*TELOS. Revista de Estudios Interdisciplinarios en Ciencias Sociales*
*UNIVERSIDAD Rafael Belloso Chacín*
*ISSN 1317-0570 ~ Depósito legal pp: 199702ZU31*
*Vol. 14 (3): 439 - 446, 2012*

# La revolución en la toma de decisiones estadísticas: el p-valor

*Nelson Romero Suárez**

## Introducción

El concepto de p-valor como resultado de un tratamiento estadístico, es un concepto debido a Fisher, es una contraposición del concepto de tamaño á del test en la teoría de Neyman-Pearson. La idea es conseguir eliminar la posibilidad de que dos investigadores que informen del resultado de un test estadístico, si utilizan dos tamaños diferentes lleguen a resultados distintos con la misma evidencia estadística, lo que no puede ocurrir con el p-valor. Se pretende seguir el nacimiento del concepto en el trabajo de Jacob Bernoulli y ver su evolución hasta el concepto tal y como es utilizado en nuestros días.

## Breve reseña histórica

De acuerdo con Gómez (2011), Laplace hace una contribución hacia el concepto del p-valor, que en su memoria de 1823 utiliza el método de los mínimos cuadrados, que él había llamado el "método más ventajoso" para estudiar el efecto de la luna en las mareas terrestres. Sigue Gómez, en esta memoria contrasta la hipótesis de que los cambios barométricos no son influidos por las fases de la luna y compara los cambios en la media estimada en cada una de dos series de 792 días; una sujeta a la atracción lunar, con otra del mismo tamaño, cuando ésta atracción no es tan pronunciada. En terminología moderna, Laplace establece que la diferencia observada entre las medias será significativa al nivel 0,01 si se hubieran estimado las medias a partir de datos de 72 años, es decir, Laplace se anticipa determinando no solo el p-valor sino también éste en función del tamaño muestral.

Considerando lo arriba descrito, Gómez (2011) indica que, la conclusión de Laplace es correcta, en el sentido de que al haber escogido París, donde la marea lunar existe pero tiene un valor muy pequeño, no fue, sino hasta 1945 cuando

* Doctor en Ciencias de la Educación. MSc en Educación Mención Matemática. Licenciado en Educación Mención Matemática y Física. Profesor Asociado del Departamento de Matemáticas de la Facultad Experimental de Ciencias, LUZ. Maracaibo, Venezuela. Correo electrónico: nromero1512@gmail.com

se pudo determinar correctamente; es decir, que Laplace se anticipa en 122 años a la resolución del problema. También debemos citar la contribución de Poisson, en relación a la estimación de la probabilidad de que un jurado de un veredicto correcto. Esto lo hace Poisson en 1837, donde aproxima la distribución binomial por la normal y calcula el p-valor correspondiente a la aproximación realizada.

Otro aporte importante surge por parte de Pearson, sobre la familia de distribuciones asimétricas que significó un paso más, en la dirección señalada, sobre la distribución simétrica de Laplace. Pearson suponía que este sistema de curvas podría describir cualquier tipo de colección de números. Cada distribución de esta familia se identifica con cuatro números: media, desviación estándar, asimetría y kurtosis. No obstante las críticas realizadas por Fisher (muchos de los métodos eran menos que óptimos) y Neyman (no cubría el universo de las posibles distribuciones), el sistema de curvas de Pearson sigue vigente en nuestros días (Gómez, 2011).

Por otro lado, Pearson desarrolló una herramienta estadística básica: la *prueba chi cuadrado de bondad de ajuste*. Esta prueba permite determinar si un conjunto de observaciones responde a cierta función matemática de distribución. Demostró que la distribución de la prueba es la misma cualquiera sea el tipo de datos usados. Esto significa que pudo tabular la distribución de probabilidad de este estadístico y usar el mismo conjunto de tablas para cada una de las pruebas. En un trabajo Fisher en 1922, demostró que en el caso de comparación de dos proporciones el valor del parámetro de Pearson era errado. Este error no invalida la importancia de esta prueba utilizada hasta nuestros días (Gómez, 2011).

La prueba de bondad de ajuste de Pearson fue el disparador de la componente principal del análisis estadístico moderno: el contraste de hipótesis o prueba de significación (Salsburg, 2001).

Hacia fines de la década de los 20 y principios de la de los 30 Egon Pearson (1895-1980), hijo de Karl, y Jerzy Neyman (1894-1980) afirmaron que las pruebas de significación no tendrían sentido si no hubiera, al menos, dos hipótesis posibles que llamaron: hipótesis nula (la de Fisher) y a la otra, hipótesis alternativa. Esto es la conocida teoría de pruebas de hipótesis (hypothesis testing) de Neyman-Pearson (Gómez, 2011).

## El P-Valor

En las pruebas de significación y diseño de experimentos Fisher utilizó el p-valor (p-value) que es la probabilidad que permite declarar la significación de una prueba. El término significación en los primeros desarrollos de esta idea se usaba para indicar que la probabilidad es suficientemente pequeña como para rechazar la hipótesis planteada. Este es el concepto que aún perdura, Fisher no tenía dudas acerca su importancia y utilidad del p-valor (Salsburg, 2001).

Gran parte de su *Statistical Methods for Research Workers* (Fisher, 1925) está dedicado a mostrar cómo se calcula el p-valor. En el libro Fisher no describe de donde derivan estos test y nunca dice exactamente qué p-valor puede considerarse

significativo. En su lugar presenta ejemplos de cálculos y notas si el resultado es o no significativo. En un ejemplo que muestra el p-valor menor que 0.01 dice: Sólo un valor en cien excederá (el test estadístico calculado) por casualidad, entonces la diferencia entre los resultados es claramente significativa. Para Fisher un test de significación tiene sentido sólo en el contexto de una secuencia de experimentos referidos a un tratamiento específico. De la lectura de los trabajos de aplicación de Fisher se puede deducir que usó los test de significación para una de tres posibles conclusiones:

- Si el p-valor es muy pequeño (usualmente menos de 0.01) declara que un efecto ha sido demostrado.
- Si el p-valor es grande (usualmente mayor que 0.20) el declara que si hay un efecto es tan pequeño que ningún experimento de ese tamaño es capaz de detectarlo.
- Si el p-valor está entre esos dos valores discute como diseñar un nuevo experimento para tener una idea mejor del efecto.

Recordemos que para Fisher la hipótesis a contrastar es que no existe diferencia entre los tratamientos. Para distinguir entre la hipótesis usada por Fisher para calcular el p-valor y otras posibles hipótesis Neyman y Pearson llamaron *hipótesis nula* a la hipótesis a contrastar y a la otra, *hipótesis alternativa*. En esta formulación, el p-valor es calculado para contrastar la hipótesis nula pero la potencia de la prueba se refiere a como este p-valor funcionará si la alternativa es, en los hechos, verdadera. La potencia de la prueba es una medida de cuan buena es la prueba. Dadas dos pruebas la de mayor potencia sería la mejor a usar (Salsburg, 2001).

De modo muy sintético recordemos que la Teoría de Neyman-Pearson, cuya estructura matemática es aceptada hasta nuestros días, establece, dos hipótesis posibles: la nula y la alternativa. De acuerdo a ciertos autores, existen dos fuentes de error: rechazar la hipótesis nula cuando es verdadera (nivel de significación, $\alpha_o$ error de tipo I) y no rechazarla cuando es falsa ($\beta_o$ error de tipo II). Sus contrapartidas, en sentido probabilístico, son las decisiones correctas de no rechazar una hipótesis cuando es verdadera ($1-\alpha$) y rechazarla cuando es falsa ($1-\beta$), esto último es la potencia de la prueba.

Lo ideal sería que minimizáramos ambos tipos de errores. Pero, por desgracia, para cualquier tamaño muestral, no es posible minimizar ambos errores de manera simultánea, el planteamiento clásico de este problema, incorporado en los trabajos de los estadísticos Neyman y Pearson, consiste en suponer que es más probable que un error de tipo I sea más grave, en la práctica, que uno de tipo II. Por tanto, deberíamos intentar mantener la probabilidad de cometer error de tipo I a un nivel bastante bajo, como 0.01 ó 0.05, y después minimizar el error de tipo II todo lo que se pueda. La única forma de reducir un error de tipo II sin aumentar un error de tipo I es aumentar el tamaño de la muestra, lo que no siempre resulta fácil (Gujarati, 2006).

Siguiendo a Salburg (2001) admitimos que el uso de pruebas de significación de Fisher produce un número que llamó p-valor. Es una probabilidad calculada, una probabilidad asociada a los datos observados bajo el supuesto de que la hipótesis nula es verdadera. **El p-valor es una probabilidad, y así es como se calcula.**

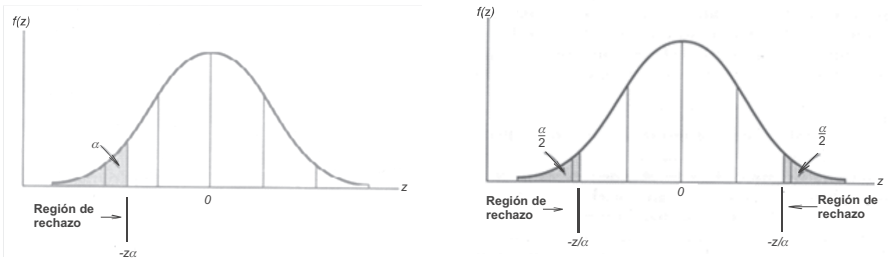
## Ejemplo: Cálculo del P-Valor

El p-valor nos proporciona el grado de credibilidad de la hipótesis nula: si el valor de *p* fuese "muy pequeño" (inferior a 0,001), significaría que la hipótesis nula es del todo increíble (en base a las observaciones obtenidas), y por tanto la descartaríamos; si el valor de *p* oscilase entre 0,05 y 0,001 significaría que hay fuertes evidencias en contra de la hipótesis nula, por lo que la rechazaríamos o no en función del valor que hubiésemos asignado (a priori) a á. Finalmente, si elvalor de *p* es "grande" (superior a 0,05), no habría motivos suficientes como para descartar la hipótesis nula, por lo que la tomaríamos como cierta.

## Criterios de decisión

- Si la hipótesis alternativa $H_1$ contiene " >" $\Rightarrow$*p-valor = P(Z>z)*
- Si la hipótesis alternativa $H_1$ contiene " <" $\Rightarrow$*p-valor = P(Z<z)*
- Si la hipótesis alternativa $H_1$ contiene "≠" $\neq$*p-valor = P(Z< -ó Z>) = 2 P(Z<)*

**Gráficos 1 y 2**
**Regiones de rechazo para pruebas de una y dos colas**




**Fuente:** Mendenhall y Sincich (1997, p. 431).

Tomemos un contraste de dos extremos, (para varianza conocida y n>30) ahora tenemos que calcular el p-valor correspondiente al valor del estadístico de prueba, por ejemplo $z_{cal}$ = 2,98, es decir el área que hay por debajo de $z_{cal}$ = -2,98 más el área que hay por encima de $z_{cal}$ = 2,98, es decir, el área en las dos colas. Donde los dos primeros dígitos (-2,9) se encuentran en la primera columna de la tabla normal y el último dígito (8) se busca en la primera fila de la misma, luego de la ubicación de ambos valores se busca la intersección entre la fila y la columna.

Si observamos la tabla de la distribución normal estándar, podemos comprobar que el área que hay a la izquierda de $z_{cal}$ = -2,98 es 0,0014 y el área que hay a la derecha de 2,98 es también 1 - 0,9986 = 0,0014 por lo que el p-valor = 2*0,0014 = 0,0028

Cuando la muestra es pequeña (n<30) y no se conoce ó$^2$debe hacerse un supuesto acerca de la forma de la distribución a fin de obtener un procedimiento para la prueba, el más razonable es el supuesto de normalidad y la distribución recomendada por sus características es la t–student (Montgomery y Runger, 2010).

La distribución t es similar a la z, por cuanto ambas distribuciones son simétricas y unimodales, y el valor máximo en las ordenadas se alcanza cuando $\mu$=0. Sin embargo, la distribución t tiene colas de mayor peso que la normal, es decir, tiene más probabilidad en las colas que la distribución normal. Cuando los grados de libertad $\nu \to \infty$, la forma límite de la distribución t es la distribución normal estándar (Montgomery y Runger, 2010).

Tomemos el contraste dos extremos (con varianza desconocida y n<30).

Supongamos que: tamaño de la muestra sea 19, los grados de libertad son 18, hipótesis $H_0$: $\mu$ = 87 y $H_1$: $\mu \neq 87$ entonces tenemos un contraste de dos colas, luego, el estadístico calculado es $t_{cal}$ = -2,6747 el p-valor será la probabilidad de estar por encima de 2,6747 más la probabilidad de estar por debajo de t =-2,6747. Cuando no aparece en la tabla de la t–studentel valor exacto del estadístico del cual se quiere calcular su p-valor, se toma como referencia el valor más cercano, en este caso t = -2,552. Por tanto el p-valor = P(t > 2,552)+P(t < -2,552) = 0,01 + 0,01 = 2*0,01 = 0,02, porque a la derecha de 2,5524 hay la misma probabilidad que a la izquierda de -2,5524 Así que el p-valor de t=-2,6747 será menor a 0,02 porque a mayor valor del estadístico menor área por encima como se puede ver en la tabla t–student.

Cuando los grados de libertad no aparezcan en la tabla de la t–student, se toma los grados de libertad más cercanos al cual se quiere tener en cuenta.

Si el contraste hubiese sido de una cola, bien por la derecha o bien por la izquierda, $H_1$: $\mu$ > 87 o $H_1$: $\mu$ < 87, entonces el p–valor del estadístico (supongamos que el estadístico es $t_{cal}$ = 2,6747) si el contraste es de cola derecha, es decir (mayor que), sería la probabilidad de estar por encima de t = 2,5524 que sería 0,01, por lo que el p-valor de $t_{cal}$= 2,6747 sería menor que 0,01.

Si es por la cola izquierda (es decir menor que), el p-valor del estadístico (supongamos que el estadístico vale $t_{cal}$ = -2,6747) sería la probabilidad de estar por debajo de $t_{cal}$ = -2,5524 que sería 0,01, por lo que el p-valor de $t_{cal}$ = - 2,6747 sería menor que 0,01.

## El P-Valor: consideraciones de algunos autores

Si bien tratamiento del p-valor se le asigna a Fisher entendemos que Karl Pearson lo usó en su Prueba de chi cuadrado para la bondad de ajuste que es ante-

rior a la denominación de p-valor según Fisher. Según sigamos el punto de vista de Fisher o el de Neyman-Pearson, en su origen, el p-valor tenía significados teóricos levemente diferentes. Sin embargo, con el avance de la tecnología y la difusión de software estadísticos su diferencia teórica, en apariencia, se desdibuja.

Una selección arbitraria de los libros de texto editados en la década pasada, nos ayudan a avalar esta idea:

De acuerdo con Walpole, Myers, Myersy Ye (2007, Pp.334-335):

> "El valor p se puede ver simplemente como la posibilidad de obtener este conjunto de datos dado que las muestras provienen de la misma distribución…., la aproximación del valor p como ayuda en la toma de decisiones es bastante natural, ya que casi todos los paquetes computacionales que ofrecen el cálculo de prueba de hipótesis dan valores p junto con valores del estadístico de prueba adecuado. Un valor p es el nivel (de significancia) más bajo donde es significativo el valor observado del estadístico de prueba"

Para Daniel (2009, p. 216):

> "El informe de valores p como parte de los resultados de una investigación proporciona más información al lector que afirmaciones como la hipótesis nula se rechaza con un nivel de significación de 0,05 o los resultados no fueron significativos en el nivel 0,05. Al informar el valor p asociado con una prueba de hipótesis se permite al lector saber con exactitud qué tan extraño o qué tan común es el valor calculado de la estadística de prueba dado que $H_0$ es verdadera".

Según Montgomery y Runger (2010, p. 37):

> "El valor p es la probabilidad de que el estadístico de prueba asuma un valor que sea al menos tan extremo como el valor observado del estadístico cuando la hipótesis nula $H_0$ es verdadera. Por lo tanto, un valor p transmite mucha información acerca del peso de la evidencia en contra de $H_0$ y, por consiguiente, el responsable de la toma de decisiones puede llegar a una conclusión con cualquier nivel de significación especificado".

No siempre es sencillo calcular el valor p exacto de una prueba. Sin embargo, la mayoría de los programas de computadora modernos para el análisis estadístico reportan valores p, e incluso pueden obtenerse también en algunas calculadoras portátiles.

En opinión de Gujarati (2006, p. 120):

> "El talón de Aquiles del planteamiento clásico para la contrastación de hipótesis es la arbitrariedad en la elección de á. Aunque 1, 5, y 10 por ciento en los valores comúnmente utilizados para á, no hay nada inviolable en estos valores…En la práctica, es preferible encontrar el valor p (es decir, el valor de probabilidad), también conocido como *nivel exacto de significancia del estadístico de prueba*. Este valor se puede definir como el *menor nivel de significancia al que se puede rechazar una hipótesis nula*".

Según Lind, Marchal, y Mason (2004, p. 347):

> "En años recientes, debido a la disponibilidad de los programas de cómputo (software), se proporciona con frecuencia información adicional relativa a la fuerza del rechazo"… "El valor p es la probabilidad de observar un valor muestral tan extremo, o más extremo, que el valor observado, dado que la hipótesis nula es cierta".

Para Levine, Krehbiel y Berenson (2006, p. 281):

> "La mayoría de los programas de cómputo moderno, incluyendo Excel, Minitab y SPSS calculan el valor-p al realizar una prueba de hipótesis….El valor-p es la probabilidad de obtener un estadístico de prueba igual o más extremo que el resultado de la muestra, dado que la hipótesis nula Ho es cierta…. El valor-p, que a menudo se denomina nivel de significación observado, es el nivel más pequeño en el que se puede rechazar Ho".

**Comentarios finales**

- El p-valor se puede definir como el menor nivel de significación al que se puede rechazar una hipótesis nula cuando es verdadera.
- El discutido p-valor se puede interpretar de distinta forma según el enfoque de Fisher o la Teoría de Neyman-Pearson.
- El avance de la tecnología permitió que los paquetes estadísticos reportaran el p-valor.

Desde el punto de vista de la tarea cotidiana, disponer del p-valor no implica inconsistencias. En efecto, el investigador podrá fijar de antemano el nivel de significación según lo establece la Teoría de Neyman-Pearson y, con el resultado que reporta el software decidir sobre el rechazo, o no, de la hipótesis nula.

## Referencias Bibliográficas

Daniel, Wayne. (2009). **Bioestadística, base para el análisis de las ciencias de la salud**. Limusa Wiley. 4ta Edición. México.

Fisher, Ronald. (1925). Statistical Methods for Research Workers. Originally published in edinburgh by oliver and boyd. Extraído de http://es.scribd.com/doc/58873576/Fisher-R-a-1925-Statistical-Methods-for-Research-Workers, Consulta08/08/2012.

Gujarati, Damodar. (2006). **Principios de Econometría**. 3era Edición en español. McGraw-Hill. España.

Gómez, María. (2011). **Sobre el Concepto del P-Valor**. Universidad Complutense de Madrid. España.

Levine, David; Krehbiel, Timothy y Berenson, Mark. (2006). **Estadística para Administración**. Pearson. 4ta Edición. México.

Lind, Douglas; Marchal, William; Mason, Robert. (2004). **Estadística para Administración y Economía**. 11va Edición. Alfaomega México D.F.

Mendenhall, William y Sincich, Terry. (1997). **Probabilidad y Estadística, para ingeniería y ciencia**s. 4ta Edición. Prentice-Hall Hispanoaméricana. México.

Montgomery, Douglas y Runger, George. (2010). **Probabilidad y Estadística aplicada a la Ingeniería**. Limusa Wiley. 2da Edición. México.

Salsburg, David. (2001). **The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century**. Henry Halt and Company LLC. New York.

Walpole,Ronald; Myers, Raymond; Myers, Sharon yYe, Keying. (2007). **Estadística y Probabilidad para Ingeniería y Ciencias**. 8va Edición. Pearson. México.