

Hacia un uso responsable de la Estadística en la investigación biomédica y epidemiológica

Towards a Responsible Use of Statistics in Biomedical and Epidemiological Research

Dr. Cs. Jorge Bacallao Gallestey

[0000-0001-7961-9970](https://orcid.org/0000-0001-7961-9970)

Centro de Investigaciones y Referencia de Aterosclerosis de la Habana (CIRAH), Cuba.

Correspondencia: jbacallag@gmail.com

RESUMEN

Aunque la aplicación de métodos numéricos al estudio de los problemas típicos de la investigación en las ciencias médicas y de la salud se remonta a la década de 1830 con los aportes del médico francés Pierre A. Louis, la formalización de la Estadística matemática como recurso analítico se asocia históricamente a las contribuciones de Karl Pearson, Ronald Fisher y Austin Bradford Hill, quienes sentaron las bases para lo que hoy conocemos como medicina basada en la evidencia, desde mediados del siglo XX. Desde entonces la presencia de alguna herramienta analítica de la estadística inferencial, se ha considerado una marca indispensable e insustituible de calidad en la investigación. Sin negar un ápice del valor instrumental del recurso estadístico, el propósito del presente ensayo es argumentar que su aplicación irresponsable y superficial ha contribuido a fomentar una actitud de pereza intelectual en los investigadores del ámbito médico, y al establecimiento de normativas rígidas, contrarias al rigor y a la calidad del pensamiento científico. El tema tiene interés y vigencia universales, pero los argumentos del artículo remiten básicamente a la experiencia del autor en escenarios nacionales, y se refieren de modo particular a dos recursos analíticos cuyo uso irresponsable es particularmente frecuente: las pruebas de significación estadística y los modelos predictivos.

Palabras clave: estudios observacionales; ensayos clínicos controlados y aleatorizados; tamaño muestral; variables confusoras; modelos lineales; predicción vs. explicación; validación de hipótesis.



ABSTRACT

Although the application of numerical methods to the study of typical research problems in the medical and health sciences dates back to the 1830s with the contributions of the French physician Pierre A. Louis, the formalization of mathematical statistics as an analytical tool is historically associated with the contributions of Karl Pearson, Ronald Fisher, and Austin Bradford Hill, who laid the foundations for what we know today as evidence-based medicine, beginning in the mid-20th century. Since then, the presence of some analytical tool from inferential statistics has been considered an indispensable and irreplaceable mark of quality in research. Without denying the instrumental value of statistical resources, the purpose of this essay is to argue that their irresponsible and superficial application has contributed to fostering an attitude of intellectual laziness among medical researchers and to the establishment of rigid norms that are contrary to the rigor and quality of scientific thought. The topic has universal interest and relevance, but the arguments in the article refer mainly to the author's experience in national scenarios, and refer in particular to two analytical resources whose irresponsible use is particularly frequent: statistical significance tests and predictive models.

Keywords: observational studies; randomized controlled clinical trials; sample size; confounding variables; linear models; prediction vs. explanation; hypothesis validation.

Recibido: 16/10/2025

Aprobado: 05/11/2025

Desarrollo

En un espectro amplio de usos y aplicaciones, las técnicas y procedimientos de la Estadística Matemática aportaron rigor y posibilidades analíticas a todas las investigaciones de las ciencias médicas y de la salud. Si bien los comienzos de una Estadística científica en salud se remontan al siglo XVII con las tablas de mortalidad, y un poco más próximos en el tiempo, a los trabajos de Pierre C. Alexandre Louis en la cuantificación de variables clínicas, la “bioestadística” como término y como disciplina científica se consolidó en los siglos XIX y XX con los aportes de figuras como Karl Pearson y Ronald Fisher, que extendieron la aplicación de sus métodos a los ámbitos de la epidemiología, la genética y la salud pública.^{(1),(2)} De entonces a la fecha, los artículos científicos en revistas médicas y en otros soportes, abundan en usos instrumentales de la Estadística matemática, que se ha instituido prácticamente como un requisito de rigor en las publicaciones.



Sus aplicaciones cubren un vasto campo dentro del cual cabe distinguir la descripción y la inferencia, esta última entendida como un recurso para trascender los datos que son materia prima de operación racional en un conjunto (la muestra), y realizar deducciones o generalizaciones a un conjunto mayor (la población).

De ese afán inferencial a ultranza, de la búsqueda de respuestas taxativas y de permanentes replicaciones de los hallazgos, se deriva una práctica rayana en el culto, que en no pocos casos ha degradado el uso de la Estadística matemática a una mera formalidad que bien poco tiene que ver con la buena práctica científica.

Las pruebas de significación estadística (PSE), en franco descrédito entre la comunidad estadística de avanzada en el mundo, aportan el ejemplo tal vez más aleccionador de estas desviaciones del rigor científico en el uso de los recursos analíticos de la Estadística.⁽³⁻⁵⁾ El siguiente ejemplo ficticio es harto elocuente en relación con esas desviaciones.

Las PSE: una historieta ficticia pero aleccionadora

El conocimiento como único propósito final es inconcebible en la sociedad moderna en la que la investigación también cuesta, como la salud y la educación; por consiguiente es menester garantizar que cada investigación tenga su rédito como insumo para diseñar estrategias, programas, y en general, para tomar decisiones. Al menos quienes se encargan de financiar la investigación, así lo exigen.

Un funcionario recién estrenado en la dirigencia del deporte, en desleal emulación con otros en el ámbito de la economía, la agricultura y otras esferas, decidió que su desempeño profesional debía basarse en el uso y aplicación de la ciencia.

En esa cuerda, se dedicó a acopiar y procesar estadísticas sobre finales olímpicas en natación y “descubrió” que los ganadores de esas finales se habían ubicado en las carrileras centrales de la piscina (3, 4 y 5) con mucha mayor frecuencia que en las restantes. Luego de una acuciosa recolección de los registros de esas finales, practicó una prueba de significación estadística para comparar dos grupos de registros, los correspondientes, respectivamente a nadadores de los carriles centrales y del resto de los carriles. Esa prueba arrojó diferencias significativas para un umbral preestablecido de $p=0.05$.

Satisfecho con la confirmación de sus expectativas, y desoyendo las reservas de varios de sus funcionarios subordinados, a los que no dio oportunidad de exponer sus razones, puso ese conocimiento científico en función de las decisiones propias de su cargo y ordenó que, en lo sucesivo, se reconstruyeran los fondos de las piscinas de competencia



en todo el territorio nacional. Esa medida –argumentó– contribuiría a una selección no sesgada de los mejores atletas.

...

Recién removido de su cargo, junto a otros varios colegas con las mismas aficiones por la ciencia en función de las decisiones, se enteró de que en las finales olímpicas es práctica habitual ubicar a los nadadores con mejor desempeño en su historia competitiva reciente, en las carrileras centrales.

Esta historieta ilustra uno de los problemas que habitualmente no se tienen en cuenta en la evaluación crítica de las PSE, cuando se aplican a estudios observacionales y no a ensayos aleatorizados y controlados. El problema consiste en interpretar la prueba como una contienda entre dos únicos actores, el efecto que se desea evaluar y el azar, ignorando la influencia de otros actores posibles. Descartado el azar, debido a una probabilidad por debajo del umbral arbitrario, sólo resta un responsable: así lo creyó inocentemente el protagonista de esta historieta, y así se conducen con frecuencia quienes desvalorizan cualquier investigación que no se acompañe del ornamento de una prueba de significación.

Para extraer un corolario de la anterior historieta, nada mejor que la cita textual del artículo de Hernán y Robbins: “¿Es posible garantizar que los resultados de un estudio observacional están libres del efecto de “confusores” no observados? La única respuesta de un epidemiólogo o un salubrista es: “NO”. No importa que el diseño sea óptimo y las mediciones completamente fiables. El supuesto de que no existe “confusión” no observada es completamente inverificable”.⁽⁶⁾ Esta cita expone de forma clara uno de los argumentos irrefutables que descalifican el uso de las PSE en estudios observacionales. No es posible descartar la existencia de variables que sin tener un rol intermediario entre ambos, se asocian simultáneamente con la presunta causa y el presunto efecto, protagonistas clave en toda prueba de significación estadística.

No todos los casos son tan previsibles y obvios como los que ilustra la siguiente tabla 1, que corresponde a un estudio observacional ficticio con el que se desea validar la efectividad de una nueva unidad de cuidados intensivos neonatales. Los resultados de la tabla ilustran un efecto de confusión típico, que habría sido fácil prever y evitar.



Tabla 1- Mortalidad en la nueva unidad de cuidados intensivos y en otros hospitales maternos.

Hospital	Casos atendidos	Número de fallecimientos	Mortalidad x 1000
Nueva UCIN	2500	25	10,0
Hospitales maternos	6000	47	7,8

UCIN: nueva unidad de cuidados intensivos neonatales.

Sin otro análisis, la tabla 1 estaría mostrando una realidad (mayor mortalidad en la nueva UCIN) contraria a las expectativas lógicas que se derivarían de una atención de mejor calidad en el nuevo servicio instalado. Los datos así presentados pasan por alto el hecho de que a la nueva UCIN se suelen remitir los casos de mujeres en mayor situación de riesgo. Estratificados los resultados según riesgo materno, la realidad cambia diametralmente de apariencia, como muestran las siguientes tablas 1A y 1B.

Tabla 1A- Mortalidad en recién nacidos de madres con alto riesgo.

Hospital	Casos atendidos	Número de fallecimientos	Mortalidad x 1000
Nueva UCIN	650	19	29,2
Hospitales maternos	425	15	35,3

UCIN: nueva unidad de cuidados intensivos neonatales

Tabla 1B- Mortalidad en recién nacidos de madres con bajo riesgo.

Hospital	Casos atendidos	Número de fallecimientos	Mortalidad x 1000
Nueva UCIN	1850	6	3,2
Hospitales maternos	5575	32	5,7

UCIN: nueva unidad de cuidados intensivos neonatales.

El nivel de riesgo de la madre es un confusor de la asociación entre el tipo de unidad hospitalaria y el desenlace del caso del recién nacido. En efecto, las embarazadas con mayor riesgo se remiten en una mayor proporción a la unidad especializada, pero entre los recién nacidos de estas madres hay lógicamente una mortalidad más alta. Si no se controla este efecto confusor, la Tabla 1 estaría mostrando mayores cifras de mortalidad, en contra de las expectativas. La estratificación según riesgo materno muestra todo lo contrario y estaría aportando evidencias favorables a la creación de la nueva UCIN, que es la intervención objeto de evaluación

Otro argumento letal contra un uso irreflexivo de las PSE es que su resultado, que se contrae a la decisión dicotómica de rechazar o no una hipótesis estadística, se subordina estrictamente al tamaño de la muestra. Para expresarlo con una breve sentencia: si la muestra es suficientemente grande, todo resultado (todo efecto, toda diferencia, toda asociación) es significativo. Esta circunstancia invalidante permite refutar la sinonimia entre “significativo” y “relevante” con que suelen interpretarse los resultados de las PSE. (7-8)



La tabla 2 es harto elocuente. Muestra la asociación entre dos variables X e Y, generadas como números aleatorios comprendidos entre los límites 0 y 100. El modo de generación de dichas variables descarta la existencia de alguna base teórica o empírica para suponer que puedan estar asociadas.

Tabla 2- Tamaños muestrales, coeficientes de correlación y valores de “p”.

Tamaño muestral	Coeficiente de correlación de Pearson	Significación estadística (valor de p)
100	-0,054	0,591
200	-0,054	0,444
500	-0,054	0,225
1000	-0,054	0,086
1500	-0,054	0,035

La PSE acerca de la asociación entre ambas variables se refiere a la prueba de hipótesis

$$H_0: \rho = 0 \text{ vs. } H_A: \rho \neq 0$$

relativa al coeficiente de correlación de Pearson (ρ) en la que la práctica habitual consiste en enfrentar una hipótesis nula de no asociación ($H_0: \rho = 0$), a una alternativa de asociación ($H_A: \rho \neq 0$).

Cada conjunto de datos se ha obtenido mediante el simple expediente de replicar (copiar y pegar) la base de datos original de 100 observaciones: 2, 5, 10 y 15 veces. Como es obvio, el coeficiente de correlación se mantiene invariante, pero la “p” asociada disminuye desde un valor inicial de 0,591, hasta un valor final de 0,035, ya inferior al umbral de significación de 0,05 prefijado en el estudio, que muy probablemente satisface las expectativas del investigador, quien, alcanzado ese umbral o punto de corte, da por terminada la investigación, corroboradas sus expectativas y cumplidos sus propósitos.

¿Conducen estos argumentos al funeral inminente de las PSE? La respuesta es NO, porque hay un espacio de aplicación que les garantiza tal vez larga supervivencia. En armonía con el razonamiento anterior, en ese espacio no tendrían validez las dos objeciones básicas a las PSE. Se trata de los ensayos clínicos controlados y aleatorizados, que se acompañan de un cálculo estricto del tamaño muestral. En otras palabras, la antípoda de los estudios observacionales. La aleatorización es un recurso que tiende a balancear los grupos que se comparan con respecto a todos los posibles confusores, observables o no. En la práctica, no obstante, es necesario verificar que ese balance se cumple y aplicar, por ejemplo, un recurso de estratificación con el posible confusor responsable del desbalance.

En cuanto al número de casos, el tamaño muestral se calcula a partir de un algoritmo riguroso que toma en cuenta, la dispersión de las variables que son objeto de estudio y



los riesgos tolerables por el investigador. Estos son de dos tipos: no detectar un efecto relevante (error de primer tipo), o proclamar la existencia de un efecto, aunque éste no sea relevante (error de segundo tipo).

Es fácil advertir que el cálculo del tamaño muestral se rige por criterios rigurosos que implican: (a) definir *a priori* lo que se considera un efecto relevante, (b) fijar cotas explícitas para la magnitud de los errores que se está dispuesto a tolerar y (c) aportar estimaciones sobre la magnitud de la dispersión de la o las variables a considerar. En la práctica, sin embargo y con mucha frecuencia, se exige exponer el tamaño muestral (¿de qué otro modo podrían elaborarse los presupuestos para los proyectos de investigación?) sin que la información anterior esté rigurosamente disponible, con lo cual el cálculo se transforma en un ritual sin fundamento.

Antes de pasar al otro tema, la siguiente cita, que data de 1997, resume con economía de recursos expresivos la opinión generalizada entre los expertos. “Las PSE constituyen con toda seguridad el procedimiento más irracional que se haya institucionalizado jamás en el entrenamiento habitual de los estudiantes de ciencia”.⁽⁹⁾

¿Predecir o explicar? Un dilema ignorado en el ajuste irreflexivo de modelos

Para simplificar, y porque los argumentos básicos son igualmente válidos para otros tipos de modelos, consideraremos dos casos, los que con más frecuencia se presentan en la práctica, y que se clasifican en el grupo de los llamados modelos lineales: la regresión lineal y la regresión logística. Ambos tienen en común la presencia de una variable dependiente, continua en la regresión lineal, y binaria o dicotómica en la regresión logística, y una o más variables independientes. El error más común consiste en aplicar un modelo sin distinguir si el propósito es predecir o explicar.

El término “predecir” sugiere de inmediato una relación de precedencia temporal que con mucha frecuencia se pasa por alto, entre las variables independientes “predictoras” y la variable dependiente “a predecir”. El diseño del estudio observacional debe garantizar esa precedencia temporal. Si todas las variables se han obtenido simultáneamente como parte de un diseño transversal, es imposible concebir un escenario de predicción.

En el caso simple de la regresión logística, para que sea legítimo hablar de predicción, es necesario garantizar que la ocurrencia del evento representado por la variable dependiente dicotómica no se haya producido en el momento de la obtención de los datos. Un ejemplo ilustrativo son los estudios longitudinales prospectivos basados en el



seguimiento de personas vacunadas con los tres tipos fundamentales de vacunas anti-COVID-19 (ARN mensajero, vacunas de vector viral y vacunas de subunidades proteicas) y en el registro riguroso de complicaciones o eventos patológicos de interés. Este tipo de estudio permitiría estimar el riesgo asociado a cada tipo de vacuna, con la posibilidad de incluir en el modelo predictivo otros posibles factores predisponentes como la edad y otras condiciones personales de las personas vacunadas. Sin que sea necesario analizar otros detalles colaterales, el propósito fundamental de un estudio de este tipo es obtener una estimación del riesgo. De esa estimación para los tres tipos de vacuna, podrían extraerse conclusiones simples para una comparación esencialmente descriptiva de los riesgos.

Mucho más problemático es el uso con fines predictivos de los diseños longitudinales retrospectivos. Por ejemplo, consideremos el caso de un estudio realizado con la finalidad de evaluar el riesgo de malformaciones de miembros en los hijos de madres que han consumido anticonvulsivantes durante el embarazo. El recurso elemental de incorporar al estudio niños con malformaciones y controles pareados, podría ser útil para concluir si los anticonvulsivantes incrementan o no el riesgo, pero no para estimarlo, ya que las proporciones relativas de malformados y no malformados no se corresponden con lo que ocurre a escala poblacional.

No obstante, en situaciones en las que por razones éticas o prácticas no sea factible utilizar diseños longitudinales prospectivos, pudieran aplicarse diseños retrospectivos. Un caso bien conocido es la estimación del riesgo de enfermedades respiratorias obstructivas en trabajadores expuestos a sustancias químicas específicas. En estos casos, se identifican cohortes retrospectivas de trabajadores con y sin el diagnóstico de la enfermedad y se revisan sus expedientes clínicos laborales para reconstruir el historial de exposición. Las frecuencias relativas de enfermos y no enfermos se corresponden con las frecuencias poblacionales reales y la secuencia temporal garantiza la precedencia de la exposición con respecto a la enfermedad.

Predecir, en esencia, se reduce a la estimación de un parámetro de interés, y no entraña otras pretensiones inferenciales, como las que implica el propósito de explicar. Las mismas objeciones ya expuestas en el caso de las PSE son válidas en el caso de los modelos lineales, que llevan implícito el uso de las PSE para las inferencias propias relativas a los parámetros de estos modelos. El uso de los modelos lineales con fines explicativos supone, además: (a) en la regresión lineal:

- ✓ Linealidad en la relación entre la variable dependiente y las variables explicativas
- ✓ Independencia estadística, homocedasticidad (varianza constante) y distribución normal de los residuos.
- ✓ Ausencia de multicolinealidad fuerte entre las variables explicativas.



(b) en la regresión logística:

- ✓ La relación entre la variable dependiente (la que define el evento de interés) debe ser lineal en términos del logit, es decir, del logaritmo de los odds, estos últimos entendidos como el cociente entre el riesgo de ocurrencia y el de no ocurrencia del evento.
- ✓ Observaciones independientes.
- ✓ Ausencia de multicolinealidad fuerte entre las variables explicativas.

Esta distinción teórica básica entre los propósitos de explicar o de simplemente predecir, no suele tomarse en cuenta, y un estudio descriptivo que sólo alcanza para predicciones, se utiliza para poner a prueba hipótesis sobre la relevancia de las variables mediante la realización de PSE.

¿Y entonces?

El giro de este subtítulo describe sintéticamente la reacción habitual de quienes, sin los recursos y con el enfoque que han venido aplicando –y viendo aplicar– durante años, desearían disponer de otras recetas para responder las mismas preguntas.⁽¹⁰⁾ De lo que se trata, por el contrario, es de encontrar y formular las preguntas adecuadas. No hay que empeñarse obstinadamente en respuestas dicotómicas del tipo: “hay - no hay”, “sí - no”, “significativo - no significativo”, que conducen inevitablemente al empobrecimiento de los réditos de un análisis estadístico, que no es un sucedáneo, sino un complemento del análisis de la evidencia acumulada y del razonamiento científico.⁽⁵⁾

Procede, no obstante, una breve reflexión adicional. La ciencia discurre por un dilatado proceso, en el que las acciones de explorar y confirmar tienen una importancia vital. En estas etapas la relevancia de los procedimientos descriptivos que aporta la Estadística matemática es crucial. En la etapa confirmatoria, el centro de gravedad del análisis son las conjeturas razonables y el camino a seguir implica la acumulación y la evaluación crítica de las evidencias, tanto las que parecen respaldar como las que parecen contradecir dichas conjeturas.

Toda conjetura debe tener un respaldo empírico y/o teórico razonable y la actitud del investigador, no en el orden psicológico, sino en el orden práctico, debe ser intentar refutarlas. Cada tentativa fallida por refutar una conjetura, no implica haberla probado, sino simplemente haberla fortalecido. Esta es la esencia de la propuesta filosófico-metodológica de Karl Popper, en su libro Conjeturas y refutaciones.^(11,12) Un tema apasionante y de gran vigencia que desborda el alcance y las pretensiones del presente texto.



Dos enfáticas sugerencias pueden derivarse de las consideraciones de este artículo: (1) fuera del reducido ámbito de los ensayos clínicos controlados y aleatorizados, abandonar el uso de las PSE, lo cual no implica la abolición taxativa de los valores de “p”, como una pieza de evidencia más, pero sí evitar el uso de umbrales o puntos de corte subjetivos que conducen a las dicotomías denunciadas, y (2) evitar la práctica de utilizar modelos sin una clara distinción entre los propósitos de predecir o explicar; la predicción es mucho menos exigente en términos de requisitos formales y puede utilizarse como un recurso descriptivo con la debida validación posterior.

CONCLUSIONES

Si bien el uso de la estadística matemática es un recurso analítico que ha abierto grandes posibilidades a la investigación y que ha aportado un incuestionable rigor científico en el campo de la investigación clínica y epidemiológica, su uso irresponsable, sin reparar en los fundamentos teóricos bien establecidos, ha propiciado una actitud de acomodamiento o indiferencia, no sólo entre los investigadores, sino también entre revisores y editores que a menudo establecen como normativo el uso de procedimientos estadísticos sin el debido fundamento. En este, como en otros ámbitos de la ciencia aplicada, se impone con urgencia una toma de conciencia.

Referencias

1. Barreto Villanueva A. El progreso de la Estadística y su utilidad en la evaluación del desarrollo. Papeles de población. 2012; 18 (73), 241-71.
2. Quintero KDS, Melo JSO. Bioestadística: evolución y aplicaciones. Revista Sigma [Internet]. 2022; 18(1), 1-7. Disponible en: <https://dialnet.unirioja.es/>
3. Blakeley B, McShane DG, Gelman A, Christian R, Tackett JL. Abandon Statistical Significance. The American Statistician [Internet]. 2019; 73 (sup1):235-45. Disponible en: <https://doi.org/10.1080/00031305.2018.1527253>
4. Stuart H, Hurlbert RA, Jessica Utts. Coup de Grâce for a Tough Old Bull: “Statistically Significant” Expires. The American Statistician [Internet]. 2019; 3(sup1):352-7. Disponible en: <http://doi.org/10.1080/00031305.2018.1543616>
5. Amrhein V, Trafimow D, Greenland S. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication. The American Statistician [Internet]. 2019;73(sup1):262-70. Disponible en: <http://doi.org/10.1080/00031305.2018.1543137>
6. Hernán MA, Robins JM. Instruments for causal inference. An epidemiologist’s dream? . Epidemiology. 2006; 17: 360-372.



7. Joachim I. Krueger P, Heck R. Putting the P-Value in its Place. The American Statistician [Internet]. 2019; 73 (sup1): 122-8. Disponible en: <http://doi.org/10.1080/00031305.2018.1470033>
8. Ronald L. Wasserstein, Allen L. Schirm, Nicole A. Lazar Moving to a World Beyond “ $p < 0.05$ ”. The American Statistician [Internet]. 2019; 73(sup1): 1-19. Disponible en: <http://doi.org/10.1080/00031305.2019.1583913>
9. Rozeboom WW. Good science is abductive, not hypothetico-deductive. En: Harlow LL, Mulaik SA, Steiger JH, eds. What if there were no significance tests?. Hillsdale, NJ: Erlbaum; 1997. pp. 335–91.
10. Steven N. Goodman. Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. The American Statistician [Internet]. 2019; 73 (sup1): 26-30. Disponible en: <http://doi.org/10.1080/00031305.2018.1558111>
11. Popper Karl R. Conjeturas y refutaciones. El desarrollo del conocimiento científico. Barcelona: Ediciones Paidós; 1991.
12. Gattei Stéfano. Karl Popper’s Philosophy of Science. Rationality Without Foundations. Routledge Studies in the Philosophy of Science. 2009; 5.

Conflicto de interés

El autor declara que no existen conflictos de interés.

Declaración de autoría

El autor es único responsable del contenido del trabajo.

