COMUNICACIONES A LA PONENCIA IV

LA TEORÍA DE LA INFORMACIÓN EN EL ANÁLISIS DE DATOS CUALITATIVOS

Xabier de Salvador González
Universidade da Coruña

A partir de los trabajos de Shannon-Weaver y McMillan a mediados de este siglo, se desarrolló la llamada **Teoría matemática de la Información** con origen en el concepto termodinámico de *desorden* o *entropía*. Los conceptos y resultados de esta teoría han tenido un amplio dominio de aplicaciones en el campo de las ciencias humanas y sociales (lingüística, psicología,...). Si bien ha sido R. A. Fisher quien introduce en 1925 el término de *información* en estimación estadística (cantidad de información proporcionada por los datos sobre un parámetro no conocido), las ideas de Shannon-Weaver, más comprometidas con la probabilidad, también fueron asumidas en estadística.

Nos interesa aquí esta teoría en su aplicación al análisis de datos cualitativos. Fueron varios los trabajos realizados bajo este objetivo. Entre ellos destacan los de McGill (1954) y Kullback (1968) en análisis multivariable de tablas de contingencia para medidas y tests de asociación, Garner (1956) en análisis de varianza, Krippendorff (1982,1986) en análisis de regresión y modelos estructurales.

Todos estos estudios parten de los conceptos básicos de entropía y cantidad de información que una variable transmite a otra definidos por Shannon-Weaver:

Para una variable discreta X con m categorías, se define la **entropía** de X, representada por **H**(X), como medida de incertidumbre, por la expresión

$$H(X) = \sum_{i=1}^{m} (P_i \cdot \log_2 P_i)$$

siendo P_i la probabilidad de que la variable X tome el valor i P[X=i], con i=1,...,m. Esta medida del grado de indeterminación que presenta una variable es no-negativa y presenta su valor máximo en una distribución equiprobable de sus valores (distribución uniforme), aumentando con el número de éstos. Se recoge así el *principio de razón insuficiente* (o principio de indiferencia) utilizado ya en el siglo XVII por Bernouilli y establecido por Laplace en el siglo XIX: supuesto que no hay fundamento para preferir una modalidad a otra de un acontecimiento, debemos considerar que tienen la misma probabilidad de ocurrencia.

Considerando dos variables discretas X (con categorías i) e Y (con categorías k) y su tabla de contingencia X*Y, podemos definir, aparte de sus respectivas entropías (entropías marginales), las entropías condicionales (bajo las distribuciones de X condicionada a Y, e Y condicionada a X) y la entropía total (bajo la distribución conjunta de ambas variables):

entropía condicionada:

$$H(Y|X) = \sum_{i=1}^{m} \sum_{k=1}^{n} (P_{ik} \cdot \log_2 P_{k|i})$$

entropía total:

$$H(XY) = \sum_{i=1}^{m} \sum_{k=1}^{n} (P_{ik} \cdot \log_2 P_{ik})$$

siendo P_{ik} la probabilidad conjunta P[X=i, Y=k] y P_{kli} la probabilidad condicionada P[Y|X=i], con $P_{ik} = P_{kli} \cdot P_i$.

La propiedad de la disminución de la entropía

$$\mathbf{H}(Y|X) \leq \mathbf{H}(Y)$$

pone de manifiesto para Gil Álvarez (1981, p. 29) que «el conocimiento de una experiencia sólo puede disminuir nuestra incertidumbre, nunca aumentarla. Además, la reducción será efectiva salvo en el caso en que ambas experiencias no tengan ninguna relación».

Es precisamente esta propiedad la que da lugar a Shannon a definir la **cantidad de información** que una variable transmite a otra, representada por **I**(XY), como la diferencia entre la entropía marginal y su entropía condicionada:

$$\mathbf{I}(XY) = \mathbf{H}(Y) - \mathbf{H}(Y|X)$$

es decir, la disminución de la incertidumbre de una variable que se consigue con el conocimiento de otra. Esta cantidad se mide en «bits» [1 bit= $-\log_2(1/2)$] que para McGill (1954, p. 99) «representa la información aportada por la elección entre dos alternativas igualmente probables».

Si consideramos la variable X como variable antecedente (transmisor) de Y (receptor), la ecuación anterior da pie a descomponer aditivamente la entropía de Y en dos componentes: la información transmitida por el emisor, **I**(XY), y la entropía condicionada **H**(Y|X) que se la denomina *ruido*. La dependencia total entre las variables podemos asociarla a una *transmisión sin ruido*: cada código de entrada (categorías de X) está asociado biunívocamente a un código de salida (categorías de Y). En el otro caso extremo, la independencia entre variables, viene dada por un *canal inútil*, incapaz de transmitir información: cada código de entrada tiene la misma probabilidad de estar asociado a cualquiera de los códigos de salida.

Considerando un sistema de tres variables X, T e Y podemos atender a diferentes tipos de cantidades de información:

— *informaciones bivariadas*: las producidas entre dos variables,

— *informaciones condicionales*: la información transmitida entre dos variables condicionada a una tercera, y definida por la expresión

$$\mathbf{I}(XY|T) = \mathbf{H}(Y|T) - \mathbf{H}(Y|XT)$$

— *información total*: la cantidad de información que se transmite en el sistema y definida como la entropía máxima de éste (cuando se produce la independencia entre todas las variables) y su entropía observada (la de la distribución conjunta de todas las variables)

$$\mathbf{I}(XTY) = \mathbf{H}(X) + \mathbf{H}(T) + \mathbf{H}(Y) - \mathbf{H}(XTY)$$

Esta información total se puede descomponer en informaciones binarias:

$$\mathbf{I}(XTY) = \mathbf{I}(XT) + \mathbf{I}(TY) + \mathbf{I}(XY|T)$$

Las diferentes cantidades de información cuando se calculan sobre una muestra representativa de la población pueden ser probadas en significación a partir de la razón de verosimilitud (Miller and Madow, 1954):

$$L^2 = 1.3863 \cdot N \cdot I(XY)$$

siendo N el tamaño muestral. Para muestras grandes, L^2 se distribuye aproximadamente como χ^2 con (i-1) · (k-1) g.l.

Como señalamos al principio, estas ideas fueron utilizadas para realizar diferentes tipos de análisis multivariables con datos cualitativos (variables nominales). Así para un modelo tridimensional de transmisión de información con variable consecuente Y y antecedentes X, T, se puede descomponer la entropía de Y (McGill, 1954) en:

$$\mathbf{H}(Y) = \mathbf{H}(Y|XT) + \mathbf{I}(XY) + \mathbf{I}(TY) + \mathbf{A}(XTY)$$

donde la entropía condicionada de Y a sus variables antecedentes X y T, **H**(Y|XT), supondría la variabilidad residual o no-explicada en Y (término de error) y **A**(XTY) el término de interacción entre las tres variables que representaría la ganancia (o pérdida) en la transmisión de la información en una de las informaciones bivariadas, **I**(XY) o **I**(TY), debido al conocimiento adicional de la tercera variable (T o X respectivamente).

Si bien los análisis de la varianza en variables nominales con la utilización de la teoría de la información fueron desplazados por los modelos log-lineales, tanto el análisis de regresión como el análisis causal realizados con dicha teoría presentan ciertas ventajas sobre los modelos logit:

— permite trabajar con politomías en la variable dependiente frente a la necesaria dicotomía en regresión logística,

— asigna un único valor a los caminos causales cuando se trabajan con variables politómicas frente a los múltiples valores que asignaría el análisis causal a partir de modelos logit.

Hoy en día, el uso de las diversas técnicas de análisis para este tipo de datos (análisis de correspondencias, modelos log-lineales, análisis multivariables no-lineales,...) suponen un importante reto en el campo de la investigación educativa.

BIBLIOGRAFÍA

ASH, R. (1965): *Information theory*. New York: John Wiley & Sons.

GARNER, W. R. (1956): «The relation between information and variance analyses». *Psychometrika, 21* (3); 219-228.

GIL ÁLVAREZ, P. (1981): *Teoría matemática de la información*. Madrid: Edic. ICE.

GUIASU, S. et THEODORESCU, R. (1966): *La Théorie mathématique de l'information*. París: Dunod, 1968.

KRIPPENDORFF, K. (1982): «Regression analysis using information theory». En L. TRONCALE (Ed.): *A Survey of Systems Methodology*. Louisville, KY: Society for General System Research; 1.007-1.012.

KRIPPENDORFF, K. (1986): *Information theory. Structural models for qualitative data.* Beverly Hills, CA: Sage.

KULLBACK, S. (1968): *Information theory and statistics.* Gloucester, Mass.: Peter Smit, 1978.

KULLBACK, S. and KEEGEL, J. C. (1984): «Categorical data problems using information theory». En P. R. KRISHNAIAH & P. K. SEN (Edt.): *Handbook of Statistics, 4: Nonparametric methods.* Netherland: North-Holland; 831-871.

McGILL, W. J. (1954): «Multivariate information transmission». *Psychometrika, 19* (2), 97-116.

MILLER, G. A. and MADOW, W. G. (1954): On the maximun likelihood estimate of the Shannon-Weaver measure of information. Report No. TR-54-75. Washintong, DC: Air Force Cambridge Research Center.

SHANNON, C. E. and WEAVER, W. (1949): *The matematical theory of communication.* Urbana: University of Illinois Press.