresarse de 20 a 1). Fuente: elaboración propia.

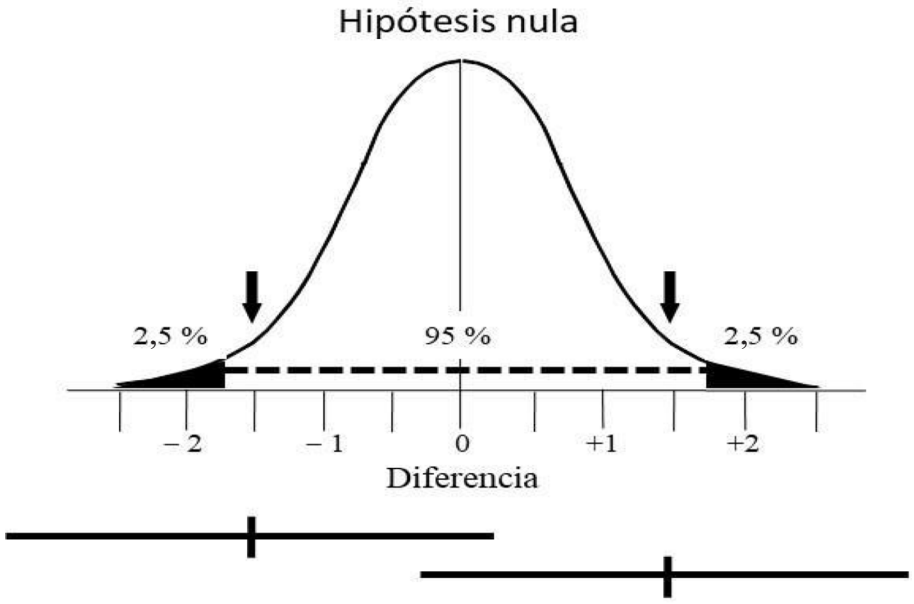

Figura 2. Curva de distribución de frecuencia (curva de Gauss) de la hipótesis nula. Ambas flechas indican el resultado de un ensayo en el que la diferencia hallada entre las muestras (positiva o negativa) no alcanzó una diferencia significativa, ya que se ubica en el sector de la curva correspondiente al 5% para ambas colas. (La línea de puntos delimita el 95% del área de dicha curva). Otra forma de expresar el resultado se representa en la parte inferior donde ambas líneas continuas grafican los intervalos de confianza construidos a partir de las diferencias halladas, indicando en este caso, por no alcanzar el cero, que el estudio fue negativo. Fuente: elaboración propia.

Si el resultado del ensayo demuestra que la diferencia hallada está entre −1,5 y +1,5 (*ambas colas*), como indican las flechas de la figura 2, el estudio fue negativo, ya que no se "alcanzó" la significación al "ubicarse" esta en el área de p ≥ 95%, (no "llega" a < 2, 5%).

En investigación clínica habitualmente se construye un intervalo de confianza (IC) del 95% con la diferencia hallada, −/+ 1,5, el cual –al superar el cero de la curva de hipótesis nula– nuevamente indica que la diferencia no es significativa (Fig. 2).

En el ejemplo anterior se empleó la diferencia de porcentajes (DR) entre, por ejemplo, intervención y control. Esto podría expresarse también como un cociente entre ambos, es decir, como *riesgo relativo* (RR). Para el caso, la media de la curva será 1y la hipótesis nula se rechaza cuando el IC alcanza o sobrepasa la unidad.

De este modo, se *acepta la hipótesis nula* y se concluye que el punto final (criterio diagnóstico/ pronóstico o intervención) no difiere entre las muestras en comparación.

Ahora bien, si se repitiera el estudio 100 veces, en ≥ 95 la diferencia hallada se ubicaría dentro del IC –*es decir no significativo– pero en < 5 por fuera de esos límites, por lo que el resultado podría considerarse falsamente como positivo.*

Comparaciones múltiples: fundamentos estadísticos

Resulta claro –según lo anteriormente expuesto– que la posibilidad de hallar una diferencia significativa aumenta si se repiten las comparaciones. El valor de p hallado luego de múltiples comparaciones suele denominarse *p nominal*, significando una determinación numérica no ubicada en el contexto de la probabilidad real. Existen diferentes procedimientos estadísticos para calcular la *p corregida* al número de comparaciones efectuadas.

Uno de ellos es corregir el valor de *p* en función del número de comparaciones efectuadas, utilizando procedimientos de ajuste por comparaciones múltiples.

En primer lugar se aplicó la corrección de Bonferroni, en el cual el nivel de significación corregido (αc) se obtiene dividiendo el valor nominal de α por el número de comparaciones (κ), según la expresión: $\alpha c = \alpha/\kappa$.

Asimismo, se consideró el cálculo del nivel de significación requerido en función del número de comparaciones, estimado mediante la expresión: $\alpha c = 1 - (1 - \alpha)\ \kappa$, donde α representa el valor nominal de *p* y κ el número de comparaciones realizadas.

Los decimales llevan coma, no punto.

La tabla 3 muestra, para una p nominal de 0,05, los valores reales de p y los corregidos resultantes al aplicar la fórmula detallada que deberían emplearse en una investigación si se efectuara más de una comparación.

Que de cada 100 comparaciones 5 puedan ser falsamente positivas parece ser una situación poco probable en investigación clínica. Pero si esa proporción se define como 20 a 1, resulta más aceptable. Ahora bien, regresando a la tabla 1, obsérvese que –con solo dos o tres comparaciones– se modifica notablemente la p requerida para alcanzar una diferencia significativa.

Tabla 1. Valores de p nominales y valores ajustados según el número de comparaciones, calculados mediante la corrección de Bonferroni y el método de ajuste del nivel de significación.

α	κ	α_c	α / κ
0,05	1	0,05	0,05
0,05	2	0,10	0,025
0,05	3	0,14	0,017
0,05	4	0,19	0,013
0,05	5	0,22	0,010

Nota: α: valor nominal de p; αc: p no corregida; k: número de comparaciones; α/k valor de p corregido.
Fuente: elaboración propia.

Comparaciones múltiples en investigación clínica

En investigación clínica, las comparaciones múltiples pueden surgir en diferentes escenarios. En los ensayos clínicos, el investigador suele definir una hipótesis general y un objetivo primario, ya sea para evaluar el efecto de una intervención o el valor pronóstico de un determinado factor. Sin embargo, cuando la comparación inicial no muestra una asociación estadísticamente significativa, puede surgir la tentación de explorar objetivos alternativos o modificar el punto final, repitiendo los análisis hasta obtener un resultado favorable[1,5].

Aun sin modificar el punto final primario, las comparaciones múltiples pueden aparecer cuando dicho punto se evalúa en distintos momentos del seguimiento, por ejemplo a los 3, 6 o 12 meses, seleccionando posteriormente aquel que muestra un resultado estadísticamente significativo[3].

Otra fuente frecuente de comparaciones múltiples es el análisis de subgrupos poblacionales. Ante un resultado global negativo, la muestra original puede fraccionarse para explorar asociaciones en distintos subgrupos, conservando únicamente aquellos con resultados favorables[7].

Asimismo, es posible aplicar diferentes pruebas estadísticas sobre los mismos datos e informar solo aquella que arroja un valor de *p* significativo. Estas estrategias, aisladas o combinadas, incrementan sustancialmente la probabilidad de obtener asociaciones espurias[8].

Estrategias de corrección en investigación clínica

La principal estrategia para otorgar credibilidad a los resultados de una investigación clínica consiste en definir de manera precisa, *a priori*, la hipótesis que se va a evaluar, el objetivo del estudio y el punto final primario, así como la prueba estadística por emplear y el tamaño muestral requerido[1]. Los hallazgos adicionales deben presentarse como puntos finales secundarios, aclarando explícitamente que responden a un análisis *post hoc* y que el valor de *p* no ha sido corregido por comparaciones múltiples[5,6].

En el caso de los análisis secundarios, es posible ajustar el valor de *p* en función del número de comparaciones realizadas mediante pruebas estadísticas apropiadas, como la corrección de Bonferroni u otros métodos de ajuste[4,6].

Otra alternativa consiste en establecer un análisis jerárquico de los puntos finales secundarios. En este enfoque, una vez alcanzada la significación estadística para el punto final primario, se define un orden jerárquico preespecificado para los puntos secundarios. Cuando se obtiene un valor de *p* no significativo, los análisis subsiguientes deben interpretarse como valores nominales, aun cuando sean inferiores a 0,05[1].

Finalmente, en estudios observacionales, un hallazgo obtenido luego de múltiples comparaciones puede considerarse exploratorio y ser presentado como resultado de una cohorte de derivación, requiriendo su validación posterior en una cohorte independiente[2,4].

Ampliando el escenario

La problemática de las comparaciones múltiples es abordada con diferentes criterios según los diferentes escenarios de investigación.

En los *ensayos de intervención, controlados y aleatorizados*, con un punto final primario predefinido, lo habitual para los puntos finales secundarios es aplicar diferentes criterios de ajuste o simplemente indicar que el resultado no está corregido por múltiples comparaciones.

Los *estudios observacionales derivados de grandes bases de datos* que tienen finalidad pronóstica o de intervención (Fase IV) suelen ser más o menos rigurosos, ya que en ocasiones carecen de las correcciones necesarias que ajusten los múltiples análisis, lo que es habitual en este tipo de investigaciones.

Finalmente en *ensayos unicéntricos* son habituales los análisis *post hoc* no explicitados, que exploran –dentro de la formulación general de una hipótesis– múltiples objetivos o puntos finales, sin el correspondiente ajuste por las correcciones realizadas.

De la inteligencia artificial a las comparaciones múltiples

¿Corrige la inteligencia artificial (IA) la problemática de las asociaciones probablemente ficticias consecuencia de comparaciones múltiples? La respuesta al interrogante planteado es no.

Al contrario, la capacidad de la inteligencia artificial de explorar miles de asociaciones a partir de millones de datos puede amplificar el riesgo de identificar relaciones espurias si no se aplican controles estadísticos adecuados[9].

CONCLUSIÓN

Regresando a la anécdota con que se inició esta comunicación,

"...ocurre que había estado apostando toda la noche, fallando repetidamente en cada una de ellas, hasta que

gané con el 18" (no sabemos si nuestro personaje recuperó todo su dinero, pero –en todo caso– esto no cambia la historia).

Ahí precisamente reside el problema: no contar toda la historia, y presentar el resultado como el objetivo original y no como el alcanzado luego de varios intentos que alteran indefectiblemente el valor de p.

Una conclusión sencilla es afirmar que no es igual decir –encontré lo que buscaba– que, –miren qué encontré–. Una sentencia tan simple como esta define el objetivo central de esta presentación.

En este caso, nada mejor que recordar: –si se torturan suficientemente los datos, usted obtendrá lo que quiera oír y, bajo estas circunstancias, el perpetrador puede examinar los datos hasta encontrar una asociación «significativa» entre variables, para luego elaborar una hipótesis que biológicamente se ajuste a esa asociación–). *("If you torture your data long enough they will tell you whatever you want to hear"; ...the perpetrador simply pores over the data until a "significant" association is found between variables and then devises a biologically plausible hypothesis to fit the association*[8,10].

O tal vez, que "una p no sustituye al cerebro" (*A p value is no substitute for a brain*. Anonymous), como ha sido señalado en discusiones contemporáneas sobre el uso e interpretación del valor de *p*[11]. Aunque una mejor consideración al respecto es la del físico Richard Feynman cuando concluye que: "es mucho más interesante vivir con cierta incertidumbre que vivir con respuestas que pueden estar equivocadas" ("It is much more interesting to live with uncertainty than to live with answers that might be wrong")[12].

En medicina esta última aseveración requiere una pequeña corrección: aun con incertidumbre, siempre debemos tomar decisiones ya sea por acción u omisión. El problema es ser conscientes del escenario estrecho al que nos impone aquella condición.

Contribuciones de los autores: Conceptualización, Supervisión, Validación (AMC). Curación de datos, Análisis formal, Captación de fondos, Investigación, Metodología, Administración del proyecto, Recursos, *Software* (MEK, MB, AMC).

Conflictos de intereses: los autores declaran no poseer conflictos de intereses relacionados con el contenido del presente trabajo.

Financiamiento: los autores declaran que este estudio no recibió financiamiento de ninguna fuente externa.

REFERENCIAS

1. Pocock SJ, Stone GW. The primary outcome fails - what next? N Engl J Med. 2016;375(9):861-870. https://doi.org/10.1056/NEJMra1510064.
2. Sugitani T, Bretz F, Maurer W. A simple and flexible graphical approach for adaptive group-sequential clinical trials. J Biopharm Stat. 2016;26(2):202-216. https://doi.org/10.1080/10543406.2014.972509.
3. Freemantle N, Calvert MJ. Interpreting composite outcomes in trials. BMJ. 2010;341:c3529. https://doi.org/10.1136/bmj.c3529.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
5. Bauer P. Multiple testing in clinical trials. Stat Med. 1991;10(6):871-889; discussion 889-890. https://doi.org/10.1002/sim.4780100609.
6. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-1238. https://doi.org/10.1136/bmj.316.7139.1236.
7. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Statist. Med.2002;21:2917-2930. https://doi.org/10.1002/sim.12968. Bender R, Lange S. Adjusting for multiple testing--when and how? J Clin Epidemiol. 2001;54(4):343-349. https://doi.org/10.1016/s0895-4356(00)00314-0.
9. Westphal M, Zapf A, Brannath W. A multiple testing framework for diagnostic accuracy studies with co-primary endpoints. Stat Med. 2022;41(10):1848-1866. doi:10.1002/sim.9308.10.
10. Mills JL. Data torturing. N Engl J Med. 1993;329(16):1196-1199. https://doi. org/10.1056/NEJM199310143291613.
11. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat. 2019;73(Suppl 1):1-19. https://doi.org/10.1080/00031305.2019.1583913.
12. Feynman RP. The pleasure of finding things out. Cambridge (MA): Perseus Books; 1999. 270 p.