

REVIEW ARTICLE

Epistemological scope and limitations of contemporary statistical inference: a systematic review

Alcances y limitaciones epistemológicas de la inferencia estadística contemporánea: una revisión sistemática

Lisette Beatriz Villavicencio Cedeño¹  

Yoiler Batista Garcet²  

Sandra Elizabeth Tenelanda Cudco¹  

Deysi Margoth Guanga Chunata¹  

Darwin Casaliglla Ger¹  

¹ Universidad Nacional de Tumbes, Perú.

² Universidad San Gregorio de Portoviejo, Ecuador.

Citar como: Villavicencio, L.B, Batista, Y., Tenelanda, S.E., Guanga, D.M., & Casaliglla, D. (2026). Epistemological scope and limitations of contemporary statistical inference: a systematic review. *Revista San Gregorio*, 1(Especial_3), 131-141. http://dx.doi.org/10.36097/rsan.v1iEspecial_3.4198

Received: 28-01-2026

Accepted: 13-08-2026

Published: 31-08-2026

ABSTRACT

Statistical inference is fundamental to producing scientific knowledge, interpreting evidence, and guiding decision-making under uncertainty. However, its validity depends on study design, underlying assumptions, analytical transparency, and the appropriate interpretation of results. The objective of this study was to qualitatively synthesize the epistemological scope and limitations attributed to statistical inference in recent scientific literature. A qualitative systematic review guided by PRISMA 2020 was conducted. The PubMed and Europe PMC databases were searched to identify publications issued between 2021 and 2025. Thirteen studies were included and analyzed through a thematic, comparative, and interpretive synthesis. The results were organized into five themes: reinterpretation of the p-value and intervals as measures of compatibility; coexistence of frequentist and Bayesian approaches; reproducibility and quality of statistical reporting; uncertainty quantification in artificial intelligence; and the relationship between causal inference and machine learning. The reviewed literature questions dichotomous decisions based exclusively on statistical significance and emphasizes the need to consider effect sizes, assumptions, context, and practical relevance. It also indicates that inferential validity is shaped by institutional practices, transparency, and study design quality. The study concludes that statistical inference remains essential to scientific reasoning but does not provide definitive certainty. Its rigorous use requires making assumptions explicit, calibrating uncertainty, ensuring analytical transparency, and clearly distinguishing among association, prediction, and causality.

Keywords: Statistical inference; epistemology; statistical significance; causal inference.

RESUMEN

La inferencia estadística es fundamental para producir conocimiento científico, interpretar evidencia y orientar decisiones bajo incertidumbre. Sin embargo, su validez depende del diseño, los supuestos, la transparencia analítica y la adecuada interpretación de los resultados. El objetivo del estudio es sintetizar cualitativamente los alcances y las limitaciones epistemológicas atribuidos a la inferencia estadística en la literatura científica reciente. Se realizó una revisión sistemática cualitativa orientada por PRISMA 2020. Se consultaron las bases de datos de PubMed y Europe PMC para identificar publicaciones aparecidas entre 2021 y 2025. Se incluyeron 13 estudios, analizados mediante una síntesis temática, comparativa e interpretativa. Los resultados se organizaron en cinco ejes: reinterpretación del valor p y los intervalos como medidas de compatibilidad; coexistencia de los enfoques frecuentista y bayesiano; reproducibilidad y calidad del reporte estadístico; cuantificación de la incertidumbre en inteligencia artificial; y relación entre inferencia causal y aprendizaje automático. La literatura revisada cuestiona las decisiones dicotómicas basadas exclusivamente en la significación estadística y destaca la necesidad de considerar tamaños del efecto, supuestos, contexto y relevancia práctica. También señala que la validez inferencial está condicionada por prácticas institucionales, transparencia y calidad del diseño. Se concluye que la inferencia estadística continúa siendo esencial para el razonamiento científico, pero no proporciona certezas definitivas. Su uso riguroso requiere explicitar supuestos, calibrar la incertidumbre, garantizar transparencia analítica y diferenciar claramente asociación, predicción y causalidad.

Palabras clave: Inferencia estadística; epistemología; significación estadística; inferencia causal.



INTRODUCCIÓN

En las ciencias biomédicas y sociales, la epidemiología y la inteligencia artificial, la inferencia estadística permite evaluar intervenciones, analizar asociaciones, estimar efectos y cuantificar la incertidumbre para sustentar decisiones clínicas, poblacionales y automatizadas. Esta presencia transversal confirma su relevancia metodológica, pero también hace visibles sus límites interpretativos, sus fundamentos filosóficos y las condiciones institucionales que influyen en su validez.

La confianza científica en los resultados inferenciales no depende únicamente de la aplicación correcta de procedimientos estadísticos. En esta misma línea, la discusión metodológica contemporánea ha señalado que los resultados estadísticos deben interpretarse junto con la calidad del diseño, la medición, la plausibilidad teórica y el contexto de producción científica (Altman, 1994; Ioannidis, 2005).

Durante gran parte del siglo XX, la inferencia estadística fue presentada como un instrumento para decidir entre hipótesis rivales. El umbral convencional de $p < .05$ se consolidó como criterio operativo para clasificar resultados como estadísticamente significativos o no significativos. Sin embargo, la literatura metodológica reciente ha cuestionado el uso dicotómico de este indicador. Diversos autores advierten que el valor p no debe sustituir el razonamiento científico ni utilizarse como umbral automático de decisión (Wasserstein & Lazar, 2016; Wasserstein et al., 2019). Otros han propuesto abandonar la significación estadística como categoría rígida o redefinir sus umbrales, aunque estas propuestas mantienen debates abiertos sobre sus implicaciones (Amrhein et al., 2019; Benjamin et al., 2018; McShane et al., 2019).

La literatura reciente recomienda interpretar el valor p junto al tamaño del efecto, los intervalos de confianza, los supuestos del modelo y la relevancia sustantiva del hallazgo. Boscardin et al. (2024) sostienen que los valores p solo resultan informativos cuando se acompañan de contexto interpretativo, hipótesis explícita y análisis de sensibilidad. Homa-Bonell (2023) señala que el valor p , por sí mismo, no informa sobre magnitud del efecto, relevancia clínica ni utilidad práctica. Sin embargo, otros autores señalan que los valores p y los intervalos deben interpretarse como indicadores de compatibilidad entre los datos, los modelos y sus supuestos, no como pruebas directas de verdad o falsedad (Greenland et al., 2016; Mansournia et al., 2023; Rafi & Greenland, 2020).

Otra tensión relevante se relaciona con la coexistencia de los enfoques frecuentista y bayesiano. Ambos paradigmas atribuyen significados diferentes a la probabilidad y a la incertidumbre. En el enfoque frecuentista, la probabilidad se vincula con frecuencias relativas en repeticiones hipotéticas. En el enfoque bayesiano, expresa grados racionales de creencia que se actualizan con nueva evidencia. Fornacon-Wood et al. (2022) y Goligher et al. (2024) muestran que el desarrollo computacional ha incrementado la aplicabilidad de los métodos bayesianos, especialmente en investigación clínica. Sin embargo, el enfoque bayesiano también exige justificar las distribuciones previas y evaluar la sensibilidad de los resultados (Gelman et al., 2013; Goodman, 1999a, 1999b).

La reproducibilidad constituye otro eje clave para comprender los alcances de la inferencia estadística. Diversos trabajos han mostrado que la baja potencia estadística, el sesgo de publicación, la flexibilidad analítica y el reporte incompleto pueden comprometer la credibilidad de los hallazgos (Button et al., 2013; Gelman & Loken, 2014; Munafò et al., 2017). La preregistración y la apertura de materiales se han propuesto como estrategias para distinguir análisis confirmatorios y exploratorios, reducir decisiones analíticas posteriores a los resultados y fortalecer la trazabilidad de la evidencia (Nosek et al., 2018).

En años recientes, la expansión de la inteligencia artificial y de los modelos predictivos complejos ha ampliado la discusión sobre la inferencia estadística. El aprendizaje automático interpretable ofrece oportunidades para el descubrimiento científico, aunque enfrenta retos de validación, estabilidad y cuantificación de la incertidumbre (Allen et al., 2024). Asimismo, la inferencia conformal depende de supuestos específicos y requiere ajustes cuando se aplica a datos no intercambiables (Oliveira et al., 2024), mientras que la cuantificación explícita de la incertidumbre puede mejorar los análisis posteriores en transcriptómica espacial (Sun et al., 2024). Además, un sistema de inteligencia artificial puede mostrar un buen rendimiento promedio y, al mismo tiempo, presentar altos niveles de incertidumbre para individuos o grupos específicos (Chakraborti et al., 2025).

La discusión contemporánea también ha sido reconfigurada por el desarrollo de la inferencia causal. Brand et al. (2023) señalan que la integración entre inferencia causal y aprendizaje automático abre posibilidades para estimar heterogeneidad, mediación e interferencia, pero exige diferenciar capacidad predictiva y explicación causal. Por su parte, Tejada-Lapuerta et al. (2025) indican que el aprendizaje causal aplicado a la genómica de célula única enfrenta problemas de generalización, interpretabilidad e inferencia de dinámicas celulares. Desde la perspectiva de la inferencia causal, la predicción no sustituye la explicación de mecanismos ni la evaluación de intervenciones (Hernán & Robins, 2020; Pearl, 2009).

A partir de estas tensiones, el presente estudio da respuesta a la siguiente interrogante: ¿cómo caracteriza la literatura reciente los alcances y las limitaciones epistemológicas de la inferencia estadística en la ciencia contemporánea? El objetivo general fue sintetizar cualitativamente los alcances y las limitaciones epistemológicas atribuidos a la inferencia estadística en la literatura científica reciente. El planteamiento central sostiene que la validez epistémica de la inferencia estadística depende de la explicitación de márgenes

de error, la formulación de decisiones revisables, la organización de desacuerdos mediante procedimientos públicos y la vinculación de los datos con modelos susceptibles de crítica, refinamiento o sustitución.

METODOLOGÍA

Diseño del estudio

Se realizó una revisión sistemática orientada por la declaración PRISMA 2020 (Page et al., 2021). La revisión fue cualitativa en su lógica de síntesis y analítica en su propósito, debido a que buscó identificar líneas argumentales, núcleos de controversia y cambios conceptuales relevantes.

Fuentes de información y estrategia de búsqueda

La búsqueda se realizó en las bases de datos PubMed y Europe PMC. Se emplearon combinaciones en inglés debido a que la literatura recuperada sobre estos temas se concentra predominantemente en ese idioma. La ecuación base fue: (“statistical inference” OR “inferential statistics”) AND (epistemology OR “p-value” OR compatibility OR Bayesian OR frequentist OR reproducibility OR “causal inference” OR “uncertainty quantification” OR “machine learning”). A partir de ella se generaron variantes específicas para localizar trabajos sobre reporte estadístico, inferencia conformal, investigación clínica y aprendizaje causal. También se revisaron referencias cruzadas en los artículos inicialmente elegibles para ampliar la sensibilidad de la búsqueda.

Criterios de inclusión y exclusión

Se incluyeron documentos que cumplieran simultáneamente los siguientes criterios: haber sido publicados entre 2021 y 2025; contar con revisión por pares o estar publicados en una revista académica reconocida; abordar de manera explícita uno o más de los ejes temáticos definidos; presentar contenido suficiente para identificar implicaciones epistemológicas; y disponer del resumen o del texto completo necesario para la síntesis. En cuanto al tipo de publicación, se consideraron artículos de revisión, perspectivas, editoriales metodológicas, y estudios empíricos, siempre que su relevancia para el análisis inferencial estuviera claramente establecida.

Por otra parte, se excluyeron los trabajos puramente técnicos que no incorporaran una discusión inferencial sustantiva; los artículos publicados fuera del periodo establecido; los registros duplicados; los documentos centrados exclusivamente en el rendimiento algorítmico, sin análisis de incertidumbre o causalidad; y los textos en los que la estadística apareciera de manera periférica. Asimismo, se descartaron las piezas divulgativas, las notas docentes no arbitradas y las entradas de opinión sin respaldo editorial científico.

Proceso de selección

Como se encuentra representado en la Figura 1, la selección se realizó en cuatro etapas que establece la guía PRISMA: identificación, depuración, elegibilidad y síntesis. Se recuperaron 56 registros mediante búsqueda estructurada y revisión de referencias. Después de eliminar duplicados o registros claramente irrelevantes, quedaron 35 documentos para evaluación por título y resumen. De ellos, 19 pasaron a revisión de elegibilidad mediante texto completo o resúmenes ampliados, y 13 estudios fueron incorporados a la síntesis cualitativa final.

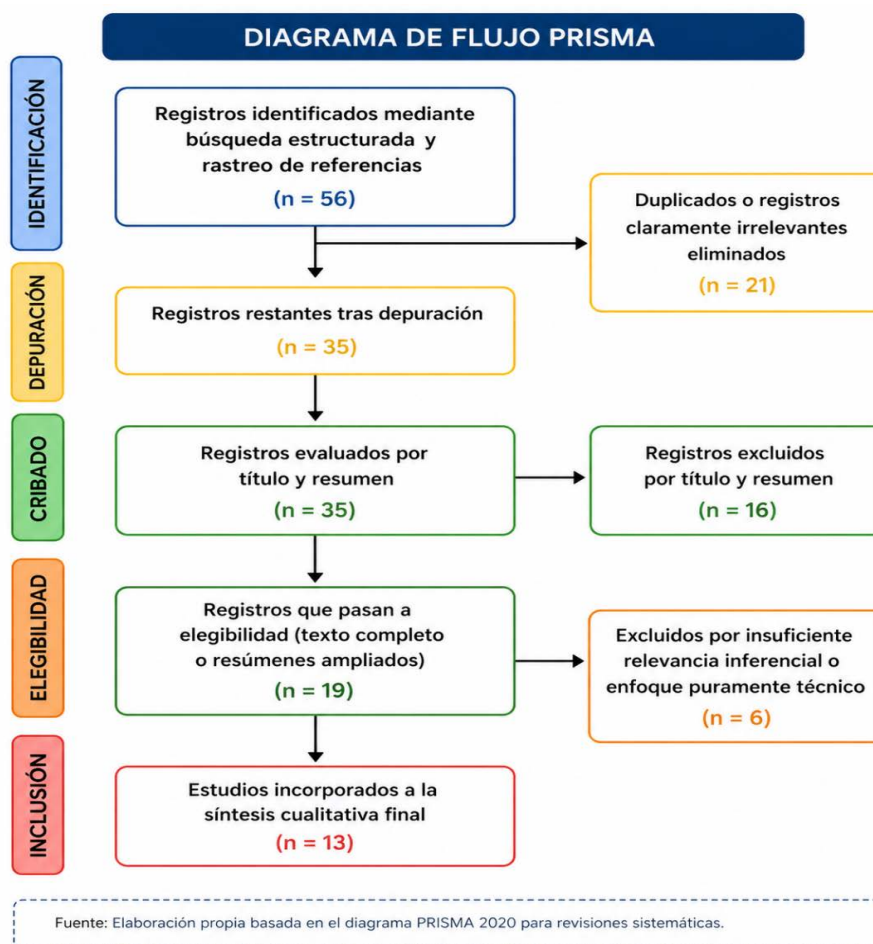


Figura 1. Diagrama de selección de estudios incluidos en la revisión sistemática.

Análisis de contenido

Para cada estudio incluido se registraron autoría y año, tipo de publicación, campo disciplinar, problema inferencial central, alcance atribuido a la inferencia, limitación principal identificada y propuesta de mejora o relectura conceptual. La codificación fue temática y comparativa. Primero se realizó una lectura analítica de cada texto; luego se agruparon los hallazgos en categorías emergentes; finalmente se construyó una síntesis interpretativa transversal.

RESULTADOS Y DISCUSIÓN

Caracterización general del corpus

La Tabla 1 presenta las principales características de los estudios incluidos en la revisión, considerando el autor y año de publicación, el tipo de documento, el tema central abordado y su principal aporte al análisis de la inferencia estadística.

Tabla 1. Características de los estudios incluidos en la revisión sistemática.

Estudio	Tipo	Tema central	Aporte principal
Boscardin et al. (2024)	Guía metodológica	Uso y reporte de valores p	Recomienda contextualizar p con tamaño del efecto, intervalos e hipótesis.
Homa-Bonell (2023)	Editorial	Interpretación de valores p	Plantea que el valor p debe complementarse con relevancia práctica y clínica.

Estudio	Tipo	Tema central	Aporte principal
Mansournia et al. (2023)	Artículo conceptual	Compatibilidad y S-value	Reformula p e intervalos desde una lógica de compatibilidad.
Fornacon-Wood et al. (2022)	Tutorial/revisión	Bayesiano vs. frecuentista	Expone diferencias conceptuales y operativas entre paradigmas.
Goligher et al. (2024)	Revisión	Bayesianismo clínico	Describe la investigación clínica como reducción progresiva de incertidumbre.
Cobey et al. (2024)	Encuesta internacional	Reproducibilidad	Muestra que la crisis es percibida como amplia y vinculada a incentivos institucionales.
Mansournia & Nazemipour (2024)	Correspondencia	Reporte estadístico	Resume deficiencias frecuentes en manuscritos y propone mejores prácticas.
Allen et al. (2024)	Revisión	Interpretable machine learning	Subraya retos de validación, estabilidad y cuantificación de incertidumbre.
Oliveira et al. (2024)	Artículo metodológico	Inferencia conformal	Extiende garantías a datos no intercambiables con penalización controlada.
Sun et al. (2024)	Artículo de métodos	TISSUE	Integra incertidumbre calibrada en predicción transcriptómica espacial.
Brand et al. (2023)	Revisión	Causalidad y machine learning	Propone integrar identificación causal con estimación flexible.
Chakraborti et al. (2025)	Perspective	Incetidumbre personalizada en IA	Advierte que el rendimiento promedio no basta en decisiones de alto riesgo.
Tejada-Lapuerta et al. (2025)	Perspective	Aprendizaje causal en genómica	Identifica retos de generalización, interpretabilidad y dinámica celular.

La distribución temática del corpus permitió identificar cinco núcleos. El primero agrupó estudios dedicados a la reinterpretación del valor p, de la significación estadística y de los intervalos de confianza. El segundo reunió trabajos sobre pluralismo frecuentista y bayesiano. El tercero concentró investigaciones y orientaciones sobre reproducibilidad y calidad del reporte. El cuarto incluyó trabajos sobre cuantificación de incertidumbre y calibración predictiva en inteligencia artificial. El quinto se articuló en torno a inferencia causal y aprendizaje automático. Esta organización evidencia que el análisis de los alcances de la inferencia se sitúa en la intersección entre metodología, prácticas institucionales y tecnologías de decisión.

Del valor p como umbral al valor p como compatibilidad

Boscardin et al. (2024) reconocen que el valor p continúa siendo frecuente en la investigación cuantitativa, pero advierten que su uso solo resulta informativo cuando se acompaña de contexto interpretativo, hipótesis explícita, tamaño del efecto, intervalo de confianza y, cuando corresponde, análisis de sensibilidad. Esta propuesta no implica eliminar el valor p, sino evitar que un indicador parcial se convierta en criterio único de validez científica.

Homa-Bonell (2023) refuerza esta crítica desde una perspectiva aplicada. Señala que investigadores y clínicos suelen otorgar al valor p un peso interpretativo excesivo. La interpretación adecuada debe incorporar estimaciones de efecto, intervalos de confianza y relevancia contextual. Esta observación tiene implicaciones

epistemológicas porque muestra que la inferencia no termina con el cálculo de una probabilidad condicionada bajo la hipótesis nula, sino con la inserción de ese resultado en un juicio sustantivo informado por el diseño, el tamaño muestral, la plausibilidad teórica y las consecuencias prácticas.

Por su parte, Mansournia et al. (2023) profundizan esta revisión mediante la noción de compatibilidad. En su formulación, el valor p puede interpretarse como un índice de compatibilidad entre los datos observados y el modelo que incluye la hipótesis contrastada y sus supuestos de fondo. Los intervalos representan, a su vez, el conjunto de valores paramétricos más compatibles con los datos bajo esas condiciones.

En este bloque temático se observa que el alcance de la inferencia estadística no consiste en autorizar decisiones binarias, sino en graduar la compatibilidad entre datos y modelos. Su limitación principal también es clara: cuando el modelo es inadecuado, los supuestos no se cumplen o el diseño presenta deficiencias, la precisión aparente del resultado puede conducir a interpretaciones erróneas. Un análisis estadístico formalmente correcto no compensa deficiencias en el diseño, la medición o el reporte de una investigación.

Frecuentismo y bayesianismo: pluralismo epistemológico de la incertidumbre

El segundo núcleo del corpus mostró que la relación entre enfoques frecuentista y bayesiano sigue siendo un punto central de la inferencia contemporánea. Fornacon-Wood et al. (2022) explican que ambos paradigmas operan con concepciones distintas de probabilidad. En el enfoque frecuentista, los parámetros son fijos y desconocidos, mientras la aleatoriedad corresponde a los datos hipotéticos que podrían observarse bajo repetición. En el enfoque bayesiano, la incertidumbre se representa directamente sobre los parámetros mediante distribuciones previas y posteriores.

Goligher et al. (2024) sitúan esta diferencia en investigación clínica y muestran por qué el enfoque bayesiano ha ganado relevancia. Al comparar la investigación con un proceso diagnóstico, los autores describen el análisis bayesiano como reducción progresiva de incertidumbre mediante acumulación de información. Mientras el frecuentismo evalúa el comportamiento de un procedimiento en repeticiones hipotéticas, el bayesianismo actualiza racionalmente la creencia sobre una hipótesis a partir de datos nuevos.

La revisión no muestra una sustitución simple del frecuentismo por el bayesianismo, sino un pluralismo metodológico activo. El frecuentismo conserva ventajas de estandarización y uso extendido. El bayesianismo ofrece flexibilidad para incorporar conocimiento previo y formular inferencias más directamente interpretables, pero exige justificar las distribuciones previas y evaluar la sensibilidad a supuestos. Ninguno de los dos paradigmas elimina la incertidumbre; ambos la formalizan de manera distinta.

Desde una perspectiva epistemológica, este pluralismo tiene dos implicaciones. Primero, impide comprender la estadística como un campo monolítico y conceptualmente neutro. Segundo, exige explicitar la concepción de incertidumbre que guía el análisis. No basta con aplicar un procedimiento estadístico y reportar resultados; es necesario justificar el marco inferencial adoptado y sus condiciones de interpretación.

Reproducibilidad, reporte y ecología institucional de la inferencia

La tercera línea temática evidenció que una parte importante de la crisis inferencial contemporánea no se origina únicamente en los procedimientos matemáticos, sino en la organización social de la ciencia. Cobey et al. (2024), a partir de 1.630 respuestas utilizables en una encuesta internacional, informaron que el 72% de los participantes considera que existe una crisis de reproducibilidad en biomedicina; el 62% identifica la presión por publicar como una causa frecuente de irreproducibilidad; solo el 16% percibe que su institución posee procedimientos establecidos para mejorarla; y el 67% considera que se valora más la investigación novedosa que los estudios de réplica. Estos datos proponen que el problema de la inferencia incluye incentivos, culturas editoriales y economías de prestigio.

El artículo de Mansournia & Nazemipour (2024) complementa esta lectura desde el ámbito editorial. Basados en la revisión estadística de más de mil manuscritos enviados a revistas del grupo The Lancet, los autores identificaron deficiencias frecuentes de reporte y formulan recomendaciones para una comunicación estadística más precisa. La recurrencia de problemas de especificación, reporte y justificación metodológica indica que las fallas inferenciales no son hechos aislados, sino dificultades sistemáticas de la comunicación científica.

La dimensión epistemológica de esta línea temática es relevante. La validez científica no depende solo de la corrección interna del método, sino también de las condiciones institucionales que rodean su aplicación. Una inferencia puede estar correctamente formulada y, aun así, contribuir a conocimiento frágil si se produce en un contexto de flexibilidad analítica no declarada, publicación selectiva, presión por novedad, insuficiencia de réplica o bajo nivel de transparencia.

Desde esta perspectiva, la inferencia estadística puede funcionar como un procedimiento de control metodológico cuando se inserta en prácticas de apertura, preregistro, análisis de sensibilidad y reporte

completo. Sin embargo, no compensa por sí sola incentivos institucionales que privilegian la novedad sobre la robustez. El uso riguroso de la estadística requiere prácticas editoriales e institucionales orientadas a la transparencia, la reproducibilidad y el reporte completo.

Cuantificación de incertidumbre: del contraste clásico a la calibración predictiva

Este núcleo temático se concentró en la expansión de la inferencia hacia problemas de predicción compleja y sistemas algorítmicos. En este contexto, la pregunta central no es únicamente si una asociación resulta compatible con una hipótesis nula, sino cuán confiable es una predicción específica y bajo qué margen de error puede emplearse responsablemente. Oliveira et al. (2024) ofrecen un ejemplo relevante: la inferencia conformal, utilizada para construir intervalos predictivos con garantías formales, depende del supuesto de intercambiabilidad de los datos. Cuando ese supuesto no se cumple, como ocurre en series temporales o datos espaciotemporales, las garantías deben reformularse mediante ajustes teóricos y penalizaciones controladas.

Sun et al. (2024) trasladan esta discusión a una aplicación biomédica concreta. Su método TISSUE utiliza inferencia conformal para estimar incertidumbre en la predicción de expresión génica espacial a nivel unicelular y reporta intervalos calibrados en once conjuntos de datos de referencia. Del mismo modo, Chakraborti et al. (2025) amplían este argumento al campo de la inteligencia artificial de alto impacto. Los autores señalan que los modelos pueden ser precisos en promedio y, al mismo tiempo, presentar alta incertidumbre para individuos específicos o grupos particulares. Por ello proponen avanzar hacia una cuantificación personalizada de incertidumbre.

Allen et al. (2024) debaten el aprendizaje automático interpretable orientado al descubrimiento científico e identifican retos persistentes en validación, estabilidad, consistencia de selección y cuantificación de incertidumbre. La interpretabilidad no equivale automáticamente a credibilidad epistémica. Un modelo interpretable puede ser inestable, depender de particiones específicas de datos o ser vulnerable a sobreajuste. Por tanto, la inferencia contemporánea debe evaluar si los mecanismos identificados por el algoritmo sostienen descubrimientos reproducibles y confiables.

En este escenario, la inferencia estadística contribuye a que las predicciones complejas se traduzcan en decisiones más prudentes mediante la cuantificación de incertidumbre y la identificación de áreas de menor fiabilidad. Su límite también se incrementa: cuanto más complejos son los sistemas y más heterogéneos los datos, más difícil resulta mantener garantías sólidas sin aumentar los supuestos o reducir la interpretabilidad.

Inferencia causal y aprendizaje automático: entre asociación, intervención y explicación

El quinto núcleo temático mostró una tensión central para la ciencia contemporánea: la diferencia entre predecir y explicar. Brand et al. (2023) revisaron desarrollos recientes en inferencia causal y aprendizaje automático y sostienen que ambos campos pueden complementarse si se mantienen claras sus funciones. La inferencia causal se ha concentrado históricamente en la identificación; el aprendizaje automático, en la estimación flexible y el desempeño predictivo. Su combinación ofrece ventajas para estimar heterogeneidad, mediación e interferencia espacial o temporal, pero exige preservar la lógica de las preguntas causales.

Tejada-Lapueta et al. (2025) ofrecen un caso relevante en genómica de célula única. Su perspectiva identifica tres problemas abiertos para el aprendizaje causal en ese campo: generalización a condiciones experimentales no observadas, interpretabilidad de modelos aprendidos y dificultad para inferir dinámicas celulares. Esta discusión muestra que el paso desde asociación hacia causalidad no depende solo de una mayor cantidad de datos o de modelos más complejos, sino de supuestos sustantivos sobre mecanismos, intervenciones plausibles y estabilidad entre contextos.

Esta línea temática confirma que la inferencia estadística contemporánea responde a dos demandas simultáneas. Por una parte, debe atender el interés por el rendimiento predictivo. Por otra, debe preservar la distinción entre predicción, explicación e intervención. La predicción puede ser suficiente cuando el objetivo es clasificar, anticipar o priorizar riesgos. La causalidad es necesaria cuando la pregunta adopta la forma de una intervención: qué ocurriría si se actuara de otro modo. Un modelo con alta capacidad predictiva no proporciona, por sí mismo, una explicación causal ni una base suficiente para intervenir.

Desde una perspectiva epistemológica, este bloque muestra que la acumulación de correlaciones no sustituye a la teoría. La literatura revisada sugiere que, a medida que aumenta el poder predictivo, se vuelve más importante diferenciar entre patrones útiles, mecanismos plausibles y relaciones causales defendibles. El aprendizaje automático no elimina la necesidad de inferencia; por el contrario, exige criterios inferenciales explícitos.

Alcances y limitaciones epistemológicas de la inferencia estadística

La lectura transversal del corpus permite sintetizar que la inferencia estadística contemporánea no se comprende adecuadamente como un procedimiento binario de aceptación o rechazo. El debate actual privilegia lenguajes de compatibilidad, estimación, rangos plausibles y sensibilidad contextual. Se reconoce su pluralismo filosófico, expresado en la coexistencia de enfoques frecuentistas y bayesianos, pero a crisis de reproducibilidad ha mostrado que la validez inferencial depende del análisis, pero también del sistema de incentivos, del reporte y de la cultura de investigación. Del mismo modo, el desarrollo de la inteligencia artificial ha trasladado parte de la discusión hacia la calibración de predicciones, la incertidumbre individualizada y la evaluación de sistemas complejos. Finalmente, el avance de la inferencia causal ha reafirmado la diferencia entre asociación, predicción e intervención.

Como se muestra en la Tabla 2, los hallazgos convergen en cinco dimensiones: la reinterpretación del valor p , el pluralismo frecuentista-bayesiano, la reproducibilidad, la cuantificación de la incertidumbre en inteligencia artificial y la inferencia causal. Para cada dimensión se presentan su contribución epistemológica y sus principales restricciones.

Tabla 2. Alcances y limitaciones epistemológicas de la inferencia estadística identificadas en los estudios revisados.

Categoría	Hallazgos principales	Alcance epistemológico	Limitación principal
Valor p y compatibilidad	Se cuestiona la dicotomía significativo/no significativo.	La inferencia gradúa compatibilidad entre datos y modelos.	Sobreinterpretación del valor p como criterio único de validez.
Pluralismo frecuentista/bayesiano	Conviven formalizaciones distintas de la incertidumbre.	La estadística no es conceptualmente monolítica.	Riesgo de interpretaciones opacas si no se justifican los supuestos.
Reproducibilidad y reporte	Persisten presiones institucionales y fallos de comunicación estadística.	La validez inferencial depende del ecosistema científico.	Incentivos editoriales que privilegian novedad sobre robustez.
Cuantificación de incertidumbre en inteligencia artificial	Aumenta la demanda de calibración predictiva e incertidumbre individualizada.	La inferencia se expande a sistemas algorítmicos complejos.	Dependencia de supuestos y dificultad de interpretación.
Inferencia causal y aprendizaje automático	Se integran identificación causal y estimación flexible.	Se distinguen mejor predicción, explicación e intervención.	Persisten retos de generalización y supuestos causales.

Durante décadas, se asumió que un procedimiento correcto y un resultado situado por debajo de un umbral convencional eran suficientes para autorizar una conclusión. La literatura revisada cuestiona esa simplificación y propone comprender la inferencia como una práctica formal de razonamiento público bajo incertidumbre. El primer desplazamiento identificado es semántico y epistemológico: pasar de la significación a la compatibilidad. Esta transición reduce la sobrecarga interpretativa atribuida al valor p . Cuando los resultados se presentan como compatibles o incompatibles, en cierto grado, con un modelo y sus supuestos, la inferencia adquiere mayor precisión conceptual.

El segundo desplazamiento es filosófico: la visibilidad del pluralismo entre frecuentismo y bayesianismo. La revisión mostró que la estadística actual opera con más de una concepción de probabilidad y, en consecuencia, con más de una forma legítima de representar incertidumbre. Este hecho desaconseja posiciones metodológicas rígidas y exige una formación estadística más reflexiva. La enseñanza de la estadística no debe limitarse a reglas de decisión o uso de software, sino que debe explicar las opciones conceptuales sobre las que descansan los resultados.

Del mismo modo, la validez inferencial depende del ecosistema institucional de la ciencia. La encuesta de Cobey et al. (2024) y las observaciones editoriales de Mansournia & Nazemipour (2024) sugieren que parte del problema contemporáneo no reside en la falta de herramientas estadísticas adecuadas, sino en estructuras de incentivos que favorecen la sobreinterpretación, el reporte incompleto y la publicación rápida.

La discusión sobre inteligencia artificial añade un componente ético relevante. La inferencia ya no se aplica solo a estudios y poblaciones, sino también a sistemas automatizados que pueden incidir sobre personas específicas. La incertidumbre personalizada propuesta por Chakraborti et al. (2025) muestra que no basta con que el promedio esté bien calibrado si los errores se concentran en grupos particulares. Por ello, la epistemología de la inferencia se vincula con debates sobre equidad, responsabilidad y distribución del error en sistemas técnicos.

Asimismo, el desarrollo de métodos como la inferencia conformal aporta una consideración metodológica relevante. Las nuevas técnicas reformulan la necesidad de supuestos. Oliveira et al. (2024) muestran que, cuando la intercambiabilidad falla, las garantías deben revisarse. El uso de métodos algorítmicos avanzados no elimina la necesidad de explicitar los supuestos, las condiciones de validez y los límites interpretativos del análisis.

La predicción es indispensable en muchos contextos operativos y puede ofrecer beneficios prácticos importantes. Sin embargo, cuando la pregunta científica implica intervención, política pública o decisión clínica orientada al cambio, la explicación causal sigue siendo necesaria. El error metodológico consiste en asumir que un modelo con alta capacidad predictiva proporciona, por sí mismo, una explicación causal o una base suficiente para la intervención.

A nivel pedagógico, la formación estadística suele oscilar entre la abstracción formal desconectada de la práctica y la enseñanza procedimental reducida a la selección de pruebas. Ambos enfoques son insuficientes. Formar en inferencia exige enseñar interpretación, reporte, sensibilidad a supuestos, distinción entre asociación y causalidad, incertidumbre calibrada y conciencia de los sesgos institucionales que moldean la investigación. La formación estadística debe superar la enseñanza meramente procedimental y fortalecer la interpretación crítica de los resultados, los supuestos y los límites de cada procedimiento.

También es necesario ubicar la inferencia dentro de marcos más amplios de diseño, medición y teoría. Ningún procedimiento inferencial puede corregir una medición deficiente, variables mal conceptualizadas o diseños incapaces de distinguir mecanismos plausibles. Los procedimientos estadísticos pueden identificar limitaciones analíticas, pero no corrigen deficiencias originadas en el diseño, la medición o la conceptualización de variables.

Dentro de las limitaciones del presente trabajo se declara que el corpus documental privilegió textos conceptualmente densos y de alta visibilidad, por lo que puede subrepresentar discusiones regionales o aplicadas fuera del circuito anglófono. Además, la interpretación fue deliberadamente filosófica; otros trabajos podrían enfatizar implicaciones técnicas, regulatorias o disciplinares. Estas limitaciones delimitan el alcance de la contribución: se trata de una síntesis analítica de debates recientes, no de un censo universal de toda la producción relevante.

CONCLUSIONES

La inferencia estadística constituye un componente esencial para la producción de conocimiento científico, pero no proporciona conclusiones definitivas, su valor está en permitir la interpretación de la evidencia, estimar la incertidumbre y formular conclusiones condicionadas por el diseño, los supuestos del modelo y la calidad del reporte. En este sentido, el valor p debe analizarse junto con el tamaño del efecto, los intervalos de confianza y la relevancia sustantiva del hallazgo, mientras que los enfoques frecuentista y bayesiano deben asumirse como marcos complementarios para formalizar la incertidumbre.

La validez inferencial depende de factores metodológicos, analíticos e institucionales. Asimismo, en contextos como la inteligencia artificial y el aprendizaje automático, resulta imprescindible distinguir entre asociación, predicción y causalidad, a fin de evitar interpretaciones excedidas. Por tanto, la ciencia contemporánea requiere una inferencia estadística aplicada con rigor metodológico, claridad conceptual y responsabilidad científica.

REFERENCIAS

- Allen, G. I., Gan, L., & Zheng, L. (2024). Interpretable machine learning for discovery: Statistical challenges and opportunities. *Annual Review of Statistics and Its Application*, 11, 97-121. <https://doi.org/10.1146/annurev-statistics-040120-030919>
- Altman, D. G. (1994). The scandal of poor medical research. *The BMJ*, 308(6924), 283-284. <https://doi.org/10.1136/bmj.308.6924.283>
- Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748), 305-307. <https://doi.org/10.1038/d41586-019-00857-9>
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P.,

- Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10. <https://doi.org/10.1038/s41562-017-0189-z>
- Boscardin, C. K., Sewell, J. L., Tolsgaard, M. G., & Pusic, M. V. (2024). How to use and report on p-values. *Perspectives on Medical Education*, 13(1), 250-254. <https://doi.org/10.5334/pme.1324>
- Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49, 81-110. <https://doi.org/10.1146/annurev-soc-030420-015345>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365-376. <https://doi.org/10.1038/nrn3475>
- Chakraborti, T., Banerji, C. R. S., Marandon, A., Hellon, V., Mitra, R., Lehmann, B., Braeuninger, L., McGough, S., Turkay, C., Frangi, A. F., Bianconi, G., Li, W., Rackham, O., Harbron, C., Parashar, D., & MacArthur, B. D. (2025). Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7, 522-530. <https://doi.org/10.1038/s42256-025-01024-8>
- Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., et al. (2024). Biomedical researchers' perspectives on the reproducibility of research. *PLOS Biology*, 22(11), e3002870. <https://doi.org/10.1371/journal.pbio.3002870>
- Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the differences between Bayesian and frequentist statistics. *International Journal of Radiation Oncology, Biology, Physics*, 112(5), 1076-1082. <https://doi.org/10.1016/j.ijrobp.2021.12.011>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460-465. <https://doi.org/10.1511/2014.111.460>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.
- Goligher, E. C., Heath, A., & Harhay, M. O. (2024). Bayesian statistics for clinical research. *The Lancet*, 404(10457), 1067-1076. [https://doi.org/10.1016/S0140-6736\(24\)01295-9](https://doi.org/10.1016/S0140-6736(24)01295-9)
- Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130(12), 995-1004. <https://doi.org/10.7326/0003-4819-130-12-199906150-00008>
- Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12), 1005-1013. <https://doi.org/10.7326/0003-4819-130-12-199906150-00019>
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350. <https://doi.org/10.1007/s10654-016-0149-3>
- Homa-Bonell, J. K. (2023). Interpreting P values in 2023. *Journal of Patient-Centered Research and Reviews*, 10(3), 102-103. <https://doi.org/10.17294/2330-0698.2064>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Mansournia, M. A., & Nazemipour, M. (2024). Recommendations for accurate reporting in medical research statistics. *The Lancet*, 403(10427), 611-612. [https://doi.org/10.1016/S0140-6736\(24\)00139-9](https://doi.org/10.1016/S0140-6736(24)00139-9)
- Mansournia, M. A., Nazemipour, M., & Etminan, M. (2023). P-value, compatibility, and S-value. *Global Epidemiology*, 5, 100085. <https://doi.org/10.1016/j.gloepi.2022.100085>
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235-245. <https://doi.org/10.1080/00031305.2018.1527253>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606. <https://doi.org/10.1073/pnas.1708274114>
- Oliveira, R. I., Orenstein, P., Ramos, T., & Romano, J. V. (2024). Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225), 1-38. <https://www.jmlr.org/papers/v25/23-1553.html>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
- Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20, 244. <https://doi.org/10.1186/s12874-020-01105-9>
- Sun, E. D., Ma, R., Navarro Negredo, P., Brunet, A., & Zou, J. (2024). TISSUE: Uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *Nature Methods*, 21, 444-454. <https://doi.org/10.1038/s41592-024-02184-y>
- Tejada-Lapueta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y., & Theis, F. J. (2025). Causal machine learning for single-cell genomics. *Nature Genetics*, 57(4), 797-808. <https://doi.org/10.1038/s41588-025-02124-2>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129-133. <https://doi.org/10.1080/00031305.2016.1154108>

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond $p < .05$. *The American Statistician*, 73(sup1), 1-19. <https://doi.org/10.1080/00031305.2019.1583913>

Conflicts of Interest:

The authors declare no conflicts of interest.

Author Contributions:

The authors were responsible for all aspects of the study, including conceptualization, methodology, analysis, and writing.

Disclaimer/Publisher's Note:

The statements, opinions, and data contained in all publications are solely those of the authors and individual contributors and not of Revista San Gregorio or the editors. Revista San Gregorio and/or the editors disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.