Santiago Inzunsa Cazares, José Vidal Jiménez Ramírez

CARACTERIZACIÓN DEL RAZONAMIENTO ESTADÍSTICO DE ESTUDIANTES UNIVERSITARIOS ACERCA DE LAS PRUEBAS DE HIPÓTESIS

The characteristics of college students' statistical reasoning on hypothesis testing

RESUMEN

En este artículo se presentan resultados de una investigación sobre el aprendizaje de la inferencia estadística en estudiantes de la carrera de matemáticas, en particular sobre su nivel de razonamiento estadístico acerca de los conceptos y el proceso que involucran las pruebas de hipótesis. Los resultados se analizaron de acuerdo con el modelo taxonómico SOLO (Structure of Observed Learning Outcomes) y muestran que los estudiantes se ubican principalmente en los niveles preestructuraly uniestructural, es decir, que poseen información aislada de los diversos conceptos que intervienen en una prueba de hipótesis y/o toman en cuenta algún aspecto relevante del proceso pero sin lograr comprender lo que hacen; además, se observaron concepciones erróneas sobre la naturaleza de las pruebas de hipótesis y los principales conceptos involucrados como el nivel de significancia, valor de p y planteamiento de hipótesis.

PALABRAS CLAVE:

- *Razonamiento estadístico*
- *Inferencia estadística*
- *Pruebas de hipótesis*
- *Concepciones erróneas*
- *Estudiantes universitarios*

ABSTRACT

This article shows the results from a research on college students' statistical inference learning process from the major in mathematics. It was specially focused on the students' statistical reasoning on the concepts and processes that involved hypothesis testing. The results were analyzed based on the Structure of Observed Learning Outcomes (SOLO). They showed that students are primarly in the pre-

KEY WORDS:

- *Statistical reasoning*
- *Statistical inference*
- *Hypothesis testing*
- *Misconceptions*
- *College students*

Revista Latinoamericana de Investigación en Matemática Educativa (2013) *16* (2): 179-211.
Recepción: Abril 5, 2011 / Aceptación: Febrero 13, 2013. DOI: 10.12802/relime.13.1622

structural and unistructural level, which means that they have isolated information about the different concepts involved in hypothesis testing and/or that they take into consideration some relevant aspect of the process without having full awareness of what they are doing. Moreover, it was noticed that there are misguided conceptions on the nature of hypothesis testing and the main concepts involved such as the level of significance, value p and the formulation of the hypothesis.

RESUMO

Neste artigo, apresentam-se resultados de uma pesquisa sobre a aprendizagem da inferência estatística em estudantes do curso de matemática, em particular sobre seu nível de raciocínio estatístico sobre os conceitos e o processo que envolve as provas de hipótese. Os resultados são analisados de acordo com o modelo taxonômico SOLO (Structure of Observed Learning Outcomes) e mostram que os estudantes estão localizados principalmente nos níveis pre-estrutural e uni-estrutural , ou seja, que possuem informação isolada dos diversos conceitos que intervêm em uma prova de hipótese e/ou têm em conta algum aspecto relevante do processo mas sem conseguir compreender o que fazem; além disso, foram observadas concepções errôneas sobre a natureza das provas de hipótese e os principais conceitos envolvidos como o nível de significância, valor p e abordagem hipotética.

PALAVRAS CHAVE:

- *Raciocínio estatístico*
- *Inferência estatística*
- *Provas de hipótese*
- *Concepções errôneas*
- *Estudantes universitários*

RÉSUMÉ

Dans cet article on présente les résultats d'une recherche sur l'apprentissage de l'inférence statistique des étudiants de la licence en mathématiques, plus en particulier, leur niveau de raisonnement statistique par rapport aux concepts et au processus des tests d'hypothèses. Les résultats ont été analysés avec le modèle taxonomique SOLO (Structure of Observed Learning Outcomes) et montrent que les étudiants se trouvent principalement aux niveaux pre-structural et uni-structural, c'est-à-dire, qu'ils possèdent information isolée des différents concepts des tests d'hypothèses ou son processus, mais ils n'arrivent pas à tout comprendre. On observe aussi que les étudiants ont des conceptions erronées sur la nature des tests d'hypothèses et sur les principaux concepts associés comme le rapport de vraisemblance, p-valeur et formulation d'hypothèses.

MOTS CLÉS:

- *Raisonnement statistique*
- *Inférence statistique*
- *Tests d'hypothèses*
- *Conceptions erronées*
- *Étudiants universitaires*

1. Planteamiento del problema

1.1. *Introducción*

En los últimos años la estadística se ha convertido en una disciplina científica que ocupa un lugar muy importante en el currículo de muchas carreras universitarias debido al papel que juega como herramienta metodológica para el estudio de diversos fenómenos y por el desarrollo del razonamiento y pensamiento estadístico para interpretar adecuadamente información de diversos acontecimientos cotidianos, de las profesiones y de las ciencias. En particular, la inferencia estadística –campo al que corresponde el tema de esta investigación–, constituye una de las áreas de mayor utilidad de la estadística, ya que mediante la utilización de sus métodos, se pueden obtener conclusiones acerca de una población con base en la información que proporcionan los datos de una muestra.

Los principales métodos de inferencia estadística son la *estimación de parámetros* y las *pruebas de hipótesis* (también llamadas *contraste de hipótesis* o *pruebas de significancia*). La estimación de parámetros puede ser *puntual* o mediante *intervalos de confianza*, en ambos casos se tiene como propósito estimar el valor de un parámetro desconocido de una población (por ejemplo la media o el coeficiente de correlación) a partir de los datos de una muestra. En una estimación puntual se proporciona un valor único como estimación del parámetro, mientras que en una estimación por intervalo de confianza se establece un intervalo aleatorio y un nivel de confianza $1-\alpha$ que mide la probabilidad de contener al parámetro. Por su parte, una prueba de hipótesis, es un método que permite verificar una aseveración acerca del valor de un parámetro poblacional; dado que los datos son proporcionados por una muestra, los resultados pueden estar sujetos a variaciones aleatorias, por lo que una prueba de hipótesis permite decidir si pequeñas desviaciones observadas respecto al resultado que idealmente debería haber ocurrido según nuestra hipótesis, son atribuibles al azar o efectivamente los resultados no se corresponden con la hipótesis que se ha planteado sobre el valor del parámetro. Por otra parte, la sociedad actual requiere ciudadanos estadísticamente cultos que puedan comprender argumentos estadísticos de moderada complejidad y comprender conceptos y vocabulario estadístico sobre las pruebas de hipótesis que les permitan interpretar adecuadamente resultados de diversos estudios (McLean, 2002). En este mismo sentido, Garfield y Ben-Zvi (2008) señalan: "hacer inferencias a partir de los datos es ahora parte de la vida cotidiana y la revisión crítica de los resultados de inferencia estadística a partir de estudios de investigación es una capacidad importante para todos los adultos" (p. 262).

Sin embargo, la experiencia de muchos profesores y resultados de investigación (Liu & Thompson, 2005; Vallecillos, 1997; Williams, 1998; Good & Hardin, 2009; Cumming, 2010; Yañez & Behar, 2010; Grings & Viali, 2011) muestran que la comprensión de la lógica que subyace a los métodos de inferencia estadística y la interpretación de sus resultados es compleja para muchos estudiantes y profesores, incluso para profesionales que aplican la estadística en su desempeño profesional. Entre las causas que se ofrecen como explicación de tal dificultad está la diversidad de conceptos abstractos que intervienen para realizar una inferencia (Chance, delMas & Garfield, 2004; Lipson, 2002), así como el enfoque formal deductivo a través del cual es abordada su enseñanza (Moore, 1992; Lipson, 2000). 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.

1.2. *Origen y controversias sobre las pruebas de hipótesis*

Los elementos lógicos que dieron origen a las pruebas de hipótesis fueron presentados en artículos científicos a principios del siglo XVIII (Stigler, 1986). Sin embargo, formalmente las pruebas de hipótesis surgen en las décadas de 1920 y 1930 como resultado del trabajo de dos grupos o escuelas de pensamiento: por un lado, Ronald Fisher (1890-1962), y por el otro, Jerzy Neyman (1894-1981) y Egon Pearson (1895-1980). Los enfoques conceptuales sobre el significado y desarrollo de las pruebas de hipótesis en los cuales se basaron estos dos grupos de investigación parten de posiciones filosóficas distintas, por lo que la historia de las pruebas de hipótesis no ha estado exenta de controversias y desacuerdos desde su origen, factor que ha conducido a diversas dificultades para su aplicación e interpretación (Levin, 1998; Kirk, 2001).

De acuerdo con Kline (2004), el enfoque de Fisher se caracteriza por definir únicamente una hipótesis (hipótesis nula) y a partir de la ella y con base en la distribución muestral del estadístico de prueba, se estima la probabilidad de una muestra de datos para decidir sobre el rechazo o no rechazo de la hipótesis. Los datos solo permiten rechazar la hipótesis pero no pueden confirmarla. El enfoque de Neyman-Pearson se caracteriza por la adición de una hipótesis alternativa en contraposición con la hipótesis nula, lo que conduce a la definición de regiones (de rechazo y no rechazo) y errores asociados a la decisión sobre H_0 denominados errores Tipo I y Tipo II. En este sentido, en el enfoque de Neyman-Pearson una

prueba de hipótesis es una regla de comportamiento inductivo que permite elegir entre una hipótesis nula y una hipótesis alternativa. La evidencia de los datos obtenidos puede conducir a no rechazar la hipótesis nula, lo cual no implica que ésta sea cierta (Vallecillos, 1996). Al respecto, Batanero (2000) considera que la principal diferencia entre estas dos teorías no radica en los cálculos, sino en las concepciones y el razonamiento subyacente.

Sin embargo, la integración de los dos modelos por parte de estadísticos, investigadores y autores de libros de texto se hizo práctica común desde 1935 (Huberty, 1993). Es decir, al aplicar las pruebas de hipótesis comúnmente se utilizan elementos de los dos enfoques de forma ecléctica; Gigerenzer (1993) se refiere a este modelo integrado como *lógica híbrida de la inferencia estadística*. Esta lógica híbrida en la aplicación de las pruebas de hipótesis ha sido una de las fuentes de controversias y críticas desde su surgimiento hasta nuestros días (Carver, 1978; McLean & Ernest, 1998; Morrison & Henkel, 1970; Levin, 1998; Triola, 2009). Entre las críticas que se hacen a las pruebas de hipótesis se encuentra el hecho que no reportan mayor información que la significancia estadística, la cual puede ser insuficiente para tomar una decisión sobre los efectos de una variable o tratamiento en una investigación.

En tanto las pruebas de hipótesis han constituido uno de los principales métodos estadísticos para el análisis de datos utilizados en las ciencias experimentales y de la conducta, en los años recientes se ha generado un debate entre investigadores y organizaciones como la Asociación Americana de Psicología (APA por sus siglas en inglés) para discutir si las críticas sobre las pruebas de hipótesis tienen el suficiente mérito, y de ser así, buscar alternativas adicionales que las complementen al ser utilizadas en la investigación. En este contexto, la APA realizó en 1999 un reporte de investigación titulado *Task Force on Statistical Inference* (TFSI por sus siglas en inglés) en el cual, entre otros aspectos, se concluye que debe haber cambios en las formas de presentar los resultados para publicarlos en las revistas de investigación. En general, tanto el reporte TSFI como investigadores (Mittag & Thompson, 2000; Xitao, 2001; Robinson & Levin, 1997) recomiendan complementar las pruebas de hipótesis con el reporte de tamaño de los efectos y cálculo de intervalos de confianza. Estas recomendaciones se encuentran incorporadas en la quinta edición del Manual de Publicaciones de la Asociación Americana de Psicología[1] y de igual forma están siendo requeridas en revistas de investigación de otras disciplinas.

[1] APA-American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th. ed.). Washington, DC: Autor.

1.3. *Fundamentos y conceptos de las pruebas de hipótesis*

Las pruebas de hipótesis involucran diversos conceptos y términos que son necesarios comprender para su correcta aplicación e interpretación. Nos parece muy ilustrativo el mapa conceptual elaborado por Lipson (2000), donde se muestran los conceptos que intervienen en una prueba de hipótesis y la forma como estos se relacionan, desde una perspectiva híbrida de los enfoques de Fisher y Neyman-Pearson (ver Figura 1).

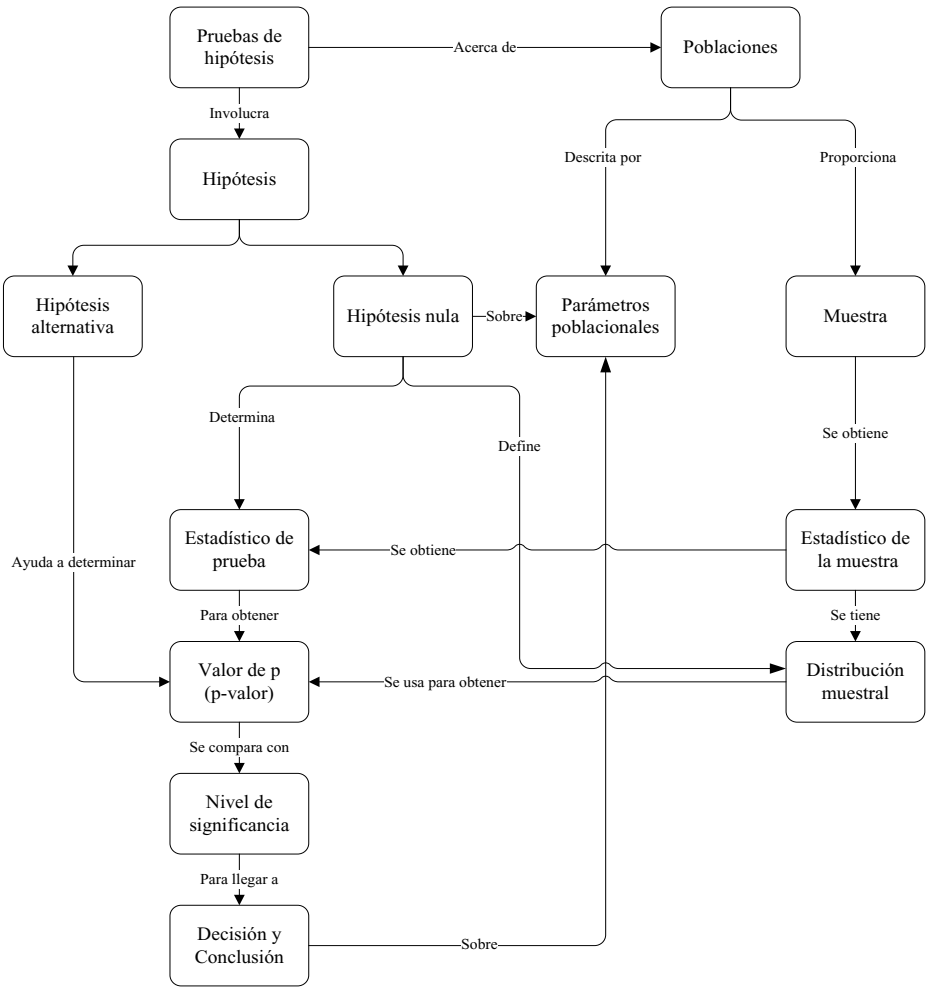


Figura 1. Mapa conceptual de una prueba de hipótesis (Lipson, 2000, p. 41)

Para comprender la lógica que subyace a las pruebas de hipótesis estadísticas es importante partir del concepto de hipótesis de investigación. Una hipótesis de investigación es un enunciado que establece un investigador en el cual se ofrece una posible respuesta a una pregunta de investigación. Por ejemplo, un investigador educativo puede plantear la siguiente hipótesis: el uso de software educativo con representaciones dinámicas de los datos (simbólicas, gráficas, numéricas) acompañado del uso de datos reales, facilita la exploración y contribuye a una mejor comprensión del análisis descriptivo de datos. Dicho enunciado como tal tiene dos valores de "verdad": verdadero o falso. En este sentido, la investigación se realiza para determinar el valor de verdad que corresponde a la hipótesis de investigación; es decir, se realiza un estudio para obtener los datos que contengan la información necesaria para dar respuesta a la pregunta de investigación y decidir si la hipótesis de investigación se rechaza o no (Monterrey & Gómez-Restrepo, 2007).

Por su parte, las hipótesis estadísticas son afirmaciones acerca de los parámetros, por lo que para probar la validez de una hipótesis de investigación es necesario plantearla en término de hipótesis estadísticas. De esta manera y teniendo en cuenta el ejemplo del software educativo mencionado anteriormente, consideremos que se selecciona una muestra de estudiantes a partir de la cual se forman dos grupos en forma aleatoria con condiciones iguales en todas las variables sobre el conocimiento del tema para evitar que los resultados puedan deberse a otros factores que no son de interés; uno toma la clase de análisis descriptivo de datos de forma tradicional (grupo control) y otro que toma la clase usando el software educativo (grupo experimental). Al final del curso se aplica un mismo cuestionario para determinar los puntajes promedio de cada grupo. La decisión involucra el planteamiento de hipótesis sobre los puntajes promedio de ambas poblaciones teóricas de estudiantes (los que recibieron la enseñanza tradicional y los que recibieron la enseñanza con software educativo).

En el enfoque de Fisher se plantea una sola hipótesis estadística (hipótesis nula), que suele ser expresada en términos de no diferencia; para nuestro ejemplo podemos partir de que el aprendizaje con los dos métodos de enseñanza es igual de efectivo, lo que nos conduce a la siguiente hipótesis nula $\mu_s = \mu_t$; es decir, de ser cierta la hipótesis, no debe haber diferencia entre los puntajes promedio en el cuestionario de ambas poblaciones de estudiantes, sólo la que pudiera deberse a la aleatoriedad del muestreo. Si el investigador encuentra una diferencia positiva entre los puntajes promedio de los dos grupos, esto es, $\mu_s - \mu_t > 0$, es un resultado que apoya su hipótesis de investigación. En el caso 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 de p, que nos informa la probabilidad que tienen los datos de la muestra de acuerdo con la distribución muestral del estadístico de prueba, misma que es determinada bajo el supuesto de que la hipótesis nula es cierta. De esta manera, valores muy pequeños del valor de p representan una evidencia fuerte contra la hipótesis nula pues significan que los datos obtenidos son muy improbables, por lo que se rechaza ésta hipótesis ante la falta de evidencia experimental. Los límites más comunes para rechazar la hipótesis nula, popularizados por el mismo Fisher, son valores menores de 0.05 y menores que 0.01. Sin embargo, la elección de estos valores depende de las características del problema y de la magnitud del error que desea asumir el investigador.

Por su parte, en el enfoque de Neyman-Pearson, las pruebas de hipótesis se planten como un proceso de decisión entre dos hipótesis. En él se consideran una hipótesis alternativa (H_1) que es la negación o complemento de la hipótesis nula (H_o). Se definen regiones de rechazo y no rechazo sobre la distribución muestral del estadístico de prueba y los siguientes tipos de errores que se pueden cometer:

1. Error tipo 1: Rechazar H_o dado que H_o es cierta. Su probabilidad se denota por α, más formalmente P (*rechazar* $H_o | H_o$ es cierta)= α, también se conoce como *nivel de significancia.*
2. Error tipo 2: No rechazar H_o dado que H_o es falsa. Su probabilidad se denota por β, esto es P (*no rechazar* $H_o | H_o$ es falsa)=β.

El nivel de significancia se fija previo a la prueba y permite a su vez delimitar las regiones de rechazo y no rechazo de la hipótesis nula. Si el valor del estadístico de prueba cae en la región de rechazo, la hipótesis nula es rechazada; en caso contrario, la hipótesis nula no es rechazada. En el contexto del problema del software educativo para el análisis de datos, las hipótesis podrían quedar de la siguiente forma:

$$H_o : \mu_s \leq \mu_t \quad \text{vs} \quad H_1 : \mu_s > \mu_t$$

Sin embargo, como señalamos anteriormente, es común que investigadores y autores de libros de texto utilicen una lógica híbrida de los enfoques de Fisher y Neyman-Pearson en el proceso de una prueba de hipótesis, y usualmente se compara el valor de p con el nivel de significancia α para decidir sobre el rechazo de la hipótesis nula. Para asegurar que se cumpla con el nivel máximo de error tipo I definido por el nivel de significancia, el criterio consiste en que si $p < \alpha$ se rechaza la hipótesis nula. De esta manera se mezclan los enfoques de Fisher y Neyman-Pearson al comparar el valor de p que es un cálculo *a posteriori* y distintivo del enfoque de Fisher, con el valor de α que es un valor definido *a priori* distintivo del enfoque de Neyman-Pearson.

Para ejemplificar lo anterior, en la Figura 2 se muestra una distribución muestral del estadístico de prueba[2] para ciertos datos, la cual ilustra la región de rechazo y no rechazo de una prueba de hipótesis de cola derecha.

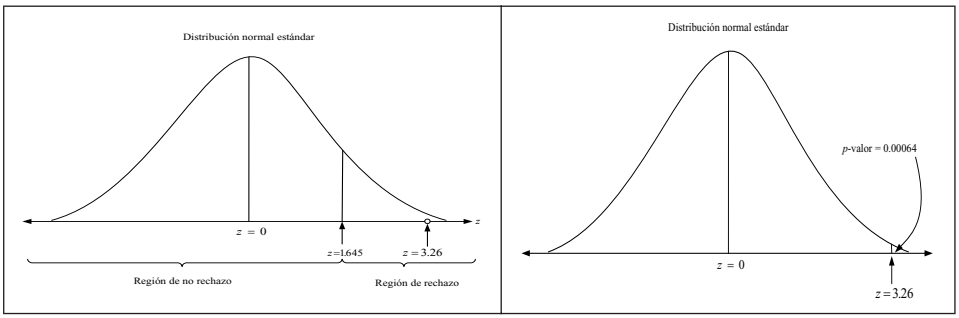


Figura 2. Distribución muestral de un estadístico de prueba con regiones de rechazo y aceptación y valor de p.

El valor $Z_{0.95}=1.645$ que corresponde a un $\alpha=0.05$ representa el cuantil que define el límite de ambas regiones y para los datos del problema que se representan en dicha distribución se tiene un valor calculado del *estadístico de prueba* $z=3.26$ el cual cae en la región de rechazo de la hipótesis nula. Es decir, la muestra que se ha seleccionado ha producido un estadístico que tiene una pequeña probabilidad de ocurrir ($p=0.00064$) partiendo de que es cierta la hipótesis nula, y es precisamente por ello que se le rechaza.

1.4. *Algunos resultados de investigación sobre concepciones erróneas en las pruebas de hipótesis*

Los problemas de la enseñanza y aprendizaje de la inferencia estadística, particularmente de las pruebas de hipótesis, han sido poco estudiados hasta ahora por la investigación educativa, no obstante la complejidad y la importancia que representan en los cursos de estadística universitarios y en la aplicación en diversas disciplinas científicas. Castro Sotos, Vanhoof, Van den Noortgate y Onghena (2007) realizaron una investigación minuciosa en importantes fuentes bibliográficas para conocer las principales dificultades y

[2] El estadístico de prueba es una variable aleatoria que comúnmente se denota mediante una mayúscula (Z). El valor del estadístico de prueba con los datos de una muestra es su valor calculado y es usual representarlo mediante letra minúscula (z). El valor crítico que delimita una región de la distribución muestral de rechazo o no rechazo corresponde a un determinado cuantil de la distribución del estadístico de prueba (por ejemplo: $Z_{0.95}=1.645$ con un $\alpha=0.05$). Como puede verse se trata de tres conceptos diferentes que no deben confundirse.

concepciones erróneas que tienen los estudiantes universitarios sobre los conceptos de inferencia estadística. En dicha investigación encontraron sólo diecisiete estudios que proporcionaban evidencia empírica en un total de más de 500 artículos que fueron publicados sobre el tema en el período de 1990 a 2006. Dichos autores concluyen que muchas de las concepciones erróneas sobre las pruebas de hipótesis son derivadas de los libros de texto y que los profesores, incluso algunos estadísticos, comparten las mismas concepciones erróneas de estudiantes. En la Tabla I se muestran las concepciones identificadas por dichos autores acerca de los conceptos que intervienen en el proceso de una prueba de hipótesis.

TABLA I
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 resumen, las principales concepciones erróneas asociadas al aprendizaje y uso de las pruebas de hipótesis derivadas de la investigación de Castro Sotos y otros (2007), en general son atribuidas a dos aspectos fundamentalmente: la filosofía que subyace a las pruebas derivadas del manejo híbrido de los enfoques de Fisher y Neyman-Pearson y a la interpretación de conceptos y resultados.

Con base en lo anterior, en el presente trabajo se plantea investigar sobre la caracterización y nivel de razonamiento estadístico que estudiantes universitarios de matemáticas han alcanzado sobre las pruebas de hipótesis después de tomar un curso de inferencia estadística. En particular, la pregunta que guía esta investigación es: ¿cuál es el nivel de desarrollo y cómo se caracteriza el razonamiento estadístico de estudiantes universitarios de matemáticas acerca de las pruebas de hipótesis estadísticas?

2. Marco teórico

2.1. *Significado de razonamiento estadístico*

El razonamiento y el pensamiento estadístico son procesos a los que en años recientes se les ha prestado mucha atención por parte de los investigadores en educación estadística. En particular, Garfield (2002) señala que el razonamiento estadístico puede ser definido como la manera en la cual las personas razonan con ideas estadísticas y el sentido que le dan a la información estadística, lo cual implica hacer interpretaciones basadas en conjuntos de datos y sus representaciones; incluso el razonamiento estadístico puede implicar conectar un concepto con otro y combinar ideas sobre datos y azar; en suma, razonar estadísticamente significa entender y explicar los procesos estadísticos e interpretar completamente los resultados estadísticos. De esta manera, el razonamiento estadístico acerca de las pruebas de hipótesis implica comprender que la muestra que aporta la evidencia para probar la hipótesis del valor previamente definido de un parámetro de la población, es solo una del conjunto de posibles muestras que pueden ser extraídas de una población, y que por lo tanto, existe el riesgo de cometer errores con cualquiera de las dos decisiones que se tomen (rechazo o no rechazo de la hipótesis nula). En el proceso de prueba, los estudiantes deben de ser conscientes de la importancia de conocer la distribución muestral del estadístico de prueba y la definición de un criterio a partir del cual se decide el rechazo o no rechazo de la hipótesis nula.

Para evaluar y caracterizar el razonamiento estadístico acerca de un concepto o grupo de conceptos se pueden definir categorías que delimitan particularidades de dicho razonamiento, los cuales a su vez permiten clasificar por niveles a los sujetos que desarrollan cierta tarea o resuelven un problema estadístico. En este contexto, el modelo taxonómico SOLO (*Structure of Observed Learning Outcomes*) desarrollado por Biggs y Collis (1982) ha sido utilizado para definir categorías de desarrollo cognitivo de diversos conceptos estadísticos (por ejemplo Pfannkuch, 2005; Vizcarra & Inzunsa, 2009; Reading & Reid, 2006). En el modelo SOLO los conceptos y procesos utilizados por los sujetos al dar respuesta a las preguntas o tareas planteadas se pueden clasificar en un determinado nivel de los cinco niveles que se contemplan. En la segunda columna de la Tabla II se describe cada nivel del modelo en forma genérica, mientras que en la tercera columna describimos el modelo adaptado para el caso de las pruebas de hipótesis.

TABLA II
Niveles del modelo SOLO aplicados al proceso de una prueba de hipótesis

Nivel	*Descripción genérica*	*Descripción en contexto de las pruebas de hipótesis*
Nivel Preestructural	Los estudiantes no se enfocan en aspectos relevantes de la tarea que les ha sido planteada. Los conceptos o procesos son utilizados de forma simplista que los conduce a cometer errores. Pueden dejar la tarea sin resolver por falta de comprensión.	Los estudiantes cometen errores debido a que poseen información aislada y superficial de los conceptos que intervienen en una prueba de hipótesis. En la etapa de identificación y formulación de las hipótesis pueden no tener claro que éstas hacen referencia a parámetros de una población. Pueden no distinguir cuál es la hipótesis nula y cuál es la hipótesis alternativa a partir del enunciado del problema y tienen dificultades para simbolizarlas y plantearlas correctamente. Una vez establecidas las hipótesis pueden tener dificultades en el procedimiento de prueba, cómo identificar y calcular el estadístico de prueba y el valor de p como consecuencia de no deducir correctamente la información proporcionada del problema. Pueden desconocer que el signo de la hipótesis alternativa ($>,<,\neq$) y el nivel de significancia determinan las regiones de rechazo y no rechazo de la hipótesis nula y cometer errores en su delimitación. Pueden no tener claro los criterios de decisión para rechazar o no rechazar la hipótesis nula y tomar decisiones erróneas. No distinguen la diferencia entre una prueba matemática y una prueba de hipótesis ni tienen claro el significado de conceptos como nivel de significancia y valor de p.

Nivel Uniestructural	Los estudiantes se enfocan en algún aspecto relevante de la tarea planteada o en alguna etapa de la tarea, realizan alguna conexión de un concepto o proceso con otro. Son capaces de desarrollar procesos simples.	Los estudiantes se enfocan en algún aspecto relevante de alguna etapa del proceso de prueba de hipótesis, son capaces de realizar conexiones sencillas entre conceptos y procesos simples de cálculo. En la etapa de identificación y formulación de hipótesis pueden identificar las hipótesis involucradas correctamente pero cometer errores en su planteamiento tanto a nivel simbólico como conceptual. En la etapa de cálculo de los elementos que determinan la decisión sobre la prueba pueden seleccionar correctamente el estadístico de prueba y calcularlo, de igual manera calcular correctamente el valor de p, pero no definir correctamente las regiones de rechazo o no rechazo y/o desconocer la regla de decisión para el rechazo o no rechazo de la hipótesis nula. Distinguen la diferencia entre una prueba matemática y una prueba de hipótesis y tienen claro el significado de conceptos como nivel de significancia y valor de p.
Nivel Multiestructural	Los estudiantes se enfocan en más de un aspecto relevante de la tarea, pero no logran integrarlos para obtener una solución correcta.	Los estudiantes se enfocan en más de un aspecto relevante en cada etapa de una prueba de hipótesis. Los estudiantes pueden identificar y formular las hipótesis correctamente y relacionar el signo de la hipótesis alternativa ($>,<,\neq$) para determinar las regiones de rechazo y no rechazo de la hipótesis nula en la distribución muestral del estadístico de prueba, así como el cálculo del valor de p; sin embargo pueden no tener claro los criterios de decisión, lo cual los conduce a una decisión errónea. O bien, pueden cometer errores en la identificación y planteamiento de hipótesis pero no tener claro como se calcula el valor de p, estadísticos de prueba y regiones de rechazo, lo cual también los conduce a errores de decisión. En cualquier caso, el no poder integrar de forma coherente todos los conceptos y procesos que intervienen en una prueba de hipótesis los conduce a cometer errores y no concluir con el proceso exitoso de la prueba.

Nivel Relacional	Los estudiantes integran todos los aspectos relevantes de la tarea como un todo coherente con estructura y significado.	Los estudiantes tienen la capacidad para integrar todos los conceptos que intervienen en una prueba de hipótesis y resolver los problemas que se les plantean de forma correcta, con una comprensión del proceso de la prueba. Identifican y formulan las hipótesis de forma correcta. Utilizan la forma de la hipótesis alternativa y el nivel de significancia para determinar las regiones de rechazo y no rechazo en la distribución muestral del estadístico de prueba. Identifican y calculan el estadístico de prueba y el valor de p correctamente con la información proporcionada del problema. Tienen claro los criterios de decisión para rechazar o no rechazar la hipótesis nula e interpretan correctamente los resultados utilizando el lenguaje estadístico adecuado.
Nivel Abstracto Extendido	Los estudiantes cumplen con las exigencias cognitivas del nivel relacional, pero además son capaces de transferir los conceptos y procesos fuera del contexto en que fueron aprendidos.	El estudiante debe ser capaz de realizar lo que exige el nivel relacional, además debe ser capaz de aplicar las pruebas de hipótesis en contextos más generales de donde fueron aprendidos.

3. Metodología

3.1. *Sujetos de estudio y escenario de la investigación*

Los sujetos de estudio que participaron en la investigación fueron once estudiantes voluntarios del último grado de la Licenciatura en Matemáticas en la Universidad

Autónoma de Sinaloa que tomaron como parte de su formación un curso de estadística matemática que contenía el tema de las pruebas de hipótesis. El investigador verificó con el profesor que imparte la materia que todos los estudiantes hubiesen tomado el curso completo y que los conceptos y aspectos a evaluar en el estudio sobre las pruebas de hipótesis habían hubiesen sido cubiertos.

3.2. *Instrumentos de recolección y análisis de los datos*

Se diseñó un cuestionario para evaluar el razonamiento estadístico con ítems tomados de otras investigaciones sobre el tema y de libros de texto (ver Anexo 1). El análisis de los datos ha sido básicamente cualitativo, para lo cual cada ítem requería una justificación de la respuesta. No todos los ítems fueron diseñados para contemplar los cinco niveles del modelo SOLO y ningún ítem fue diseñado para el nivel abstracto extendido. Con base en el análisis de las respuestas y la actividad matemática desarrollada por los estudiantes se ubicó a cada uno de ellos en alguno de los niveles del modelo SOLO que fueron considerados. La Tabla III muestra el nivel máximo esperado de cada ítem y los porcentajes de estudiantes ubicados en cada nivel de acuerdo con la descripción que se hace en la Tabla II. Los conceptos que se evaluaron fueron: lógica global del proceso de prueba de hipótesis, definición de nivel de significación, formulación de hipótesis, relación del valor de p con el tamaño de muestra y significancia estadística, procesos de prueba en contexto de una distribución *t* de Student y muestras con datos apareados.

4. Resultados y discusión

Los resultados obtenidos del análisis de los cuestionarios teniendo en cuenta la descripción de los niveles del modelo SOLO para el caso del razonamiento con pruebas de hipótesis se muestran en la Tabla III. Posteriormente se hace un análisis de respuestas que proporcionaron algunos estudiantes para algunos ítems, con el propósito de mostrar la comprensión o dificultades que tuvieron en su razonamiento.

TABLA III
Modelo SOLO para el razonamiento sobre pruebas de hipótesis con porcentajes de estudiantes por ítem y nivel

Ítem	*Nivel del Modelo SOLO y Porcentaje de Estudiantes*			
	Preestructural	*Uniestructural*	*Multiestructural*	*Relacional*
1 Significado de una prueba de hipótesis.	Eligen la opción incorrecta debido a que no tienen en cuenta que sólo una prueba matemática establece la verdad de una hipótesis. (64%)	Eligen la opción correcta señalando que una prueba no establece la verdad de la hipótesis nula, sino que se rechaza o no se rechaza por la evidencia que proporcionan los datos de la muestra, a un cierto nivel de significancia. (36%)		
2 Comprensión del nivel de significación.	Eligen la opción incorrecta debido a que no tienen en cuenta el orden en que intervienen los conceptos que se involucran. Invierten el condicional del nivel de significancia. (82%)	Eligen la respuesta correcta y señalan que el nivel de significancia cuantifica el porcentaje de casos en los que la hipótesis nula, siendo cierta, es rechazada. Se identifica como uno de los dos errores que se pueden cometer en una prueba de hipótesis. (8%)		
3 Formulación de hipótesis	Eligen una respuesta incorrecta debido a que no tienen claro que las hipótesis hacen referencia a parámetros poblacionales. (45%)	Eligen la respuesta correcta como consecuencia de comprender que la hipótesis nula se refiere a un parámetro poblacional. (55%)		

4 Formulación de hipótesis	Eligen una respuesta incorrecta porque desconocen que una hipótesis nula se plantea como una igualdad. (9%)	Eligen la respuesta correcta porque identifican que una hipótesis nula se plantea como una igualdad, para definir con ello la distribución muestral del estadístico de prueba. (91%)		
5 valor de p y significancia estadística	Eligen la respuesta incorrecta porque desconocen que el significado de significancia estadística está determinado por el valor de p. (27%)	Eligen la respuesta correcta porque comprenden que la significancia estadística está determinada por el valor de p, y que a valores más pequeños de p se hace menos factible la hipótesis nula. (73%)		
6 valor de p y su relación con el tamaño de muestra.	No responden el ítem o señalan que se debe rechazar la hipótesis nula. (64%)	Centran su atención en la comparación del valor de p con el nivel de significancia igual a 0.05 (valor que asumen por si mismos), lo que los lleva a considerar que el tratamiento no tiene efecto. (9%)	Señalan que existe la posibilidad de que exista un efecto del tratamiento, pues el tamaño de muestra es pequeño para concluirlo. Con ello los estudiantes comparan el valor de p con valores preestablecidos pero son conscientes que el tamaño de la muestra es pequeño para concluir la existencia del efecto. (27%)	

7 Lógica y aplicación de proceso de una prueba de hipótesis.	No responden el ítem. (64%)	Presentan argumentos incompletos. Confunden las regiones de aceptación y rechazo. Confunden la hipótesis nula y la alternativa. (27%)	Los argumentos muestran que comprenden el proceso general del contraste de hipótesis y realizan bien el proceso. Es decir, elige la hipótesis adecuada, realiza bien el procedimiento de prueba y elige la solución correcta. (9%)	
8 Proceso de prueba de hipótesis para muestras con datos apareados.	No responden al cuestionamiento sobre el mejor método de prueba o señalan que no hay diferencia en ambos métodos. Realizan la prueba considerando muestras independientes cuando no lo son (método de Roger) (18%)	Realizan la prueba utilizando el método de datos apareados (método de Annete) con lo cual identifica la dependencia en las muestras, solo utilizan el valor de p como criterio de decisión. (73%)	Realizan la prueba utilizado el método de datos apareados (método de Annete) correctamente mediante los criterios de regiones de rechazo y el valor de p, para lo cual toman en cuenta todos los datos que les proporciona el enunciado del problema. (9%)	
9 Prueba de hipótesis aplicando *t* de Student e interpretación de resultados con software.	No responden la tarea o responden con argumentos erróneos, lo que evidencian falta de su incapacidad de interpretación de los resultados del software, o no utilizan el estadístico de prueba correcto. (73%)	Realizan la prueba correctamente utilizando el valor de p como criterio de decisión. Todos los estudiantes utilizan un nivel de significancia del 5%. (18%)	Realizan la prueba correctamente utilizando el criterio del valor de p. Hacen una interpretación adecuada de los resultados. (9%)	

10 Proceso de prueba de hipótesis aplicando *t* de Student.	No responden la tarea. Plantean mal las hipótesis. Cometen errores en el procedimiento de prueba, en particular en el planteamiento de hipótesis. (27%)	Realizan cálculos correctos de los estadísticos. Confusión entre las distribuciones muestrales. 55%	Establecen de forma correcta las hipótesis. Realizan cálculos correctos pero no usan bien los criterios de decisión y fallan en la interpretación. 9%	Realizan en forma correcta la prueba de hipótesis e interpretan en contexto. 9%

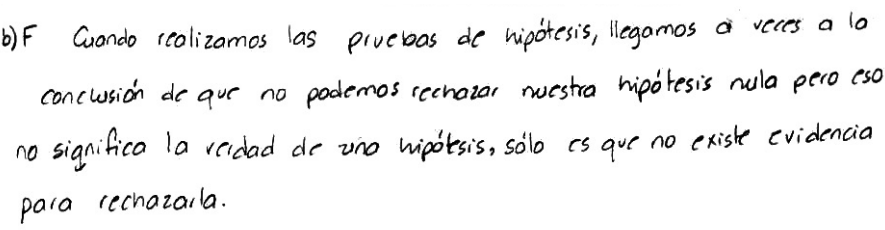
El análisis de la Tabla III nos muestra que un alto porcentaje de estudiantes (del 45 al 82%, según el ítem) se encuentran en el nivel más bajo de razonamiento estadístico (nivel preestructural), en un grupo de ítems que evaluaban conceptos como nivel de significancia, valor de p y su relación con el tamaño de muestra, significado de una prueba de hipótesis e interpretación de resultados de una prueba de hipótesis. De este grupo de ítems, los ítems 1, 2 y 3 tenían como nivel máximo el nivel uniestructural mientras que los ítems 6, 7 y 9 el nivel multiestructural. En el otro grupo de ítems (4, 5, 8 y 10) que evaluaban conceptos como efecto del tamaño de muestra en el valor de p, elección adecuada del proceso de prueba con datos apareados e identificación y proceso que involucra una prueba *t* de Student, el porcentaje de estudiantes que se ubicaron en el nivel preestructural fue del 9 al 27%, lo que significa que tuvieron una ligera mejoría en el nivel de razonamiento estadístico en estos conceptos respecto a los del primer grupo de ítems. En este grupo de ítems, los ítems 4 y 5 tenían como nivel máximo el nivel uniestructural, el ítem 8 el nivel multiestructural y el ítem 10 el nivel relacional.

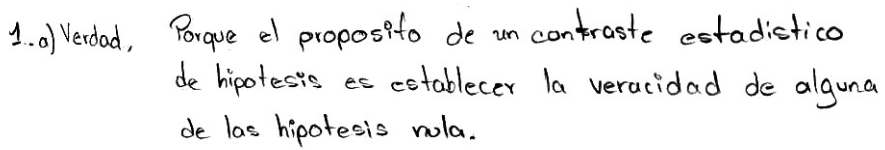
En la misma Tabla III se observa que en el primer grupo de ítems (1, 2, 3, 6, 7 y 9) se ubicaron del 9 al 36% de los estudiantes en el segundo nivel de razonamiento (nivel uniestructural), mientras que en el segundo grupo de ítems (4, 5, 8 y 10) se ubicaron del 60 al 90% de los estudiantes. En los ítems que fueron diseñados para mayores niveles de razonamiento (6, 7, 8, 9 y 10) se ubicaron muy pocos estudiantes en los niveles multiestructural y relacional. Es importante señalar que en los ítems 3, 4 y 5 se ubicaron del 54 al 72% de estudiantes en el nivel uniestructural, nivel máximo evaluado para estos ítems, por lo que fueron los ítems relativos a los conceptos de formulación de hipótesis y relación de valor de p con la significancia estadística los que tuvieron mejor desempeño.

Esto significa que, de acuerdo al modelo, los estudiantes que se ubicaron en el nivel preestructural poseen información aislada de los conceptos que

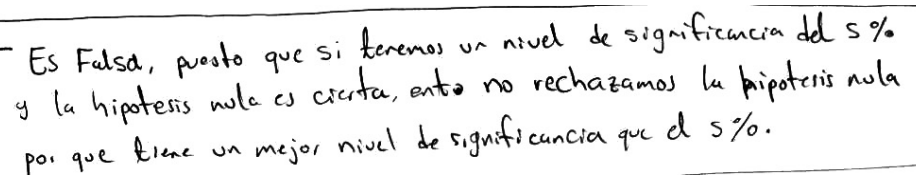
intervienen en una prueba de hipótesis, cometen errores en el planteamiento, tienen dificultades para realizar el procedimiento de prueba y no interpretan correctamente resultados de pruebas realizadas mediante un software. No se tiene clara la diferencia entre una prueba matemática y una prueba de hipótesis, ni se comprenden conceptos de suma importancia como el nivel de significancia y el valor de p. Los estudiantes que se ubicaron en el nivel uniestructural se enfocaron en algún aspecto relevante de las pruebas de hipótesis y son capaces de realizar conexiones sencillas entre conceptos y procesos simples de cálculo. Por ejemplo, pueden identificar las hipótesis involucradas correctamente pero comenten errores en su planteamiento, realizan cálculos correctamente pero desconocen las reglas de decisión o tienen errores de cálculo. Pueden distinguir entre una prueba matemática y una prueba de hipótesis y tienen claro el significado de conceptos como nivel de significancia y valor de p. Los estudiantes que se ubicaron en el nivel multiestructural se enfocaron en más de un aspecto relevante en el proceso de una prueba de hipótesis, pero no lograron integrar de forma coherente todos los conceptos y procesos que intervienen, lo que los condujo a cometer errores y a no concluir el procedimiento de prueba de forma correcta. Finalmente, los estudiantes que se ubicaron en el nivel relacional tuvieron la capacidad para integrar todos los conceptos que intervienen en una prueba de hipótesis y resolvieron los problemas que se les plantearon en forma correcta, con una comprensión del proceso de la prueba. En general, los estudiantes mostraron dificultades en todas las etapas del proceso de una prueba de hipótesis, desde el planteamiento hasta la interpretación de los resultados, mostrando falta de dominio de procedimientos y sobre todo de aspectos conceptuales.

Un análisis conceptual y procedimental más detallado nos permite ver que 63% de los estudiantes encuestados incurrieron en la concepción errónea –ya documentada en otros estudios–de considerar una prueba de hipótesis como prueba matemática que establece la verdad. Se esperaba que estos estudiantes, dada su formación matemática, tuvieran claro que las pruebas de hipótesis no son una prueba matemática. A continuación se muestran dos ejemplos de argumentaciones de los alumnos, una correcta y la otra incorrecta:

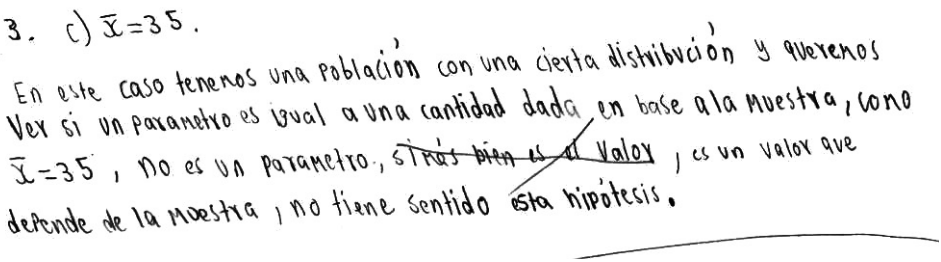




En cuanto al nivel de significancia, los estudiantes lo utilizaron en su mayoría en forma adecuada en los procesos de prueba de hipótesis como elemento de decisión para rechazar o no rechazar la prueba, pero fallaron notablemente en el ítem 2 donde se solicitaba su comprensión. Un ejemplo de razonamiento incorrecto lo mostramos a continuación:



Dos aspectos relevantes en el planteamiento de una hipótesis nula consisten en lo siguiente: la hipótesis nula hace referencia a un valor numérico del parámetro poblacional y la hipótesis nula se define como una igualdad respecto al valor supuesto del parámetro. El primer caso resultó más complicado pues sólo el 55% de los estudiantes lograron responder correctamente el ítem 3, en el segundo caso (ítem 4) sólo un estudiante tuvo dificultades para identificar la relación de igualdad de la hipótesis con el valor supuesto del parámetro. A continuación se muestra un ejemplo de respuesta correcta en el ítem 3:



Por su parte, en cuanto al significado del valor de p y sus implicaciones en la significancia estadística, el 73% de los estudiantes mostraron una comprensión adecuada, no así en el caso de relación entre el tamaño de la muestra y el valor de p que resultó ser difícil para los sujetos de estudio (ítems 5 y 6 respectivamente). Un ejemplo de respuesta correcta se muestra a continuación:

5. b), si el p_valor es menor o igual al nivel de significancia, entonces los resultados son significativos.

Un ejemplo de respuesta incorrecta a estos ítems es el siguiente:

5) a) Un gran p-valor.
Tenemos que un p-valor mayor que 0.05 indica que es estadísticamente significativo.

Finalmente, en cuanto a la aplicación de procedimientos para realizar correctamente una prueba de hipótesis, los resultados muestran que se utilizaron tanto los criterios del valor de p como los de regiones de rechazo y no rechazo sobre la distribución muestral del estadístico de prueba. Sin embargo, se observaron diversas dificultades para realizar en forma completa y correcta el proceso de prueba. Las principales dificultades consistieron en confundir términos como el estadístico de prueba y el valor de Z definido por el nivel de significación, no tener claro el criterio de decisión, no elegir el proceso de prueba adecuado para datos apareados e interpretación de resultados.

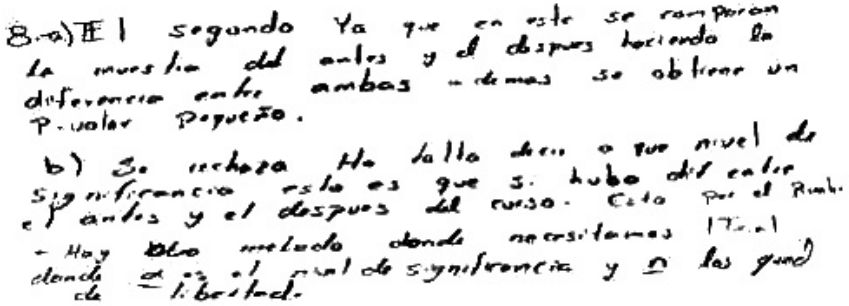
Por ejemplo, en el ítem 7, un estudiante calcula el estadístico de prueba correctamente pero no tiene claro el criterio de decisión y decide incorrectamente no rechazar la hipótesis nula a pesar de que el estadístico de prueba cae en la región de rechazo.

7- a)
Ya que $Z = \frac{\bar{X}-\mu_x}{\sigma_x} = \frac{60-50}{4} = 2.5 > 1.96 = z$ cuantil de la dist. normal estándar.
y por esto no se rechaza H_0.

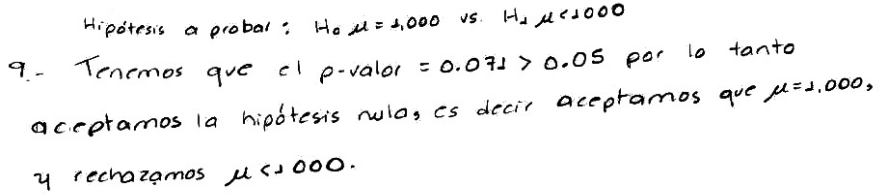
Otro estudiante, por su parte, realiza todo el proceso correctamente. Es la única respuesta que se ubicó en el nivel multiestructural de este ítem.

7- b), p_valor $= Pr(\bar{z} > \frac{\bar{x}-\mu_x}{\sigma_x}; H_0)$
$= Pr(z > \frac{60-50}{4})$
$= Pr(z > \frac{10}{4}) = Pr(z > 2.5)$
$= 1 - Pr(z \leq 2.5) = 1 - \Phi(2.5)$
$= 1 - 0.9938$
$= 0.0062 < 0.05$

Por su parte en el ítem 8 que consistía en elegir uno de los dos procedimientos de prueba mostrados (la opción correcta es la de datos apareados), de acuerdo con la dependencia de las muestras. El estudiante selecciona el método de forma correcta en el inciso a) y rechaza correctamente la hipótesis del inciso b).

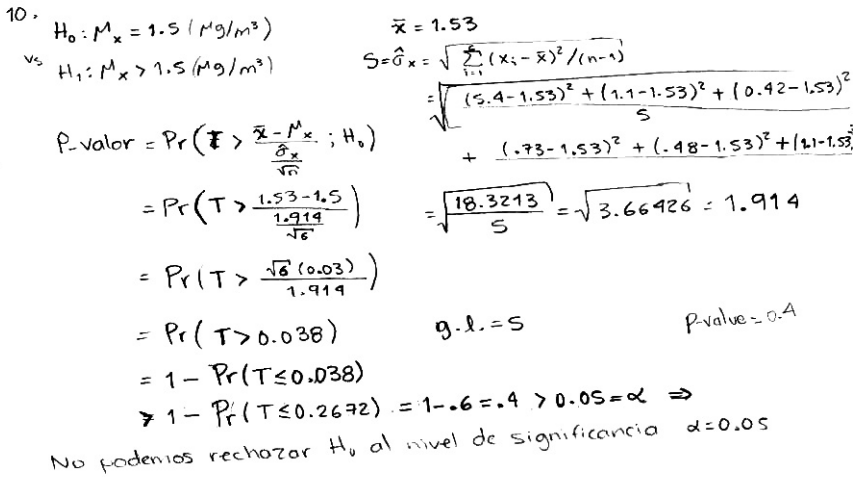


En el ítem 9 nos proponemos evaluar la capacidad de interpretar correctamente los resultados que se obtienen cuando se utiliza software para el desarrollo de una prueba de hipótesis. El ítem no requiere cálculos, sólo análisis de la información que proporcionan los datos del problema y los resultados del software.



La respuesta anterior muestra que el estudiante utiliza el valor de p ($p=0.071$) que produce el software y lo compara con el nivel de significancia $\alpha=0.05$. Decide correctamente sobre el resultado de la prueba utilizando el lenguaje de aceptarla en lugar de utilizar el término apropiado: no rechazarla.

Finalmente, en el ítem 10 que trata de un problema abierto, nos propusimos obtener información de todo el proceso de prueba, desde el planteamiento de las hipótesis hasta la interpretación de los resultados. Sólo un estudiante desarrolló el proceso completo como se muestra a continuación:



Se observó que los estudiantes cometieron diversos errores que van desde cálculos incorrectos de la media, la desviación y el error estándar, hasta errores en el planteamiento de hipótesis y en el uso de estadísticos de prueba.

5. Conclusiones

Los resultados de la investigación muestran que las pruebas de hipótesis son un concepto complejo para los estudiantes universitarios, aún cuando han tomado cursos de estadística matemática y fundamentos de teoría de la probabilidad. Un alto porcentaje de estudiantes se ubicaron en el nivel preestructural, en ítems que evaluaban conceptos clave tales como nivel de significancia, valor de p y su relación con el tamaño de muestra, significado de una prueba de hipótesis e interpretación de resultados de una prueba de hipótesis. En los ítems que contemplaban los niveles superiores del modelo SOLO, muy pocos estudiantes se ubicaron en los niveles multiestructural y relacional. Los ítems de mejor desempeño abordaron conceptos de formulación de hipótesis y relación de valor de p con la significancia estadística.

En este sentido, el razonamiento estadístico de los estudiantes que participaron en la investigación se caracteriza por ser aislado en relación con los diversos conceptos que se involucran en las pruebas de hipótesis, esto derivado de la falta de comprensión y por creencias erróneas sobre diversos

conceptos involucrados. Ello trae como consecuencia que de acuerdo con las categorías del modelo SOLO se ubiquen en un nivel de razonamiento estadístico de preestructural a uniestructural principalmente. En la forma que Garfield (2002) define al razonamiento estadístico, estos estudiantes tienen dificultades para comprender y explicar la lógica y los procesos estadísticos que subyacen a las pruebas de hipótesis, así como dar sentido a información estadística que se les presente en medios de comunicación o en la lectura de algún reporte de investigación que presente resultados estadísticos basados en esta metodología estadística.

Entre las implicaciones que se derivan del presente estudio está la sugerencia para que los cursos de inferencia estadística hagan mayor énfasis en aspectos conceptuales y procedimentales que consideren los dos enfoques de las pruebas de hipótesis, para desarrollar el razonamiento estadístico inferencial de los estudiantes. Por ello es importante que los estudiantes comprendan los fundamentos y la lógica que subyace a las pruebas de hipótesis, así como la relación que guardan entre sí los diversos conceptos que intervienen en el proceso de prueba; de tal forma que tengan sentido para ellos los procedimientos que utilizan y estén conscientes de las limitaciones de los resultados.

En esta dirección, la literatura en educación estadística ya reporta estudios y propuestas de diversos investigadores (Rossman & Chance, 1999; Inzunsa, 2010; Lipson, 2002; Garfield & Ben-Zvi, 2008) para el desarrollo de ambientes de aprendizaje que promuevan en los estudiantes un razonamiento estadístico adecuado sobre la inferencia estadística. En dichos estudios y propuestas, la tecnología computacional aparece como un elemento importante a considerar en la enseñanza, dada la multiplicidad de representaciones y el potencial cognitivo que proporcionan algunas herramientas.

Referencias bibliográficas

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning: An International Journal 2*(1-2), 75-97.

Biggs, J. B.; Collis, K. F. (1982). *Evaluating the Quality of Learning: The Solo Taxonomy*. New York: Academic Press.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review 48*(3), 378-399.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W.; Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review 2*(2), 98-113.

Chance, B., del Mas, R.; Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295-323). Netherlands: Kluwer Academic Publishers.

Cumming, G. (2010). Understanding, teaching and using p values [CD-ROM]. In Ch. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics*. Ljubljiana Slovenia: University of Slovenia.

Garfield, J. (2002). The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education 10*(3). Recuperado el 08 de octubre de 2010 de http://www.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J.; Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning. Connecting Research and Teaching Practice.* The Netherlands: Springer.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (Vol. 1, pp. 311-339). Hillsdale: Erlbaum.

Good, P. I. & Hardin, J. W. (2009). *Common Errors in Statistics (and how to avoid them)*. (3th ed.). New Jersey: John Wiley and Sons, Inc.

Grings, R.; Viali, L. (2011). Teste de Hipóteses: uma análise dos erros cometidos por alunos de engenharia. *Boletim de Educação Matemática* 24(40), 835-854.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in the textbooks. *Journal of Experimental Education 61*(4)*,* 317-333.

Inzunsa, S. (2010). Entornos virtuales de aprendizaje: un enfoque alternativo para la enseñanza y aprendizaje de la inferencia estadística. *Revista Mexicana de Investigación Educativa 15*(45), 423-452.

Kirk, R. (2001). Promoting good statistical practices: some sugestions. *Educational and Psychological Measurement 61*(2), 213-218.

Kline, R. B. (2004). *Beyond Significance Testing*. *Reforming Data Analysis Methods in Behavioral Research.* Washington: American Psychological Association.

Levin, J. R. (1998). What If There Were No More Bickering About Statistical Significance Tests? *Research in Schools 5*(2), 43-53.

Lipson, K. (2000). *The role of the sampling distribution in developing understanding of statistical inference*. Doctoral Thesis unpublished, Swinburne University of Technology, Swinburne, Australia.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution [CD-ROM]. *Proceedings of the 6th International Conference on Teaching of Statistics*. Cape Town South Africa: South African Statistical Association.

Liu, Y.; Thompson, P. W. (2005). Teachers' understanding of hypothesis testing [CD-ROM]. In S. Wilson (Ed.), *Proceedings of the 27thAnnual Meeting of the International Group for the Psychology of Mathematics Education*. Virginia: Virginia Tech University.

McLean, A. (2002). Statistacy: Vocabulary and Hypothesis Testing [CD-ROM]. In B. Phillips (Ed.), *Proceedings of the 6th International Conference on Teaching of Statistic*. Cape Town South Africa: South African Statistical Association.

McLean, J. E; Ernest, J. M. (1998). The Role of Statistical Significance Testing In Educational Research. *Research in Schools 5* (2), 15-22.

Mittag, K. C.; Thompson, B. (2000). A National Survey of AERA Member´s Perceptions of Statistical Significance Tests and Other Statistical Issues. *Educational Researcher 29*(4)*,* 14-20.

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.

Moore, D. S. (1992). ¿What is Statistics? In D. C. Hoaglin & D. S. Moore (Eds.), *Perspectives on Contemporary Statistics* (pp. 1-17). Washington: Mathematical Association of America.

Morrison, D. E. & Henkel, R. E. (1970). *The Significance Test Controversy.* New Brunswick: Transactions Publishers.

Pfannkuch, M. (2005). Characterizing year 11 student´s evaluation of a statistical process. *Statistics Education Research Journal 4(2)*, 5-25. Recuperado el 08 de dieciembre de 2010. de http://www.stat.auckland.ac.nz/serj.

Reading, Ch.; Reid, J. (2006). An emerging hierarchy of reasoning about distribution: from a variation perspective. *Statistics Education Research Journal 5(2)*, 46-68. Recuperado el 15 de dieiembre de 2010 de http://www.stat.auckland.ac.nz/serj.

Robinson, D. H.; Levin, J.R. (1997). Reflections on statistical and substantive significance with a slice of replication. *Educational Researcher 26*(5), 21-26.

Rossman A.; Chance, B. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *College Mathematical Journal 30(4),* 297-305. Recuperado el 20 de octubre de 2010 de http://rossmanchance.com/papers/topten.html

Stigler, S. M. (1986). *The history of statistics: The measurement of the uncertainty before 1900*. Cambridge, MA: Belknap.

Triola, M. F. (2009). *Estadística*. México: Pearson Educación.

Vallecillos, A. (1996). *Inferencia estadística y enseñanza: un análisis didáctico del contraste de hipótesis estadísticas.* Granada: Editorial Colmenares.

Vallecillos, A. (1997). El papel de las hipótesis estadísticas en los contrastes: Concepciones y dificultades de aprendizaje. *Educación Matemática 9*(2), 5-20.

Vizcarra, F.; Inzunsa, S. (2009). Un estudio sobre la caracterización del razonamiento estadístico de estudiantes de preparatoria: el caso de los promedios y las gráficas [CD-ROM]. *Memorias del XXII Congreso Nacional de Enseñanza de las Matemáticas*, Tuxtla Gutiérrez Chiapas, México: Asociación Nacional de Profesores de Matemáticas.

Williams, A. M. (1998). Students' understanding of the significance level concept. In L. Pereira Mendoza, L. Seu Kea, T. Wee Kee & W. Wong ((Eds.), *Proceedings of the 5th International Conference on Teaching Statistics.* Singapur Korea: University of Korea.

Xitao, F. (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin. *The Journal of Educational Research 94*(5), 275-282.

Yañez, G.; Behar, R. (2010). The confidence intervals: a difficult matter, even for experts [CD-ROM]. In Ch. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics*. Ljubljiana Slovenia: University of Slovenia.

Anexo 1: Cuestionario

1. Significado de una prueba de hipótesis

Un contraste estadístico de hipótesis, correctamente realizado, establece la verdad de una de las dos hipótesis nula o alternativa. Justifica tu respuesta. V / F

El propósito del ítem es determinar si el estudiante comprende la diferencia entre una prueba de una hipótesis estadística y la demostración de una hipótesis matemática.

(Tomado de Vallecillos, 1997).

2. Definición del nivel de significación (α)

Un nivel de significación del 5% significa que a la larga, 5 de cada 100 veces que la hipótesis nula sea cierta, la rechazaremos. Justifica tu respuesta. V / F.

El objetivo de este ítem es investigar si los estudiantes comprenden correctamente la definición del nivel de significación. (Tomado de Vallecillos, 1997)

3. Formulación de hipótesis

¿Cuál de las siguientes no es una hipótesis nula legítima? Justifica tu respuesta.

a) $\mu_x = 10$ *b)* $\sigma_x = 3$ *c)* $\bar{x} = 35$ *d)* $\mu_1 = \mu_2$

El ítem tiene el propósito de ver si los estudiantes comprenden que la hipótesis nula involucra a un parámetro de una población. También se pretende ver si confunden el estadístico con el parámetro.

(Tomado de Vallecillos, 1997)

4. Formulación de hipótesis

En una encuesta electoral, se desea investigar si hay más ciudadanos estadounidenses a favor de la política económica del presidente que en contra de la misma. Supongamos que p representa la proporción de habitantes que están de acuerdo con dicha política económica y que q = 1-p es la proporción de que no estén de acuerdo con dicha política económica. ¿Cuál de las siguientes hipótesis elegirías como hipótesis nula? Justifica tu respuesta

a) $p > q$ *b)* $p = q = 1/2$ *c)* $p \neq q \neq 1/2$ *d)* $q > p$

Con este ítem nos propusimos investigar si los estudiantes identifican, en el enunciado del problema, la información que le permitirá plantear adecuadamente la hipótesis nula.

(Tomado de Vallecillos, 1997).

5. El p-valor de una prueba de hipótesis y su relación la significancia estadística

Si soy un investigador con la esperanza de demostrar que los resultados de un experimento fueron estadísticamente significativos, ¿qué prefiero? Justifica tu respuesta.

a) Un gran p-valor
b) Un pequeño p-valor
c) Los p-valores no están relacionados con la significación estadística

Con este ítem se propone evaluar si los estudiantes comprenden el significado del valor de p en una prueba de hipótesis estadística y cómo está relacionado con la expresión "estadísticamente significativo", de uso corriente en las pruebas de hipótesis.

(Tomado de Proyecto ARTIST: https://app.gen.umn.edu/artist/index.html)

6. El p-valor de una prueba de hipótesis y su relación con el tamaño de la muestra

Un investigador realiza un experimento sobre la memoria humana y recluta a 15 personas para participar en su estudio. Realiza el experimento y analiza los resultados. Obtiene un p-valor de 0.17. ¿Cuál de las siguientes opciones es una interpretación razonable de sus resultados? Justifica tu respuesta.

a) Esto demuestra que su tratamiento experimental no tiene ningún efecto sobre la memoria.
b) Podría haber un efecto del tratamiento, pero el tamaño de la muestra es demasiado pequeño para descubrirlo.
c) Se debe rechazar la hipótesis nula.
d) Hay pruebas de un efecto pequeño sobre la memoria de su tratamiento experimental.

El propósito de este ítem es investigar el razonamiento de los estudiantes acerca del valor de p en una prueba de hipótesis y el efecto que en él puede tener el tamaño de la muestra.

(Tomado de Proyecto ARTIST: https://app.gen.umn.edu/artist/index.html)

7. Lógica y aplicación del proceso

Supongamos las siguientes hipótesis:

$H_o: \mu_x = 50$
$H_1: \mu_x > 50$
$H_2: \mu_x < 50$

Con un nivel de significación $\alpha = 0.05$, $z = 1.96$, *un valor de* $\bar{x} = 60$, $\sigma_x = 4$, *un valor del estadístico* $z = \frac{\bar{X} - \mu_x}{\sigma_x}$ *y una población normalmente distribuida, entonces podríamos:*

a) No rechazar $H_o: \mu_x = 50$
b) No rechazar $H_1: \mu_x > 50$
c) Necesitar más información.
d) No rechazar ambas hipótesis alternativas y H_1 y H_2.

Tomado de Vallecillos (1996).

Este es un ítem de aplicación en el que establecida una hipótesis relativa a la media de una población, un nivel de significación y dado el valor y fórmula del estadístico a emplear, se requiere realizar el procedimiento de prueba para aceptar o rechazar la hipótesis nula. Es importante mencionar que el planteamiento de hipótesis sólo debe contener una hipótesis nula y una hipótesis alternativa, sin embargo, lo que se pretende al colocar dos hipótesis alternativas a la vez es que los estudiantes seleccionen la hipótesis correcta y realicen el análisis. Un inciso solicita la aceptación de ambas hipótesis con el propósito de ver si los estudiantes son conscientes que no pueden ser utilizadas ambas a la vez.

8. Prueba de hipótesis con datos apareados

Los siguientes datos se generaron en un estudio de la eficacia de la formación del personal de enfermería. En este estudio, a 15 enfermeras se les aplicó un examen de conocimientos sobre el cáncer antes de asistir a un seminario de un día sobre el tema. Posteriormente se les aplicó el mismo examen sobre el tema al término de éste. Los resultados se muestran en la tabla siguiente:

Enfermera	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*Ñ*
Antes	29	20	24	32	33	19	17	32	16	28	35	28	18	45	19
Después	35	41	33	41	39	20	29	42	36	37	36	33	35	42	19

El organizador del curso quería saber si había habido un aumento en el conocimiento sobre el cáncer después de asistir al seminario. Dos asistentes, Roger y Annette, se dieron a la tarea de analizar los datos utilizando el software Minitab. Roger realizó una prueba t para con los siguientes resultados:

Prueba t.
Después vs Antes

	n	*Media*	*Desviación Estándar*	*Error Estándar de la Media*
Después del Seminario	15	34.53	7.14	1.8
Antes del Seminario	15	26.33	8.29	2.1

$$H_o : \mu_{Después} = \mu_{Antes}$$
$$vs$$
$$H_1 : \mu_{Después} > \mu_{Antes}$$

t = 2.90; p = 0.0036; grados de libertad = 27. Annette convirtió las dos filas de datos en una sola colocando solamente las diferencias (d_i) de las puntuaciones y llevó a cabo una prueba t, obteniendo los siguientes resultados:

Prueba
$\mu d_i = 0 \;\; vs \;\; \mu d_i > 0$

	n	*Media*	*Desviación Estándar*	*Error Estándar de la Media*	*t*	*p-valor*
Diferencias	15	8.20	7.15	1.85	4.44	0.003

a) ¿Cuál análisis es el más apropiado y por qué?

b) Usando la información generada en el análisis que consideraste más apropiado, prueba la hipótesis de que se ha producido un incremento en las puntuaciones de las enfermeras que asistieron al taller. Menciona los valores de los estadísticos relevantes.

En este ítem se propone evaluar si los estudiantes saben plantear e interpretar los resultados de una prueba que involucra muestras dependientes con datos apareados. El estudiante debe elegir la mejor manera de resolver el problema y justificar su elección, además debe de escoger correctamente los valores adecuados que se deben considerar para poder estar en condiciones de emitir un juicio sobre el problema.

(Tomado de Lipson 2000).

9. Prueba de hipótesis aplicando la distribución t de Student e interpretación con software

Por el costo que implica, las pruebas de choques de automóviles suelen utilizar muestras pequeñas. Cuando se chocan cinco automóviles BMW en condiciones estándar, se emplean los costos de reparación (en dólares) para probar la aseveración de que el costo medio de reparación de todos los automóviles BMW es menor que $1,000. Los resultados del software Minitab de esta prueba de hipótesis se presentan abajo.

$$\mu = 1000 \ \ vs \ \ \mu < 1000$$

Variable	*n*	*Media*	*Desviación Estándar*	*Error Estándar de la Media*
Costo	5	767	285	127

Variable	*Límite Superior del 95%*	*t*	*p-valor*
Costo	1039	-1.83	0.071

Con base en los resultados de esta prueba de hipótesis, ¿se justificaría que BMW anunciara que, en condiciones estándar, el costo promedio de reparación es menor de $1,000? Justifica tu respuesta.

El propósito de este ítem es evaluar si los estudiantes tienen la capacidad de interpretar correctamente los resultados que se obtuvieron con el software Minitab y poder emitir su juicio sobre el contraste de hipótesis que se presenta.

Tomado de Triola (2009).

10. Procedimental. Usando la distribución t de Student

La Enviroment Protection Agency (EPA) ha establecido un estándar de calidad del aire para el plomo: 1.5 μg/m³. Las siguientes mediciones: 5.40, 1.10, 0.42, 0.73, 0.48 y 1.10, se registraron en el edificio cinco del World Trade Center en diferentes días, inmediatamente después de la destrucción causada por los ataques terroristas del 11 de septiembre de 2001. Después del colapso de los dos edificios del World Trade Center surgió una preocupación sobre la calidad del aire. Utilice un nivel de significancia de 0.05 para probar la aseveración de que la muestra proviene de una población con una media mayor que el estándar de la EPA, de 1.5 μg/m³. Justifica tu respuesta.

Tomado de Triola (2009).

Con este ítem se pretende investigar si los estudiantes identifican adecuadamente en el enunciado de un problema la información correcta para plantear y resolver adecuadamente una prueba de hipótesis y concluir con una explicación contextualizada.

Autores

Santiago Inzunsa Cazares. Universidad Autónoma de Sinaloa, México. sinzunza@uas.edu.mx

José Vidal Jiménez Ramírez. Universidad Autónoma de Sinaloa, México. vidaljr@uas.edu.mx