ESTADÍSTICA Y PSICOLOGÍA: ANÁLISIS HISTÓRICO DE LA INFERENCIA ESTADÍSTICA[1]

Dr. Enerio Rodríguez Arias
Universidad Autónoma de Santo Domingo
erodriguez27@uasd.edu.do

RESUMEN

Después de enunciar brevemente los principales aportes de los pioneros ingleses de la ciencia estadística, se examina en una perspectiva histórica el proceso de la inferencia estadística como el mecanismo fundamental para el manejo de los errores variables en una investigación. Se identifican los componentes distintivos de los dos enfoques conceptuales que se amalgaman en la exposición establecida de los diferentes pasos de la inferencia estadística, a saber, el enfoque de Fisher y el enfoque de Pearson y Neyman. Las enemistades personales entre estadísticos ilustran el papel de las pasiones humanas en las controversias intelectuales. Se alude brevemente a las consecuencias negativas del híbrido representado por el matrimonio de conveniencia entre puntos de vista claramente opuestos.

Palabras clave: Historia, estadística, inferencia, significación, controversia.

El presente artículo es de interés tanto para el estadístico como para el psicólogo. Las técnicas de análisis de datos más frecuentemente utilizadas por los psicólogos en sus investigaciones, fueron creadas por un pequeño grupo de estadísticos ingleses: Francis Galton, Karl Pearson, William Gosset (Student), Ronald Fisher y Egon Pearson (hijo de Karl Pearson); este último (Egon Pearson) trabajó en colaboración con el matemático polaco Jerzy Neyman, quien vivió por un tiempo en Inglaterra y más tarde se estableció en Los Estados Unidos de América. Las ideas de correlación y regresión provienen de Galton; el primer Pearson, además de producir la fórmula para el cálculo de la correlación, es el creador de la prueba de la ji cuadrada; Gosset creó la prueba t en su forma original, Fisher desarrolló aún más la prueba t, bautizándola con el nombre de "la t de Student" y no la t de Gosset, porque éste, debido a los términos del contrato laboral suscrito entre él y la cervecería Guinness de Dublín, Irlanda, sólo podía firmar con su verdadero nombre los informes y documentos preparados para la empresa, y por esa razón usaba el pseudónimo de "Student" para firmar sus artículos sobre estadística.

Además de desarrollar la famosa t de Student, Fisher creó el análisis de varianza, que luego fue bautizado con el nombre de prueba F en su honor. Pero el legado más controversial de Fisher es la prueba de la hipótesis nula como la estrategia de inferencia inductiva que debe guiar el análisis estadístico de los datos en una investigación científica. Es en este punto donde han intervenido Egon Pearson y Jerzy Neyman (de aquí en adelante, Pearson & Neyman), contradiciendo la posición de Fisher y generando un debate, desconocido para la mayoría de estadísticos y psicólogos, que a través de los libros de texto hemos heredado una estrategia de análisis que aparentemente ha disuelto la contradicción. (Aron & Aron, 2001; Gigerenzer et al., 2004).

Además de informar al lector acerca de cómo la enemistad personal puede afectar una controversia científica, la meta principal del presente artículo es dejar claramente establecido qué es de Fisher y qué

1- Trabajo publicado en *Perspectivas Psicológicas, Vol. 5,* Año 6, 2005.

es de Pearson & Neyman en la prueba de la hipótesis nula, tal como el procedimiento es expuesto en los libros de texto y rutinariamente practicado en la investigación psicológica publicada.

REFRESCANDO LA MEMORIA DEL LECTOR

Utilizando una analogía de indudable valor didáctico, Hyman (1964) comparaba el análisis estadístico de los datos de una investigación con el proceso de la digestión. Cuando el alimento, que en esta analogía corresponde a los datos, es ingerido por un organismo, no puede ser utilizado por éste mientras no haya sido desintegrado y reconstituido en una forma asimilable. Diferentes sustancias alimenticias requieren para este proceso diferentes períodos de tiempo y combinaciones distintas de la actividad digestiva. El proceso digestivo ha de adaptarse en una forma altamente específica a cada tipo de alimento. Y en una forma análoga, el investigador ha de adaptar sus procedimientos de análisis de los datos a la naturaleza originaria de sus observaciones. Algunos instrumentos de reunión de datos arrojan datos iniciales en forma de números, a los cuales se pueden aplicar directamente técnicas estadísticas. Otros instrumentos, como los cuestionarios de respuesta libre y las entrevistas, arrojan datos iniciales que requieren de tratamiento y codificación adicionales, antes de que se pueda comenzar a realizar con ellos cualquier análisis estadístico. Diferentes clases de datos requieren tipos diferentes de análisis antes de quedar reducidos a una forma en la que puedan ser interpretados. (Girden, 1996; Stern & Kalof, 1996).

En la situación más sencilla de investigación, hay una variable dependiente y por lo menos una variable independiente con dos valores. La tarea del investigador consiste en determinar el grado en que los datos de la investigación reflejan una relación entre las variables independiente y dependiente; o dicho en otros términos, el análisis estadístico de los datos persigue determinar si dos grupos que difieren en el lugar que ocupan en la variable independiente, difieren también en la variable dependiente. Tanto la relación entre las variables como la ausencia de relación entre las mismas, pueden resultar enmascaradas por dos clases de errores: los errores constantes y los errores variables. Los errores constantes son producidos por variables extrañas que afectan de manera constante los resultados de una investigación; por ejemplo afectan siempre de la misma manera la relación (o la falta de relación) entre dos variables (independiente y dependiente), o afectan siempre de la misma manera (favorable o desfavorablemente) a los grupos de la investigación. Los errores variables, en cambio, son producidos por variables extrañas que afectan de manera variable los resultados de la investigación; se trata de variables extrañas que actúan aleatoriamente sobre todos los sujetos de la investigación, y que por esa misma razón, sus efectos tienden, en el largo plazo, a anularse mutuamente.

No existen fórmulas simples para prevenir los errores constantes. A través de su entrenamiento y experiencia, el investigador aprende a anticipar muchas de las fuentes más graves de errores constantes; pero nunca tiene una garantía absoluta de que ha eliminado toda posibilidad de que alguna variable extraña pueda estar produciendo un error constante. En realidad, una gran parte del progreso alcanzado en cualquier campo de la investigación científica depende del descubrimiento de las fuentes de tales errores constantes. Salvo algunos descubrimientos científicos excepcionales, puede decirse que en general la mayoría de los descubrimientos científicos, eliminan alguna fuente de error constante.

En cuanto a los errores variables, además de los esfuerzos que puede hacer el investigador para minimizar la varianza de error a través del diseño de la investigación (Kerlinger & Lee, 2002), una forma de enfrentar dichos errores es la inferencia estadística (Hyman, op. cit.). Esta expresión se aplica a un conjunto de procedimientos utilizados para determinar el grado en que la relación observada entre dos variables (o la diferencia observada en una variable dependiente entre dos o más grupos de sujetos) puede explicarse como resultado del azar, es decir, atribuirse a fuentes de errores variables. Para verificar la posibilidad de que la relación observada entre las variables independiente y dependiente de una investigación

Perspectivas Psicológicas/Santo Domingo (Rep. Dom.)
/Vols. 6-7/pp. 165-171/2010/ISSN 1992-5131

pueda explicarse como resultado del azar, el investigador realiza una operación designada con el nombre de ritual de la prueba de significación estadística. El término "ritual" describe adecuadamente dicha operación, porque sucede que la mayoría de los investigadores realizan la prueba de significación estadística de una manera rutinaria, ejecutando cada paso de la operación de modo automático, tal como están formulados en los libros de texto. Dicho ritual, conocido también como la prueba de la hipótesis nula fue creado por Ronald Fisher. Veamos los pasos de la prueba de la hipótesis nula, tal como aparecen en cualquier libro de texto de estadística. Valga la advertencia previa de que en dicha exposición se combinan las posiciones de Fisher y de Pearson-Neyman, pero en cada caso se identifica la posición a la que pertenece cada elemento.

PASOS DEL RITUAL DE LA PRUEBA DE SIGNIFICACIÓN ESTADÍSTICA:

1. El investigador formula la hipótesis nula. En términos generales, la hipótesis nula afirma que no existe ninguna relación real o verdadera entre las variables independiente y dependiente de una investigación, y que, por tanto, si alguna relación es observada entre dichas variables en los datos de la investigación, la misma podría explicarse como resultado del azar. Es por eso que a la hipótesis nula se le llama la hipótesis del azar. Dicho de otra manera, la hipótesis nula expresa que si se repitiera la investigación un número suficiente de veces, siempre con una muestra distinta extraída aleatoriamente de la misma población, las diferencias en la variable dependiente entre los grupos de la investigación tenderían a neutralizarse y terminarían siendo cero. El razonamiento implícito en la hipótesis nula es el siguiente: Suponiendo que el resultado de una investigación particular constituye una selección al azar de entre una multitud de resultados posibles, el investigador se pregunta cuál sería la probabilidad de obtener por azar la diferencia que él ha encontrado entre los grupos de su investigación.

   Si esa probabilidad es igual o menor que un nivel de probabilidad convencional previamente establecido, entonces el investigador concluye que los resultados por él observados no se deben al azar, y, por tanto, rechaza la hipótesis nula. Si, en cambio, la probabilidad de que la diferencia observada entre los grupos se pueda explicar como resultado del azar es superior al nivel de probabilidad convencional previamente establecido, entonces no se puede descartar el azar, es decir, no se rechaza la hipótesis nula. Esta formulación es puramente fisheriana.

2. Es obvio que la decisión sobre la hipótesis nula requiere de que se haya establecido previamente un nivel de significación estadística, es decir, un criterio que sirva de base a la decisión de rechazar o no rechazar la hipótesis nula. Al establecer un criterio de decisión sobre la hipótesis nula, el investigador puede ponderar los errores que podría cometer en su decisión sobre la hipótesis nula. Una primera forma de error (se conoce como el error tipo I) consiste en rechazar una hipótesis nula verdadera, es decir, descartar el azar como explicación cuando los resultados podrían explicarse razonablemente con base en el mismo. Este es el error que comete el investigador que ve más que lo que hay en los datos; es decir, el investigador concluye que existe una relación real o verdadera entre las variables independiente y dependiente de la investigación, cuando en realidad la relación observada se puede explicar razonablemente como resultado del azar. El llamado error tipo I es el error del investigador que se apresura a concluir a favor de su hipótesis de investigación. Fisher no habló de ningún otro error, pues la prueba de la hipótesis nula para él no era otra cosa que un freno a la tendencia natural de un investigador a creer que una hipótesis ha sido confirmada por el simple hecho de que los resultados de la investigación siguen la misma dirección de la hipótesis.

   En la estrategia de Fisher, sólo hay un error posible: rechazar una hipótesis nula verdadera. Una segunda forma de error (se conoce como el error tipo II), introducida por Egon Pearson y Jerzy Neyman, consiste en no rechazar una hipótesis nula falsa, es decir, no descartar el azar aun cuando éste no constituye una explicación razonable de los datos. Este es el error que comete el investigador que ve menos que lo que hay en los datos; por miedo a rechazar incorrectamente el azar, el investigador puede exponerse al riesgo de pasar por alto una

relación real o verdadera entre las variables de su investigación. Fueron Pearson y Neyman los que, al introducir un segundo tipo de error, bautizaron como error tipo uno al error de que había hablado Fisher.

En la perspectiva fisheriana, el nivel de significación estadística es el punto que separa las probabilidades que nos conducen a rechazar la posibilidad de que la relación observada entre las variables de una investigación se deba completamente a errores variables (errores de azar) de aquellas probabilidades que nos conducen a no rechazar esa posibilidad.

Según Fisher, el nivel de significación estadística equivale a la magnitud del riesgo que está dispuesto a correr el investigador, de cometer el error de rechazar una hipótesis nula verdadera (el llamado error tipo I). Para la mayoría de los propósitos, el nivel de significación previamente establecido suele ser de 0.05, aunque en áreas de investigación más rigurosas se trabaja con un nivel de significación de 0.01. Suponiendo que se trabaja con un nivel de significación de 0.05, se rechazaría la hipótesis nula siempre que la probabilidad de explicar los resultados obtenidos en una investigación como si fueran obra del azar sea igual o menor que 0.05.

En la perspectiva de Pearson y Neyman, para establecer el nivel de significación estadística habría que atender al impacto de cada tipo de error en el objetivo del investigador, y a partir de ahí se decidiría cuál de ellos es preferible minimizar. Pearson y Neyman llamaron alfa al error tipo I y beta al error tipo II; a partir de este último tipo de error, introdujeron el concepto de "poder de una prueba estadística", el cual se refiere a su capacidad para evitar el error tipo II, y está definido por 1 menos beta, y en estrecha relación con éste se ha desarrollado el concepto de "tamaño del efecto" que algunos han propuesto como sustituto de los valores p en los informes de investigación científica. (Cohen, 1990, 1994; Kraemer & Thiemann, 1987; Murphy & Myors, 2004).

3. El tercer paso del llamado ritual de la prueba de significación estadística consiste en la elección de la prueba estadística que se utilizará para someter a prueba la hipótesis nula. Hay dos clases de pruebas estadísticas: Las paramétricas y las no paramétricas. Se llama paramétricas a aquellas pruebas estadísticas que exigen que los datos a los que se aplican cumplan con los siguientes requisitos: Que los valores de la variable dependiente sigan la distribución de la curva normal, por lo menos en la población a la que pertenezca la muestra en la que se hizo la investigación; que las varianzas de los grupos que se comparan en una variable dependiente sean aproximadamente iguales (homoscedasticidad, u homogeneidad de las varianzas); y que la variable dependiente esté medida en una escala que sea por lo menos de intervalo, aunque este último requisito no es compartido por todos los estadísticos (McGuigan, 1993; Siegel, 1956). Cuando los datos cumplen con los requisitos indicados, especialmente con los dos primeros, las pruebas estadísticas paramétricas exhiben su máximo poder, es decir, su máxima capacidad para detectar una relación real o verdadera entre dos variables, si es que la misma existe. Las pruebas paramétricas más conocidas y usadas son la prueba t de Student, la prueba F, llamada así en honor a Fisher, y el coeficiente de correlación de Pearson, simbolizado por r. Cuando estas pruebas estadísticas se aplican a datos que violan los dos primeros de los requisitos señalados, pierden parte de su poder. Las pruebas estadísticas no paramétricas, en cambio, no hacen a los datos ninguna de las exigencias que les hacen las pruebas estadísticas paramétricas; por eso se les denomina "pruebas estadísticas libres de distribución". Las más conocidas y usadas de estas pruebas son la ji cuadrada de Pearson, la prueba de la probabilidad exacta de Fisher, los coeficientes de contingencia de Pearson y Cramer, la prueba U de Mann & Whitney, el coeficiente de correlación de rangos de Spearman, y el coeficiente de asociación ordinal de Goodman y Kruskal (coeficiente gamma), (Conover, 1999; Leach, 1979; Siegel, op. cit.). Todas estas pruebas poseen menos poder que las pruebas paramétricas correspondientes, pero han demostrado ser muy útiles como alternativas cuando no se considera apropiado el uso de pruebas paramétricas.

4. El último paso del ritual de la prueba de significación estadística consiste en comparar el valor arrojado por la prueba estadística aplicada a los datos, con el valor que en circunstancias comparables puede ocurrir por azar con una

Perspectivas Psicológicas/Santo Domingo (Rep. Dom.)
/Vols. 6-7/pp. 165-171/2010/ISSN 1992-5131

probabilidad de 0.05 ó 0.01, según el valor de la probabilidad que se haya adoptado como nivel de significación estadística. Si al consultar la tabla de los resultados de la prueba estadística que pueden ocurrir por azar con diferentes niveles de probabilidad, se observa que el resultado de la investigación tiene una probabilidad de ocurrir por azar igual o menor que la probabilidad adoptada como nivel de significación estadística, entonces se rechaza la hipótesis nula. Si, en cambio, el resultado de la investigación tiene una probabilidad de ocurrir por azar mayor que la probabilidad adoptada como nivel de significación estadística, entonces no se rechaza la hipótesis nula. Esto es todo cuanto diría Fisher al terminar la prueba de la hipótesis nula. Pearson & Neyman, en cambio, incorporaron la idea de simetría entre el rechazo y la confirmación de la hipótesis nula; es a partir de ellos que los libros de texto de estadística han incorporado la expresión "se acepta la hipótesis nula", pues para Fisher sólo era posible rechazar o no rechazar la hipótesis nula.

LA HISTORIA DETRÁS DE UNA DIFERENCIA CONCEPTUAL

En un pequeño libro dedicado al estudio de los pioneros de la estadística, Tankard (1984) se refiere al hecho de que las cuatro técnicas estadísticas más comunes fueron creadas por cuatro ingleses nacidos dentro de un período de sesenta y ocho años. El primero de ellos, Francis Galton (1822-1911) es el precursor inmediato de la ciencia estadística en Inglaterra. Primo de Charles Darwin, fue el primero en hablar de biometría (término que usaba para referirse a lo que luego sería la estadística. Creó la cátedra de Eugenesia en la Universidad de Londres. Sus estudios de la relación entre la altura de los hijos y sus padres le llevaron a los conceptos de correlación y regresión. Karl Pearson (1857-1936) ha sido considerado como el fundador de la ciencia estadística. Seguidor entusiasta de la teoría de la evolución, e influido por las ideas de Galton, creyó encontrar en la correlación, cuya fórmula de cálculo desarrolló, el instrumento adecuado para convertir la psicología, la antropología y la sociología en ciencias tan respetadas como la física y la química. Su contribución más famosa a la estadística es la prueba ji cuadrada, aunque es más común escuchar "la correlación de Pearson". Fundó la revista especializada en Estadística *Biometrika***,** y contribuyó de manera notable a elevar el prestigio de la estadística como un instrumento de gran valor para el método científico. Como toda persona controversial, fue capaz de provocar tanto amistades devotas como enconadas enemistades. William Gosset (1876-1937) fue uno de sus mejores amigos, mientras que Ronald Fisher (1890-1962) fue uno de sus peores enemigos. En realidad Gosset era amigo de ambos, y siempre trató de suavizar los problemas entre ellos. Pero la enemistad entre Pearson y Fisher era tan profunda que Fisher sólo pudo publicar un artículo en la revista *Biometrica* que dirigía Pearson. Eso ocurrió en 1915, cuando Fisher tenía 25 años y Pearson 39. A partir de ese año, Pearson apeló a las más inverosímiles excusas para negarle las páginas de *Biometrica* a Fisher; ni siquiera algunas notas técnicas breves, enviadas por Fisher para aclarar asuntos conceptuales de estadística, le fueron publicadas en *Biometrica*. (Cowles, 2001).

Karl Pearson se retiró en 1933, y le correspondió a Ronald Fisher sustituirlo en la cátedra de Eugenesia, originalmente creada por Galton en la Universidad de Londres. Pearson murió en 1936, precisamente el año en que él y Fisher comenzaban su más punzante discusión.

Ronald Fisher fue probablemente el más brillante y productivo de los miembros del pequeño grupo de estadísticos ingleses. Publicó alrededor de 300 trabajos y siete libros, en los cuales desarrolló muchos de los conceptos de la estadística: La importancia de la aleatorización, la varianza, el análisis de varianza, la distinción entre estadística (medida de muestra) y parámetro (medida de población), hipótesis nula, niveles de significación, y las ideas fundamentales del diseño de investigación. De temperamento difícil, se vio involucrado en profundas enemistades. Se dice de él que cuando le hablaban en broma, él contestaba en serio; cuando los demás estaban serios, entonces él bromeaba.

Karl Pearson tenía un hijo, Egon Pearson (1895-1980) quien también era estadístico y trabajaba en el laboratorio Galton bajo las órdenes de su padre.

En 1925, el joven Egon inició una perdurable amistad con Jerzy Neyman (1894-1981), un joven matemático de la Universidad de Varsovia que acababa de llegar al laboratorio Galton. Cuando Karl Pearson se retiró en 1933, para que su hijo Egon Pearson no estuviera bajo la dirección de Ronald Fisher, se creó un nuevo Departamento de Estadística, a ser dirigido por Egon Pearson. De nada valieron los esfuerzos del binomio Pearson-Neyman por evitar la continuación de la vieja enemistad de Fisher, pues este desplazó su hostilidad hacia ellos, reaccionando de manera enfurecida frente a las extensiones y elaboraciones que de la posición de Fisher trataban de hacer Pearson & Neyman.

Fisher consideraba que su manera de someter a prueba la hipótesis nula era absolutamente objetiva y rigurosa y la única forma de inferencia inequívoca. Cuando Jerzy Neyman terminó de pronunciar su discurso de ingreso a la *Royal Statistical Society* en Londres, Fisher comentó sarcásticamente que Neyman debería haber elegido un tema "sobre el cual pudiera hablar con autoridad". Neyman, por su parte, declaró que los métodos de prueba de Fisher eran "en un sentido matemáticamente especificable, peores que inútiles". (Aron & Aron, op. cit, pp. 580-581). Esta historia revela de qué manera, aun tratándose de una ciencia como la estadística, que en su naturaleza parece totalmente ajena a los sentimientos humanos, las creaciones del intelecto humano se pueden ver condicionadas por factores temperamentales y pasiones deleznables.

Aunque estadísticos y psicólogos hemos sido formados a partir de la hibridación de dos enfoques diferentes de la inferencia estadística, y al parecer existe cierta conformidad en el campo con el híbrido resultante, algunos psicólogos, entre los que se cuentan Gigerenzer y Murray (1987), sostienen que los puntos de vista de Fisher y de Pearson & Neyman son fundamentalmente opuestos, y que su errónea combinación no fue más que un matrimonio por conveniencia, basado en el deseo de presentar tanto a la estadística como a la psicología como ciencias basadas en un método de toma de decisiones unificado, mecánico y sin defectos.

El resultado de ese proceso, según Gigerenzer & Murray, es el abandono de la controversia y los métodos alternativos, al igual que textos de estadística "repletos de confusión conceptual, ambigüedad y errores" (Gigerenzer & Murray, op. cit. p. 23).

El autor del presente artículo espera haber contribuido a que el lector conozca la combinación de enfoques opuestos en el proceso de la decisión estadística y la historia oculta detrás de una controversia en apariencia puramente conceptual.

REFERENCIAS

Aron, A. & Aron, E. N. (2001). *Estadística para Psicología*. Buenos Aires: Pearson Education, S.A.

Cohen, J. (1990). Things I Have Learned (so far). *American Psychologist*. *45*, 12, 1304-1312.

Cohen, J. (1994). The Earth is Round. ($p<0.05$). *American Psychologist, 49*, 997-1003.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd Ed.). New York: John Wiley & Sons, Inc.

Cowles, M. (2001). *Statistics in Psychology* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Gigerenzer, G., Krauss, S. Vitouch, O. (2004). The Null Ritual: What you always wanted to know about significance testing but were afraid to ask. En David Kaplan (ed), *The Sage Handbook of Methodology for the Social Sciences*. (cap. 21), (pp. 391-408). Thousand Oaks, CA: Sage Publications, Inc.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Girden, E. R. (1966). *Evaluating Research Articles: From Start to Finish*. Thousand Oaks, CA: Sage Publications, Inc.

Hyman, R. (1964). *The Nature of Psychological Inquiry*. Englewood Cliffs, NJ: Prentice-Hall.

Kerlinger, F. N. & Lee, H. B. (2002). *Investigación del Comportamiento*. (4. Ed.). México, D. F.: McGraw-Hill.

Kraemer, H.C. & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications, Inc.

Leach, C. (1979). *Introduction to Statistics: A Nonparametric Approach for Social Sciences*. New York: John Wiley & Sons, Inc.

McGuigan, F. J. (1993). *Psicología Experimental: Metodos de Investigación* (Sexta edición). México, D.F.: Prentice-Hall.

Murphy, K. R. & Myors, B. (2004). *Statistical Power Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill Book Company.

Stern, P. & Kalof, L. (1996). *Evaluating Social Science Researc*h (2nd Ed.). New York: Oxford University Press.

Tankard, J.W. Jr. (1984). *The Statistical Pioneers*. Cambridge, MA: Schenkman.