ALGUNOS FACTORES QUE PUEDEN AFECTAR AL P-VALOR

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

Palabras clave: nivel de significancia, prueba de hipótesis, prueba de significancia
Key words: significance level, hypothesis test, significance test

RESUMEN

Una medida de evidencia estadística que aparece comúnmente en los resultados inferenciales que despliega cualquier software estadístico es el p-valor. Su interpretación no siempre es clara para el usuario de este software y es muy común confundir el p-valor con el nivel de significancia de una prueba estadística, o bien considerar solamente su magnitud para emitir una conclusión estadística. Ambas situaciones pueden afectar dicha conclusión. Es por ello que surge la necesidad de implementar algunas actividades didácticas que permitan al estudiante mostrar las diferencias entre p-valor y nivel de significancia, y otras que muestren cómo al variar ciertos factores se puede afectar la magnitud del p-valor. La diferencia entre p-valor y nivel de significancia puede explicarse con experimentos sencillos, llevados a cabo en el salón de clase, y el mostrar cómo ciertos factores pueden afectar la magnitud del p-valor, se observa fácilmente mediante simulaciones que pueden ser efectuadas con diversos tipos de software.

ABSTRACT

The p-value is a measure of statistical evidence that commonly appears in the inferential output that provides most of the statistical software. Its interpretation is not always clear for statistical software users and it is very usual to confound the p-value with the significance level of a test, or to consider just the p-value magnitude to issue a statistical conclusion. Both aspects can affect a statistical decision. For that reason we consider necessary to carry out some didactic activities that allow the students to distinguish some basic differences between p-value and significance level, and some others that illustrate how some factors variations can affect the p-value magnitude. Some simple classroom experiments can be used to show the difference between p-value and level of significance, and software simulations are very helpful to illustrate how some factors variations affect the p-value magnitude.

Introducción

Una parte fundamental en la mayoría de las investigaciones que se realizan en diferentes disciplinas científicas lo constituye el análisis estadístico que las acompaña, y un concepto que aparece comúnmente en los resultados inferenciales que despliega cualquier software estadístico, que se use para tal fin, es una probabilidad conocida como p-valor, cuya interpretación no siempre es clara para muchos de quienes hacen uso de software estadístico. Es muy común observar una confusión entre el concepto de p-valor y el de nivel de significancia de una prueba estadística y ello puede llevar a conclusiones equivocadas y una toma de decisiones no adecuada. Comúnmente, el análisis de los resultados que involucran la interpretación de un p-valor se realiza de una manera mecánica y considerando sólo la magnitud del p-valor obtenido, sin reflexionar que tal valor puede ser influido por diversos factores, entre los que se encuentran cambios en la hipótesis nula, características de la muestra, como por ejemplo cambios en sus medidas descriptivas y tamaños de ésta, entre otros aspectos.

Por lo general, durante clase no se analiza cómo la variación de ciertos factores puede resultar en un p-valor muy pequeño (o muy grande) y con ello cambiar rotundamente una conclusión estadística, la cual resulta una cuestión muy delicada si uno se olvida de analizar los resultados obtenidos, dentro del contexto del problema. Es necesario entonces integrar, dentro de la enseñanza de la estadística, algunas estrategias que permitan contextualizar los contenidos, en situaciones que resulten interesantes para el alumno, como sugieren Batanero y Díaz (2005).

Para diferenciar los conceptos de p-valor y nivel de significancia se realiza una actividad durante, clase con la cual se pretende que el estudiante, de una manera sencilla, recuerde la diferencia esencial entre estos dos conceptos. Una vez logrado ello, se realiza una actividad que consiste en efectuar algunas corridas de simulación que permiten al estudiante observar cómo la variación de ciertos factores puede afectar la magnitud del p-valor, y por lo tanto la conclusión estadística, ya una vez fijado, previamente, el nivel de significancia de la prueba estadística.

Esta última actividad se puede realizar utilizando software diverso, como Matlab, R, Octave, etcétera, con el cual se simula una gran cantidad de muestras aleatorias de distribuciones específicas (aquí sólo se ilustra con la distribución normal) y examinando la distribución de los p-valores que resultan, considerando diferentes aspectos. Actividades de este tipo permiten a los estudiantes desarrollar habilidades que les facilitan resolver no sólo los problemas de aplicación planteados en clase, sino también el analizar cuidadosamente aquellos a los que pudieran enfrentarse en su vida profesional.

Referente teórico

Como referente teórico para el planteamiento de las actividades didácticas propuestas, se retomó de Batanero y Díaz (2005), el desarrollo del razonamiento estadístico, por considerarlo como una componente esencial en la enseñanza de la estadística, (Wild y Pfannkuch, 1999; citado por Batanero y Díaz, 2005), el cual incluye las siguientes componentes: 1) Reconocer la necesidad de los datos, esto es, que muchas de las situaciones de la vida real sólo pueden comprendidas a partir del análisis de los datos recolectados de manera correcta; 2) Transnumeración, o sea, la comprensión que puede surgir al cambiar la representación de los datos; 3) Percepción de la variabilidad, esto es, la recolección adecuada de los datos y los juicios correctos a partir de los mismos, requiere la comprensión de la variabilidad que

hay, y se transmite en los datos, así como la incertidumbre debida a la variabilidad no explicada; 4) Razonamiento con modelos estadísticos, éstos pueden ser cualquiera que resulte útil, por ejemplo un gráfico sencillo o un resumen, que pueda usarse para representar la realidad; aquí lo importante es diferenciar el modelo de los datos y a la vez relacionar el modelo con los datos; por último 5) Integración de la estadística con el contexto, la cual es fundamental para el razonamiento estadístico. Estas referentes se encuentran reflejadas en la descripción de las actividades didácticas que se presentan en el siguiente apartado.

El concepto de p-valor es una medida de la evidencia estadística que aparece en la mayoría de los trabajos de investigación y es simplemente *la probabilidad de observar la muestra obtenida, considerando que la hipótesis nula es cierta.* El p-valor es de hecho, una variable aleatoria que toma valores entre 0 y 1, y bajo ciertas condiciones puede distribuirse de manera uniforme. La utilidad del p-valor es indiscutible, tal como lo ejemplifica Lew (2013), donde señala que seis de los doce artículos de investigación e informes, de la edición del 14 de diciembre 2012 de la revista Science, y 20 de 22 artículos de la edición de diciembre de 2012 del Journal of Pharmacology and Experimental Therapeutics, utilizan los p-valores para describir sus resultados experimentales.

A pesar de ello, su interpretación no siempre es correcta y es muy común confundirlo con el nivel de significancia de una prueba estadística, es decir, confundirlo con *una probabilidad fija, acordada de antemano, de rechazar la hipótesis nula* H_0 *cuando ésta se supone cierta*. Nótese pues, que para establecer el nivel de significancia de una prueba no es necesario contar con la muestra observada.

Hubbard y Bayarri (2003) señalan que la mayoría de los usuarios de pruebas estadísticas, en las ciencias aplicadas, desconocen la distinción entre nivel de significancia y p-valor. Una de las razones de que al estudiante no le quede claro la diferencia entre éstos, proviene de no distinguir la diferencia entre los enfoques de prueba de hipótesis y pruebas de significancia. Ronald Fisher (1966) popularizó el uso del p-valor a través de las pruebas de significancia y éste sólo requiere el planteamiento de una hipótesis nula H_0.

Por su parte, Jerzy Neyman y Egon Pearson fueron quienes propusieron el enfoque de prueba de hipótesis, donde se plantea tanto una hipótesis nula H_0, como una hipótesis alternativa H_1; y equivocarse al rechazar H_0 cuando ésta es cierta, es lo que se fija de antemano y se conoce como nivel de significancia de la prueba estadística. Esto es, los conceptos de p-valor y nivel de significancia están asociados a los enfoques de pruebas de significancia y pruebas de hipótesis, respectivamente.

El no distinguir entre dichos conceptos origina lo que señalan Blocker y otros (2006), cuando explican que una de las preguntas más frecuentes dirigidas al comité de estadística, es la que se refiere al cálculo e interpretación de los valores de significación, esto es, “los valores p".

Un problema adicional a la confusión entre conceptos como p-valor y nivel de significancia de una prueba de hipótesis, es el comportamiento generalizado de basar una decisión estadística en tal sólo la magnitud del p-valor, sin tomar en cuenta los factores que pueden influir en su magnitud y el entorno

que rodea a la investigación. Lo usual es que si el p-valor es menor que el nivel de significancia establecido, se rechazará la hipótesis nula.

Mingfeng, Lucas y Galit (2013) sugieren que cuando el tamaño de la muestra es grande y un p-valor indica significación estadística, es recomendable ignorar este resultado, y en su lugar utilizar intervalos de confianza para el parámetro en estudio. Es importante considerar esta sugerencia pues por ejemplo, dos estudios con el mismo p-valor no proporcionan la misma evidencia acerca de la hipótesis nula, o bien, al comparar dos muestras, una diferencia no significativa no necesariamente implica que no existe diferencia entre los dos grupos.

■ Metodología

Las actividades didácticas que se proponen para abordar la problemática planteada son básicamente dos, pero ilustran de una manera sencilla cómo distinguir entre p-valor y nivel de significancia de una prueba estadística, así como el comprender que el p-valor es una variable aleatoria que no se puede fijar de antemano, a diferencia del nivel de significancia.

Por último, se retoma una problemática particular con la cual se muestra, mediante simulaciones, cómo el p-valor puede ser afectado al variar ciertos factores y con ello cambiar fácilmente una conclusión estadística, deduciendo con esto que es necesario considerar el contexto del problema antes de emitir una conclusión.

■ Primera Actividad didáctica

Una manera sencilla de distinguir entre p-valor y nivel de significancia consiste en efectuar un pequeño experimento en el grupo. Se solicita la ayuda de un estudiante para que trate de adivinar el resultado del lanzamiento de una moneda, antes de que ésta caiga. La moneda, proporcionada por el maestro o alguno de los estudiantes, será lanzada por el maestro, quien emitirá la palabra “correcta” cuando el estudiante adivine el resultado e “incorrecta” cuando no sea así.

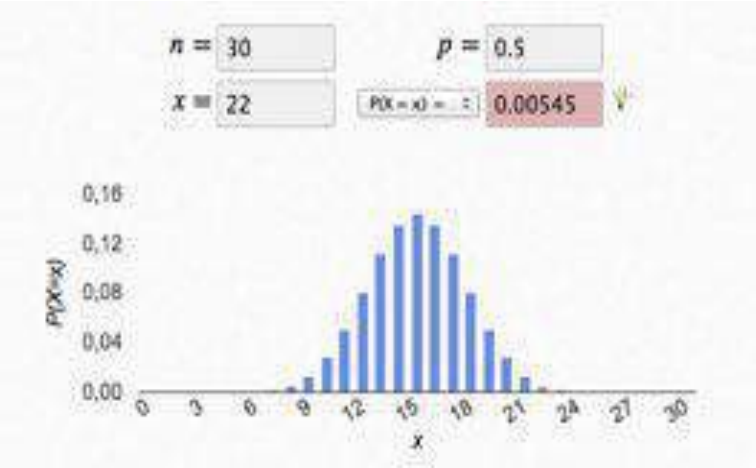
El propósito de este experimento es que el voluntario “atine” a un número inusual de lanzamientos. Se recomienda hacer algunos ensayos para establecer la forma en que se va a tirar la moneda, la altura desde donde se va a lanzar y el tipo de moneda a utilizar. En realidad, en este caso dichos aspectos no son muy relevantes, pero brindarán credibilidad al experimento. Se puede establecer, por ejemplo, que la moneda será lanzada 30 veces.

Se planteará la hipótesis nula $H_o: p = 0.5$, que equivaldría a afirmar que es igual de probable que el estudiante acierte o no al resultado del lanzamiento; así pues, ese será el supuesto. Como se mencionó anteriormente, el maestro fingirá que el estudiante obtiene más resultados correctos de lo que sería lo usual.

Por ejemplo, si la variable aleatoria X representa el número de aciertos en los treinta lanzamientos, los ensayos se suponen independientes y la probabilidad de ensayo a ensayo se mantiene constante, entonces $X \sim B(n = 30, p = 0.5)$, esto es, X se distribuye como una binomial con parámetros n y p. Esta

variable puede entonces tomar valores entre cero y treinta y la probabilidad de que tome cada uno de éstos se ilustra en la siguiente gráfica.

Figura 2. Posibles resultados al lanzar 30 veces una moneda legal.



Como puede observarse en la Figura 1, la probabilidad de que el estudiante acierte 22 veces es de 0.00545, una probabilidad pequeña. Esta probabilidad sigue siendo pequeña aún si calculamos la probabilidad de acertar 22 o más veces el resultado del lanzamiento.

Si el experimento se maneja de manera que algo como este resultado sea lo que ocurra, y durante el transcurso del mismo se mantiene una observación constante sobre los estudiantes del grupo, con el fin de detectar el momento a partir del cual algunos manifiesten, como generalmente ocurre, que el estudiante debe tener "poderes" para adivinar tal cantidad de resultados, entonces el experimento resulta exitoso ya que permitirá hacer la distinción entre p-valor y nivel de significancia.

El p-valor será simplemente la probabilidad de observar lo ocurrido, si el supuesto establecido, $p = 0.5$, es cierto. Esto es, $p\_valor = P(X \geq 22)$. ¿Dónde queda entonces el nivel de significancia α ?

Para aclarar este punto basta con recordar a partir de cuántos aciertos del estudiante voluntario, mostraron incredulidad los estudiantes espectadores. Por ejemplo, si 20 aciertos del estudiante voluntario causó asombro en algún estudiante, entonces la probabilidad de que la variable tome valores a partir de 20 sería el equivalente a su probabilidad fija, acordada de antemano, de rechazar la hipótesis nula H_0, cuando ésta es cierta. Esto es, esa probabilidad sería su nivel de significancia particular.

Como por lo general varios estudiantes expresan asombro con diferente número de aciertos del voluntario, queda claro que el nivel de significancia se fija de antemano y éste puede variar de persona a persona, mientras que el p-valor se podrá calcular hasta que se cuente con los datos observados, y una vez observada la muestra éste será único.

Este sencillo experimento del cual pueden hacerse variaciones utilizando dados, barajas, etcétera, permite diferenciar estos dos conceptos, que por lo general causan confusión en el estudiante.

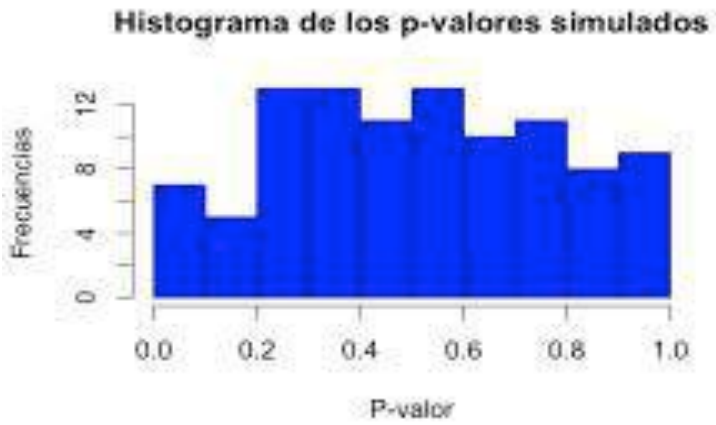
Segunda Actividad didáctica

Una vez efectuada la actividad anterior, se realiza una segunda actividad que consiste en tomar un ejemplo sencillo de cualquier libro o artículo, y por medio de simulación de varias muestras, se ilustrar el comportamiento del p-valor resultante bajo diferentes características de la muestra, como pueden ser cambios en los valores resultantes de la media muestral (se varía la media de la distribución de donde se simulan los datos) y variaciones en los tamaños de muestra. Un ejemplo de problema a considerar podría ser el siguiente.

Un fabricante afirma que el peso de las bolsas de dulces que vende es una variable aleatoria que sigue una distribución normal con una media de 5 libras y una desviación estándar de 0.05 libras. Supongamos que diariamente se toma una muestra de 20 bolsas para verificar el proceso. El simular 100 muestras de tamaño 20 de una distribución normal con dichos parámetros y realizar una prueba de significancia, en la cual se plantee $H_0: \mu = 5$, equivaldría a suponer que durante cien días se estuvo verificando la media de un proceso que se está bajo control.

Para cada una de estas muestras se puede calcular el p-valor resultante y graficar con estos un histograma. A continuación se muestra un histograma resultado de la simulación anteriormente explicada, la cual se efectuó en el software R.

Figura 2. P-valores de muestras de tamaño 20 de un proceso bajo control.



Dado que se está simulando de una distribución con media $\mu = 5$ y se está probando la hipótesis nula $H_0: \mu = 5$, podemos ver que en la mayoría de las veces el p-valor será mayor de 0.05, como se muestra

en la Figura 2. Ello significa que sólo en muy contadas ocasiones se concluirá que el proceso está fuera de control, cuando el proceso en realidad está bien.

Si ahora simulamos de un proceso que está fuera de control, que podría ser el caso de simular cien muestras aleatorias de tamaño 20 de un proceso que llena bolsas de dulces, siguiendo una distribución normal, con media 5.05 libras y una desviación estándar de 0.05 libras, y nuevamente probamos la hipótesis nula $H_0: \mu = 5$, tendremos un histograma de p-valores como el mostrado en la Figura 3, que corresponde a los p-valores obtenidos al suponer que un proceso está controlado, cuando en realidad no es así. Estos resultados podemos compararlos con los obtenidos en la Figura 4, donde sólo se varió el tamaño de muestra a 10; esto es, se simularon cien muestras de tamaño 10, del mismo proceso que se encuentra fuera de control.

Figura 3. P-valores de muestras de tamaño 20 de un proceso fuera de control.

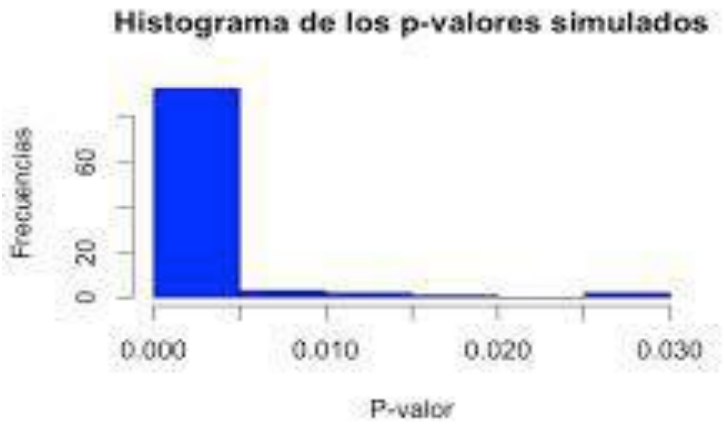


Figura 4. P-valores de muestras de tamaño 10 de un proceso fuera de control

Es muy interesante para el alumno el observar que el cambio en el tamaño de muestra, de 10 a 20 bolsas, implica cambios notables en el p-valor y por lo tanto en la conclusión estadística; esto es, resulta más fácil distinguir que el proceso está fuera de control cuando se toman muestras de tamaño 20, en lugar de las muestras de tamaño 10. Si los escenarios considerados se analizan desde el contexto de pruebas de hipótesis, debe enfatizarse que el nivel de significancia se mantendrá fijo, y planteado de antemano; puede acordarse, por ejemplo, trabajar con α=0.05.

Con base en el análisis de los resultados obtenidos en estas simulaciones, los estudiantes deben responder una serie de preguntas como las siguientes:

- ¿Se puede calcular el p-valor asociado a la hipótesis nula, antes de tener la muestra?
- ¿Cómo es la forma de los histogramas construidos con los p-valores generados aleatoriamente, en los diferentes escenarios?
- ¿Se debe analizar con el mismo cuidado un p-valor pequeño y un p-valor grande?
- ¿Cómo cambia el p-valor al variar el tamaño de muestra?
- ¿Cómo cambiaría el p-valor si se modifica lo establecido en la hipótesis nula?

Después de planteadas estas preguntas se sugiere promover una discusión grupal con la finalidad de que los estudiantes observen que: los p-valores pudieron calcularse hasta contar con la muestra observada; la magnitud del p-valor resultante puede verse afectada por las características de la muestra simulada, como es el muestrear de una población con parámetros diferentes a lo planteado en la hipótesis nula o bien variar el tamaño de muestra; ya que ambos influyen, en gran medida, en el p-valor resultante.

En este último punto se pueden retomar algunos aspectos señalados por Mingfeng, Lucas, y Galit (2013), con respecto a ser más cuidadosos al concluir una significancia estadística cuando se analizan estudios o experimentos con tamaños de muestra muy grande y consideramos también muy conveniente la sugerencia de verificar la robusticidad del p-valor, trabajando con múltiples submuestras.

Conclusiones

Las actividades anteriormente descritas han sido implementadas en varios cursos de estadística y se ha observado que la estrategia de trabajo resulta motivante para el estudiante, participando de manera activa en el desarrollo de las mismas, lo cual cambia radicalmente la forma tradicional de enseñanza donde el estudiante suele ser sólo espectador del discurso del profesor. Por otro lado, el análisis que se realiza después de éstas permite al estudiante diferenciar entre p-valor y nivel de significancia, así como interpretar el p-valor de manera más adecuada, lo cual consideramos como un gran logro dadas las dificultades y errores, ya documentadas, que se suelen cometer en torno a esos dos conceptos. Muy particularmente, el uso de la simulación para modelar mediante histogramas resultó de gran utilidad para el logro de los objetivos de las actividades didácticas.

Referencias bibliográficas

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.

Blocker, C., Conway, J., Demortier, L., Heinrich, J., Junk, T., Lyons, L., Punzi, G. (2006). *Simple Facts about p-values.* Recuperado de http://physics.rockefeller.edu/luc/technical_reports/cdf8023_facts_about_p_values.pdf

Fisher, R. (1966). *The design of experiments* (8 ed ed.). Edinburgh: Oliver and Boyd.

Hubbard , R., Bayarri , M. J. (2003). Confusion over measures of evidence $(p's)$ versus errors $(\alpha's)$ in classical statistical testing. *The American Statistician, 57*, 171-178.

Lew, M. J. (2013). *To P or not to P: on the evidential nature of P-values and their place in scientific inference.* Recuperado de Cornell University Library: http://arxiv.org/abs/1311.0081

Mingfeng, L., Lucas, H. C., Galit , S. (2013). *Information system research.* Obtenido de http://dx.doi.org/10.1287/isre.2013.0480.