

71

Fecha de presentación: octubre, 2021

Fecha de aceptación: diciembre, 2021

Fecha de publicación: enero, 2022

EL USO

DE LAS PRUEBAS DE HIPÓTESIS EN LA INVESTIGACIÓN EDUCATIVA

THE USE OF HYPOTHESIS TESTING IN EDUCATIONAL RESEARCH

Fermín Campos Solorzano¹

E-mail: fermin_campos@unu.edu.pe

ORCID: <https://orcid.org/0000-0002-6237-9234>

Harold Abraham Choquetico Apaza²

E-mail: hchoqueticoa@unia.edu.pe

ORCID: <https://orcid.org/0000-0003-3889-7266>

Beker Maraza Vilcanqui²

E-mail: bmarazav@unia.edu.pe

ORCID: <https://orcid.org/0000-0002-0155-4094>

Carlos Manuel Zacarias Mercado²

E-mail: czacariasm@unia.edu.pe

ORCID: <https://orcid.org/0000-0003-4326-9977>

Aniceto Elías Aguilar Polo²

E-mail: aaguilarp@uladech.edu.pe

ORCID: <https://orcid.org/0000-0002-0474-3843>

¹ Universidad Nacional de Ucayali. Perú.

² Universidad Nacional Intercultural de la Amazonía. Perú.

Cita sugerida (APA, séptima edición)

Campos Solorzano, F., Choquetico Apaza, H. A., Maraza Vilcanqui, B., Zacarias Mercado, C. M., & Aguilar Polo, A. E. (2022).

El uso de las pruebas de hipótesis en la investigación educativa. *Revista Universidad y Sociedad*, 14(1), 701-711.

RESUMEN

En las últimas décadas se ha producido un gran desarrollo de la investigación empírica, tanto a nivel nacional como internacional. Predominan en los estudios de humanidades, especialmente entre los de investigación educativa. Los estudios estadísticos de este campo de investigación suelen incluir varios procedimientos para verificar las hipótesis estadísticas, normalmente se utilizan las pruebas chi-cuadrado (y similares), las pruebas t y F, así como otras de bondad de ajuste. Los informes y la interpretación de las pruebas mencionadas pueden ser diferentes en distintas situaciones de investigación y los errores teóricos y de concepto resultan comunes en dichos informes. La intención de los autores del presente artículo es identificar las principales carencias metodológicas y de concepción teórica en investigadores de las carreras de las ingenierías vinculadas a las ciencias de la computación y la información. A partir de la aplicación de una encuesta no evaluativa se pudo constatar dificultades relacionadas con el planteamiento de las hipótesis, el uso incorrecto de los valores p y la escasa utilización de análisis complementarios al nivel de significación.

Palabras clave: Prueba de hipótesis, errores de presentación, significación, valor p.

ABSTRACT

In recent decades there has been a great development of empirical research, both nationally and internationally. They predominate in humanities studies, especially among those of educational research. Statistical studies in this field of research usually include several procedures to verify statistical hypotheses, usually chi-square (and similar) tests, t-tests and F-tests as well as other goodness-of-fit tests are used. The reporting and interpretation of the above tests may be different in different research situations, and theoretical and conceptual errors are common in such reports. The intention of the authors of this article is to identify the main methodological and theoretical conception deficiencies in researchers of engineering careers related to computer and information sciences. The application of a non-evaluative survey revealed difficulties related to the formulation of hypotheses, the incorrect use of p-values and the scarce use of complementary analyses at the significance level.

Keywords: Hypothesis testing, presentation errors, significance, p-value.

INTRODUCCIÓN

Casi todos los estudios, especialmente los empíricos, contienen hipótesis. Son respuestas no verificadas al problema de investigación o a una pregunta específica de investigación. “La construcción de hipótesis es un procedimiento que consiste en la construcción de una teoría implícita sobre la naturaleza de un fenómeno, los vínculos entre sus elementos, su tamaño, proporción, etc.” (Eastwell, 2014). Las hipótesis suelen basarse en la teoría existente. En el modelo de investigación clásico, la pregunta general de investigación o el problema de investigación se definen al principio. Esta etapa suele ser la que más tiempo consume, ya que requiere una larga búsqueda bibliográfica. Sin embargo, cuando se hace bien, marca un gran paso en el curso de la investigación. En el momento en que se ha definido la pregunta general (problema de investigación), la respuesta a la misma puede suponerse de antemano. Esta respuesta se llama hipótesis (Gyllenpalm & Wickman, 2011).

Dado que las leyes más importantes de la pedagogía son las que se refieren a las relaciones y vínculos entre los fenómenos educativos, la mayoría de las veces tanto las preguntas como las hipótesis de investigación se refieren a estas relaciones y vínculos. Definitivamente, con menos frecuencia, las hipótesis hablan del estado de cosas en la población encuestada (por ejemplo, cuántos profesores utilizan el ordenador cuando preparan las clases, cuántos alumnos tienen su propio pupitre, etc.). Sin embargo, la cuestión de la formulación de hipótesis debe abordarse con cierta cautela e incluso con dudas y con la pregunta: ¿qué sentido tiene la formulación de hipótesis cuando la respuesta propuesta por la hipótesis es el objeto de la búsqueda (es decir, el objetivo de la investigación)? Todas las actividades de la investigación están dirigidas a encontrar una respuesta a la pregunta de investigación. Por lo tanto, cabe preguntarse por qué se plantea una hipótesis de investigación al principio de un estudio si sólo se trata de encontrar una respuesta (Gyllenpalm & Wickman, 2011).

El sentido de las hipótesis radica en que deben orientar la investigación hacia los elementos y aspectos esenciales del problema estudiado. La ausencia de una hipótesis conduciría, por consiguiente, a una situación en la que en cada estudio habría que empezar de cero, como si el investigador no supiera absolutamente nada sobre el problema investigado o como si no hubiera más teorías generales relativas a los problemas investigados (Reilly, 2019; Ioannidou & Erduran, 2021).

En la investigación más cuantitativa, las hipótesis suelen formularse de forma deductiva, es decir, al principio del

estudio. Mientras que en la investigación más cualitativa las hipótesis se forman más a menudo de forma inductiva, durante la propia investigación. Una situación en la que el investigador sigue inflexiblemente sólo las hipótesis que se forman por medio de la deducción resulta ser una solución débil o incluso mala. Igualmente, incorrecta es la orientación del investigador exclusivamente hacia las hipótesis procedentes de la inducción (Rao, 2019).

La búsqueda de respuestas a las preguntas de la investigación supone, de hecho, la verificación de las hipótesis. Sin embargo, el término “prueba de hipótesis” se utiliza con mayor frecuencia en el caso de la verificación estadística (por ejemplo, de las hipótesis nulas). La verificación se lleva a cabo mediante diversos métodos (no sólo mediante la comprobación de hipótesis estadísticas). Simplificando este argumento, se puede afirmar que todas las actividades del estudio sirven para establecer la validez de las hipótesis adoptadas (Rao, 2019).

La investigación mediante encuestas en el ámbito de la educación suele realizarse sobre muestras. Esto significa que, de hecho, se encuesta a toda la población, pero el material empírico se recoge sólo en relación con la muestra de la investigación (por ejemplo, mediante la observación, las encuestas, diversas pruebas, escalas, etc.). A partir de los datos de la muestra, se extraen conclusiones aplicables a toda la población. En un modelo de investigación de este tipo, además de las hipótesis de investigación, aparecen también las hipótesis estadísticas. Se forman en la fase de tratamiento estadístico de los datos recogidos y sirven de base y condición necesaria para llevar a cabo los procedimientos estadísticos. Gracias a las hipótesis estadísticas es posible verificar las hipótesis de investigación (Gamboa, 2018). La prueba de hipótesis muestra si la hipótesis debe ser aceptada o rechazada. Se ha dicho con razón que *“hay que torturar los datos hasta que confiesen”* (Batabyal, 2020, p.34)

Para verificar las hipótesis de la investigación con la mayor fiabilidad posible, a partir del material de investigación recogido, deben aplicarse métodos estadísticos precisos. Los investigadores no deben guiarse por su instinto, juzgando subjetivamente si los resultados confirman o no la hipótesis. Por ello, la verificación de las hipótesis de investigación no suele ser posible sin la verificación de las hipótesis estadísticas (Reilly, 2019; Stunt et al., 2021). Sin embargo, hay que subrayar claramente que esta regla sólo se aplica a las pruebas por muestreo, es decir, cuando los resultados obtenidos en la muestra se generalizan a la población general. Tanto las hipótesis de investigación como las hipótesis estadísticas se refieren siempre a la población general (Nuzzo, 2014).

En la investigación, las hipótesis estadísticas suelen ser específicas, mientras que las hipótesis de investigación parecen más generales. Sin embargo, esta norma se aplica más bien al aspecto formal, mientras que la esencia de la relación sigue siendo diferente. En realidad, al menos en las consideraciones teóricas, cada hipótesis de investigación más general debería descomponerse en hipótesis parciales más específicas (de forma similar a las preguntas de investigación) (Ioannidou & Erduran, 2021). Gracias a ello, las hipótesis siempre aparecen por parejas: una hipótesis de investigación y una hipótesis estadística. Normalmente hay más parejas de este tipo en el estudio: una docena o incluso varias docenas (Gamboa, 2018).

Una hipótesis estadística se opone siempre a una hipótesis de investigación. La relación entre ambos (tanto en términos de generalidad - especificidad como de afirmación - negación) se ilustrará con un ejemplo de la práctica. En un estudio sobre el éxito escolar de los alumnos del primer ciclo de secundaria, se comparó a los niños que viven en una localidad con colegio con los que tienen que desplazarse a la escuela. Una de las preguntas de la investigación era: ¿Difiere el éxito escolar de estos alumnos? La hipótesis general fue: sí, el éxito escolar difiere porque los estudiantes más exitosos son los que no se desplazan a la escuela. Esta hipótesis se basa en el hecho de que los estudiantes que se desplazan a diario utilizan mucho tiempo y energía en sus trayectos de ida y vuelta a la escuela, tienen menos contacto con otros estudiantes de su clase, les resulta más difícil participar en actividades extraescolares, etc. En la verificación empírica de esta hipótesis, la variable definida como "éxito escolar" se examinó mediante todas las calificaciones recibidas por los alumnos durante el curso escolar. De este modo, el conjunto de la hipótesis se dividió en varias hipótesis específicas (por ejemplo, hay diferencias en las calificaciones de lengua española, en las calificaciones de matemáticas, etc.). En el estudio, sólo algunas notas, o incluso sólo la nota media, pueden considerarse indicadores de la variable "éxito escolar", ya que los supuestos de cada estudio resultan ser diferentes. Una de las hipótesis específicas era que la media de notas de los estudiantes que se desplazan a la escuela es inferior a la de los estudiantes que no tienen que desplazarse. La hipótesis estadística era: las puntuaciones medias de los dos grupos no difieren. Las hipótesis estadísticas casi siempre se presentan de esta forma y suponen que no hay diferencias (ni interdependencia). Son hipótesis nulas.

El primer paso en esta etapa es recoger los datos de la muestra. Si el investigador está interesado en saber si las valoraciones medias de los dos grupos difieren en

la muestra, basta con calcular las valoraciones medias y compararlas entre sí. Si hay una diferencia, esto ya responde a la pregunta de investigación. No se necesita ninguna hipótesis estadística, al igual que cualquier prueba adicional resulta innecesaria. Lo mismo ocurre cuando el investigador dispone de datos de toda la población (población general). Una simple comparación de dos puntuaciones medias es entonces suficiente para responder a la pregunta de investigación. La diferencia de estos promedios es una prueba en sí misma. En este caso concreto, hay que llamar la atención sobre el problema que se plantea al utilizar e interpretar la media aritmética para las calificaciones.

Las calificaciones escolares son una variable ordinal (aunque los valores numéricos den la ilusión de que es una variable de intervalo). Esto significa que los intervalos entre las calificaciones no son los mismos (por ejemplo, la diferencia entre una calificación de suspenso y una calificación de admisión no es la misma que la diferencia entre una calificación buena y una calificación muy buena, etc.). Por lo tanto, la suma de las calificaciones no está justificada, es decir, no se puede calcular una media aritmética. Sin embargo, la naturaleza de esta variable permite utilizar, por ejemplo, la mediana, aunque muy a menudo en el caso de las calificaciones escolares esto no lleva a ninguna parte, porque las medianas -incluso para grupos obviamente diferentes- rara vez son diferentes. Por lo tanto, la media aritmética suele ser la única solución. En tales situaciones, hay que tener mucho cuidado al interpretar los resultados: hay que pensar detenidamente en lo que muestran realmente las medias aritméticas (Thron & Miller, 2015, Gamboa, 2018).

Una situación completamente diferente se produce cuando tanto las preguntas como las hipótesis de la investigación se refieren a toda la población general y los datos recogidos proceden únicamente de una muestra. Estas son las únicas situaciones en las que es necesario y razonable aplicar pruebas estadísticas de verificación de hipótesis. En estos casos -además de la hipótesis de investigación- también se formula una hipótesis estadística y se aplican pruebas para su verificación (Thron & Miller, 2015).

En el enfoque de Fisher, sólo se plantea una hipótesis estadística (hipótesis nula), que suele expresarse en términos de ausencia de diferencias; para nuestro ejemplo podemos suponer que el aprendizaje con los dos métodos de enseñanza es igual de eficaz, lo que nos lleva a la siguiente hipótesis nula $\mu_s = \mu_t$; es decir, si la hipótesis es cierta, no debería haber diferencia entre las puntuaciones medias del cuestionario de ambas poblaciones

de estudiantes, sólo la que pudiera deberse al muestreo aleatorio.

Si el investigador encuentra una diferencia positiva entre las puntuaciones medias de los dos grupos, es decir, $\mu_s = \mu_t > 0$, es un resultado que apoya su hipótesis de investigación. En el caso de que la diferencia sea negativa, es decir, $\mu_s = \mu_t < 0$, el resultado contradice la hipótesis nula. La decisión depende del resultado del valor p, que nos informa de la probabilidad que tienen los datos de la muestra según la distribución muestral del estadístico de prueba, que se determina bajo el supuesto de que la hipótesis nula es verdadera. Así, los valores p muy pequeños representan una fuerte evidencia en contra de la hipótesis nula porque significan que los datos obtenidos son muy improbables, por lo que se rechaza esta hipótesis por falta de evidencia experimental. Los límites más comunes para rechazar la hipótesis nula, popularizados por el propio Fisher, son los valores inferiores a 0,05 y a 0,01. Sin embargo, la elección de estos valores depende de las características del problema y de la magnitud del error que el investigador desee asumir.

En el enfoque de Neyman-Pearson, la prueba de hipótesis se plantea como un proceso de decisión entre dos hipótesis. Se considera una hipótesis alternativa (H_1) que es la negación o el complemento de la hipótesis nula (H_0). Las regiones de rechazo y no rechazo se definen en la distribución de muestreo del estadístico de la prueba. El nivel de significación se fija antes de la prueba y permite delimitar las regiones de rechazo y no rechazo de la hipótesis nula. Si el valor del estadístico de la prueba cae en la región de rechazo, se rechaza la hipótesis nula; en caso contrario, no se rechaza la hipótesis nula.

En el caso de las hipótesis de investigación, lo más frecuente es que la palabra "población general" no aparezca en absoluto, porque es evidente. En cambio, en el caso de las hipótesis estadísticas, siempre se indica claramente que se refieren a poblaciones generales. ¿Por qué estas hipótesis nunca se refieren a las muestras? Porque ya se sabe todo sobre las muestras, porque se han recogido los datos necesarios sobre ellas, y no es necesario hacer hipótesis. Una situación en la que las medias aritméticas de las muestras difieren no implica ni demuestra que las medias aritméticas de las poblaciones generales también difieran. Esto también se aplica a todas las diferencias entre muestras (y, en general, se aplica a todas las inferencias de los datos de la muestra). Las pruebas sobre la

población general sólo pueden ser proporcionadas por una prueba de verificación (Harrison et al., 2020).

Existen varias herramientas estadísticas, como la prueba t, la prueba F, la prueba chi-cuadrado, la prueba D-W, los análisis de varianza y covarianza y la prueba z, para comprobar la validez de la hipótesis (Wilcox & Serang, 2017). La prueba chi-cuadrado es la más utilizada en la investigación educativa. Su popularidad se debe a que los datos sobre fenómenos educativos pertenecen en su mayoría a variables descriptivas (nominales y ordinales). Esta prueba se utiliza principalmente cuando se examinan las interdependencias entre los fenómenos. En estos casos, la hipótesis de investigación suele suponer que los fenómenos están relacionados, mientras que la hipótesis estadística supone que los fenómenos en la población general son independientes (hipótesis de independencia) (Gamboa, 2018).

Aunque las pruebas de hipótesis han sido uno de los principales métodos estadísticos para el análisis de datos utilizados en las ciencias experimentales y del comportamiento, múltiples estudios llaman la atención sobre errores comunes de conceptualización y procedimiento de uso de las pruebas de hipótesis, mostrado ser un concepto expuesto a falacias, confusiones, errores, malos usos y posiblemente abusos (Gyllenpalm & Wickman, 2011; Gamboa, 2018; Rao, 2019; Harrison et al., 2020; Stunt, 2021).

Thron & Miller (2015), indican que los autores de algunos artículos informan incorrectamente las pruebas de hipótesis utilizadas, reportan cálculos estadísticos aparentemente inadecuados, utilizan la prueba de 1-cola cuando debe utilizarse una prueba de 2-colas, o simplemente los autores no indican la prueba de cola utilizada. Entre las causas que se ofrecen como explicación de tal dificultad está la diversidad de conceptos abstractos que intervienen para realizar una inferencia, así como el enfoque formal deductivo a través del cual es abordada su enseñanza y el no representar claramente y sin ambigüedades la naturaleza condicional de los conceptos y cantidades probabilísticas que intervienen en la comprobación de hipótesis. En la tabla 1 se muestra un resumen de las concepciones erróneas identificadas por dichos autores acerca de los conceptos que intervienen en el proceso de una prueba de hipótesis.

Tabla 1. Concepciones erróneas sobre las pruebas de hipótesis.

Tipo de concepción errónea	Descripción
Sobre los diferentes enfoques de las pruebas de hipótesis.	– Falta de distinción del paralelismo entre pruebas de hipótesis y procesos de decisión derivado de los enfoques de Fisher y Neyman-Pearson.
Sobre la definición de hipótesis.	– Confusión en la definición de hipótesis nula e hipótesis alternativa. – Confusión de la hipótesis nula con la región de aceptación. – Creencia de que una hipótesis puede referirse tanto a la población como a la muestra.
Sobre la naturaleza condicional del nivel de significancia.	– Invertir el condicional del nivel de significancia. – Interpretar el nivel de significancia como la probabilidad de que una de las hipótesis sea cierta. – Interpretar el nivel de significancia como la probabilidad de cometer un error. – Interpretar el valor de p como la probabilidad de que el evento sucedió por azar.
Sobre la interpretación de valores numéricos del valor de p .	– Interpretar el valor numérico del valor de p como un indicador de la intensidad del efecto del tratamiento o la variable bajo prueba.
Sobre la naturaleza de las pruebas de hipótesis.	– Considerar una prueba de hipótesis como una demostración matemática. – Considerar una prueba de hipótesis como una demostración probabilística de una de las hipótesis. Esta concepción también es llamada ilusión de demostración probabilística por contradicción o ilusión de lograr la improbabilidad.
Sobre la interpretación de la significancia estadística.	– Confundir significancia práctica y significancia estadística.

En particular, en el caso de las pruebas de hipótesis se requiere comprender la integración y la relación que guardan entre sí en el proceso de prueba conceptos como población, muestra, estadístico de prueba, distribución muestral del estadístico de prueba, nivel de significancia, hipótesis nula, hipótesis alternativa, valor de p , regiones de rechazo y regiones de no rechazo, entre otros (Inzunza Cazares & Jiménez Ramírez, 2013; Ioannidou & Erduran, 2021).

Los docentes investigadores de temas educativos universitarios poseen formaciones diversas y experiencias de trabajos en disímiles ámbitos. A partir de esta realidad, es posible plantear la hipótesis de que su visión de las pruebas de hipótesis es distinta y que esto se refleja en el diseño y el análisis de los resultados de sus investigaciones. El objetivo del presente estudio es identificar las principales carencias metodológicas y de concepción teórica en investigadores de las carreras de las ingenierías vinculadas a las ciencias de la computación y la información.

MATERIALES Y MÉTODOS

Con el fin de comprender cómo los profesores abordan la comprobación de hipótesis, su relación con la ciencia y sus componentes históricos, se realizó una encuesta entre un grupo de profesores investigadores de temas educativos de la Universidad Nacional de Ucayali, Pucallpa-Perú, pertenecientes a las carreras de las ciencias de la computación, específicamente; Ingeniería en Sistemas Computacionales; Ingeniería de Software y Tecnología de la Información.

Se aplicó un muestreo intencionado de casos críticos y por cuotas, para el cual, el principal criterio de inclusión se basó en el número de publicaciones realizadas en los últimos 5 años (con aplicación de, al menos, una prueba de hipótesis en alguno de ellos). Fueron escogidos en cada carrera, los 10 investigadores (cuota por punto de saturación teórica) con mayores resultados en este aspecto (casos críticos), para fundamentar las implicaciones del análisis de los resultados, desde la concepción lógica de selección de dicho tipo de muestreo. El criterio de exclusión aplicado consistió en no encuestar a docentes de la asignatura estadística matemática, los cuales son considerados como especialistas en la materia y fueron consultados para validar la calidad del cuestionario aplicado.

Para determinar si la muestra puede considerarse como crítica o decisiva, Rai & Thapa (2015), plantean que, debe responder con solidez a la pregunta “¿Si ese grupo tiene problemas, podemos estar seguros de que todos los grupos tienen problemas?” (p. 8). Para los autores de este artículo, resulta adecuado pensar que, si los profesores con mayor

número de publicaciones en las que se aplican pruebas de hipótesis, presentan deficiencias teóricas o metodológicas, es lógico asumir que el resto tendrá dificultades similares o carencias aún más profundas en el tema.

La encuesta aplicada fue diseñada a partir de la propuesta por Ponteville & Crespo (2016), la cual permite, a través de las respuestas obtenidas, analizar los diferentes tipos de argumentos utilizados en el uso de las pruebas de hipótesis como objeto de validación científica y acceder al análisis de las prácticas investigativas, así como los elementos conceptuales de las pruebas que los investigadores asocian a las mismas. Se estableció y comunicó previamente a los participantes el carácter no evaluativo de la encuesta.

Las preguntas que se realizaron a los encuestados fueron las siguientes:

P1) ¿Es imprescindible el uso de las pruebas de hipótesis en la validación científica?

P2) ¿Para qué tipo de investigaciones utiliza las pruebas de hipótesis?

P3) ¿Es imprescindible plantear las hipótesis nula y alternativa?

P4) ¿Qué nivel de significación utiliza generalmente? ¿En qué se basa para su selección?

P5) ¿Cómo influye el tamaño de la muestra en la probabilidad de cometer los errores tipo I y tipo II?

P6) ¿Cómo interpreta y utiliza el p valor para tomar una decisión respecto a la hipótesis nula?

P7) ¿Utiliza Ud. algún otro análisis complementario en la aplicación de las pruebas de hipótesis? ¿Por qué?

Para estimar la pertinencia de las preguntas seleccionadas, se aplicó el método IADOV como herramienta de consulta al criterio de ocho profesores de la asignatura Estadística de la Universidad Nacional de Ucayali, Pucallpa-Perú, con más de 12 años de experiencia a los cuales se les aplicó el cuestionario antes de ponerlo a disposición de los miembros de la muestra seleccionada. Aunque el método se aplica generalmente para medir satisfacción de usuarios, se puede adaptar de forma sencilla para medir pertinencia como equivalente de la satisfacción de los especialistas al respecto. El cuadro lógico de IADOV elaborado para esta investigación se muestra en el anexo 1 de este artículo. Para el cálculo del índice de satisfacción global se aplicaron la escala que se muestra en la tabla 1 y la fórmula (1), como aparecen en Castro-Fabre et al. (2020).

Tabla 1. Escala para el cálculo del índice de satisfacción global.

Escala	Resultado
1	Máximo de satisfacción
0,5	Más satisfecho que insatisfecho
0	No definido y contradictorio
-0,5	Más insatisfecho que satisfecho
-1	Máxima insatisfacción

Fuente: Castro-Fabre et al. (2020).

Ecuación para el cálculo del ISG (F1).

$$ISG = \frac{A(+1)+B(+0.5)+C(0)+D(-0.5)+E(-1)}{N} \quad (F1)$$

RESULTADOS Y DISCUSIÓN

La aplicación del método IADOV permitió verificar la pertinencia del cuestionario elaborado a partir de la satisfacción de los especialistas con los resultados de su aplicación. En la Figura 1 se muestra las frecuencias de los niveles de satisfacción (conformidad) parcial de los especialistas.

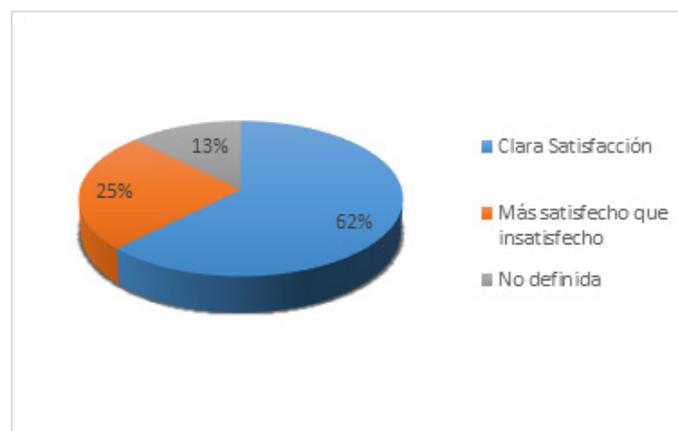


Figura 1. Frecuencias de los niveles de satisfacción parcial de los especialistas.

Solo uno de los especialistas mostró una postura indefinida respecto a la pertinencia del cuestionario aplicado, y se observó que más del 50% de los profesores de estadísticas mostraron una clara satisfacción respecto a las preguntas incluidas. Mediante (1), se obtuvo un ISG de 0,75, lo que indica que los especialistas consideran pertinente el cuestionario diseñado como herramienta para recopilar la información necesaria para dar cumplimiento al objetivo planteado.

De acuerdo a los resultados obtenidos en cada pregunta, se muestran a continuación los resultados generales

sobre el dominio predominante en los individuos en estudio sobre cada ítem, a partir de los aspectos más relevantes de las respuestas obtenidas. Igualmente se presentan las respuestas más representativas o que sobresalen por aspectos positivos o negativos, según se considere o no relevante. Cada resultado se muestra acompañado por algunas reflexiones que refuerzan el criterio de los autores del presente artículo.

Acerca de cuan imprescindible puede resultar el uso de las pruebas de hipótesis en la validación científica

Las respuestas analizadas se pueden catalogar como similares y giran alrededor de la idea de la importancia de las pruebas de hipótesis para la verificación de proposiciones respecto de la población, principalmente para decidir si estas proposiciones deben ser aceptadas o rechazadas. Sin embargo, en dos de las respuestas no se hace alusión a la población, sino que se presenta la prueba de hipótesis como la verificación de una afirmación respecto a la/s muestra/s.

“...para demostrar relaciones entre las muestras”.

“...para verificar si se cumple lo planteado para una muestra”.

Esto no es totalmente desacertado cuando se trata de investigaciones cualitativas en las que no se deben hacer inferencias poblacionales, salvo excepciones en las que se realice un muestreo no probabilístico intencionado que garantice una representatividad significativa o una fiabilidad suficientemente alta.

Por otra parte, las pruebas de hipótesis paramétricas, que son fundamentales para la investigación cuantitativa en las ciencias pedagógicas, estas suelen presentarse con un sólido fundamento teórico sobre la asociación positiva o negativa entre los valores de dos variables, a partir de parámetros poblacionales, e informar de una medida de asociación entre esas variables. Al redactar cualquier conclusión a que se arribe, siempre debe tenerse en cuenta el control de otros factores potencialmente relevantes y presentar la decisión tomada sobre alguna prueba de la hipótesis de nulidad, en la cual se enuncia que la asociación es nula. Constituye un error, asumir en este tipo de pruebas un alcance muestral del análisis de los resultados.

Otro uso importante de las pruebas de hipótesis, al que la mayoría de los entrevistados no hicieron alusión (salvo 4 investigadores), y que ha recibido mayor atención en la teoría estadística, es evaluar las proposiciones basadas en la teoría sobre el valor de un parámetro (Wallpole et al., 2012). Pero realmente este caso no es común que se aplique en las ciencias sociales, debido a que, si una teoría

predice un valor exacto para un parámetro, es posible probar la hipótesis nula de que tiene ese valor. Aceptar la hipótesis nula es entonces un éxito para la teoría en el sentido de que no es refutada, aunque no proporciona pruebas positivas a favor de ella. En las ciencias educativas, es raro que una teoría prediga un valor específico no nulo de algún parámetro, en la mayoría de los casos, predican que un parámetro no será cero.

Acerca del tipo de investigaciones en las que utilizan las pruebas de hipótesis

Los criterios sobre este ítem se centran en la alusión a la inferencia estadística en investigaciones cuantitativas y cualitativas. Se hace una referencia acertada respecto a los estudios de tipo deductivos, relacionales. En la mayoría de los casos los encuestados refieren la utilización de las pruebas de hipótesis para validar el efecto positivo de la implementación de las propuestas de su investigación para la solución del problema planteado. Algunos hacen referencia al uso de las pruebas de hipótesis para verificar el cumplimiento de supuestos de normalidad y/o homogeneidad de varianzas antes de la aplicación de pruebas paramétricas, y cinco de los encuestados se refieren a su aplicación para la evaluación de modelos predictivos de diferente índole, principalmente para los modelos de regresión lineal múltiple.

Esto indica un uso poco habitual de dichos modelos en las investigaciones realizadas por los profesores seleccionados. Si bien es cierto que, para su correcta aplicación e interpretación de las múltiples salidas que ofrecen los softwares estadísticos en la actualidad, se requiere de conocimientos avanzados de estadística matemática, su utilidad puede permitir la cuantificación y sentido de la relación entre múltiples variables, así como la predicción de valores de la variable independiente a partir del modelo obtenido.

Se puede observar, en su defecto, la aplicación de pruebas no paramétricas alternativas que implican menos restricciones en cuanto al muestreo y se aplican en los estudios cualitativos, que resultan predominantes en los estudios realizados por estos investigadores. En una misma línea de pensamiento, resulta más común la aplicación de métodos estadísticos, pre y cuasi experimentales, que favorecen la aplicación de dichas pruebas.

Acerca de si resulta imprescindible plantear las hipótesis nula y alternativa

En este aspecto todos los encuestados coinciden en que es necesario el uso de ambas hipótesis, aunque solo el 40% de ellos manifestó que suelen plantear y mostrar las hipótesis específicas de las pruebas que utilizan en

sus investigaciones. Las justificaciones expuestas para mostrar las hipótesis que se plantean son varias, pero destaca:

“...considero innecesario mostrar el planteamiento de las hipótesis pues están recogidas en la bibliografía referenciada”

Estas omisiones metodológicas atentan contra la comprensión del lector y resulta errado pretender que no es necesario adecuar los enunciados de las hipótesis planteadas mediante su contextualización a las proposiciones específicas de la investigación. Por otra parte, el uso extendido del enfoque ecléctico de los métodos tradicionales o clásicos de prueba de hipótesis de Fisher y Neyman-Pearson, ha mantenido la inclusión de ambas hipótesis (como proponen Neyman y Pearson), cuando no necesariamente debe utilizarse ambas para todas las pruebas aplicadas; pues como plantean Thron & Miller (2015), en el enfoque de Fisher solo se requiere del uso de la hipótesis nula.

Acerca de la selección del nivel de significación

Respecto a este ítem, el 100% de los encuestados refirió que utiliza un nivel de significación del 5% y las justificaciones se basan fundamentalmente en que es el nivel que más se utiliza según la bibliografía consultada o bien porque es un valor estandarizado para las pruebas de hipótesis en el ámbito de la investigación educativa.

“... porque es un valor universalmente aceptado”

“...es uno de los valores estándares de significación que más se utiliza en las ciencias sociales”

“...es el valor predeterminado en la mayoría de los softwares que utilizo”

Al respecto deben realizarse varios señalamientos dirigidos hacia el análisis de los estándares de significación y su flexibilidad. Si bien es cierto que algunos autores recomiendan el cumplimiento estricto de los niveles convencionales, ya que así se reduce la influencia de las esperanzas y expectativas del investigador en las conclusiones; también debe tenerse en cuenta que en ocasiones un investigador puede proponer un nivel menos estricto que el clásico 0,01 (que podría resultar inalcanzable en algunas ciencias), pero más exigente que el universalmente utilizado 0,05, debido al rigor exigido para su estudio. Por tanto, los autores de este trabajo recomiendan que las normas sean tratadas de forma flexible; por ejemplo, puede resultar razonable considerar que un valor p de 0,03 proporciona la evidencia requerida contra la hipótesis nula.

Téngase en cuenta que, independientemente de los méritos de los argumentos que defienden la estandarización de los valores de significación, la convención del 5% no se deriva ciertamente de la teoría estadística. Simplemente resultó conveniente, ya que permitió, en los años iniciales de esta ciencia y previos a la accesibilidad a los softwares para el procesamiento estadístico, reducir ciertas tablas necesarias para desarrollar una cantidad tan diversa de pruebas. Por lo demás, no existe ninguna fundamentación teórica sólida para su utilización. Incluso para muestras grandes, un valor de p de 0,05 puede sugerir evidencias a favor de la hipótesis nula. Por tal motivo algunos autores recomiendan utilizar valores críticos de estadígrafo cada vez mayores en relación con el tamaño de la muestra.

Acerca de la influencia del tamaño de la muestra en la probabilidad de cometer los errores tipo I y tipo II

En este tema, igualmente se mostró un consenso respecto a la posición de asumir que las muestras grandes reducen la probabilidad de cometer ambos errores. Esta posición está fundamentada en los argumentos mediante el criterio de relación positiva entre tamaño de muestra y significatividad de la prueba, registrada en la mayoría de las bibliografías relacionadas con el muestreo y la estadística inferencial.

“... las muestras grandes contribuyen a la significatividad de la prueba”

“...entre mayor y más representativa resulte la muestra, menor será la probabilidad de cometer los errores tipo I y II”

Este es un razonamiento que pasa por alto otros elementos que resultan claves respecto a los errores mencionados. De hecho, es el error tipo I el que influye en el tamaño de la muestra, por lo que la relación se establece en sentido contrario a lo planteado. Resulta acertado pensar que el incremento del tamaño de muestra reduce la probabilidad de aceptar, por error, la hipótesis nula, pero no necesariamente la del error tipo I. O sea, si se utiliza una norma constante de significación estadística, la probabilidad de errores de tipo II disminuye a medida que aumenta el tamaño de la muestra, pero la probabilidad de errores de tipo I sigue siendo la misma.

Sin embargo, ya que ambos tipos de error son importantes, es deseable hacer que las posibilidades de ambos disminuyan. Esto significaría utilizar una norma generosa para la significación estadística en muestras pequeñas y hacerla cada vez más estricta a medida que aumenta el número de casos. Este proceso debe desarrollarse con la mayor ética y rigor científico, para evitar introducir sesgos

derivados del deseo del investigador por corroborar su planteamiento inicial.

Por otra parte, debe tenerse en cuenta igualmente la significación sustantiva para la interpretación de los resultados, ya que, una muestra grande puede resultar estadísticamente significativa, pero demostrar un efecto poco significativo, incluso despreciable, desde el punto de vista práctico o científico. Por tanto, excederse en el dimensionamiento de la muestra en base a la reducción del error puede, en algunos casos, atentar contra los objetivos de la investigación.

Acerca de la interpretación y utilización del p valor para tomar una decisión respecto a la hipótesis nula

En cuanto al uso del valor p, el 100% lo utiliza como principal indicador para rechazar o no la hipótesis nula a partir de su comparación con el nivel de significación. La respuesta más recurrente expresa la regla de decisión:

“...si el p valor es menor que el α , se rechaza la hipótesis nula”

Aunque esta práctica es bastante común en los informes de investigación, resulta conveniente aclarar que, en tanto la significación constituye un valor predeterminado de probabilidad de cometer el error tipo I, el p valor representa la probabilidad de que una estadística de prueba sea mayor o igual que el valor observado si la hipótesis nula fuera cierta, y es un resultado obtenido a posteriori, por tanto, no tiene sentido comparar ambos valores y “*se pierde el sentido del valor de P como evidencia a posteriori contra H_0* ” (Monterrey & Gómez-Restrepo, 2007, p. 199)

En cuanto a su interpretación se comprobó una mayor diversidad de criterios, entre los que sobresalen los errores citados por Thron & Miller (2015), en el que los investigadores asumen que el p valor es una probabilidad de que el resultado se deba al azar o una medida de la eficiencia de la propuesta.

Acerca del uso de análisis complementarios en la aplicación de las pruebas de hipótesis

Respecto a este último ítem se evidenció una pobre utilización de otros criterios enriquecedores para el análisis de los resultados de las pruebas de hipótesis y predomina el uso indiscriminado del valor p para la toma de decisiones. Solo en los casos de los investigadores que aplican la estimación de modelos, se hizo referencia a: “la revisión del valor de los estadísticos t como medida de la significatividad estadística de los parámetros”.

Sin embargo, en ningún caso se hizo referencia la aplicación de los intervalos de confianza, los cuales resultan

más informativos que una indicación de significación, la cual reduce la información que se proporciona a la interpretación estadística. De hecho, las pruebas de hipótesis basadas en niveles estándar de significación no añaden información relevante más allá de la contenida en los intervalos de confianza, por lo que, cuando se trata de parámetros de interés teórico, sería posible mostrar simplemente un intervalo de confianza. Esto no lleva a conclusiones diferentes, pero sería más comprensible y dirigiría la atención al rango de incertidumbre en lugar de fijar la atención en la estimación puntual, aunque tampoco deja sin lugar el uso de los valores p o de los errores estándar calculados a partir de estos.

CONCLUSIONES

Las pruebas estadísticas de hipótesis son una herramienta importante y eficaz en la investigación empírica. Las distintas pruebas cubren todas las situaciones de investigación posibles. En cualquier caso, estas pruebas resultan indispensables en la investigación basada en muestras a la hora de generalizar los resultados de éstas a poblaciones enteras. Sin embargo, a la hora de utilizar las pruebas, es necesario comprender con precisión su esencia y ser consciente de sus posibilidades y limitaciones. La interpretación de los resultados es un procedimiento complejo porque la frontera entre las conclusiones correctas y las incorrectas sigue siendo muy borrosa para el investigador poco preparado.

En los profesores encuestados en esta investigación, se pudieron identificar varias limitaciones relacionadas con la presentación e interpretación de los resultados de las pruebas de hipótesis que aplican. Las principales deficiencias se detectaron en la presentación de las hipótesis estadísticas, la selección del nivel de significación y en el uso e interpretación del p valor como elementos más utilizados como criterio de decisión de la prueba.

Los autores de la presente investigación consideran que los errores identificados coinciden con los citados en la bibliografía, por lo que no se puede afirmar que sean exclusivos de los profesores muestreados, sino que son consecuencia del escaso tratamiento teórico que se da a las limitaciones y aspectos complejos y flexibilidades de las pruebas de hipótesis, en la bibliografía especializada.

REFERENCIAS BIBLIOGRÁFICAS

Batabyal, B. (2020). *A Basic Concept of Research Scholar Students: A Protocol of Research*. Orangebooks Publication.

- Castro-Fabre, A., Sánchez-Ortega, N., & Reyes Farrat, Y. (2020). El proceso de validación mediante la Técnica de ladov en cursos por encuentros. *Revista Ingeniería Agrícola*, 10(1), 66-70.
- Eastwell, P. (2014). Understanding Hypotheses, Predictions, Laws, and Theories. *Science Education Review*, 13(1), 16-21.
- Gamboa Graus, M. E. (2018). Estadística aplicada a la investigación educativa. *Dilemas Contemporáneos: Educación, Política y Valores*, 5(2).
- Gyllenpalm, J., & Wickman, P. O. (2011). The uses of the term hypothesis and the inquiry emphasis conflation in Science Teacher Education. *International Journal of Science Education*, 33(14), 1993-2015.
- Harrison, A. J., McErlain-Naylor, S. A., Bradshaw, E. J., Dai, B., Nunome, H., Hughes, G. T., ... & Fong, D. T. (2020). Recommendations for statistical analysis involving null hypothesis significance testing. *Sports biomechanics*, 19(5), 561-568.
- Inzunza Cazares, S., & Jiménez Ramírez, J. V. (2013). Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipótesis. *Revista latinoamericana de investigación en matemática educativa*, 16(2), 179-211.
- Ioannidou, O., & Erduran, S. (2021). Beyond hypothesis testing. *Science & Education*, 30(2), 345-364.
- Monterrey, P., & Gómez-Restrepo, C. (2007). Aplicación de las pruebas de hipótesis en la investigación en salud: ¿estamos en lo correcto? *Universitas Médica*, 48(3), 193-206.
- Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, 506(7487).
- Ponteville, C., & Crespo, C. (2016). Las pruebas de hipótesis en el discurso matemático escolar. (Ponencia). *XI Congreso Argentino de Educación Matemática*. Buenos Aires, Argentina.
- Rai, N., & Thapa, B. (2015). A study on purposive sampling method in research. *Kathmandu: Kathmandu School of Law*. https://www.academia.edu/download/48403395/A_Study_on_Purposive_Sampling_Method_in_Research.pdf
- Rao, U. (2019). Hypothesis-driven Research. *Journal of Oral and Maxillofacial Pathology*, 23(2), 168-168.
- Reilly, F. E. (2019). III The stages of the Method (i): Experience and Hypothesis. En, *Charles Peirce's Theory of Scientific Method*. (pp. 23-56). Fordham University Press.
- Stunt, J., Van Grootel, L., Bouter, L., Trafimow, D., Hoekstra, T., & de Boer, M. (2021). Why we habitually engage in null-hypothesis significance testing: A qualitative study. *PLOS ONE*, 16(10), 1-23.
- Thron, C., & Miller, V. (2015). Persistent confusions about hypothesis testing in the social sciences. *Social Sciences*, 4(2), 361-372.
- Wilcox, R. R., & Serang, S. (2017). Hypothesis testing, p values, confidence intervals, measures of effect size, and Bayesian methods in light of modern robust techniques. *Educational and psychological measurement*, 77(4), 673-689.

ANEXOS

Anexo 1. Cuadro lógico de IADOV.

Pregunta 5: ¿Cómo califica el cuestionario elaborado para esta investigación?	Pregunta 1: ¿Considera que debió utilizarse otro cuestionario que permitiera evaluar otros conocimientos relacionados con las pruebas hipótesis?								
	NO	NO SÉ			SI				
	Pregunta 3: ¿Cree que debieron utilizarse preguntas cerradas para la aplicación de alguna escala?								
	SI	NO SÉ	NO	SI	NO SÉ	NO	SI	NO SÉ	NO
Muy pertinente	1	2	6	2	2	6	6	6	6
Pertinente	2	2	3	2	3	3	6	3	6
No merece mi criterio	3	3	3	3	3	3	3	3	3
Poco pertinente	6	3	6	3	4	4	3	4	4
Nada pertinente	6	6	6	6	4	4	6	4	5
No estoy en condiciones de opinar	2	3	6	3	3	3	6	3	4