Historia de la Probabilidad y la Estadística [IV] (J. Basulto y J. J. García, eds.) Huelva: Servicio de Publicaciones de la Universidad de Huelva
(2009), 351-356

Karl Pearson, el Creador de la Estadística Matemática

M. A. Gómez Villegas
Dpto. de Estadística e Investigación Operativa
Fac. de CC Matemáticas
Universidad Complutense de Madrid

Karl Pearson fue historiador, escribió sobre folklore, fue un socialista convencido, abogado, matemático aplicado, biómetra, estadístico, maestro y biógrafo. Pero sin duda su contribución más importante es al nacimiento de la *Estadística Aplicada*. Es por lo que le debemos el mayor crédito, en frase de él mismo *"Hasta que los fenómenos de cualquier rama del conocimiento no hayan sido sometidos a medida y número, no se puede decir que se trate de una ciencia"*.

Introdujo el *método de los momentos* para la obtención de estimadores, el *sistema de curvas de frecuencias* para disponer de distribuciones que pudieran aplicarse a los distintos fenómenos aleatorios, desarrolló la *correlación lineal* para aplicarla a la teoría de la herencia y de la evolución. Introdujo el *método de la* χ^2 para dar una medida del ajuste entre datos y distribuciones, para contrastar la homogeneidad entre varias muestras, y la independencia entre variables. Fundó los *Anales de Eugenesia* y en 1900, junto con Galton y Weldon, fundó la revista *Biometrika* de la que fue editor hasta su muerte. En una descripción autobiográfica decía *"una explicación para mi vida, se debe a una combinación de dos características que he heredado: capacidad para trabajar mucho y capacidad para relacionar las observaciones de los demás"*.

Datos biográficos

Nace en Londres en 1857 y muere en 1936, su familia es originaria de Yorkshire. Hijo de un abogado, estudia en el University College School. En 1873, a la edad de 16 años fué retirado de la escuela por motivos de salud, y pasa el año siguiente con un preceptor privado. En 1875 obtuvo una beca para el King's College, en Cambridge. El decía que Cambridge le dió, placer en las amistades, placer en las polémicas, placer en el estudio, placer en la búsqueda de nuevas luces, tanto en las matemáticas como en la filosofía y la religión; así como ayuda para mantener su radicalismo científico dentro de límites moderados y razonables. Con 22 años marcha a Alemania y estudia leyes, física y metafísica. Entre 1880 y 1884 es profesor de matemáticas en el King College y en el University College. En 1911 fué el primer profesor de Galton de Eugenesia, la naciente parte de la Biología encargada de los estudios encaminados a conseguir la mejora de las especies. Era un darwinista convencido.

En el año 1890 se producen dos sucesos importantes para la trayectoria científica de Pearson; Galton publica su *Herencia Natural* donde incluye sus trabajos sobre correlación y regresión y

Weldon se incorpora a la cátedra de zoología en el University College de Londres. Los primeros trabajos le van a dotar de una herramienta, con la que cuantificar las medidas de dependencia con la que va a poder contrastar, con resultado positivo, la teoría de la evolución introducida por Darwin. La figura de Weldon le va a permitir trabajar con un biólogo que compartía sus ideas de la evolución y que sería una fuente inagotable de cuestiones, que obligarían a Pearson a ir obteniendo técnicas estadísticas que le permitieran responder a los problemas que Weldon le planteaba. Entre 1891 y 1892 imparte conferencias sobre la geometría de la estadística en el Gresham College, y en ellas introduce los estigmogramas, entigramas, histogramas, cartogramas, stereogramas, etc. Estas lecturas marcan el comienzo de una nueva época en la teoría y en la práctica de la estadística.

Entre 1893 y 1906 publica unos 100 artículos sobre la teoría estadística y sus aplicaciones. La capacidad de investigación de Pearson es asombrosa, a lo largo de su vida publicó más de 650 artículos, fundó junto con Galton y Weldon, en 1901, la revista *Biometrika* para publicar artículos de estadística aplicada a la biología, ese mismo año publica sus *Tablas para Estadísticos y Biometristas* para ayudar en los ajustes de curvas. En 1905 publica el artículo *Sobre la teoría general de la correlación asimétrica y la regresión no lineal.* En 1914 Fisher empieza la polémica con él cuando trata de publicar un artículo en *Biométrika*, sobre el coeficiente de correlación muestral para muestras de una población normal bivariante. El artículo fue referenciado por Weldon como biólogo y por K. Pearson como estadístico y fue rechazado. Posteriormente Fisher diría que su artículo había sido referenciado por un biólogo que no sabía estadística y por un estadístico que no sabía biología.

Para completar la personalidad de K. Pearson, decir que en su primera época, cuando descubre que los valores de la ruleta no son aleatorios, escribe al gobierno francés para que cierre los casinos y dedique el dinero a la Academia de Ciencias, para que se funde un laboratorio de probabilidad, que aplique ésta al problema de la evolución biológica.

Contribuciones de K. Pearson

La primera contribución de K. Pearson que me interesa citar, sobre todo en este contexto, es su serie de conferencias sobre la Historia de la Estadística que dió en el University College de Londres entre los años de 1921 y 1933. Las conferencias fueron recogidas por su hijo Egon Pearson, catedrático de Estadística en el University College también, y que aunque algunas personas no eran partidarias de su publicación sin ser revisadas, constituyen un valioso documento para la historia.

Para hacerse una idea del tipo de trabajo que entraña transcribimos la siguiente cita de la introducción de las conferencias, tomada del prefacio de las conferencias dadas por K. Pearson.

> *Lleva mucho tiempo leer las fuentes originales. En la historia de la estadística muy poca gente se ha tomado la molestia de hacerlo. Yo podría dar muchos ejemplos, de la cantidad de errores que ha propiciado esta conducta, pero me contentaré con poner tres o cuatro.*
>
> 1. *Muchos alemanes llaman a Achenwall el "padre de la estadística", cuando no es así. El aplicaba el término con un significado distinto al que se aplica actualmente.*
> 2. *Hay una curva fundamental en estadística que lleva el nombre de Gauss. Laplace la descubrió diez años antes y su descubridor real fué De Moivre medio siglo*

antes.

3. *Hay un teorema fundamental en estadística que es el teorema de Bernoulli, cuando su descubridor fué también De Moivre.*
4. *Más recientemente, y yo soy en parte culpable, el coeficiente de correlación lineal ha sido atribuido a Bravais, cuándo debiera haberlo sido a Galton.*

La segunda contribución es la familia de curvas de K. Pearson.

La siguiente contribución fue el método de la distancia de la χ^2 para dar una medida del ajuste entre una distribución teórica y una experimental.

El cuarto procedimiento que nos legó Pearson, fue la concrección de la definición del coeficiente de correlación lineal para el estudio de la dependencia estadística y el método de los momentos para determinar los parámetros desconocidos de una distribución, cuando se dispone de una muestra aleatorio simple de la misma.

La familia de distribuciones asimétricas

K. Pearson introduce la familia de distribuciones asimétricas como una alternativa a la distribución normal, que había sido la protagonista ya desde el tiempo de Quetelet. Llega a la familia de distribuciones razonando sobre una mixtura de dos distribuciones normales y concluye que puede haber situaciones en las que los errores de las observaciones no sean normales y por lo tanto se consigan mejores ajustes a situaciones prácticas mediante las mixturas. Los problemas técnicos en los que se ve envuelto son de envergadura, para la determinación de los parámetros se ve forzado a resolver una ecuación de grado 9. Esto es lo que le llevó a Galton a dudar de la corrección del método. No obstante fue, la resolución del problema de la mixtura lo que le hizo abordar el problema de la obtención de distribuciones que permitieran sustituir a la normal para modelizar la incertidumbre.

Introduce la familia de distribuciones en su publicación K.Pearson (1985), mediante la solución de la ecuación diferencial

$$\frac{1}{y}\frac{dy}{dx} = \frac{-x}{c_1 + c_2 x + c_3 x^2}$$

obtiene, para valores convenientes de las constantes, la distribución beta simétrica, la distribución beta asimétrica, la gamma y la normal.

Además para ajustar los parámetros introduce el método de los momentos.

El método de la distancia de la χ^2

Está contenido en una memoria de 1900 y lo introduce para dar una medida del ajuste entre una distribución de probabilidad y una muestra.

La idea es, dada la muestra $(x_1, ..., x_n)$ y la distribución $f(x|\theta)$ construir el estadístico

$$\sum_{i=1}^{k} \frac{(Y_i - y_i)^2}{y_i}$$

que se distribuye χ^2_{k-1}, si la muestra proviene de la distribución. Donde se supone realizada una partición de k elementos en el recorrido de la distribución, con lo que los valores Y_i, las frecuencias observadas de los x_i en el elemento i de la partición, pueden suponerse con distribución

multinomial, e y_i son las frecuencias observadas bajo la hipótesis de que la distribución de la muestra es $f(x|\theta)$.

El procedimiento sería generalizado a los problemas de homogeneidad y a las tablas de contingencia, por el propio K. Pearson y por sus discípulos, Edgeworth y Yule, hasta culminar en los trabajos posteriores de Fisher. Información relevante de esta evolución, puede verse en Stigler (1986), el desarrollo de los métodos puede verse en Gómez Villegas (2005).

El coeficiente de correlación lineal

La medida de la independencia entre dos variables ha tenido una larga historia y ha preocupado, básicamente por su utilidad práctica, a bastantes científicos. Es Galton, el que consigue concretar su definición, aunque todavía incorrecta, pero es K. Pearson el que en dos memorias consigue precisarlo. La primera titulada "Regresión, herencia y panmixia" es de 1896; la segunda escrita en colaboración con Filon "Sobre los errores probables de las frecuencias y su influencia en la selección aleatoria, la variación y la correlación" es de 1898.

En la primera memoria, está incluida con precisión la definición del coeficiente de correlación muestral cómo

$$r = \frac{S_{xy}}{S_x S_y}$$

con $S_{xy} = \frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})$, $S_x^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$ y $S_y^2 = \frac{1}{n}\sum_1^n (y_i - \bar{y})^2$ y también incluye la distribución del coeficiente de correlación poblacional ρ en el caso de una distribución normal bivariante. Curiosamente aplica un razonamiento bayesiano para determinar la distribución del coeficiente de correlación poblacional.

En la diferenciación entre el coeficiente de correlación muestral y poblacional, afirma que r es el estimador más probable de ρ, en concreto enuncia sin demostrarlo, que el valor que maximiza la distribución de probabilidad final que ha obtenido para ρ es el coeficiente de correlación muestral, con lo que anticipa el método de estimación de la máxima verosimilitud que posteriormente desarrollará Fisher.

En el verano de 1933 renuncia a su cátedra y se retira, el University College de Londres divide su cátedra en tres; una de Eugenesia que fue desempeñada por Fisher, una de Estadística que fue desempeñada por Egon Pearson, el hijo de K. Pearson, y una de Biometría. Puede decirse que en ese momento ha sido creada la estadística aplicada cómo un procedimiento para tratar la incertidumbre y para ser aplicada a todas y cada una de las ciencias experimentales.

Un estudio más detallado de la vida y del trabajo de K. Pearson puede consultarse en E. Pearson (1938).

Agradecimientos

Este trabajo ha sido realizado en parte con ayudas del Ministerio de Educación y Ciencia proyecto MTM2005-05462 y de la Comunidad de Madrid-Universidad Complutense proyecto 910395.

BIBLIOGRAFIA

Gómez Villegas, M.A. (2005) *Inferencia Estadística*, Madrid: Díaz de Santos.

Pearson, E.S. (1938) *An Appreciation of Some Aspects of His Life and Work*, Cambridge: Cambridge University Press (existe una traducción de A.Eidlicz (1948) Pearson Creador de la Estadística Aplicada, Buenos Aires: Espasa-Calpe).

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine* 5 th series, **50**, 157-175.

Pearson, K. (1978) *The History of Statistics in the 17 th and 18 th Centuries*, Edited by E.S. Pearson. New York: MacMillan.

Pearson, K. (1895) Contributions to the mathematical theory of evolution, II: skew variation. *Philosophical Transactions of the Royal Society of London*, A, **186**, 343-414.

Pearson, K. (1896) Contributions to the mathematical theory of evolution, III: regresion. heredity and panmixia, *Philosophical Transactions of the Royal Society of London*, A, **187**, 253-318.

Pearson, K. and Filon, L.N.G. (1898) Contributions to the mathematical theory of evolution, IV: on the probable errors of the frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London*, A, **191**, 229-311.

Stigler, S.M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge: Belknap Harvard.