¿PRUEBA DE HIPÓTESIS, INTERVALO DE CONFIANZA O PRUEBA DE SIGNIFICANCIA?

Gudelia Figueroa Preciado, Irma Nancy Larios Rodríguez y María Elena Parra Ramos
Universidad de Sonora México
gfiguero@gauss.mat.uson.mx, nancy@gauss.mat.uson.mx, meparra@gauss.mat.uson.mx

Resumen. En estadística, técnicas inferenciales como son intervalos de confianza, prueba de hipótesis y pruebas de significancia, son ampliamente utilizadas para hacer inferencias acerca de ciertas características de una población. Aunque este material ocupa gran parte de los cursos básicos de estadística de nivel universitario, es común que en la práctica se haga un uso indistinto de estas técnicas. El presente trabajo surge de advertir esta problemática, pues es usual observar que se plantean hipótesis nulas y alternativas durante el transcurso de un análisis, se malinterpreta la información que proporciona un intervalo de confianza y se considera el p-valor como un número mágico, cuya sola magnitud permite emitir conclusiones, sin la necesidad de información adicional. En este trabajo se sugiere realizar un análisis exhaustivo de las similitudes y diferencias entre estas técnicas inferenciales y se propone una sencilla metodología que se apoya en el uso de simulaciones realizadas en el software R, que facilitan un análisis detallado de los resultados obtenidos.
Palabras clave: Significancia, p-valor, nivel de confianza
Abstract. In statistics, some inferential procedures like confidence intervals, test of hypothesis and significance tests are widely used to make inferences about some characteristics of a population. Although this material constitutes a fundamental part of an undergraduate course, it is common to observe, in practice, a misuse of these techniques. Our proposal comes up when observing that is usual to set up a null and an alternative hypothesis in the course of an analysis, that is very common to misunderstand the information that a confidence interval provides, and also it is very usual to consider a p-value like a magic number that will allow us to take a decision with no need of an aditional information. In this work we suggest to perform an exhaustive analysis concerning the differences and simmilarities between these inferential techniques and we propose a simple methodology that relies on the use of computer simulations made in the R sofware, because these facilitate to make a detailed analysis of the obtained results.
Key words: Significance, p-value, confidence level.

Introducción

El análisis inferencial es una herramienta fundamental en las investigaciones de muchas disciplinas científicas. Sin embargo, es común observar que se desconoce cuándo o porqué utilizar alguna técnica en particular. Por ejemplo, se usa indiscriminadamente tanto un intervalo de confianza como una prueba de hipótesis, o bien una prueba de significancia. Más aún, estas técnicas se llevan a cabo de manera mecánica y los resultados obtenidos muchas veces se interpretan de una manera ligera, lo que conduce a tomar decisiones equivocadas. Pocas veces se advierte que el uso de una u otra técnica depende del objetivo del estudio y que los alcances y limitaciones de los resultados obtenidos, están estrechamente ligados con la técnica utilizada.

Al finalizar un curso de estadística, es común observar que muchos estudiantes plantean hipótesis estadísticas tan sólo porque quieren mostrar si los resultados observados son "estadísticamente significativos" con respecto a cierta hipótesis. Consideran que el p-valor, constituye un resultado mágico que permite distinguir entre efectos reales y efectos que pudieran ser resultado del azar y es usual que terminen el curso sin asociar la relación que existe entre los valores que toma un

intervalo de confianza y el rechazo o no de una hipótesis nula. Por otra parte, es en realidad preocupante las interpretaciones erróneas que muchas veces dan a un intervalo de confianza o al p-valor de una prueba. Algunos creen que éste último representa la probabilidad de que la hipótesis nula sea cierta, dada la muestra observada; cuando en realidad representa la probabilidad de observar esa muestra, si la hipótesis nula fuera cierta.

Advertir esta problemática conduce a preguntarnos si como docentes hemos dejado de lado aspectos importantes, como el inducir en el estudiante una exploración y análisis entre las semejanzas y diferencias de estos métodos de inferencia estadística y el saber cuándo y porqué utilizarlos. Una discusión sobre algunas diferencias y similitudes entre pruebas de significación y pruebas de hipótesis, la expone Batanero (2001) y es material fundamental que debiera incluirse en los cursos de estadística, pues esclarecería la problemática filosófica de la inferencia estadística. Por otro lado, el analisis sugerido debiera realizarse sobre la base de problemas reales que pueden trabajarse como proyectos de clase; con ellos se puede mejorar la enseñanza y el aprendizaje de la estadística (Batanero & Díaz, 2005), pues permiten rescatar qué se quiere probar, qué pregunta debe contestarse, qué técnica es la más adecuada al problema, etcétera. Así pues, se pretende que el estudiante pueda resolver no sólo los problemas de aplicación planteados durante el curso, sino también aquellos a los que pudiera enfrentarse en su quehacer profesional. Actualmente es necesario que se incluya el uso de software estadístico en el análisis de datos, pues por una parte es cada vez más común el trabajar grandes conjuntos de datos y por otra, el estudiante debe saber identificar qué resultados arroja el software utilizado y poder interpretarlos correctamente.

Para abordar la problemática planteada se proponen dos actividades en las que se utiliza el software estadístico R, tanto en el proceso de simulación de las muestras como en la aplicación de la técnica inferencial seleccionada. Con estas actividades se pretende que el estudiante pueda elegir la técnica inferencial adecuada y logre hacer una correcta interpretación de los resultados que ésta arroje.

Referente teórico

Es cómún observar publicaciones científicas donde se hace un uso indiscriminado de pruebas de hipótesis, sin reflexionar acerca de si son en realidad el método más adecuado para analizar los datos en estudio. Según señala Clark (2004), una vez que el estudiante o investigador aprende a utilizar las pruebas de hipótesis, no vacila en aplicarlas libremente. Cuando una prueba de hipótesis es en realidad lo más adecuado, debe cuidarse el no basar la conclusión en la simple observación de la magnitud de una variable aleatoria, como es el p-valor. Este puede indicar que existen resultados significativos que lleven a tomar una conclusión, para la cual debiera integrarse más información que la proporcionada por ese único valor. Es común observar que muchas veces se

memoriza una regla consistente en observar si el p-valor es mayor o menor que el nivel de significancia establecido para la prueba y éste es el único sustento para tomar la decisión. El hacer eso y plantear hipótesis sin fundamentación alguna, suele llevar a conclusiones que pueden carecer de sentido, como el afirmar que la esperanza de vida de un actor es mayor cuando éste gana un Oscar (Wood, 2012), o bien que la probabilidad de procrear hijos con piel oscura es mayor de 0.3, si la madre bebe más de 15 tazas de café al día.

Debe aclararse que una prueba de hipótesis no proporciona información acerca de la magnitud de un efecto, esto es, puede haber dos muestras con la misma media muestral y sólo una de ellas arrojar resultados significativos, debido al tamaño de la muestra. Es decir la distancia entre la media poblacional y la media muestral puede ser la misma, pero las muestras conducen a p-valores diferentes. Kochanski (2005) enfatiza que los intervalos de confianza son el equivalente de encapsular varias pruebas de hipótesis, ya que si el intervalo de confianza no incluye el valor establecido en H_0, entonces en la prueba de hipótesis se rechazará H_0, y viceversa. Es importante analizar que un intervalo de confianza explica, con su amplitud, la incertidumbre debida al error muestral. Abdelhamid (2005) señala también que una característica importante del intervalo de confianza es que puede ser utilizado como una prueba de hipótesis e inferir el valor de p asociado a ésta.

Garfield y Ben-Zvi (2008) abordan de manera muy amplia las dificultades que tienen los estudiantes en la interpretación de técnicas inferenciales e indican que investigaciones recientes sugieren que se puede comprender mejor la inferencia estadística mediante el uso de herramientas tecnológicas y actividades de simulación. Sin embargo, añaden que el uso de estas herramientas de simulación no serán suficientes si no están vinculadas a conceptos de inferencia estadística.

Metodología

Para abordar la problemática planteada, se proponen dos actividades. En la primera de ellas se simulan muestras de dos diferentes tamaños, con las cuales se calculan intervalos del 90 y 95 por ciento de confianza. La actividad pretende mostrar que a mayor tamaño de muestra, se espera mejor cobertura del intervalo y que la amplitud de éste será más pequeña. Por otra parte, a mayor nivel de confianza, la amplitud del intervalo aumentará. La segunda actividad plantea un problema de aplicación con el cual se pretende mostrar la relación existente entre intervalo de confianza y prueba de hipótesis, así como la diferencia entre nivel de significancia y p-valor. Con este problema se muestra que un mismo conjunto de datos puede sustentar muchas hipótesis nulas y de la igual manera, un mismo conjunto de datos permite también rechazar varias hipótesis nulas. La comprensión de esto último se facilita calculando los p-valores correspondientes de cada una de las pruebas efectuadas y graficando éstos.

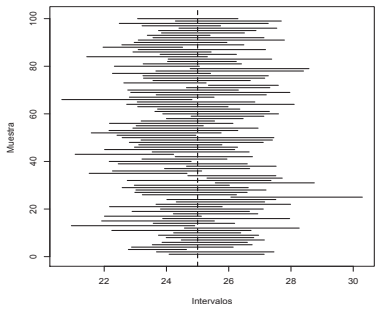
Las actividades anteriores deben estar complementadas con un análisis detallado, donde para cada problema en particular se analicen preguntas como: ¿qué utilizar? ¿Un intervalo de confianza, una prueba de hipótesis o bien una prueba de significancia? Si se selecciona una prueba de hipótesis, ¿qué plantear en la hipótesis nula? Si acaso se selecciona un intervalo de confianza ¿que ventajas tiene sobre el p-valor de una prueba de significancia?

El desarrollo de las dos actividades propuestas se puede realizar en aproximadamente dos horas clase. La experiencia ha mostrado resultados satisfactorios, pues al final del curso el estudiante es más analítico al tomar la decisión de cuál técnica inferencial utilizar para un problema particular y al mismo tiempo está más consciente en cuanto los alcances y limitaciones de los resultados obtenidos.

Aunque la propuesta presentada en este trabajo ha sido probada con un grupo de 25 estudiantes del área de Ciencias Exacta y Naturales durante el semestre 2012-2, no se ha realizado aún una experimentación estructurada metodológicamente, puesto que estamos en la etapa de retroalimentación de la misma, sin embargo esta va a ser experimentada durante el 2013 mediante un estudio de casos, utilizando diversas técnicas de recopilación de información como son hojas de trabajo, archivos del software del trabajo realizado por los estudiantes, videograbaciones, entre otras. Los análisis que se pretenden realizar son de corte cualitativo.

Primera actividad

Se simularon 100 muestras aleatorias de tamaños 10 y 50, de una distribución normal con media $\mu = 25$ y desviación estándar $\sigma = 3$. Para cada una de estas muestras se calculó un intervalo del 90 y del 95 por ciento de confianza, para la media de la distribución. Se contabilizó la cobertura empírica, esto es el número de intervalos que cubren el parámetro $\mu = 25$, esto con la finalidad de mostrar la interpretación frecuentista que proporciona un intervalo de confianza y que además el estudiante observe que, por lo general, a mayor nivel de confianza se tiene mayor cobertura ya que el intervalo de confianza es más amplio, y que al aumentar el tamaño de muestra la amplitud del intervalo de confianza disminuye. Estos resultados pueden observarse en las Figuras 1, 2, 3 y 4.



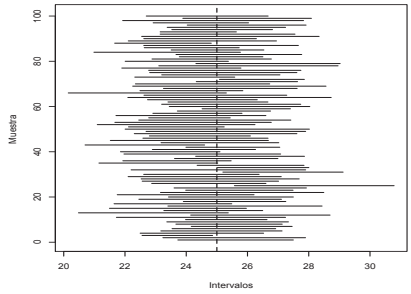


Figura 3. Intervalos del 90% confianza, n=10, cobertura=88

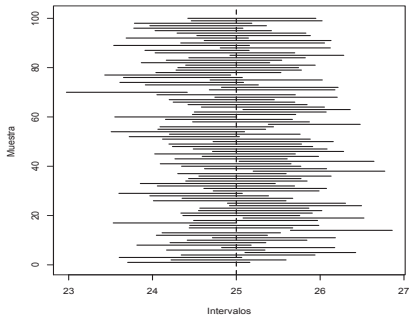


Figura 4. Intervalos del 95% de confianza, n=10, cobertura=93

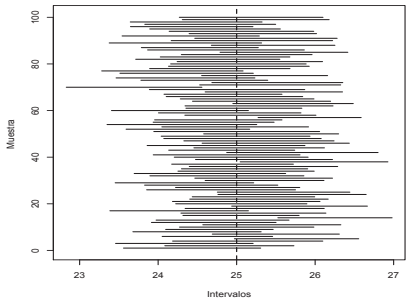


Figura 5. Intervalos del 90% de confianza, n=50, cobertura=91

Figura 6. Intervalos del 95% de confianza, n=50, cobertura=96

Segunda actividad

Primera Parte. Para esta actividad se plantea resolver un problema de aplicación, sencillo, como el siguiente:

Se midió la talla y peso de una muestra aleatoria de 20 estudiantes universitarias y con base en ello se desea estimar el índice de masa corporal (IMC). Los datos observados fueron:

26.2, 25.4, 23.4, 21.2, 20.5, 22.1, 28.4, 25.6, 19.3, 18.5,

19.7, 22.4, 26.1, 20.3, 18.8, 20.4, 22.6, 29.5, 21.3, 23.6.

La media y desviación estándar muestrales para el IMC resultan $\bar{x}$ =22.76 y $s = 3.19$, respectivamente. Al plantear este problema ante un grupo de estudiantes, es muy común que lo primero que realicen sea el cálculo de un intervalo de confianza para el índice promedio de masa corporal. Si se considera que la distribución de IMC satisface el supuesto de normalidad, entonces, un intervalo del 95% de confianza para el índice promedio de masa corporal en estudiantes universitarias se calcularía como:

$$\bar{x} \mp t_{1-\frac{\alpha}{2},n-1}\,(s/\sqrt{n})$$

donde $t_{1-\frac{\alpha}{2},n-1}$ es el valor de tablas en la distribución t-Student con $n-1$ grados de libertad, que deja un área de $1-\frac{\alpha}{2}$ a su izquierda. En este caso en particular se tiene que

$t_{0.975,19} = 2.0930$. Luego, un intervalo al 95% de confianza para μ resulta:

(21.27173, 24.25827)

Es importante señalar que al plantear ese problema de forma tan general, muchos estudiantes quieren probar alguna hipótesis nula. Se sugiere entonces, que con la ayuda de un software estadístico, se evalúe una serie de hipótesis nulas, donde la media μ tome valores entre 21.3 y

24.2, a intervalos de 0.1 (se probarán 30 hipótesis). Esto es, se asigna a μ valores que caen dentro del intervalo de confianza calculado anteriormente. Al probar cada una de estas treinta hipótesis, se guarda el p-valor resultante. Para la muestra de IMC, el valor mínimo de esta serie de p-valores fue 0.05406094 y el más grande resultó 0.9613857. El diagrama de caja que se obtiene con los treinta p-valores calculados se muestra en la Figura **7**.

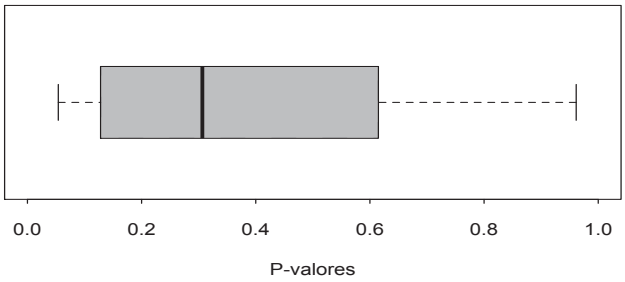


Figura 7. Diagrama de caja de p-valores resultantes de probar hipótesis nulas con valores de $\mu \in (21.1, 24.2)$

En esta parte de la actividad el estudiante analiza que todos los p-valores resultaron mayores que 0.05 y deduce que eso siempre sucederá si el valor que toma μ en la hipótesis nula es un valor comprendido dentro del intervalo de confianza.

Segunda Parte. Aprovechando lo analizado en la primer parte de la actividad, se informa que en el estudio interesa ver si la muestra aleatoria de 20 estudiantes, apoya la hipótesis de que esa muestra proviene de una distribución normal con un IMC promedio de $\mu = 25$. Se selecciona éste valor, o cualquier otro que se encuentre fuera del intervalo de confianza obtenido. En este momento se solicita a los estudiantes que emitan su opinión sobre el resultado de esta prueba, antes de realizarla, considerando un nivel de significancia $\alpha = 0.05$. El fijar ese nivel tiene la finalidad de que el estudiante observe que éste valor existe aún antes de realizar la prueba y que representa la probabilidad de rechazar la hipótesis nula, cuando ésta es cierta; concepto muy diferente del p-valor, que sólo puede calcularse después de obtenida la muestra. Un gran porcentaje de estudiantes deduce que ahora la hipótesis nula si será rechazada, como efectivamente lo es, ya que el p-valor asociado resulta 0.00548.

Con el fin de cerrar esta segunda parte de la actividad se prueban después varias hipótesis nulas, con valores de μ que fluctuan, por ejemplo, entre 24.3 y 28, o digamos, entre 18 y 21. Esto con el fin de asignar a μ, en la hipótesis nula, valores que se encuentren fuera del intervalo de confianza previamente obtenido. Al efectuar estas pruebas, con valores de μ en el intervalo (24.3,28), el p-valor más pequeño fue 5.903382e-07 y el más grande resultó 0.04451013.

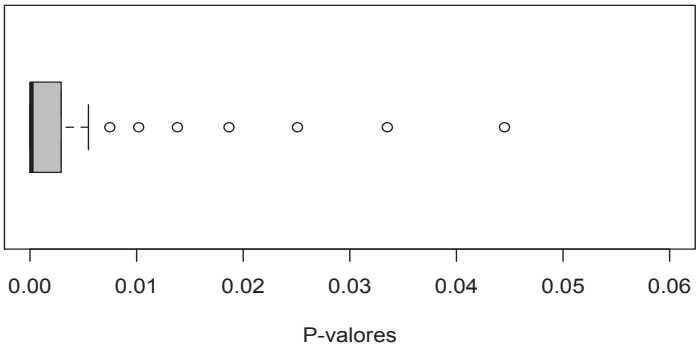


Figura 8. Diagrama de caja de p-valores resultantes de probar hipótesis nulas con valores de $\mu \in (24.3, 29)$

El diagrama de caja asociado a los p-valores resultantes se muestra en la Figura 8, donde puede observase que todos los p-valores son menores que 0.05, lo que conduciría a rechazar cualquier hipótesis nula H_0, siempre que $\mu \in (24.3, 29)$, pues μ toma valores fuera del intervalo de confianza obtenido.

Ahora, con base en la muestra de los IMC, y ya efecuadas las dos partes de esta segunda actividad, se cuestiona al estudiante lo siguiente, ¿Se rechaza $H_0: \mu = 24.3$, pero no se rechaza $H_0: \mu = 24.2$? Ellos observan que efectivamente eso ocurre cuando $\alpha = 0.05$. Esta situación debe aprovecharse para reflexionar en que la decisión de rechazo o no de una hipótesis nula, no debe tomarse a la ligera, esto es, debe haber una fundamentación relacionada con el propósito del estudio. El tomar decisiones de manera mecánica lleva a concluir que si $\alpha = 0.05$, y el p-valor=0.052, no se rechazará H_0, pero si el p-valor=0.048, se rechazará H_0. Aunque es un ejemplo algo extremo, permite ejemplificar que toda conclusión estadística no debe basarse solamente en la magnitud de un resultado. Esto es, la decisión debe también conjuntar información inherente al problema en estudio, pues con respecto al problema planteado, si en hipótesis nula se establece $\mu = 25$, decidir de manera tajante, como la mostrada anteriormente, el rechazo o no rechazo de esa hipótesis nula, equivaldría a admitir que si una de las estudiantes seleccionadas hubiera tenido quizá un kilo más de peso, entonces podremos concluir que la población de donde se tomó la muestra sufre de sobrepeso. ¿Tiene eso sentido? Así pues, este ejercicio tan sencillo permite reflexionar en que una decisión no debe fundamentarse solamente en un resultado numérico.

Conclusiones

Las actividades propuestas permiten reflexionar sobre aspectos que deben considerarse al seleccionar la técnica inferencial más apropiada para analizar un determinado conjunto de datos. Se ha observado que actividades como éstas facilitan el diferenciar conceptos como nivel de significancia y p-valor, así como el interpretar correctamente un intervalo de confianza. El abordar

problemas reales, extrapolando su análisis con la ayuda de simulaciones efectuadas por medio de algún software estadístico, permite que el estudiante identifique similitudes y diferencias entre las diferentes técnicas inferenciales analizadas y coadyuva a que su decisión no se base solamente en un resultado numérico.

Referencias bibliográficas

Abdelhamid, A. (2005). Why should researchers report the confidence interval in modern research. *Middle East Fertility Society Journal , 10* (1), 78-81.

Batanero, C. (2001). *Didáctica de la Estadística.* España: Departamento de Didáctica de la Matemática. Universidad de Granada.

Batanero, C., & Díaz, C. (2005). El papel de los proyectos de enseñanza y aprendizaje de la estadística. *I Congresso de Estatística e Investigação Operacional da Galiza e Norte de Portugal, VII Congreso Galego de Estatística e Investigación de Operacións.* Guimarães, Portugal.

Clark, M. (2004). Los valores P y los intervalos de confianza: ¿en qué confiar? *Revista Panamericana de Salud Pública , 15* (5), 293-296.

Garfield, J. B., & Ben-Zvi, D. (2008). Learning to Reason About Statistical Inference. En *Developing Students' Statistical Reasoning.* Springer.

Kochanski, G. (2005). *Confidence intervals and hypothesis testing*. Recuperado el 20 de febrero de 2012, de http://kochanski.org/gpk/teaching/0401Oxford/confidence.pdf

Wood, M. (2012). *P vlues, confidence intervals, or confidence levels for hypothesis?* Recuperado el 16 de enero de 2013, de http://arxiv.org/abs/0912.3878v4.