CAPÍTULO 1

LOS SISTEMAS DE PEARSON COMO GENERADORES DE DISTRIBUCIONES DE PROBABILIDAD. APLICACIONES ESTADÍSTICAS Y ECONÓMICAS

RAFAEL HERRERÍAS PLEGUEZUELO
JOSÉ CALLEJÓN CÉSPEDES

Departamento de Métodos Cuantitativos para la Economía y la Empresa
Facultad de Ciencias Económicas y Empresariales
Universidad de Granada

Una forma de abordar el estudio global y conjunto de un gran número de distribuciones de probabilidades, se consigue con la formulación de sistemas de distribuciones que verifican una determinada ecuación funcional, bien diferencial para las distribuciones de tipo continuo, o bien en diferencias finitas para las distribuciones de tipo discreto. Son varios los objetivos que se consiguen con esta formulación, a saber:

a) Aplicación sistemática de diversas metodologías en la generación de familias de distribuciones
b) La obtención sistemática de características comunes a cada familia: relaciones de recurrencias entre momentos, clasificación de distribuciones etc.

Los sistemas de distribuciones continuos univariantes más estudiados por sus aplicaciones son el de Pearson, y la familia exponencial, Loève (1963) Zacks, (1971). La relación entre ambas familias puede verse, con todo detalle, en Herrerías (1975)

El sistema de Pearson responde a la ecuación diferencial lineal de primer orden y homogénea, Cansado(1950), Elderton,(1969) y Jonhson-Kotz, (1970):

$$\frac{f'(x)}{f(x)} = \frac{x-a}{b_0 + b_1 x + b_2 x^2} \tag{1}$$

Es sobradamente conocido que la mayoría de las distribuciones continuas univariantes satisfacen la anterior ecuación diferencial (1) y que el conocimiento de los cuatro primeros momentos muestrales es suficicente para estimar los cuatro parámetgros que aparecen en la expresión (1) y de esta forma se dispone de una densidad, aunque por tratarse de una estimación realizada por el método de los momentos, se conoce que tales estimaciones, en general, no son eficientes.

Obsérvese que el primer miembro de (1) puede expresarse por

$$\frac{d \ln y}{dx} = D_x \ln y = \frac{D_x y}{y}$$

donde D_x representa el operador lineal derivada.

En la literatura especializada se encuentran otros sistemas que extienden y generalizan las distribuciones de (1); dos ejemplos significativos son la extensión de L.K. Roy, (1971), y la generalización de Herrerías, (1975).

La extensión de Roy se obtiene cambiando el segundo miembro de (1), que pasa a ser,

$$\frac{f'(x)}{f(x)} = \frac{a_0 + a_1 x + a_2 x^2}{x(b_0 + b_1 x + b_2 x^2)} \tag{2}$$

Como se observa, la idea es "complicar" el segundo miembro de (1), de manera que se considera una fracción en la que el numerador es un polinomio de segundo grado, en vez de primero, y el denominador es un polinomio de tercer grado sin término independiente, en vez de un polinomio de segundo grado como en (1). Con esta nueva formulación se dan cabida a cinco tipos de distribuciones de probabilidad que no satisfacen a (1) pero que, por verificar (2), tienen un tratamiento análogo a las distribuciones que cumplen (1). En particular se obtiene una relación de recurrencia entre momentos y, conocidos los siete primeros momentos muestrales, pueden estimarse los parámetros a_i y b_i de (2).

La generalización de Herrerías se obtiene cambiando el primer miembro, al utilizar el operador aleph[1], en vez del operador derivada ordinaria. Los sistemas obtenidos a partir de la generalización de Herrerías responden a la ecuación diferencial que se puede escribir[2]

$$\frac{f'(x)}{f(x)} = \frac{x - a}{b_0 + b_1 x + b_2 x^2} + \varphi(x) \tag{3}$$

que es un caso particular del sistema de Pearson generalizado, definido mediante

[1] Derivada generalizada, Aldanondo (1969), $a_f^g[y(x)] = \frac{y'(x) - g(x)y(x)}{f(x)}$. Obsérvese que el operador derivada ordinaria D, es a_1^0 .

[2] En Herrerías (1975) se hace notar que si φ(x) = 1/x entonces (3) se transforma en (2), por lo que las nuevas distribuciones de (2), que no son de (1), son un caso particular de (3).

$$\frac{a_f^{\varphi}(y)}{y} = \frac{F(x) - a}{b_0 + b_1 F(x) + b_2 [F(x)]^2}$$

donde $F(x) = \int^x f(t)dt$

El siguiente paso consiste en extender el sistema de Pearson a dos dimensiones; esto se consigue con la familia pearsoniana bivariante, definida por van Uven, (1947),

$$\begin{cases} \dfrac{\partial}{\partial x} \ln f(x,y) = \dfrac{L_1(x,y)}{Q_1(x,y)} \\ \dfrac{\partial}{\partial y} \ln f(x,y) = \dfrac{L_2(x,y)}{Q_2(x,y)} \end{cases} \tag{4}$$

donde L_1 y L_2 son formas lineales y Q_1 y Q_2 son formas cuadráticas, y donde además se ha de verificar la igualdad de Schwartz para la integrabilidad del sistema

$$\frac{\partial}{\partial y} \frac{L_1(x,y)}{Q_1(x,y)} = \frac{\partial}{\partial x} \frac{L_2(x,y)}{Q_2(x,y)}$$

Finalmente, Steyn (1960), realiza la extensión del caso bivariante al multivariante, a través de un sistema de igual número de ecuaciones que de variables componen el vector aleatorio.

Por otra parte, Fernández (1979), extiende el sistema de van Uven, utilizando funciones cuadráticas y cúbicas para el numerador y el denominador, respectivamente, del segundo miembro de las ecuaciones que definen el sistema, a la manera de lo hecho por Roy en el caso univariante.

En estudios posteriores, Herrerías, Palacios y Callejón, (1998), siguiendo la metodología pearsoniana, establecen los conceptos de función generadora de una distribución de probabilidad, y de sistema de funciones generadoras para distribuciones multivariantes.

La expresión (3) sugiere el concepto de función generadora de una distribución univariante continua. Manteniendo el primer miembro de la ecuación (3), se sustituye el segundo por una función g(x) a la que sólo se le exigen las condiciones necesarias y suficientes para que *f(x)* sea una función de densidad de una variable aleatoria, sobre un recinto determinado. Aparece de esta forma el concepto de función generadora de una distribución univariante continua.

Si *f(x)* es la función de densidad de una variable aleatoria y [a,b] es su dominio de definición, se puede calcular la correspondiente función generadora, que viene dada por:

$$g(x) = \frac{f'(x)}{f(x)}$$

Las propiedades de estas funciones generadoras permiten su utilización directa, como herramienta, en distintas aplicaciones estadísticas: estudio de la independencia entre variables aleatorias a partir de la observación directa del sistema de funciones generadoras, Herrerías, Palacios y Callejon (1997), (1998) y (1999); cantidad de información, Herrerías, Palacios, Pérez y Callejon (1998); estudio de la relación entre los métodos de estimación de los momentos y de la máxima verosimilitud, Callejon y Santos (1994), etc.

Palacios, Herrerías y Callejón (2000), realizan una generalización del sistema de Pearson mediante splines, dejando patente la flexibilidad que permite la utilización de los elementos de la familia a la hora de la modelización de los datos muestrales. La familia generalizada mediante splines viene definida por la ecuación

$$f'(x)=\frac{x-a_0+\sum_{i=1}^{k}c_i\left(x-x_i\right)_+}{\sum_{i=1}^{3}d_i x^{i-1}+\sum_{i=1}^{k}d_{i+3}\left(x-x_i\right)_+^2}\,f(x)$$

donde

$$a<x_1<x_2<\ldots\ldots<x_k<b\,,\quad 0<p_1<\ldots\ldots<p_k<1$$

y

$$\left(x-x_i\right)_+^r=\begin{cases}\left(x-x_i\right)^r & si \quad x>x_i\\ 0 & si \quad x\leq x_i\end{cases}$$

Del mismo modo que la ecuación que define la familia de Pearson sugirió la idea de función generadora de una distribución continua univariante de probabilidad, las extensiones de van Uven y Steyn permiten establecer el concepto de sistema de funciones generadoras de una distribución continua multivariante de probabilidad, que resulta estar formado por las funciones generadoras de cada una de las distribuciones condicionadas, lo que permite obtener la función de densidad conjunta como solución de una ecuación diferencial, sin el concurso de las distribuciones marginales

En el caso discreto univariante, J.K. Ord (1967), introdujo una ecuación en diferencias finitas, lineal de primer orden y homogénea, para generar funciones de cuantía de manera análoga al sistema de Pearson, que responde a la expresión:

$$\frac{\Delta p_{r-1}}{p_{r-1}}=\frac{a-r}{b_0+b_1r+b_2r(r-1)} \tag{5}$$

de la cual se deduce una relación recurrente entre los momentos factoriales recurrentes y puede comprobarse fácilmente que la mayoría de las distribuciones univariantes discretas que se estudian en los distintos textos de Estadística, satisfacen la ecuación (5).

Este sistema fue ampliado por Herrerías (1976), manteniendo el paralelismo con la extensión de Roy en el caso continuo; la ecuación que define la extensión de Herrerías en el caso discreto es:

$$\frac{\Delta p_{r-1}}{p_{r-1}} = \frac{a_0 + a_1 r + a_2 r(r-1)}{b_0 + b_1 r + b_2 r(r-1) + b_3 r(r-1)(r-2)} \tag{6}$$

El concepto de función generadora de una distribución univariante discreta de probabilidad supone que el segundo miembro de (5) es una función a la que sólo se le exigirán las condiciones necesarias para que al resolver la ecuación en diferencias se obtenga una distribución de probabilidad. Evidentemente, no se modifica el concepto de función generadora, con respecto al caso continuo, en el sentido de que se define igual al primer miembro de la ecuación funcional correspondiente, Callejón (1995).

El sistema de distribuciones bivariantes discretas de tipo Pearson, análogo al estudiado por van Uven en el caso continuo, es estudiado por Herrerías - Cobos (1984) y Herrerías – Calvete (1985) y (1986), planteando el sistema de ecuaciones en diferencias finitas que lo define, mostrando distribuciones que pertenecen a él, y encontrando una solución general basada en las relaciones de recurrencia de los momentos factoriales descendentes.

Fajardo (1985), realiza la generalización del sistema de Pearson discreto univariante; parte de una ecuación en diferencias finitas y obtiene la solución general, caracterizando las funciones generatrices de probabilidad asociadas a las soluciones de la mencionada ecuación y estudia algunas familias de distribuciones discretas pertenecientes a este sistema.

Parece oportuno resaltar en este primer capítulo de la monografía algunas de las propiedades que se pueden estudiar, a partir de los sistemas de funciones generadoras de probabilidad, en Estadística Matemática, tales como la dependencia o independencia estocástica, la regresión en distribuciones con condicionadas simétricas, la cantidad de información de Fisher o la relación entre los métodos de estimación de la máxima verosimilitud y de los momentos.

En cuanto al estudio de la dependencia o independencia estocástica, esta metodología prescinde de las distribuciones marginales y de las dificultades de cálculo para obtenerlas y resulta mucho más corta y sencilla que la habitual. Se pone de manifiesto que la independencia estocástica se puede observar a partir de la dependencia analítica de las funciones generadoras de las distintas variables, (Herrerías, Palacios y Callejón, 1998).

Mediante un simple análisis de la derivada del logaritmo de la función de densidad (caso continuo) o de la razón entre dos valores de la función de cuantía (caso discreto), pueden extraerse todas las relaciones de interdependencia estocástica entre las componentes de un vector aleatorio. Como consecuencia de la simplicidad del método se tiene un instrumento valioso, desde el punto de vista didáctico, para enseñar al alumno a chequear la independencia estocástica de variables aleatorias. Consideramos interesante que los enseñantes de la Estadística tengan conocimiento de estos resultados para el aprovechamiento didáctico de los mismos en su quehacer diario.

La utilización como herramienta de las funciones generadoras permiten concluir que, si una distribución continua bivariante tiene las distribuciones de y condicionadas a x simétricas, entonces, la función de regresión de y sobre x se obtiene al igualar a cero la segunda función generadora de la distribución conjunta. Es decir al resolver en y la ecuación que resulta de igualar a cero la derivada parcial con respecto a y, del logaritmo de la densidad conjunta. Análogo razonamiento se puede realizar para la función de regresión de x sobre y, Herrerías, Palacios y Callejón, (1997).

La utilización de la metodología propia del sistema de Pearson y, por tanto, de las funciones generadoras, permite ajustar varios modelos probabilísticos para la distribución de la renta, empleando dos procedimientos diferentes aunque relacionados, Herrerías, Palacios y Callejon (1996) y (1997). (a) Mediante el sistema de Pearson de distribuciones continuas univariantes, se estiman los parámetros a través del método de los momentos, para lo cual se utilizan los cuatro primeros momentos muestrales respecto al origen. (b) Utilizando una función generadora polinómica, cuyos parámetros se estiman de forma análoga al apartado anterior. El grado se selecciona mediante comparación con la función de distribución empírica, Herrerías, Palacios y Ramos (1996) y Herrerías, Palacios y Pérez (1997).

Además, la generación permite la formulación de las medidas de concentración, basada en la estructura probabilística subyacente de las variables que permiten la cuantificación de la desigualdad económica y de otros conceptos afines, como lo son la pobreza, la igualdad o el bienestar social, etc.

Tanto la curva de Lorenz como el índice de Gini son medidas habituales en el estudio de la desigualdad social y en la medida de la pobreza de una población, como indica Sen (1973), entre otros. Por otra parte, la relación entre ambas herramientas es muy fuerte como ponen de manifiesto Calot (1967), Kendall y Stuart (1977), Casas y Núñez (1987), entre otros, si bien parece que el índice de Gini es más apto como medida sintética de tipo cuantitativo, mientras que la curva de Lorenz es más descriptiva puesto que permite el estudio detallado del reparto efectuado de los recursos de un modo más pormenorizado.

Resulta interesante la idea de caracterizar la curva de Lorenz a partir de una serie de propiedades que permitan obtener una gama coherente de formas funcionales para estimarlas a partir de una distribución de frecuencias observada. En este sentido se puede reseñar el trabajo de Casas-Núñez (1987) en el que se obtiene una condición necesaria para las curvas de Lorenz, que satisfacen las curvas exponencial y potencial. Así pues, existen dos posibles líneas de investigación: la primera trataría de obtener una condición suficiente que encauce el estudio de una posible caracterización de las curvas de Lorenz, y la segunda avanzar en la búsqueda de formas distribucionales coherentes con la condición necesaria allí desarrollada, para disponer de un conjunto de ellas que permitan estimar razonablemente las curvas de Lorenz.

Casas, Herrrerías y Núñez, (1990) proponen dos formas de obtención de funciones que modelizan la curva de Lorenz, una mediante combinaciones lineales convexas de formas funcionales, utilizadas en la estimación de dicha curva por otros autores y otra, utilizando la ecuación diferencial de Pearson, que genera la familia de distribuciones continuas univariantes clásica, Lafuente, M. (1994). Posteriormente se establecen las condiciones necesarias y suficientes para que una función que que verifica la ecuación diferencial de Pearson, (1), sea generadora de una curva de Lorenz, Herrerías, Palacios y Callejon (2001).

El método de los momentos, utilizado para estimar la función de densidad de una distribución continua univariante que pertenece a la familia de Pearson, también puede utilizarse para, obtener la correspondiente curva de Lorenz.

Por otra parte, García, R. M y Herrerías, J.M. (2001) han logrado incluir curvas de Lorenz en las funciones generadoras, lo que prueba que las aplicaciones económicas de las funciones generadoras no han hecho nada más que empezar.

BIBLIOGRAFÍA[3]

ALDANONDO, I. (1969). Ecuaciones diferenciales lineales de coeficientes constantes en derivadas parciales en el operador "aleph". Publicaciones del Seminario de Matemáticas García de Galdeano, Vol. 10, pág. 47 - 70. Universidad de Zaragoza.

CALLEJÓN, J. (1995). *Un nuevo método para generar distribuciones de probabilidad. Problemas asociados y Aplicaciones*. Tesis Doctoral. Universidad de Granada. ETD Micropublicaciones S.L.

CALOT, G. (1967). *Cours de Calcul des Probabilités.* Dunot. Paris.

CANSADO, E. (1950). Exposición sistemática de las distribuciones de Pearson. *Trab. de Estadística*. Vol 1. Cuad. III, pág. 279 - 287.

[3] Todas las referencias bibliográficas señaladas con asterisco se encuentran recogidas en el texto: “Aplicaciones Estadísticas y Económicas de los Sistemas de Funciones Generadoras”. Editores Herrerías, R; Palacios, F y Callejón, J (2001). Editorial Universidad de Granada

* CASAS, J.M., y NÚÑEZ, J.J. (1987). Algunas consideraciones sobre las medidas de concentración. Aplicaciones. *Actas del I Congreso ASEPELT-ESPAÑA*. Barcelona, pp. 49-62.

* CASAS, J.M., HERRERÍAS, R. y NÚÑEZ, J.J. (1990). Familias de formas funcionales para estimar la curva de Lorenz. *IV Reunión Anual ASEPELT-ESPAÑA*. Murcia.

ELDERTON, W.P. y JOHNSON N.L.(1969), *System of frequency curves*. Cambridge University Press.

FAJARDO, M.A. (1985). *Generalizaciones de los sistemas pearsonianos discretos*. Tesis doctoral. Universidad de Extremadura.

FERNÁNDEZ, F. (1979). Una extensión del sistema de Pearson bivariante o sistema de van Uven. Publicaciones de la Facultad de Ciencias de Granada.

* GARCÍA R.M y HERRERÍAS J.M. (2001). Inclusión de curvas de Lorenz en el sistema de Pearson. *Modelos de generación de distribuciones: propiedades y aplicaciones*. Granada.

HERRERÍAS, R.(1975). Sobre las estructuras estadísticas de Pearson y exponenciales, problemas asociados. *Publicaciones de la Facultad de Ciencias de Granada*.

* HERRERÍAS R. (1976). Extensión del sistema de distribuciones discretas de Pearson. *Cuadernos de Estadística Matemática*, vol 4, pp. 30-36

* HERRERÍAS, R. y CALVETE, H. (1986). Un sistema de distribuciones discretas bivariantes. *Estadística Española*, nº 109, P pp. 15-28

* HERRERÍAS, R. y CALVETE, H. (1986). Estudio del sistema de distribuciones de probabilidad bivariantes discretas del tipo Pearson-Ord. *Cuadernos Aragoneses de Economía*, nº 10, pp. 81-86

* HERRERÍAS, R. y COBOS, J. (1984). Solución General para un tipo de sistemas de distribuciones de probabilidad bivariantes discretas. *Cuadernos Aragoneses de Economía*, nº 8, pp. 133-137

* HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J. (1996) Distribución de la renta en la provincia de Valladolid: dos métodos de estimación. Valladolid; Hoy y Mañana. Pendiente de publicación.

* HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J. (1996) Distribución de la renta en la Comunidad de Castilla y León: dos métodos de estimación. *5º Congreso de Economía Regional de Castilla y León. Comunicaciones 2*, pp. 978-990

HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J.(1997) A new condition to check the independence of random variable. *Communications in Statistics. Theory and Methods*, vol 26-2, pp. 421-428

* HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J. (1997) Aproximación a la regresión en distribuciones con condicionadas simétricas. *XI Reunión Anual de la Asociación Científica Europea de Economía Aplicada, ASEPELT-ESPAÑA. CD ROM*

HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J. (1998) Using generator function system to check independence of random variables. *Communications in Statistics. Theory and Methods*, 27-2, pp. 1211-1219

* HERRERÍAS, R., PALACIOS, F. y CALLEJÓN, J. (1999). Un método sencillo para enseñar la dependencia o independencia estocástica entre variables aleatorias

Conferencia Internacional Experiencias e Perspectivas do Ensino da Estatistica, pp. 59-67

* HERRERÍAS, R., PALACIOS, F. y PÉREZ, E. (1997). Una alternativa de cálculo para mejorar el ajuste de una distribución polinómica de la renta. *XI Reunión Anual de la Asociación Científica Europea de Economía Aplicada, ASEPELT-ESPAÑA. Actas en CD ROM*

* HERRERÍAS, R., PALACIOS, F., PÉREZ, E y CALLEJÓN J. (1998). Sistema de funciones generadoras y cantidad de información de Fisher. *XII Reunión Anual de la Asociación Científica Europea de Economía Aplicada, ASEPELT-ESPAÑA. Actas en CD ROM.*

* HERRERÍAS, R., PALACIOS, F. y RAMOS, A (1998) Una metodología flexible para la modelización de la distribución de la renta *Décima Reunión ASEPRLT-ESPAÑA, CD-ROM, fichero G 25*

JOHNSON, N.L. y KOTZ, S (1970). Distributions in Statistic: Continuous univariate distributions-1. John Wiley & Sons.

KENDALL, M. & STUART, (1977). A. *The Advanced Theory of Statistics.,* Vol. 1, 4ª ed. Griffin.

LAFUENTE, M. (1994). *Medidas de cuantificación de la Desigualdad: La Desigualdad de la Renta en España según la E.P.F. 1990-91*. Tesis Doctoral. Universidad de Murcia.

LOÈVE, M. (1963*). Probability Theory*. Van Nostrand Company. New York.

ORD,J.K. (1967). On a system of discrete distributions. *Biometrika,* 54, 649-656.

* PALACIOS, F., HERRERÍAS, R. y CALLEJÓN J. (2000). Sistema de Pearson generalizado mediante splines. *XIV Reunión Anual de la Asociación Científica Europea de Economía Aplicada, ASEPELT-ESPAÑA. Actas en CD ROM.*

ROY, L.K. (1971) An extension of the Pearson system of frecuency curves. *Trabajos de Estadística e I.O. Vol XXII. Cuad. 1 y 2.* Pág 113-123.

SEN, A. (1973). *On economic inequality*. Oxford Univ. Press.

STEYN, H.S. (1960). On regression properties of discrete systems of probability functions. *Proceedings of the Royal Academy of Sciences*, Amsterdam. Vol. 63, 302-311.

UVEN, M.J. van. (1947-48). Extensions of Pearson's probability distributions to two variable. *Proceedings of the Royal Academy of Sciences. Amsterdam*, 50, 1063-1070 and 1252-1264; 51, 41-52 and 191-196.

ZACKS, S. (1971). *The theory of statistical inference*. Wiley.