Historia de la Probabilidad y la Estadística [VI] (J. M. Arribas, A. Almazán, B. Mañas y A. Vallejos eds.) Madrid: Servicio de Publicaciones de la UNED
(2012), 211-216

SOBRE EL CONCEPTO DEL P-VALOR

M. A. Gómez Villegas
Departamento de Estadística
e Investigación Operativa
Facultad de CC Matemáticas
Universidad Complutense de Madrid
Plaza de las Ciencias, 3
28040 Madrid
913944423
ma.gv@mat.ucm.es

RESUMEN

El concepto de p-valor como resultado de un tratamiento estadístico, es un concepto debido a Fisher, es una contraposición del concepto de tamaño α 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.

INTRODUCCIÓN Y PRIMEROS PASOS

Sin duda podemos decir, que hasta la contribución de Jacob Bernoulli del *Arte de la Conjetura* en 1713, no aparece, en mi opinión, la primera aproximación al concepto de p-valor. Debe reconocerse que es únicamente de forma muy rudimentaria, pero Jacob calibra en el sentido que ahora lo hacemos las probabilidades. Así dice:

> *La probabilidad es pues un grado de certeza, y se diferencia de ella como la parte del todo...* ***"posible"*** *es pues lo que tiene una parte de certeza, por ejemplo* $1/20$ *o* $1/30$ *de certeza.* ***"Moralmente cierto"*** *es aquello cuya probabilidad equivale casi a la certeza total,...Por contra* ***"moralmente imposible"*** *es aquello que sólo tiene tanta probabilidad como la que le falta a lo moralmente cierto para ser totalmente cierto. Así pues, si lo que tiene una certeza de* $999/1000$ *se considera moralmente cierto, entonces lo que sólo tiene una certeza de* $1/1000$ *es moralmente imposible.*

Es decir aparece por primera vez el concepto de lo que es "moralmente cierto" como aquello que tiene una probabilidad próxima a 1. Podemos entonces decir que es una aproximación rudimentaria, pero aproximación al cabo, al concepto de "grado de confianza" de la teoría de Neymann-E. Pearson, tal y como es usada actualmente. De todas formas está tratado sin dar ninguna idea de su utilización para aceptar o rechazar una hipótesis.

El párrafo citado puede consultarse en Rivadulla (1993)pág. 390.

El siguiente autor que debe ser citado y que hace una contribución hacia el concepto del p-valor es Laplace, que en su memoria de 1823 utiliza el método de los mínimos cuadrados, que el había llamado el "método más ventajoso" para estudiar el efecto de la luna en las mareas terrestres. 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 sería significativa al nivel $0,01$ si se hubieran estimado las medias a partir de datos de 72 años. Es decir que Laplace se anticipa determinando

no sólo el p-valor sino también éste en función del tamaño muestral. Es notable señalar 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 hasta 1945 cuando 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 dé un veredicto correcto. Esto lo hace Poisson (1837) pág. 370 donde aproxima la distribución binomial por la normal y calcula el p-valor correspondiente a la aproximación realizada.

Un nuevo paso es dado por Arbuthnot (1667-1735) quién en su memoria de 1710 "Un argumento para la providencia divina tomado de la constante regularidad observada en los nacimientos de ambos sexos" utiliza un p-valor muy pequeño, de $1/2^{82}$, para concluir la existencia de la providencia divina.

La contribución de RONALD AYLMER FISHER (1890-1962) al p-valor

La siguiente cita histórica sobre el p-valor es la de Fisher (1925) y consiste en el celebrado libro *Statistical Methods for Research Workers* (SMRW), del que se hicieron catorce ediciones en vida de Fisher.

En concreto Fisher pretende contrastar la hipótesis de que la media de una población normal es un valor θ. Siempre se insiste en que Fisher negaba la existencia de los errores de tipo I y de tipo II y que en el SMRW se mostraba partidario del concepto de p-valor. Sin embargo ésto no es así; extraigo el siguiente párrafo del libro:

> *Si conocemos la varianza de una población, podemos calcular la varianza de la media muestral de cualquier tamaño y contrastar si difiere significativamente de cualquier valor fijo. Si la diferencia es mucho más grande que el error estándar esto es ciertamente significativo, y es una convención conveniente tomar dos veces el error estándar como límite de significación; esto es aproximadamente correspondiente a un valor de* $0,05$ *en la distribución de la* χ^2 *...*

Como se ve, calcula la discrepancia existente entre la media muestral y la media poblacional, bajo la suposición de normalidad, pero sin hacer referencia, ni al concepto de tamaño, ni al de p-valor; al menos de manera clara.

Unas páginas más adelante (pág. 119) trata el mismo problema, pero planteado como una tabla de análisis de la varianza, ya que dice:

> *Si $x_1, \ldots, x_n$ es una muestra de n valores de una variable x y si esta muestra constituye toda la información posible sobre el punto en cuestión, entonces podemos contrastar si la media de x difiere significativamente de 0 calculando el estadístico $\overline{x}$ y usando que se distribuye como una T la variable $\frac{\overline{X}}{S}\sqrt{n}$. Aritméticamente, los cálculos dependen del hecho simple de que la suma de cuadrados de las desviaciones a la media, puede ser obtenida a partir de la suma de cuadrados de las desviaciones de cero, mediante la expresión*
>
> $$\sum x_i^2 = \sum (x_i - \overline{x})^2 + \overline{x} \sum x_i$$
>
> *Esto es una subdivisión de la suma de cuadrados de x en dos partes, la primera de las cuales representa la variación dentro de la muestra, mientras la segunda es debida únicamente a la desviación de la media observada de cero, la primera parte tiene $n-1$ grados de libertad y la segunda sólamente uno*

	Grados de Libertad	*Suma de cuadrados*	*Cuadrado Medio*
Desviación	*1*	$\bar{x}^2$	T^2S^2
entre muestras	*n-1*	$\sum(x_i - \bar{x})^2$	S^2
Total	*n*	$\sum x_i^2$	

> *Los cuadrados medios se obtienen en cada caso dividiendo la suma de cuadrados entre los grados de libertad. El cociente de los cuadrados medios es en este caso T^2. Esta forma de organizar los cálculos, es aplicable a muchas más situaciones y es conocida como el Análisis de la Varianza.*

Al menos inicialmente, no conoce la distribución del cociente de los cuadrados (hoy conocida como distribución F de Fisher-Snedecor) por lo que aproxima su logaritmo por una distribución normal.

Además, emplea la Tabla de la Varianza, quizá por eso no realiza el test de hipótesis, sino que niega la existencia de la hipótesis alternativa. Por cierto que Fisher se unió a la *Rothamsted Experimental Station* en octubre de 1919 y allí desarrolló *el análisis de la varianza* y los principios del *diseño de experimentos*; como su particular contribución al esfuerzo a favor de los aliados en la segunda guerra mundial, ya que se le había impedido ingresar en el ejército, debido a su deficiente visión.

Un estudio de las contribuciones estadísticas de Fisher puede verse en Girón y Gómez Villegas (1998).

CONCLUSIÓN Y ÚLTIMAS CONTRIBUCIONES

La siguiente contribución es al contraste de hipótesis no al concepto de p-valor, y se debe al trabajo de Neyman y Egon Pearson en 1933, que ha sido continuado en muchas direcciones. En este artículo introducen lo que hoy se conoce como el tratamiento de los tests óptimos, para hipótesis nula simple frente a hipótesis alternativa simple, y pasan a hacer protagonistas del problema al error de tipo I de un test, -la probabilidad de rechazar la hipótesis nula siendo "verdadera"-, y al error de tipo II, -la probabilidad de aceptar la hipótesis nula siendo "falsa"- (puede imaginarse el nombre que hubieran dado al nuevo error, caso de que hubiera existido).

Se debe reconocer que el trabajar con los tests que mantienen acotada la probabilidad de cometer el error de tipo I por un valor, y utilizar el contraste que minimiza la probabilidad del error de tipo II, resulta fácil de usar, en especial para el que utiliza la estadística sin demasiado fundamento teórico. Pero no obstante la utilización del p-valor es un procedimiento que permite que todos los estadísticos a los que se les dé la misma informacón muestral lleguen al mismo resultado. De todas formas esta contraposición todavía no está universalmente aceptada, aunque los paquetes usuales de estadística que hay en el mercado, llevan incorporado el cálculo del p-valor.

Frecuéntemente se confunde el p-valor con la probabilidad de que la hipótesis nula sea verdadera; así un p-valor alto se interpreta cómo que la probabilidad de que la hipótesis nula sea cierta es alta. Esto es lo que se conoce como "falacia del fiscal" Lindley (2000). Sin embargo para que se

pueda hablar de estas probabilidades, se tiene que introducir la distribución inicial y moverse, por tanto, en la aproximación bayesiana a la estadística. Precisamente el autor ha trabajado en ese campo y estudiado cuándo numéricamente se puede producir acuerdo, entre el p-valor y la probabilidad final de la hipótesis nula Gómez Villegas y otros (2002, 2010).

Agradecimientos

Este trabajo ha sido subvencionado parcialmente con la ayuda del proyecto del *Ministerio de Ciencia e Innovación MTM2008 Tests Bayesianos en Microarrays y Robustez en Redes Bayesianas*, y parcialmente con la ayuda para la realización de proyectos de investigación de la *Universidad Complutense de Madrid-Santander GR35/10-A 910395 Métodos Bayesianos.*

BIBLIOGRAFIA

ARBUTHNOT, J. (1710) An argument for Divine Providence, taken from the constant regularity observd́ in the births of both sexes *Philosophical Transactions of the Royal Society of London*, 27, 186-190. Reprinted in Kendall and Placket, 1977, 30-34.

BERNOULLI, J (1713) *Ars Conjectandi*, Edit. Thurnisiorum. Basilea.

FISHER, R. A. (1925) *Statistical Methods for Research Workers* . Edited by Olyver and Boyd: Edinburgh.

FISHER, R. A. (1990) *Statistical Methods, Experimental Design and Scientific Inference*. Edited by Bennett, J. M. with a foreword by Yates, F. Oxford. Oxford University Press.

GIRÓN, F. J. & GÓMEZ VILLEGAS, M. A. (1998) *R. A. Fisher: su contribución a la Ciencia Estadística*, Editado por la Real Academia de Ciencias Exactas, Físicas y Naturales *Historia de la Matemática en el Siglo XX*. Madrid, 43-61.

GÓMEZ VILLEGAS, M. A., MAÍN, P. & SANZ, L. (2002) Asuitable Bayesian approach in testing point null hypothesis: some examples revisited, *Communications in Statistics-Theory and Methods*, 31, 2, 201-247.

GÓMEZ VILLEGAS, M. A. & GONZALEZ-PÉREZ, B. (2010) r $\times$ s tables from a Bayesian viewpoint, *Revista Matemática Complutense*, 23, 1, 19-25.

LAPLACE, P. S. (1823) *De l'action de la lune sur l'atmosphere*, Annales de Chimie et de Physique, 24, 280-294

LINDLEY, D.V.(2000) The philosophy of statistics,. *The Statistician*, 3, 293-337.

NEYMAN, J. & PEARSON, E. (1933) On the problem of the most effi-

cient tests of statistical hypotheses. *Philosophical transactions of the Royal Society A*, 231, 289-337.

POISSON, S. D. (1837) *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédés des regles générales du calcul des probabilités* Ed. Bachelier. París.

RIVADULLA, A. (1993) *Teoría de Probabilidades Ars conjectandi, parte cuarta, Basilea 1713*, *Llull 16,389-418.*