CARACTERIZACIÓN ALEATORIA DEL VALOR P Y SUS IMPLICANCIAS PARA LA INFERENCIA

Agnelli, Héctor

Universidad Nacional de Río Cuarto

hagnelli@exa.unrc.edu.ar

Categoría del trabajo: Reflexiones

Nivel Educativo: Universitario

Palabras clave: test, significación, valor-p, variabilidad

Resumen

Es una práctica rutinaria entre los usuarios de la estadística utilizar el valor p de un test como parte de un procedimiento de decisión para rechazar o no la hipótesis nula, dado que su cálculo es efectuado por algún software estadístico su aplicación está casi automatizada. Sin embargo la ausencia de reflexión acerca del significado del mismo lleva a asignarle interpretaciones incorrectas tales como: creer que el valor p entrega la probabilidad de que la hipótesis nula sea verdadera, que valores pequeños del mismo implican un efecto de tratamiento de gran magnitud o que es una evidencia a favor de la replicación de los resultados. Considerando que en la etapa de su enseñanza presentaciones alternativas del tema pueden contribuir a disipar interpretaciones erróneas como las señaladas, en el presente trabajo se hace una caracterización de la naturaleza aleatoria del valor p: como una probabilidad, como una variable aleatoria y como un estadístico. También se presentan simulaciones que procuran aclarar el verdadero alcance del valor p.

Introducción

El test de hipótesis es uno de los de los procedimientos de la inferencia estadística más utilizado por los investigadores de las ciencias naturales y sociales viéndose facilitada su aplicación por la disponibilidad de software estadístico. Sin embargo, mientras que el cálculo estadístico se halla al alcance de la mayor parte de los usuarios, no puede decirse lo mismo de lo que autores como Hawkins (1997); Rubin (1989); Ben-Zvi y Garfield (2004); Pfannkuch y Wild (2004), denominan razonamiento estadístico. Para estos autores, razonar

estadísticamente significa ser capaz de comprender, explicar procesos estadísticos y de interpretar de manera global los resultados. La significación estadística derivada de la aplicación de pruebas de hipótesis origina serios problemas de interpretación entre los investigadores usuarios de la estadística. Algunos autores (Gigerenzer, 2004; Hubbard y Bayari 2003) argumentan que la razón principal que sustenta esta problemática radica en que muchos textos de metodología estadística presentan confusiones sobre el real alcance de este concepto. No basta con aprender métodos y aplicarlos como recetas, puesto que algunas de ellas detrás de su aparente simpleza esconden una potencial fuente de interpretaciones erróneas. Un ejemplo de esta situación lo constituye el valor p: "cuando es pequeño ($p<0.05$) se rechaza la hipótesis nula, en caso contrario ésta no es rechazada". El valor p desde hace décadas ha merecido consideraciones de distintos autores Berkson (1942), Gibbons y Pratt (1975), Berger y Sellke (1987), Schervish (1996), Hung, O'Neill, Bauer y Kohnen (1997), Sackrowitz y Cahn (1999), Hubbard y Bayarri (2003), Murdoch, Tay y Duncan (2008), Boss y Stefanski (2011), sin embargo en los libros de texto poca alusión se hace a su verdadera naturaleza y problemática. Con este trabajo se intenta contribuir a partir de la caracterización del valor p como una probabilidad, una variable aleatoria y un estadístico a generar procesos que favorezcan su enseñanza y disipen las interpretaciones erróneas que son más frecuentes.

Origen del valor p

El origen del valor p se puede remontar al siglo diecinueve cuando los test se realizaban de manera bastante informal pero fue R.A. Fisher quien en su libro Statistical Methods for Research Workers (1925) creo un nuevo paradigma para los test de hipótesis (Lehmann,1993) . Para Fisher los llamados test de significación, sirven para confrontar una hipótesis postulada con los datos observados. La idea básica para conducir un test de significación está basada en la elección un estadístico apropiado que resuma la información muestral y la determinación de su distribución bajo la suposición de que la hipótesis nula sea verdadera. Realizado el experimento se calculan, a partir de los datos, el valor observado del estadístico y el valor p. Si éste valor es suficientemente pequeño se rechaza la hipótesis. En caso contrario no hay conclusión. La disyuntiva esencial que se plantea al realizar un test de significación es: o la hipótesis no es verdadera o ha ocurrido un resultado excepcionalmente raro: una medida de evidencia en contra de la hipótesis está dada por el valor p. En su libro Fisher no tabuló distribuciones completas sino algunos cuantiles en particular los correspondientes al 1% y 5% y de aquí nacen las cotas habituales que se le piden al valor p para considerarlo pequeño. En este contexto el valor del 1 % ó el 5% es arbitrario y no deducible de ninguna teoría.

A principios de 1930 J. Neyman y E. Pearson desarrollaron otro paradigma para los test, el llamado test de hipótesis en el que aparecen dos hipótesis; la nula (Ho) y la alternativa (Ha), los errores tipo I y Tipo II y sus respectivas probabilidades. En particular la probabilidad de cometer el error de tipo I (rechazar Ho siendo verdadera) se llama el nivel de significación del test. Esta probabilidad habitualmente designada con α indica la tasa de error a la larga que caracteriza al test. Fijado el nivel de significación se determinan las zonas de rechazo y aceptación y según a cual de estas zonas pertenezca el valor observado del estadístico del test se rechaza o no la Ho. Si un test es de nivel α significa que el α% de las veces podríamos rechazar la Ho siendo esta verdadera. A diferencia del valor p, que es dato dependiente y por lo tanto un concepto post-experimental, α es un valor fijo: no depende de los datos y es por lo tanto un concepto pre-experimental. Otra elemento básico que introdujo la metodología de Neyman-Pearson fue la potencia del test. La potencia, en términos de probabilidad, es la capacidad que tiene un test para rechazar correctamente la hipótesis nula. Estableciendo la potencia deseada se puede, conociendo la hipótesis alternativa, determinar el tamaño de muestra necesario.

En la actualidad se utiliza un híbrido entre los dos paradigmas y los software entregan además del valor observado del estadístico del test el valor p. Habitualmente se compara este valor p con un valor de probabilidad como 1% ó %5 y se rechaza la Ho en caso de ser p menor a estos valores. Usar el p de esta manera es tomar una decisión en el sentido de Neyman-Pearson pero si además se le concede importancia al valor numérico del p para decir cuan fuerte o débil es la evidencia en contra de Ho, se esta usando al valor p en el sentido Fisheriano como una medida de evidencia en contra de Ho.

Creencias erróneas acerca de p

La metodología de tests de hipótesis ha recibido fuertes críticas desde distintos ámbitos académicos profesionales (asociaciones científicas, editores de revistas) acerca de los errores de aplicación e interpretación en los que incurren los científicos experimentales en tanto usuarios de la estadística (Kline, 2004).. También han sido reportados resultados similares en trabajo realizados con estudiantes universitarios (Vallecillos, 2002, Rodriguez y Albert, 2007). En la práctica el principal problema que se presenta con el valor p lo constituye el asignarle por parte de los usuarios estadísticos interpretaciones que van más halla de su real significado. Algunas de estas creencias erróneas son las siguientes:

- *El valor p es la probabilidad de que la hipótesis nula sea verdadera*

- *Pequeños valores de p significan un efecto de tratamiento de gran magnitud*
- *El valor p es una evidencia a favor de la replicación de los resultados*

A continuación y a partir de distintas caracterizaciones aleatorias del valor p trataremos de aportar elementos que ayuden a superar estos errores

El valor p cómo probabilidad

El valor p es la probabilidad de observar un valor tan o más extremo que el observado, siendo cierta la hipótesis nula. En símbolos: si T es el estadístico del test y t su valor observado para el experimento analizado, entonces el valor p es

$$p = \Pr(T \geq t / Ho) \qquad (1)$$

Dado que el valor p se utiliza para realizar afirmaciones (en sentido probabilístico) acerca de la hipótesis nula, lo expresado en (1), se puede escribir como

$$p = \Pr(\text{Re}ch.Ho / Ho\ vale) \quad (2)$$

Como el evento anterior depende de los datos también se puede escribir

$$p = \Pr(D / Ho) \qquad (3)$$

Expresión que nos muestra que el valor p no es la probabilidad (en distintos grados) de la validez hipótesis nula, sino que es la probabilidad de los datos en caso de ser cierta esa hipótesis nula. La interpretación errónea acerca de que *p es la probabilidad de Ho* surge al invertir equivocadamente el evento condicionante es decir tomar $p = \Pr(Ho / D)$ en (3) y de aquí solo hay sólo un paso a pensar *p = P(Ho).*

En rigor si se está interesado en la

$$\Pr(Ho / D)$$

este problema se resuelve, disponiendo de cierta información previa, aplicando la regla de Bayes

$$\Pr(Ho / D) = \frac{\Pr(D / Ho).\Pr(Ho)}{\Pr(D / Ho).\Pr(Ho) + \Pr(D / H_1).\Pr(H_1)}$$

Ahora se hace necesario contemplar la existencia de la hipótesis alternativa H_1 (aquí como complemento de Ho), y disponer de información acerca de Ho y H_1 antes de realizar el experimento. A manera de ejemplo consideremos que a priori las probabilidades de Ho y H_1 son $\Pr(Ho) = \Pr(H_1) = 1/2$, se tendrá

$$\Pr(Ho / D) = \frac{p}{p + \Pr(D / H_1)}$$

y salvo aquellos caso en que $p + \Pr(D / H_1) \geq 1$ será $\Pr(Ho / D) > p$.

Luego si resultó $p=.05$ esto en general implica que $Pr(Ho/D) > 0.05$ (ver Fig.1) y únicamente en la particular situación cuando $p + \Pr(D / H_1) = 1$ se tendrá que $\Pr(Ho / D) = p$.

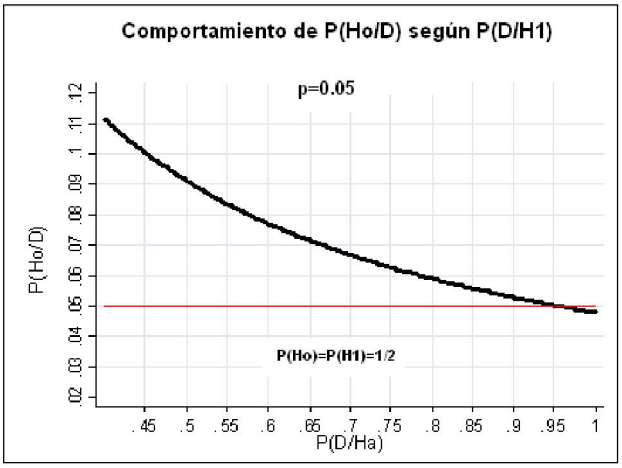

Fig.1

El valor p como variable aleatoria

Teniendo en cuenta que $p = \Pr(T \geq t / Ho) \Rightarrow P = h(T(X_1,...,X_n)$, entonces P es una variable aleatoria que asume para una muestra dada $x_1,...,x_n$ el valor $p = h(T(x_1,...,x_n)$. Por lo tanto podemos pensar que distribución tiene la variable aleatoria P. Sea T el estadístico del test y Fo su distribución bajo Ho, el valor p es

$$p = \Pr(T \geq t / Ho) = 1 - F_o(t)$$

Cuando *Fo* es continua y asumiendo que vale Ho, por la denominada *transformación integral de probabilidades* (Parzen. 1960): $Fo(T) \sim U(0,1)$, y también es cierto que

$$P = 1 - Fo(T) \sim U(0,1)$$

En síntesis, cuando vale la hipótesis nula el valor p es una variable aleatoria con distribución uniforme el intervalo (0,1). Por lo tanto es imposible que el valor p pueda ser considerado como la *P(Ho)* porque si bien este es un valor desconocido es fijo.

Para ejemplificar con una simulación, si $X \sim N(\mu, 1)$ y se plantea el test

Ho: $\mu = 0$ vs. H_1: $\mu > 0$, utilizando como estadístico del test a $T = \sqrt{n}\,\dfrac{\overline{X} - \mu_0}{\sigma} = \sqrt{n}\ \overline{X}$, y simulando 5000 realizaciones del test, el histograma de la Fig2. exhibe la distribución de los valores p. Cabe señalar que la distribución uniforme es independiente del tamaño de la muestra.

Ahora podemos suponer que vale la hipótesis la alternativa y estudiar el comportamiento de la variable aleatoria p. En este caso la distribución de p es asimétrica (Hung, O`Neill, Bauer, ,1997) y es función tanto del tamaño de muestra como del verdadero valor del parámetro en la hipótesis alternativa. A manera de ejemplo se muestran dos simulaciones en la Fig3. Estas fueron construidas simulando 5000 valores de una distribución normal con media 40.5 y desvio estandar 1 y postulando como hipótesis nula una media igual a 40. En el caso se tomaron muestras de tamaño 30 y en el caso b muestras de tamaño 5. Cabe señalar que estos gráficos ponen de manifiesto la importancia de la potencia: a mayor tamaño de muestra los valores p están más concentrados cerca de valores pequeños en cambio cuando disminuye el tamaño de muestra aumenta la variabilidad de p. En el segundo caso se observa que es alta la frecuencia relativa de valores p mayores que 0.05 y esto significa que existen muchas replicaciones simuladas del experimento en las que no se rechazará la Ho a pesar de haber construido las mismas suponiendo que Ho es falsa.

Volviendo al caso Ho: $\mu = 0$ vs. H_1: $\mu > 0$, se observa que el estadístico del test es función de la discrepancia entre la media muestral (los datos) y la media postulada multiplicado por una función creciente del tamaño de muestra:

$$T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{\sigma} = \sqrt{n}\ \ \overline{X},$$

esto implica que aún para una misma discrepancia aumentando el n el valor p se puede hacer muy pequeño conduciendo al rechazo de la hipótesis nula, sin que esto implique que *el efecto sea de una magnitud importante.* Recíprocamente si la muestra es pequeña aún siendo importante la magnitud del efecto el valor p puede no ser tan pequeño como para rechazar la hipótesis nula.

Se suele creer que *el valor p es una evidencia a favor de la replicación de los resultados* en un nuevo experimento llevado a cabo en similares condiciones experimentales El análisis de las simulaciones anteriores también nos muestra que un valor pequeño de p no garantiza que los resultados obtenidos se repitan. Para que exista reproducibilidad se necesita alta potencia en el nuevo estudio.

El valor p como estadístico

En general es una práctica habitual en la estadística cuando se reporta un estadístico hacerlo asociándole una medida de variabilidad que permita calificar su precisión. Ejemplo de esta situación lo constituye la media muestral siempre que se informa la media, con propósitos inferenciales, se informa su error estandar. En otros términos: es habitual considerar la

distribución muestral del estadístico. El valor p en tanto función aleatoria de los datos también es un estadístico y en consecuencia sería oportuno también analizar su distribución muestral para estudiar su variabilidad. Antes lo hemos hecho con simulaciones pero asumiendo conocida Ho en un caso y desconocida en otro. En la práctica, obviamente, no sabemos si Ho es o no cierta. Una manera de analizar la variabilidad en este caso es mediante la aplicación de bootstrap a los valores originales del experimento. De esta manera se podría estimar la varianza muestral y construir intervalos de confianza para p (Boos y Stefanski, 2011). Aunque la aplicación de esta técnica estadística puede no ser habitual en los cursos iniciales de estadística, sin embargo la consideración de su existencia y el planteo del valor p como estadístico puede ayudar a reforzar la idea acerca de la variabilidad del valor p.

Consideraciones finales

La estadística inferencial es una herramienta para el manejo matemático de la incertidumbre, y por su naturaleza sus afirmaciones serán siempre enunciados probabilísticos, ya que la variabilidad no puede ser suprimida. La práctica estadística tiende a automatizar la aplicación de sus procedimientos y le confiere a los mismos un rol determinístico que éstos realmente no poseen. Es en la etapa de la enseñanza donde es conveniente enfatizar los conceptos por sobre las aplicaciones rutinarias para que el futuro usuario (hoy nuestro alumno) pueda tener cabal comprensión del significado y alcance de las herramientas inferenciales que llegue a usar.

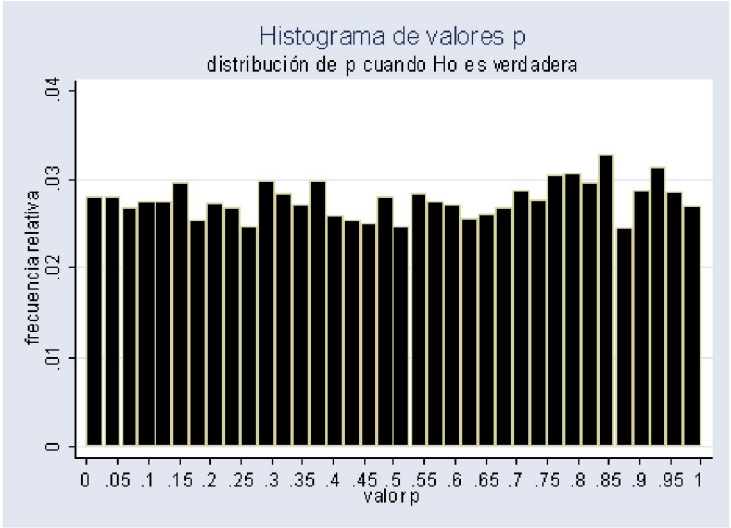

Fig.2

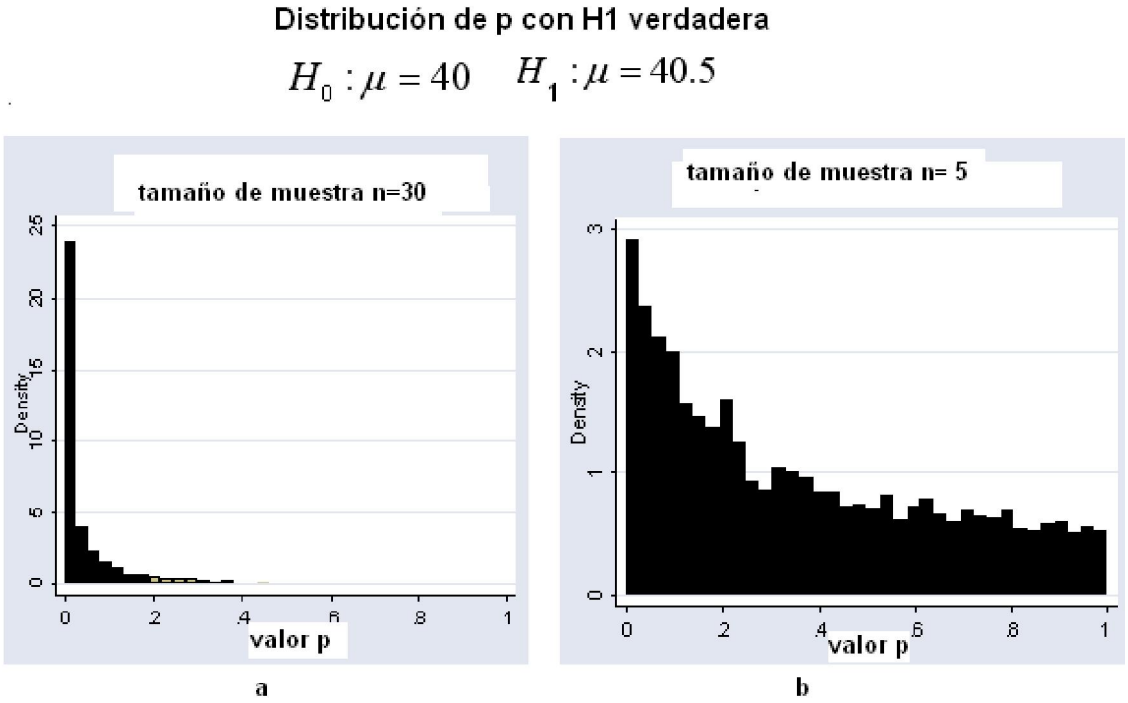

Fig.3

Bibliografía

Ben-Zvi, D. y Garfield, J. (2004). Statistical Literacy, Reasoning and Thinking: goals, definitions and challenges. En: D. Ben-Zvi y J. Garfield (eds.), *The challenge of developing statistical literacy, reasoning and thinking,* (pp. 3-15). Dordrecht: Kluwer Academic Publishers

Berger, J. O., & Sellke, T.(1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. *JASA, 82,* 112-122.

Berkson,J. (1942).Tests of Significance Considered as Evidence. *JASA*, Vol. 37, Nº. 219. pp. 325-335.

Boss, D. ; Stefanski, L.(2011). P-value precision and reproducibility. *The American Statistician*, 29, 20–25.

Gibbons, J. D., and Pratt, J. W.(1975), "P-values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587- 606

Hawkins, A. (1997). How far have we come? Do we know where we are going? En E. M. Tiit (Ed.), *Computational statistics & statistical education* (pp. 100-122). Tartu: International Association for Statistical Education e International Association for Statistical Computing

Hubbard, R., and Bayarri, M.J.(2003) "Confusion Over Measures of Evidence (*p*'s) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–182

Hung, H.M. J., O'Neill,R.T., Bauer, P., and Kohne, K.(1997). "The Behavior of the p Value When the Alternative Hypothesis is True," *Biometrics*, 53, 11–22.

Kline, R. (2004). *Beyond significance test*.American Psychological Association.Washington

Lehmann, E.(1993). The Fisher, neyman-pearson theory of testing hypotheses: one theory or two?. *JASA,* Vol.88, Nº424, pp1242-1249.

Murdoch, D; TayI, Y.; Duncan J. (2008), *P*-Values are Random Variables. *The American Statistician,* 62, 242- 245.

Parzen, E. (1971). *Teoría Moderna de probabilidades*. Limusa.

Pfannkuch, M. y Wild, C. (2004). Towards an understanding of statistical thinking. En: D. Ben-Zvi y J. Garfield (eds.) *The challenge of developing statistical literacy, reasoning and thinking, (*pp. 17–45). Dordrecht: Kluwer Academic Publishers.

Rodríguez M.I., Albert A. (2007). Prueba de hipótesis estadística:estudio de dificultades conceptuales en estudiantes de grado y de postgrado. *Memoria de la XI Escuela de Invierno de Educación Matemática. Red de Centros de Investigación en Matemática Educativa. 328-343.* ISBN: 978-970-9971-14-9.

Rubin, A (1989) Reasoning under uncertainty: Developing statistical reasoning. *Journal of Mathematical Behavior.* Vol. 8, 205-219

Sackrowitz, H., and Samuel-Cahn, E.(1999) "P Values as Random Variables- Expected p Values," *The American Statistician*, 53, 326–331.

Schervish, M.(1996) , "P Values: What They Are and What They Are Not," *The American Statistician,* 50, 203–206.

Vallecillos, A. (2002). Empirical Evidence About Understanding of the Level of significance Level Concept in Hypotheses Testing. *THEMES in Education,*3(2), 183-198.