ESTADÍSTICA ESPAÑOLA
Vol. 52, núm 175, 2010, págs. 371 a 391

Importancia de Darwin en el desarrollo de la estadística moderna

por
TONI MONLEÓN-GETINO(*)
Departamento de Estadística
Universidad de Barcelona

RESUMEN

En 2009 se celebró el segundo centenario del nacimiento de Darwin y los 150 años de la publicación del "Origen de las Especies". Darwin fue uno de los más importantes pensadores de todos los tiempos y es considerado junto a Newton el científico británico más significativo. Darwin tuvo una gran influencia en muchas otras áreas del conocimiento como el desarrollo de la correlación y la regresión, propuestas inicialmente por Francis Galton, su primo, fundador de la estadística moderna y que desarrolló estas técnicas para justificar las teorías de Darwin. Este artículo es un pequeño homenaje a Darwin que indirectamente contribuyó al avance de la estadística tal como la entendemos actualmente.

Palabras clave: Historia de la estadística, Darwin, Galton, regresión a la media, correlación, genética, eugenesia.

Clasificación AMS: 01A55.

(*) Mi más sincero agradecimiento a los Drs. Carlos Cuadras Avellana y Mª Carmen Ruiz de Villa del Departamento de Estadística de la Universidad de Barcelona por sus consejos y comentarios en la revisión del artículo.

1. 2009, EL AÑO DARWIN

El año 2009 fue declarado “Año Internacional Darwin” como conmemoración del bicentenario del nacimiento de Darwin así como del 150 aniversario de la publicación de su principal obra, “El origen de las especies” (Darwin, 1859). En ella, se expusieron por primera vez sus ideas sobre la selección natural y la teoría de la evolución de las especies por selección natural, una de las teorías científicas más importantes de la historia. Numerosas publicaciones han aparecido durante 2009 sobre los viajes de Darwin por el mundo, su vida, legado y consecuencias de su teoría.

El aniversario en sí, el 1 de julio de 1859, corresponde al día en que Charles Robert Darwin y Alfred Wallace, presentaron oficialmente sus teorías evolutivas en la Sociedad Linneana de Londres, aunque no idénticas pero sí esencialmente iguales. No lo hicieron en persona, sino a través de dos importantes científicos de la época, el geólogo Charles Lyell (1797-1875) y el botánico y explorador Joseph Hooker (1814 –1879), y a iniciativa de Darwin tras comunicarle a Wallace por carta su trabajo. Tras el interés originado, Darwin publicó el 24 de noviembre de ese año su obra "El origen de las especies", que agotó los 1.250 ejemplares impresos en el primer día de su publicación. Una versión de la obra en la que llevaba trabajando 20 años sin intención formal de publicarla hasta después de su muerte y que estuvo cuestionándose en Down House, la casa en Kent donde Darwin vivió y desarrolló muchas de sus ideas. En ese año celebró su 50 aniversario.

Alfred Russell Wallace (1823–1913), al que también se debe la teoría de la evolución, fue un gran viajero por países tropicales y tuvo la inspiración necesaria para dar forma a sus ideas durante un ataque de malaria (Wallace, 1905), un estímulo quizás comparable al que supuso para Darwin la visión de la asombrosa fauna de las Islas Galápagos más de 20 años antes durante su famoso viaje en el Beagle de 5 años alrededor del mundo (Darwin, 1845), patroneado por el capitán Fitz Roy y que constituyó una de las mayores exploraciones de la época victoriana.

2. LA MEDIDA DE LA RELACIÓN ENTRE DOS VARIABLES, UNA PRIMERA PERSPECTIVA HISTÓRICA

La necesidad de relacionar hechos, fenómenos y variables aparece de modo natural en todos los campos científicos y estudios. En matemáticas es bien conocida la relación funcional entre $y = f(x)$, que expresa una variable dependiente y en función de una variable independiente x. Esta aplicación exacta tiene múltiples aplicaciones en Física, sin embargo en muchos casos (Biología, Medicina, Economía, etc.) las variables no son controlables, sino que toman valores de acuerdo con

una cierta distribución de probabilidad, con variables estadísticas en donde la relación $y = f(x)$ no se puede plantear en términos tan estrictos, ya que si se hiciera así, sólo se tendría una burda aproximación de la realidad. Cuadras (2002) hace una buena introducción desde un punto de vista histórico a este problema que se reproduce en parte a continuación.

Desde un punto de vista histórico la medida de la relación entre variables hay que destacar un gran avance en su estudio a partir de los estudios de Legendre (1752-1833), que introdujo en el S. XVIII el método de los mínimos cuadrados utilizándolos para definir la longitud de 1 metro como una diez millonésima parte del arco meridional. Es a comienzos del XIX cuando se sientan las bases teóricas de la teoría de probabilidades con los trabajos de Lagrange (1736−1813), Laplace (1749−1827), Gauss (1777−1855) y Poisson (1781−1840), pero el nacimiento de la estadística moderna y su uso en el análisis de experimentos se debe a los trabajos de Francis Galton (1822−1911) y Karl Pearson (1857−1936). Con posterioridad a Galton, las propiedades de las técnicas de regresión fueron estudiadas por Edgeworth (1845–1926), Pearson (1857-1936) y Yule (1871-1951).

Para medir la correlación entre dos variables estadísticas X, Y, se utiliza el coeficiente de correlación poblacional ρ o r referido a una muestra (Ver apartado 5). Este coeficiente fue inicialmente propuesto por Francis Galton en 1885 (Figura 1), y perfeccionado por Karl Pearson en 1895. Hoy en día se le conoce como coeficiente de correlación lineal de Pearson. De hecho, bastante antes Friederich Gauss, al extender la distribución que lleva su nombre (Campana de Gauss) al caso de p>2 dimensiones, se encontró con r en uno de los parámetros. La existencia de r ya había sido advertida incluso antes por el astrónomo francés Auguste Bravais (1811-1863) en el caso bivariante (p=2), quien se refirió a r como “una correlation” (Cuadras, 2002).

El coeficiente de correlación más utilizado es el de Pearson, un índice estadístico que mide la relación lineal entre dos variables cuantitativas, una forma de medir la intensidad de la relación lineal entre dos variables. El valor del coeficiente de correlación puede tomar valores desde menos uno hasta uno, $-1 < r < 1$, indicando que mientras más cercano a uno sea el valor del coeficiente de correlación, en cualquier dirección, más fuerte será la asociación lineal entre las dos variables. Este coeficiente será desarrollado y utilizado posteriormente en los apartados 5 y 6.

3. APORTACIONES DE FRANCIS GALTON A LA ESTADÍSTICA

Francis Galton nació el 16 de febrero de 1822 en Sparkbrook, cerca de Birmingham (Inglaterra), siendo el menor de siete hijos de una familia muy acomodada. Su padre era banquero, un abuelo era un miembro de la Royal Society, y su otro

abuelo, Erasmus Darwin fue también el abuelo de Charles Robert Darwin. Su madre, Violetta Darwin Galton, fue la hija de Erasmus Darwin en su segundo matrimonio con Elisabeth Collier. Charles Darwin, 13 años mayor que Galton, era al igual que él nieto de Erasmus Darwin por su primer matrimonio con Mary Howard. Su padre, Samuel Tertius Galton, era el vástago de una añeja y rica familia cuáquera que se convirtió al anglicanismo, un punto importante, ya que como disidentes religiosos no eran admisible en universidades como Oxford o Cambridge. La relación entre Charles Darwin y su primo más joven fue especialmente importante en varios puntos cruciales en la vida de Galton, en particular cuando Galton comenzó a pensar seriamente en la mejora de la humanidad a través de la crianza selectiva (eugenesia). Darwin inició el contacto con Galton tras la lectura del trabajo de su primo "Narrative of an Explorer in Tropical South Africa", sobre 1853. Galton leyó posteriormente The Origin of Species, el cual le produjo un efecto turbulento en su propio pensamiento y el establecimiento de una correspondencia regular con Darwin hasta la muerte de éste (Bulmer, 2003). Esta correspondencia puede verse en http://www.galton.org/letters/darwin/correspondence.htm y un compendio de sus más de 300 publicaciones puede encontrarse en www.galton.org. (Pearson, 1922; Pearson, 1930).

Galton se mostró precoz en matemáticas, pero inicialmente asistió a la universidad para estudiar medicina como era deseo de su padre. Al igual que Darwin, Galton realizó estudios médicos y parece que se adaptó peor a ellos que Darwin. En 1840 viajó a Cambridge para estudiar matemáticas en el Trinity College de la Universidad de Cambridge hasta 1844. Después de sufrir durante los tres años de estudios en Cambridge, tuvo una crisis nerviosa y regresó al campo de la medicina. Después fue aprendiz en el Hospital General en Birmingham, para posteriormente trasladarse al King's College de Londres para continuar su trabajo matemático (Galton, 1908; Bulmer, 2003). Fue definido por Forrest (1974) como un "genio victoriano".

En 1844 murió el padre de Galton, dejándole a él y a sus hermanos una gran herencia. Sus hermanos dedicaron sus vidas a la caza y otras formas de ocio refinado como otros caballeros de la época. La gran herencia de Galton le liberó de la necesidad de trabajar, por lo que en 1845 viajó por todo el mundo. Fue a Egipto, navegó por el Nilo, y cruzó el desierto a Jartum (Sudán). A continuación, hizo él camino a Jerusalén y se instaló cerca de Damasco en Siria. Las cartas que escribió a su familia sugieren que podía haber contraído una enfermedad venérea, mientras viajaba por Oriente Medio. Regresó a Londres en el otoño de 1846 pero en 1850 fue a explorar el sur de África. Según se indica en su biografía, ayudó a resolver las guerras entre los diversos pueblos del sur de África. A lo largo de su viaje hizo diferentes mediciones geográficas y a su regreso a Inglaterra en 1850 recibió una medalla de oro de la Royal Geographical Society. Poco después de recibir el pre-

mio fue elegido miembro de la Real Sociedad. A finales de 1860, Galton concibe la desviación estándar (Pearson, 1922, 1930; Fancher, 1989; FitzPatrick, 1960; Morgan, 1969).

En 1853 se casó con Louisa Butler, quien también provenía de una distinguida familia de intelectuales, y después de una luna de miel en Florencia y Roma, se trasladó al sur de Kensington, donde permaneció casi hasta su muerte en 1911.

Se han publicado numerosos trabajos sobre la vida y obra de Galton como son los de Claves (2001), Forrest (1974, 1995), Pearson (1922, 1930), Fancher (1989), FitzPatrick (1960, 1977), Irvine (1986) y Morgan (1969). También Galton publicó su autobiografía en 1908 (Galton, 1908). Una de sus biografías más modernas puede encontrarse en Bulmer (2003). Las contribuciones estadísticas de Galton están muy bien descritas por Helen M. Walker (1975) y en “The History of Statistics. The Measurement of Uncertainty before 1900” (Stigler, 1986) donde puede encontrarse un capítulo entero dedicada a su obra.

4. INTERÉS POR DARWIN Y OBSESIÓN POR LA EUGENESIA

Eran los años de los primeros pasos de la Genética y de la consolidación de la Antropología. Galton nació el mismo año que Gregor Mendel, el descubridor de las leyes de la Genética aunque olvidado hasta 50 años después. Fue el estudio de la herencia lo que le llevó al descubrimiento de la correlación en el otoño de 1888, como se revisa en “Francis Galton's Account of the Invention of Correlation ” (Stigler, 1989) y al de la regresión en 1989. Es el propio Galton quien describe el descubrimiento de estos dos conceptos en el artículo "Kinship and Correlation” publicado en el North American Review en 1890 (Galton, 1890). Galton fue fuertemente influenciado por su primo Charles Darwin, de tal manera que es el responsable de que Galton se interesara en el estudio de la herencia de los rasgos humanos. Le cautivó especialmente el primer capítulo del “Origen de las especies” sobre la Variación bajo domesticación, relativa a la cría de animales domésticos (Darwin, 1859). Dedicó gran parte del resto de su vida al estudio de las consecuencias de esta obra para las poblaciones humanas. Durante los estudios de Galton inventó palabras tales como la eugenesia o regresión.

Su matrimonio con Louisa Butler no produjo descendencia, lo que desvió su frustración por su propia falta de niños en una obsesión con la eugenesia. Este término fue acuñado por Francis Galton en 1883 (Galton, 1884, 1908, 1865, 1988) y se puede entender como el estudio de la mejora de la raza humana, proporcionando los mecanismos para que las características que se consideran como mejores se desarrollen más rápidamente sobre las que se consideran inadecuadas. Galton acuña el término en 1883 en su libro Investigaciones sobre las facultades

humanas y su desarrollo. Su primera definición fue “El cultivo de la raza, o, como podríamos llamarlo, las cuestiones “eugénicas”, esto es, cuestiones que tratan de lo que se denomina en griego eugenes, o sea, de buena raza, dotado hereditariamente de nobles cualidades. Esta y las palabras relacionadas eugeneia, etc., son aplicables igualmente al hombre, las bestias y las plantas. Deseábamos ardientemente una palabra breve que permitiera expresar la ciencia de la mejora de la materia prima, que de ninguna manera se limita a cuestiones de emparejamientos juiciosos, sino que –y especialmente en el caso del hombre– toma conocimiento de todas las influencias que tienden, aunque sea en el grado más remoto, a dar a las razas o linajes de sangre más adecuados una mayor posibilidad de prevalecer, con más rapidez que lo que normalmente pudieran hacer, sobre los menos adecuados. La palabra eugenesia expresaría suficientemente bien la idea (Galton 1865, p. 104).

Se trata por tanto de dirigir de forma controlada la selección natural. A juicio de Galton, las características físicas, tales como altura, peso y rasgos de personalidad y habilidades son heredadas. Galton pensó que la unión de dos personas inteligentes produciría incluso una persona más inteligente. También creyó que la unión de dos personas altas produciría una persona incluso más alta. Los experimentos posteriores que realizó a lo largo de su vida demostraron que la eugenesia no iba a funcionar, debido justamente a lo que se denominó "la regresión a la media" y que se presenta a continuación. Estas dos ideas y / o teorías sobre la mejora de la raza humana, aún se están debatiendo, pero sus orígenes se remontan directamente a los estudios de Galton. Este término fue utilizado posteriormente durante principios del S. XX por el nazismo y sus teorías de superioridad de la raza aria o la limpieza étnica (Jensen, 2002). Aunque en la actualidad tendemos a asociar la eugenesia más con la genética que con la evolución, en sus orígenes la eugenesia nació al calor del desarrollo de la teoría de la evolución como se indica en el estudio entre eugenesia y evolución (Sotullo, 2006) donde se analizan las relaciones entre el pensamiento evolucionista darwiniano y la eugenesia a través de cuatro personajes especialmente destacados en estos dos campos: Francis Galton (1822-1911), fundador de la eugenesia; Charles Darwin (1809-1882), fundador de la moderna teoría da evolución; Hermann Müller (1890-1967), genetista, evolucionista y eugenista de especial relieve a mediados del siglo veinte y Edward O. Wilson (1929) fundador de la sociobiología. Una perspectiva histórica del control genético en humanos la expone Paul (1995) en su obra *Controlling Human Heredia, 1865 to the Present.*

5. ESTUDIO DE LA CORRELACIÓN Y LA REGRESIÓN COMO PRUEBA DE LAS TEORÍAS DE DARWIN

Galton tuvo la genialidad de darse cuenta que el método estadístico, a través de la teoría de la correlación y de la regresión, proporcionaba el instrumento indispensable para probar las teorías de Darwin (Stigler, 1989; Cuadras, 2002). Supo usar en definitiva la biología para interpretar adecuadamente los resultados que se obtienen de aplicar un método estadístico (Bulmer, 2003).

Algunos trabajos han revisado las aportaciones de Galton a la estadística, especialmente en cuanto a la regresión como en el trabajo de Stanton (2001), "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors", utilizados en la enseñanza habitual de estudiantes en USA.

Galton introdujo en 1889 el término regresión en Estadística. Empleó este concepto para indicar la relación que existía entre la estatura de padres y de hijos. Observó, que si los padres son altos, los hijos generalmente también lo son, y si los padres son bajos los hijos son también de menor estatura. Pero ocurría un hecho curioso: cuando el padre es muy alto o muy bajo, aparece una apreciable regresión hacia la estatura media de la población, de modo que sus hijos "retroceden" hacia la media de sus padres. Galton pudo describir el fenómeno de la regresión hacia la media en sus experimentos sobre el tamaño de las semillas de generaciones sucesivas de guisantes y los publicó en el artículo "Regression Towards Mediocrity in Hereditary Stature" (Galton, 1886). Galton en su artículo "Kinship and Correlation" (1890), describe como éste llego a descubrir el concepto de correlación.

En 1875, Galton realizó un famoso experimento con guisantes usando siete grupos de semillas. Calculó el promedio del diámetro de 100 semillas producidas por cada planta de guisante. Determinó que la más pequeña de las semillas tenía una descendencia de guisantes mayores, mientras que la variedad de semilla mayor tenía descendientes menores. En otro estudio, Galton adquirió los registros familiares de las alturas de 205 grupos de padres y sus hijos adultos. Si los padres eran bajos, sus hijos eran ligeramente mayores, por otra parte si los padres eran altos entonces los hijos eran ligeramente más bajos. Galton observó que las características extremas (por ejemplo, la altura) en los padres no estaban totalmente transmitidas a su descendencia. Por el contrario, la característica en la descendencia tiende hacia una regresión hacia el punto medio (el punto que se ha demostrado matemáticamente como la media). Al medir las alturas de cientos de personas, fue capaz de cuantificar la regresión a la media, y estimar el tamaño de su efecto. Estos dos experimentos llevaron a Galton a inventar la palabra regresión, a la que definió como el proceso de ir hacia la media. En ambos experimentos los guisantes de variedades menores y los padres más bajos tienen hijos que eran más grandes

y cercanos a la media. Los guisantes de variedad mayor y los padres de mayor altura tienen hijos menores pero cercanos a la media.

En la década de 1870 y 1880, Galton fue el pionero en el uso de la distribución normal para adaptarse a los histogramas de los datos tabulados. Inventó el Quincunx o "máquina frijol", un dispositivo concebido como una herramienta para demostrar la ley de error y la distribución normal (Bulmer 2003) que hoy en día se denomina "Pizarra de Galton" (Kacperski, 2005). También descubrió las propiedades de la distribución normal bivariada y su relación con el análisis de regresión (Pearson, 1914-1930; Fancher, 1989; FitzPatrick, 1960, 1977; Morgan, 1969).

Casi 10 años después del experimento con guisantes, Galton inició su primer gran programa de pruebas estadísticas en su laboratorio, situado en el Museo de South Kensington en 1884. Padres e hijos fueron pagados por Galton para medir su altura, peso, potencia de respiración, fuerza de audición, vista y sentido del color. Galton analizó posteriormente los datos, pero se dio cuenta de que necesitaba un método estadístico para medir la correlación entre las características de los padres e hijos. Galton fue apoyado por un joven colega, Karl Pearson de 27 años, en el desarrollo de los métodos estadísticos para el estudio de las diferencias individuales entre padres e hijos. Pearson desarrolló el método matemático de asociación entre variables mediante correlación para medir las capacidades humanas entre un padre y un hijo de las funciones sensoriales y su capacidad intelectual. Pearson realizó un estudio con más de 1.000 registros de grupos familiares observando la relación del tipo regresión lineal:

$$\text{Altura del hijo (cm)} = 85\text{cm} + 0{,}5 \times \text{altura del padre (cm)}$$

La conclusión obtenida fue que los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (regresar) a la media. Lo mismo puede decirse de los padres muy bajos (Pearson, 1922, 1930).

Tal como se ha introducido en el apartado 2, en términos matemáticos, bien fundamentados por K. Pearson dos variables estadísticas X, Y, cuyos valores pueden ser tabulados como:

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	y_n

Pueden estar asociadas linealmente y la medida de la correlación *r* mide el grado de relación lineal entre X e Y, y se define como,

$$r = \frac{s_{xy}}{s_x s_y}, \qquad [1]$$

donde $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$ es la covarianza entre X e Y, y $\bar{x}$, $\bar{y}$, s_x, s_y son las medias y desviaciones típicas de las correspondientes muestras. En realidad *r* puede tomar valores entre -1 y 1, aunque Galton sólo había contemplado el caso de $0 < r < 1$. Cuando *r* es negativo, debe interpretarse que los valores de Y están relacionados con valores bajos de X, y recíprocamente.

El coeficiente r también puede ser definido como:

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \qquad [2]$$

Si los datos proceden de una población entonces ρ puede definirse como:

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right) \qquad [3]$$

Para completar este apartado, sólo citar que en 2009 y con motivo del 150 aniversario del nacimiento de Pearson fue publicado el artículo “Kart Pearson and the Establishment of Mathematical Statistics” (Magnello, 2009) donde se estudia la conexión entre las ideas de Darwin y la fundación de la Estadística Matemática moderna. En el artículo de M. Eileen Magnello (2009) puede verse que fue el zoólogo evolutivo y biometrista inglés W. F. R. Weldon quien conectó las ideas de Darwin con los métodos estadísticos desarrollados por Pearson y Galton (Weldon, 1892a, b) a partir de datos empíricos biológicos. Así Pearson tuvo que desarrollar 1) nuevos análisis de datos, 2) rechazar la curva normal y 3) desarrollar procedimientos para ajustar curvas a datos empíricos por medio de su método de los momentos, entre otros. Las innovaciones estadísticas desarrolladas por Pearson primero con Weldon y posteriormente con Galton le permitieron sentar las bases de la Estadística Matemática moderna.

Weldon comenzó utilizando las técnicas estadísticas desarrolladas por Francis Galton (correlación y regresión) en sus estudios con cangrejos, llegando al convencimiento de que "el problema de la evolución animal es fundamentalmente un problema estadístico". Weldon comenzó a trabajar con el matemático Karl Pearson hasta su obtención de una cátedra de anatomía comparada en 1899. En los primeros años del S XX fue redescubierta la obra de Gregor Mendel, lo que desató un conflicto entre

Weldon y Pearson, por una parte, y William Bateson por otra que era contrario a los biométricos. La polémica afectó a muchos de los aspectos de la naturaleza de la teoría evolutiva formulada por Darwin y del valor del método estadístico. El debate se prolongó intensamente hasta la muerte de Weldon en 1906, aunque la polémica general entre biométricos y mendelianos continuó hasta la fundación de la Síntesis moderna hacia el año 1930 (Provine, 1971; Magnello, 2001).

6. LA FALACIA DE LA REGRESIÓN: EL FENÓMENO DE REGRESIÓN A LA MEDIA

Una vez introducido en apartados anteriores el concepto de correlación, descrito en (1), (2) y (3) como una expresión de "fuerza" entre variables (Apartado 2 y 5); la regresión da lugar a una ecuación que describe dicha relación en términos matemáticos. Así el término regresión fue introducido formalmente por Francis Galton en su libro "Natural inheritance" (1894). Partiendo de los análisis estadísticos de Karl Pearson introdujo el estudio estadístico de las variaciones biológicas y de la herencia. Este libro fue publicado casi 15 años después sus primeros experimentos con guisante donde observó el fenómeno de la regresión a la media y se gestaron los conceptos de la correlación y la regresión en la mente de Galton (Galton, 1890; Stigle, 1989). La obra "Natural inheritance", con un objetivo claramente eugenésico, se centró en la descripción de los rasgos físicos de los descendientes a partir de los de sus padres, concluyendo, como se ha comentado anteriormente, donde padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero los datos revelaban también una tendencia a regresar a la media. Bland et al. (1994a, b, b1) describen la historia de este fenómeno y reproducen algunos de los diagramas originales.

Pearson describe en su cuarto volumen sobre la biografía de Galton la génesis del descubrimiento de la pendiente de regresión (β) a partir del experimento con guisantes (Pearson 1930). Stanton (2001) describe detalladamente este experimento y la primera formulación de la regresión de Galton, facilitando los datos utilizados. Galton distribuyó en 1875 paquetes de semillas de guisante a 7 amigos; cada amigo recibió semillas de un tamaño uniforme (Galton, 1894), pero existía una gran variación de tamaño de semilla entre los diferentes paquetes. Los amigos de Galton plantaron las semillas y las cultivaron hasta obtener nuevas semillas y se las devolvieron (n= 700).

Galton representó gráficamente los pesos de los guisantes de la siguiente generación "hijos" respecto a los pesos de las semillas originales "padre". Galton se dio cuenta que las medias de los pesos de las semillas "hijo" de un tamaño particular de semilla "padre" describía aproximadamente una línea con una pendiente positiva

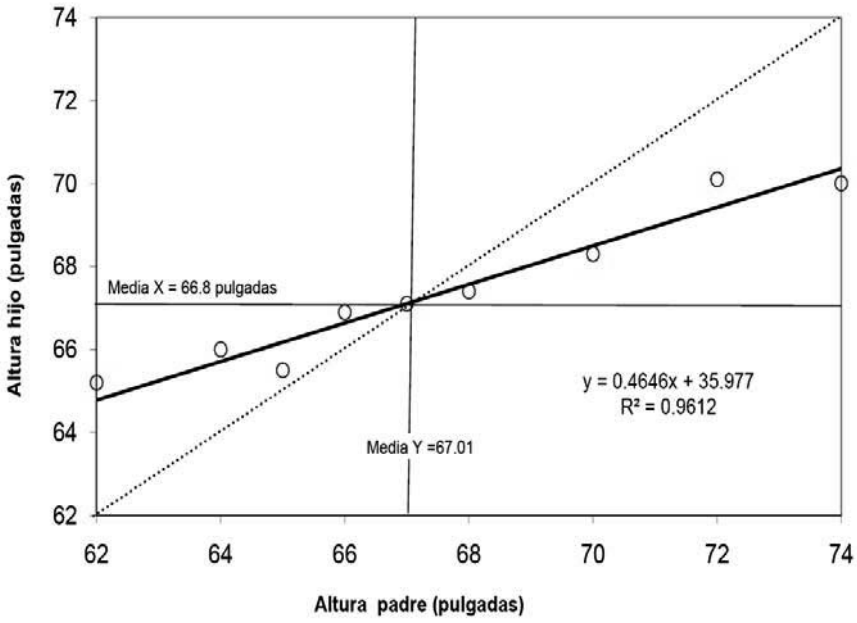
menor que 1. Así se dio cuenta de que existía una línea de regresión, y una variabilidad constante ($s_x = s_y$) entre todos los datos de un carácter con respecto al mismo carácter del segundo. Galton indica "Thus he naturally reached a straight regression line, and the constant variability for all arrays of one character for a given character of a second. It was, perhaps, best for the progress of the correlational calculus that this simple special case should be promulgated first; it is so easily grasped by the beginner" (Pearson 1930).

Galton pensando en el estudio de la heredabilidad de los caracteres entre diversas generaciones demostró que los valores extremos de guisantes en la primera generación (en el eje X) tienden a retroceder hacia la parte media de la segunda generación (en el eje Y). Los datos originales de Galton no produjeron una línea recta perfecta. Galton presentó la primera gráfica de regresión en una conferencia en 1877 (Pearson 1930). La pendiente de esta línea fue designada como "r" de retroceso (Téngase en cuenta que $\hat{\beta} = r$ cuando $s_x = s_y$). Sólo el tratamiento posterior de Pearson hizo desarrollar el coeficiente de correlación r (ver 1) y β (Pearson 1896). Como se ha comentado en el apartado 2, r es conocida como coeficiente lineal de Pearson.

La regresión a la media es un fenómeno biológico bien estudiado y descrito hace mucho tiempo, que consiste en que frecuentemente se observan valores aparentemente fuera de lo normal en las mediciones, que al repetirse la medición el valor vuelve a los límites normales y a continuación exponemos brevemente una explicación estadística del mismo con un ejemplo ilustrativo con su desarrollo matemático. Así en un ejemplo de Pearson reproducido en Ross (2007) sobre la relación entre la altura de padres e hijos, Pearson realizó la comparación entre 10 medidas de padres e hijos tomadas al azar y los resultados de la relación se presentan en la Figura 2. Así cuando la regresión se realiza de los padres a los hijos, entonces son las grandes alturas de los padres las que regresan a la media de todos ellos. De esta manera los autores Bland y Altman en el artículo "Statistic Notes: Regression towards the mean" afirman que: “This is a statistical, not a genetic phenomenon” (Bland et al, (1994a, b, b1)).

Figura 2

FENÓMENO DE REGRESIÓN A LA MEDIA OBSERVADO POR GALTON Y PEARSON ENTRE LA RELACIÓN DE LA ALTURA ENTRE PADRES E HIJOS (EN PULGADAS)



Nota: La línea punteada representa la relación Y=X y la línea completa la línea de regresión entre X e Y obtenida mediante mínimos cuadrados. La línea vertical en X y horizontal en Y representa la media de ambas variables.

Una explicación estadística al fenómeno de regresión a la media y a la formulación de la regresión y la correlación, tal como se desarrolló históricamente, sería: Si $x_1, x_2, \ldots, x_n$ es un primer conjunto de medidas (Ej. altura de los padres) e $y_1, y_2, \ldots, y_n$ (Ej. altura de los hijos) un segundo conjunto, la regresión a la media indica que para todos los *i* valores, el valor esperado de y_i (altura de los hijos) es más cercano al valor de $\bar{X}$ (media de los valores x_i) que a x_i (altura de los padres), como puede verse en la Figura 2. Esto puede ser escrito matemáticamente como:

$$E(|y_i - \bar{x}|) < E(|x_i - \bar{x}|) \qquad [4]$$

Donde E() indica la esperanza matemática. Así se propone la relación:

$$0 \leq E\left(\left|\frac{y_i - \bar{x}}{x_i - \bar{x}}\right|\right) < 1 \qquad [5]$$

[5] es más restrictivo que en la primera desigualdad propuesta [4] ya que necesita que el valor esperado de y_i esté expresado de la misma forma que la media de x_i. Para comprobar esto, si $t = E\left(\left|\frac{y_i - \bar{x}}{x_i - \bar{x}}\right|\right)$ y para n valores puede calcularse el estadístico:

$$\varphi = \sum_{i=1}^{n}\left(\left|\frac{y_i - \bar{x}}{x_i - \bar{x}}\right|\right) \qquad [6]$$

Existe un problema de cálculo en [6], ya que tomando una media aritmética puede observarse que φ no es un buen estadístico, ya que $x_i - \bar{x}$ tiende a 0. Incluso si es cercano a 0, estos puntos pueden dominar el cálculo, por ello la relación *t* no es adecuada y debe ser corregida utilizando $(x_i - \bar{x})^2$:

$$\varphi^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{x})^2} \qquad [7]$$

que puede ser escrita como:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{o como} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad [8]$$

En [8], puede verse que β es la pendiente de la fórmula de regresión con modelo $y = \beta x + \alpha + \varepsilon$. Entonces, se puede asegurar que el fenómeno de regresión hacia la media puede ser interpretado como:

$$0 \leq \beta_{x,y} < 1 \qquad [9]$$

Lo que es cierto para dos conjuntos de medidas en una misma muestra (ej.: altura de padre e hijo). Se esperará que si las desviaciones estándar s_x y s_y de dos conjuntos de medidas relacionadas son iguales, el coeficiente de regresión *β* será igual al coeficiente *r* de correlación. Será suficiente con decir que si se observa

$\beta \leq 1$ se observará un $r \leq 1$. Si la relación lineal entre medidas no es perfecta, esperaremos un coeficiente β < 1. Sin embargo, si las medidas tienen alguna información relevante, r > 0, así $\beta > 0$. r = 1 corresponde al caso de relación perfecta mientras que $r = 0$ corresponderá al caso de relación con error completo.

Se ha de tener en cuenta que $\hat{\beta} = r$ cuando $s_x = s_y$ ya que $\hat{\beta} = r\frac{s_y}{s_x}$.

Como se ve en la Figura 2 la regresión a la media ocurre cuando $0 \leq \beta_{x,y} < 1$. Para valores de x pequeños $\beta x + \alpha > x$ y para valores de x grandes $\beta x + \alpha < x$. Si se asume un modelo de regresión lineal entre la característica del ascendiente Y de los padres y la de los descendientes X, la regresión a la media ocurrirá cuando $0 \leq \beta_{x,y} < 1$ consecuentemente, $E(Y) = \beta x + \alpha$.

En el ejemplo de la Figura 2 se han representado las rectas $y = \beta x + \alpha$ e $y = x$, se ha determinado que $\hat{\beta} = 0.464$ y como $0 \leq \beta_{x,y} < 1$, la recta $y = \beta x + \alpha$ está por encima de $y = x$ para los valores pequeños de x, y está por debajo para los valores altos de x. Por lo que los datos parecen indicar que los padres más altos tienden a tener hijos más altos, también indican que los hijos de padres que son extremadamente altos o extremadamente bajos tienden a aproximarse a la media, más que a sus padres, lo que se conoce como regresión a la media (Ross, 2006; Karylowski, 1985).

Galton creía que la regresión a la media era simplemente una herencia de las características genéticas de los antepasados que no se expresan en los padres; no entendió la regresión a la media como un fenómeno estadístico. En contraste con esta opinión, ahora se sabe que la regresión a la media matemática es inevitable en datos biológicos: si hay alguna variación al azar entre la altura de un individuo y de los padres - si la correlación no es exactamente igual a 1, las predicciones tenderán hacia la media, independientemente de los mecanismos subyacentes de la herencia, la raza o la cultura. Este malentendido fundamental de un fenómeno puramente matemático es un importante factor de motivación en el desarrollo de la eugenesia durante los primeros 30 años del siglo XX. Según Ross (2006) una explicación moderna del fenómeno de regresión a la media se basa en la consideración de que un descendiente (hijo) obtiene una selección aleatoria de la mitad de los genes de cada uno de los padres, un descendiente con uno de los progenitores muy alto tendería a tener genes menos altos que los de dicho progenitor. Este fenómeno también se ha observado en situaciones en las que se dispone de dos conjuntos de datos referidos a las mismas variables (Ej.: diámetros de los guisantes entre diferentes generaciones (Galton 1894), fallecimientos por accidente de tráfico ocurridos en EEUU en 2 años consecutivos (Ross, 2006). Otros trabajos de referencia sobre la regresión a la media son los de Smith (1997) y en James (1973)

puede verse una aplicación de este fenómeno al diseño de estudios clínicos donde no existe grupo control de comparación.

La regresión a la media no se debe a una influencia exterior sino al azar y a veces se le denomina también falacia de la regresión.

La regresión hacia la media ha sido posteriormente estudiada en el contexto de distribuciones normales multivariantes (Müller et al., 2003). Actualmente se ha estudiado este fenómeno desde el punto de vista estadístico. Algunos trabajos aplicados del estudio de la regresión a la media son los de Bland et al. (1994a, b) donde se presentan algunos ejemplos y su estudio.

Un examen riguroso de las publicaciones de Francis Galton y Karl Pearson reveló que el trabajo de Galton sobre la herencia de las características de los guisantes llevó a la conceptualización inicial de la técnica de regresión lineal. Los esfuerzos posteriores por Galton y Pearson fueron por generalizar la técnica a la regresión múltiple (Stanton, 2001).

Finalmente y como se ha mencionado en el apartado 2, existe una cierta polémica histórica acerca del descubrimiento del concepto de la correlación. En 1896, Pearson publicó su primer tratamiento riguroso de la correlación y de regresión en las *Philosophical Transactions of the Royal Society of London* (Pearson, 1896). En este trabajo, no sin una cierta polémica, Pearson señaló que Bravais (1846) había descubierto el momento-producto como método para calcular el coeficiente de correlación, pero sin conseguir una demostración rigurosa que éste siempre ofrecía el mejor ajuste a los datos. Usando una prueba estadística avanzada (que implica una expansión de Taylor), Pearson demostró que los valores óptimos de la pendiente de regresión y el coeficiente de correlación [1] pueden ser calculados a partir del producto-momento (MCV o PMCC) como se transcribe en el artículo Galton, Pearson y los guisantes (Stanton, 2001):

$$r = \frac{\sum_{i=1}^{n} x'y'}{n-1} \qquad [10]$$

donde x' e y' son las “deviances” de los valores observados de sus respectivas medias y n el número de pares observados, así que llevaría a determinar la fórmula determinada por Galton en [2], aunque aquí no se presenta la equivalencia y su desarrollo matemático en profundidad si puede verse en el artículo de Stanton (2001).

CONCLUSIONES

Se ha presentado un breve resumen de la vida y obra de Francis Galton, primo de Darwin que vivió fascinado por las teorías de Darwin sobre la evolución de las especies y obsesionado con la idea de la eugenesia, pero fue capaz de idear dos métodos estadísticos como son la correlación y la regresión, durante la justificación de las teorías de Darwin.

Se recogen las conexiones entre las ideas de Darwin y la construcción de la Estadística Matemática moderna, desarrolladas por Galton y Pearson. Citando unas palabras de Sewall Wright de 1931 sobre Darwin y su influencia en el desarrollo de la estadística moderna "Darwin was the first person to effectively view evolution as primarily a statistical process in which random heredity variation merely furnished the raw material. Pearsonian mathematical statistics was thus built upon Charles Darwin's recognition that species comprised different sets of *statistical populations* underpinned by individual variation" (Magnello, 2009).

W. F. R. Weldon conectó las ideas de Darwin con los métodos estadísticos desarrollados por Pearson y Galton (Weldon, 1892a, b) a partir de datos empíricos biológicos. Weldon comenzó utilizando las técnicas estadísticas desarrolladas por Francis Galton (correlación y regresión) en sus estudios con cangrejos, llegando al convencimiento de que "el problema de la evolución animal es fundamentalmente un problema estadístico" (Magnello, 2009). Galton tuvo la genialidad de darse cuenta que el método estadístico, a través de la teoría de la correlación y de la regresión, proporcionaba el instrumento indispensable para probar las teorías de Darwin (Stigler, 1989). Supo usar en definitiva la biología para interpretar adecuadamente los resultados que se obtienen de aplicar un método estadístico (Bulmer, 2003).

Se ha presentado una breve introducción matemática a la regresión a la media como origen histórico de la teoría de regresión estadística, así como a conceptos generales como el coeficiente de correlación y la covarianza.

El coeficiente de correlación r, motivado para medir las relaciones entre variables biométricas, es posiblemente el índice más utilizado en Estadística y sorprende que su utilidad inicial siga siendo tan vigente, y que sea utilizado en otros campos científicos (Cuadras, 2002).

En España y durante 2009 se han realizado diversas actividades divulgativas para celebrar el segundo centenario del nacimiento de Darwin en 1809 y los 150 años de la publicación del "Origen de las Especies", como las realizadas por la Sociedad Española de Biología Evolutiva (www.sesbe.org).

También a nivel internacional y especialmente en Gran Bretaña, su país de origen, durante 2009 se han realizado numerosos actos, informativos, divulgativos e incluso de fomento de la investigación como pueden verse en www.darwin200.org. Se ha inaugurado el Centro Darwin del Museo de Ciencias Naturales de Londres así como la casa donde vivió, Down House en Kent, que ha sido rehabilitada y convertida en museo. Otras actividades internacionales realizadas durante el año Darwin pueden verse en darwin-online.org.uk. Aunque a nuestro entender tendría que haberse añadidos actos relacionados con las contribuciones indirectas de sus teorías a la estadística, como la regresión y la correlación, de la mano de su primo Galton y de Pearson en cuya obra "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia" (Pearson, 1896) se formuló matemáticamente los cálculos de la correlación y la regresión modernos. Ambos autores son la base de la moderna estadística y del diseño experimental que conocemos actualmente.

Según cuentan los libros de historia William Sealey Gosset ("Student") (1876-1937) trabajó durante un tiempo en el laboratorio Galton, tras contactar con Pearson, así pudo investigar sobre la correlación estadística en 1906 para posteriormente desarrollar su famoso 't-test' dos años después en 1908.

REFERENCIAS

BLAND, J. M., ALTMAN, D. G. (1994a). «Statistic Notes: Regression towards the mean». *British Medical Journal* 308: 1499. PMID 8019287. (Fuente: http://bmj.bmjjournals.com/cgi/content/full/308/6942/1499).

BLAND, J. M., ALTMAN, D. G. (1994b). «Regression towards the mean». *British Medical Journal* 308, 1499.

BLAND, J. M., ALTMAN, D.G. (1994b1). «Some examples of regression towards the mean». *British Medical Journal,* 309, 780.

BRAVAIS, A. (1846). «Analyse Mathematique sur les Probabilites des Erreurs de Situation d'un Point». *Memoires par divers Savans*, 9, 255-332.

BULMER, M. (2003). «Francis Galton: Pioneer of Heredity and Biometry». Johns Hopkins University Press, USA.

BYNUM, W. F. (2002). «The childless father of eugenics». *Science*, 296, 472.

CLAYES, G. (2001). «Introducing Francis Galton, 'Kantsaywhere' and The Donoghues of Dunno Weir». *Utopian Studies*, 12(2), 188-190.

CUADRAS, C. M. (2002). «El coeficiente de correlación y sus extensiones. Las matemáticas del mundo y el mundo de las matemáticas». Francisco R. Fernández (editor). *Edicions Universitat de Barcelona*, 117-130.

DARWIN, C. (1845). «The voyage of the Beagle». John Murray edt. London.

DARWIN, C. (1859). «The origin of species». John Murray edt. London.

DOBZHANSKY, TH. (1937). «Genetics and the Origin of Species». Columbia University Press, New York. 3rd ed., 1951.

FANCHER, R. E. (1989). «Galton on examinations: an unpublished step in the invention of correlation». *Isis,* 80(303), 446-455.

FITZPATRICK, P. J. (1977). «Leading British statisticians of the nineteenth century». M. G. Kendall and R. L. Plackett (eds.), *Studies in the History of Statistics and Probability II,* London, 180-212.

FITZPATRICK, P. J. (1960). «Leading British statisticians of the nineteenth century». *Journal of the American Statistical Association,* 55, 38-70.

FOREST, D. (1995). «Francis Galton (1822-1911)", *R. Fuller (Ed.), Seven pioneers of psychology: Behavior and mind*. London and New Cork, 1-19.

FORREST, D. W. (1974). «Francis Galton: the life and work of a victorian genius». Taplinger. London.

GALTON, F. (1865). «Talento y carácter hereditarios», *Asclepio*, 1984, vol.XXXVI: 191-223.

GALTON, F. (1988). «Herencia y eugenesia». Madrid: Alianza Editorial, S.A.

GALTON, F. (1886). «Regression Towards Mediocrity in Hereditary Stature». *Journal of the Anthropological Institute* 15: 246–263. (Fuente: http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf)

GALTON, F. (1890). «Kinship and Correlation». *North American Review*, 419-431 (1890).

GALTON, F. (1894). «Natural Inheritance». Macmillan and Company. New York.

GALTON, F. (1908). «Memories of my life». Methuen. London.

IRVINE, P. (1986). «Sir Francis Galton (1822-1911)». *Journal of Special Education*, 20(1).

JAMES, K. E. (1973). «Regression toward the Mean in Uncontrolled Clinical Studies». *Biometrics*. 29 (1): 121-130.

JENSEN, A. (2002). «Galton's legacy to research on intelligence». *Journal of Biosocial Science*, 34, 145-172.

KACPERSKI, J. L. (2005). «Galton Board with memory». *Annales UMCS Informatica AI,* 3, 205-211

KARYLOWSKI, J. (1985). «Regression Toward the Mean Effect: No Statistical Background Required». *Teaching of Psychology*, 12, 229-230.

MAGNELLO, M. E., PEARSON, K. (2009). «Establishment of Mathematical Statistics». *International Statistical Review* (77)1: 3–29.

MAGNELLO, M. E., WELDON, W. F. R. (2001). «Statisticians of the Centuries» (ed. C. C. Heyde and E. Seneta): 261-264. New York: Springer.

MORGAN, R. W. (1969). «Sir Francis Galton (1822-1910), in Some nineteenth century British scientists». *Oxford*, Oxford (UK), 65-95.

MÜLLER, H.G., ABRAMSON, I., AZARI, R. (2003). «Nonparametric regression to the mean». *PNAS* (100) 17, 9715–9720.

PAUL, D. B. (1995). «Controlling Human Heredity, 1865 to the Present». *Atlantic Highlands, Humanities Press*. N.J.

PEARSON, K. (1914-1930). «The life, letters, and labours of Francis Galton», four vols, Cambridge University Press. London.

PEARSON, K. (1896). «Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia». *Philosophical Transactions of the Royal Society of London*, 187, 253-318.

PEARSON, K. (1922). «Francis Galton: A Centenary Appreciation». Cambridge University Press. London.

PROVINE, W. B. (1971). «The Origins of Theoretical Population Genetics». University of Chicago Press.

ROSS, S. M. (2007). «Introducción a la estadística». Editorial Reverte. Madrid.

SIMONTON, D. K. (2003). «Francis Galton's Hereditary Genius: Its place in the history and psychology of Science». *The anatomy of impact: What makes the great works of psychology great.* RJ. Sternberg (Ed.), American Psychological Association. Washington DC. 3-18.

SMITH, G. (1977). «Do Statistics Test Scores Regress Toward the Mean?». *Chance* 10(4).

SOUTULLO, D. (2006). «Evolución y Eugenesia». *Ludus Vitalis*, vol. XIV, num. 25, 2006, pp. 25-42.

STANTON, J. M. (2001). «Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors». *Journal of Statistics Education* 9(3). (Fuente: http://www.amstat.org/publications/JSE/v9n3/stanton.html)

STIGLER, S. M. (1986). «The History of Statistics. The Measurement of Uncertainty before 1900». Capítulos 8-10. Harvard University Press.

STIGLER, S. M. (1989). «Francis Galton's Account of the Invention of Correlation». *Statistical Science* Vol. 4, No. 2 (May, 1989), pp. 73-79.

STIGLER, S. M. (1996). «Kinship and Correlation (1890)». *Statistical Science*, 1989, vol. 4, Noo 2, 73-86.

WALLACE, A. R. (1905). «My Life». Chapman & Hall. London.

WALTER, H. M. (1975). «Studies in the History of Statistical Method: Special Reference to Certain Educational Problems (History, philosophy, and sociology of science)». Capítulo 5, Ayer Co Publications.

WELDON, W. F. R. (27 November 1892a). Letter to Francis Galton. FG:UCL/293/A.

WELDON, W. F. R. (27 November 1892b). Letter to Karl Pearson. KP:UCL/891/A.

WELDON, W. F. R. (1902). «Mendel's law of alternative inheritance in peas». *Biometrika*, 1, 1–50.

IMPORTANCE OF DARWIN IN THE DEVELOPMENT OF MODERN STATISTICS

ABSTRACT

In 2009 we celebrate the second centenary of the birth of Darwin and 150 years of publishing the "Origin of Species." Darwin was one of the most important thinkers of all time and is considered by the British scientist Newton's most important all times. Darwin had a great influence on many other areas of knowledge as the development of correlation and regression, proposed by Francis Galton, his cousin, founder of the modern and developed statistical techniques to justify the theories of Darwin. This article is a tribute to Darwin that indirectly contributed to the advancement of statistics as we understand it today.

Keywords: History of Statistics, Darwin, Galton, regression towards the mean, correlation, genetics, eugenics.

AMS Classification: 01A55.