# MÁS ALLÁ DEL VALOR *P*: UNA MIRADA CRÍTICA PARA LA INVESTIGACIÓN MÉDICA DEL SIGLO XXI.

**AUTOR:** GUILLERMO RAMÓN FRANCO DEL RÍO.
DIVISIÓN DE CIENCIAS DE LA SALUD. UNIVERSIDAD ANÁHUAC QUERÉTARO

**RESUMEN**

El valor *p* ha sido, durante casi un siglo, el pilar de la inferencia estadística en la literatura médica. Sin embargo, su uso indiscriminado y su frecuente interpretación errónea han contribuido a decisiones clínicas cuestionables, a la crisis de reproducibilidad y a una falsa dicotomía entre "significativo" y "no significativo". En este artículo se revisa la evolución histórica del concepto, examinamos críticas contemporáneas y mostramos por qué los intervalos de confianza (IC) y las estimaciones de efecto ofrecen una alternativa más informativa. Se dan recomendaciones prácticas para autores, revisores y lectores clínicos.

**Abstract**

The p-value has been, for almost a century, the cornerstone of statistical inference in medical literature. However, its indiscriminate use and frequent misinterpretation have contributed to questionable clinical decisions, the reproducibility crisis, and a false dichotomy between "significant" and "not significant." This article reviews the historical evolution of the concept, examines contemporary criticisms, and shows why confidence intervals (CIs) and effect estimates offer a more informative alternative. Practical recommendations are provided for authors, reviewers, and clinical readers.



## 1. Introducción

En los resúmenes de ensayos clínicos se repite invariablemente una cifra: $p < 0.05$. Para muchos clínicos se ha convertido en sinónimo de "resultado verdadero" o "efecto real". Pero ¿qué significa realmente? ¿Por qué nació el umbral 0.05? ¿Sigue siendo pertinente? Comprender las limitaciones del valor *p* no es un ejercicio teórico; es esencial para interpretar correctamente la evidencia que sustenta la práctica médica basada en pruebas.

## 2. Orígenes y desarrollo histórico

### 2.1 Fisher y la "prueba de significación"

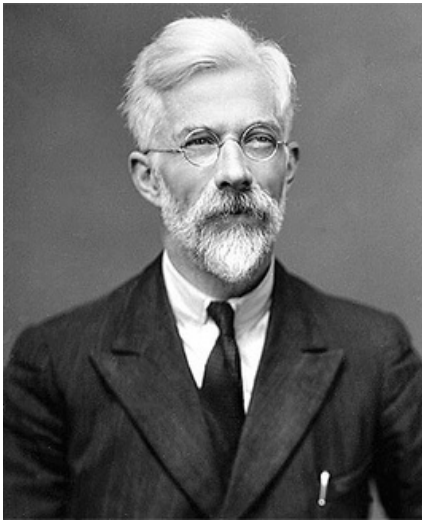
En los años veinte, Ronald A. Fisher introdujo el valor *p* como la probabilidad de observar un resultado tan extremo (o más) que el obtenido, **si** la hipótesis nula ($H_0$) fuera cierta. Fisher nunca propuso un umbral rígido; sugería evaluar la evidencia gradualmente (p. ej., $p < 0.05$ como "sugerente", $p < 0.01$ como "impresionante"). Su enfoque era **inductivo**: el valor *p* mide la "discordancia" entre datos y $H_0$ y facilita la generación de nuevas hipótesis.

**2.2 Neyman y Pearson: decisiones y errores de tipo I/II**

Jerzy Neyman y Egon Pearson (década de 1930) transformaron la prueba de significación en un **procedimiento de decisión** binario: se rechaza $H_0$ cuando $p<\alpha$, donde $\alpha$ es la tasa de error de tipo I fijada de antemano (clásicamente 0,05). Su marco se centró en el control a largo plazo de errores y la potencia ($1-\beta$), bases de los cálculos muestrales modernos.

**2.3 Fusión y banalización**

A mediados del siglo XX, los dos enfoques se fusionaron de manera informal en la "prueba de significancia nula" (NHST). Ese híbrido se popularizó porque era sencillo de aplicar, podía automatizarse en los primeros paquetes estadísticos y parecía ofrecer una respuesta clara. Con el tiempo, la comunidad médica olvidó los matices y adoptó el "0.05" como un sello de aprobación.



**RONALD A. FISHER, 1890-1962**
INTRODUCE POR PRIMERA VEZ EL VALOR DE P, COMO UNA PROBABILIDAD.



**3. ¿Qué es y qué no es un valor p?**

| MITO FRECUENTE | REALIDAD |
|---|---|
| "p = 0.03 indica que existe 97 % de probabilidad de que $H_0$ sea falsa." | El valor *p* no cuantifica la probabilidad de que la hipótesis sea verdadera o falsa; solo describe la compatibilidad de los **datos** con $H_0$. |
| "p bajo implica un efecto grande." | La magnitud del efecto es independiente del valor *p*. Con muestras enormes, un efecto clínicamente trivial puede producir $p < 0.001$. |
| "p > 0.05 demuestra que no hay diferencia." | Un resultado no significativo puede deberse a potencia insuficiente o a variabilidad alta; la ausencia de evidencia no es evidencia de ausencia. |

## 4. Críticas contemporáneas

### 4.1 Crisis de reproducibilidad

Diversas áreas biomédicas muestran tasas de replicación inferiores al 40%. Entre los factores implicados destacan el **p-hacking** (selección post hoc de análisis que producen $p<0.05$) y el **sesgo de publicación**. El uso ritual del umbral $\alpha=0.05$ amplifica el problema: estudios "negativos" no se publican y los "positivos" pueden ser falsos hallazgos.

### 4.2 Declaración de la ASA (2016) y la "marcado abuso"

La American Statistical Association emitió seis principios clave, entre ellos: (1) el valor *p* no mide la importancia del efecto, (2) conclusiones científicas deben incluir información más allá del *p*, y (3) replicabilidad y contexto científico son esenciales. En 2021 la ASA y organizaciones afines exhortaron a "acabar con la tiranía del 0.05" y privilegiar estimaciones y compatibilidad.

### 4.3 Propuestas: bajar el umbral, Bayes, S-values

- **Umbral 0.005:** Benjamin et al. (2018) sugirieron exigir $p<0.005$ para "nuevo descubrimiento" a fin de reducir falsos positivos.
- **Paradigma bayesiano:** Reemplazar p-values por factores de Bayes o probabilidades posteriores permite incorporar conocimiento previo y produce interpretaciones más directas ("Hay 85% de probabilidad de beneficio clínico ≥10%").
- **S-values** ($-\log_2 p$): traducen p a bits de evidencia contra $H_0$, evitando la semántica confusa de "probabilidad".

## 5. Intervalos de confianza: una solución (casi) olvidada

### 5.1 Definición operativa

Un intervalo de confianza del 95% es un rango de valores compatibles con los datos, tal que, si repitiésemos el estudio infinitas veces en condiciones idénticas, **95%** de los intervalos incluirían al verdadero parámetro. Aunque la interpretación precisa es frecuentemente mal entendida, los IC aportan dos dimensiones ausentes en el valor *p*: magnitud y precisión.

### 5.2 Ventajas concretas

1. **Magnitud del efecto**
   - Permite juzgar relevancia clínica. Un riesgo relativo (RR) de 0.70 con IC 95% 0.48–1.02 ($p=0.06$) sugiere un posible beneficio del 30%, clínicamente valioso, pese a no alcanzar "significancia".

2. **Dirección y plausibilidad**
   - Si el IC incluye el valor de nulo ($RR=1$), indica que efectos de ambos signos son compatibles con los datos; facilita un análisis de "zonas de indeterminación" más rico que un único bit (rechazar / no rechazar).

3. **Precisión y planificación**
   - Informan del error estándar y ayudan a diseñar estudios futuros: un IC amplio señala necesidad de muestras mayores.

4. **Comunicación interdisciplinaria**
   - Los clínicos, tomadores de decisiones y pacientes entienden mejor un rango tangible ("la reducción de HbA1c está entre 0.3% y 1.2%") que un p críptico.

5. **Menor susceptibilidad a p-hacking**

- Aunque no lo eliminan, desplazan la atención a la estimación y así desalientan la búsqueda de umbral.

**5.3 Limitaciones**

Los IC clásicos comparten supuestos frecuentistas (repetición hipotética), pueden malinterpretarse igual que el p, y, si se reportan sin contexto clínico, se vuelven meras formalidades. No sustituyen al juicio experto ni al análisis de heterogeneidad.

**Richard McElreath, 1973.**
Sugiere sustituir la "ciencia de la *p*-valor" por la "ciencia de la compatibilidad"

**6. Ejemplos prácticos**

| ESCENARIO | **RESULTADO ESTADISTICO** | **INTERPRETACION CON p- value** |
|---|---|---|
| Ensayo de un fármaco antihipertensivo (n = 800): ΔPAS = –3 mmHg, IC 95 % –5 a –1 mmHg, p = 0.004 | "Significativo". | El descenso promedio es 3 mmHg; con 95 % de confianza, el efecto real oscila entre 1 y 5 mmHg. **¿Es clínicamente relevante?** |
| Suplemento nutricional para prevenir bajo peso al nacer (n = 120): RR = 0.65, IC 95 % 0.22–1.86, p = 0.42 | "No significativo". | Los resultados muestran una reducción del 65% del bajo peso al nacer. La muestra es pequeña: gran incertidumbre. Se requiere estudio más grande. |

**7. Recomendaciones para autores y revisores**

1. **Informe siempre la estimación del efecto con su IC**, aun cuando $p<0.05$.
2. **Evite etiquetas simplistas** ("significativo/no significativo"). Discuta la plausibilidad clínica y biológica.
3. **Planifique la potencia a partir de la mínima diferencia clínicamente importante** y describa claramente supuestos y metodología estadística.
4. **Desglose análisis exploratorios** de los confirmatorios y ajuste por multiplicidad si corresponde.
5. **Comparta código y datos** para facilitar la replicación.
6. **Considere complementos bayesianos o métodos de evidencia graduada** (p. ej., *S*-values) cuando el público lector esté familiarizado.
7. **Promueva la educación estadística continua** de clínicos y decisores. El uso crítico de la estadística es un requisito ético, no opcional.

**9. Conclusiones**

El valor p no desaparecerá de la noche a la mañana, pero su papel debe cambiar. Tratarlo como el árbitro supremo distorsiona la literatura y, potencialmente, las decisiones clínicas. Pasar de la dicotomía ($p<0.05$ vs. $\geq 0.05$) a un paradigma de estimación, compatibilidad y contexto enriquece la interpretación y eleva la calidad de la evidencia.

Los intervalos de confianza constituyen una herramienta inmediata y accesible que añade magnitud y precisión al debate, al tiempo que desplaza el foco hacia la relevancia clínica. Acompañados de buenos diseños, transparencia y, cuando sea pertinente, enfoques bayesianos, ofrecen un camino para retomar la confianza en la investigación médica.

**REFERENCIAS**

1. Fisher RA. Statistical Methods for Research Workers. Oliver & Boyd; 1925. Obra fundacional que introduce el concepto de significancia.

2. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A. 1933;231:289337. Marco de decisión y errores tipo I/II.

3. American Statistical Association. The ASA Statement on pValues and Statistical Significance. Am Stat. 2016;70(2):129133. Primer pronunciamiento institucional sobre abusos del pvalue.

4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat. 2019;73(sup1):119. Recomendaciones prácticas postASA.

5. Benjamin DJ, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:610. Propuesta de umbral 0,005.

6. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020;20:244. Presenta el concepto de S-value.

7. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2.ª ed. Chapman & Hall; 2020. Enfoque bayesiano centrado en compatibilidad.

8. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. Artículo icónico sobre reproducibilidad.

..