Teoría, estadística e información

FERNANDO CORTÉS
ROSA MARÍA RUBALCAVA

I. *Consideraciones preliminares*

Hay una extensa literatura dedicada al tema de la construcción de teorías científicas en general [1] y a las ciencias sociales en particular.[2]

Tal vez no sea menos profusa la cantidad de material escrito en que se ha tratado, en detalle, los pasos involucrados en la transformación de la información en dato.[3] La literatura referida a indicadores, índices y escalas, que estuvo muy de moda durante parte de los años cincuenta y los sesenta, por razones de todos conocidas, ya no es una temática en las ciencias sociales en América Latina.

También se encuentra bastante material cuyo centro de interés consiste en establecer los pasos que ligan los conceptos con la información.[4] Se ha destacado en esta secuencia: i) la *objetivación,* que consiste en relacionar, por medio de hipótesis, un concepto teórico inobservable, con

[1] Véase, por ejemplo, la vasta bibliografía citada por Bunge, Mario, *La investigación científica,* Buenos Aires, Editorial Ariel, cuarta edición, 1975, capítulos 7 y 8.

[2] Dos textos importantes dedicados a este tema son: Dubin, Robert, *Theory Building,* New York, London, The Free Press, 1969 y Stinchcombe, Arthur, *La construcción de teorías sociales,* Buenos Aires, Ediciones Nueva Visión, 1970.

[3] Respecto a este tema se puede consultar algunos de los libros clásicos como son: Boudon y Lazarsfeld, *Metodología de las ciencias sociales*: *I. conceptos e índices,* Barcelona, Editorial Laia, 1965; Goode, William & Paul Hatt, *Methods in Social Research,* New York, Toronto, London y Tokio, Kogakusha Company Ltd., 1952, Caps. 15, 16 y 17; Selltiz C. *et al., Métodos de investigación en las relaciones sociales,* Madrid, México y Pamplona, Ediciones Rialp, 1965, Caps. 5 y 10. Un par de libros exclusivamente dedicados a medición son el de Torgenson, Warren. *Methods of Scaling,* New York, London y Sydney, John Wiley, 1958 y el de Bohrnstedt, George y Edgar Borgatta, Editores. *Social Measurement*: *current issues,* Beverly Hills, London, Sage Publications, 1980. Un libro reciente y en español donde se encuentra un buen tratamiento del tema es el Padua, Jorge *et al., Técnicas de investigación aplicada a las ciencias sociales,* México, Fondo de Cultura Económica y El Colegio de México, 1979, Caps. V y VI.

[4] El artículo que sintetiza las principales características del problema es el de Lazarsfeld, Paul, "De los conceptos a los índices empíricos" en Boudon y Lazarsfeld, *op. cit.*

un concepto observable; ii) *la operacionalización,* proceso que implica vincular, a través de un conjunto de hipótesis, el concepto observable con sus correspondientes indicadores;[5] iii) la especificación de la unidad en la cual se procederá a registrar las propiedades observables definidas por los indicadores y iv) el proceso de medición que desemboca en la información. A partir de ésta tienen lugar una serie de operaciones que implican su transformación en dato, es decir, en un *concepto ya medido.*

La relación concepto-información es doble y está constituida por una serie de hipótesis o conjeturas que eslabonan el movimiento de lo observable con el de la teoría. En el proceso de "bajada" (desde el concepto hacia la información), se establecen una serie de conjeturas, ya sea explícitas o implícitas, que definen los estadios de objetivación y de operacionalización en que los argumentos de las hipótesis vinculan conceptos inobservables con observables y a estos últimos con los indicadores. En el proceso de "subida" (desde la información hacia el concepto) tiene lugar el planteamiento de otro conjunto de hipótesis (explícitas o implícitas) en que los argumentos ligan indicadores con índices o escalas, que pueden interpretarse como *conceptos medidos.*

Nuestro interés, dentro de este campo, radica en examinar el papel que juega parte del conocimiento estadístico en la relación entre teoría[6] e información, y más particularmente, entre teoría y dato. Hemos decidido limitar nuestras consideraciones a la asociación y a la regresión debido a que: i) de entre todas las técnicas que permiten el estudio estadístico de relaciones entre variables (contraparte observable de las relaciones entre conceptos que forman parte de una teoría), son las más usadas en la investigación social y ii) plantean requisitos diferenciales en relación con las transformaciones del lenguaje, derivadas básicamente del nivel de medición de las variables.[7] En este último aspecto centraremos los desarrollos que siguen.

La mayoría de las veces se usa el análisis estadístico, en general, y la asociación y regresión, en particular, para dar apoyo a las generalizaciones empíricas que alcanzan el estatuto de proposiciones teóricas. Sin embargo, esta utilización empirista del instrumental no es ineluctable. El conocimiento generado dentro de la disciplina estadística es lo suficientemente

[5] En nuestra perspectiva, un indicador tiene tres atributos: *i*) vincularse a un concepto teórico, *ii*) referirse a una propiedad observable de una unidad de registro y *iii*) variar. La variable estadística sólo recoge las dos últimas, es decir, lleva implícita una ruptura con la noción teórica.

[6] Toda teoría comporta, a lo menos, un conjunto de conceptos ligados y un conjunto de proposiciones articuladas.

[7] Esta afirmación es válida sólo en principio por cuanto dentro de la econometría y del campo de la estadística social se han desarrollado esfuerzos por extender la aplicación del análisis de regresión a variables cualitativas. Véase, por ejemplo Winship Cristopher y Robert D. Mare, "Structural Equations and Path Analysis for Discrete Data", en *American Journal of Sociology,* Vol. 89, Núm. 1, Julio de 1983, Págs. 55 a 107.

plástico como para servir a una óptica que invierta los términos; también presta utilidad en una perspectiva que recurre al análisis estadístico para establecer el vínculo entre el desarrollo teórico y los datos, lo que a su vez permite repensar algunos de los problemas habituales que se presentan en los estudios con base estadística.

Nos interesa tratar el papel que puede jugar el análisis de asociación y de regresión en el proceso que media las relaciones entre conceptos, teóricamente justificadas, y la relación entre sus correspondientes observables. Hemos señalado los pasos involucrados en la constitución del referente empírico de un concepto (objetivación, operacionalización, indicador y construcción de índices y escalas); nos preocupa examinar las operaciones que median entre lo teórico y lo empírico, pero, esta vez para *relaciones*[8] entre conceptos.

Para focalizar con mayor precisión el punto central, se supondrá que el vínculo teórico involucra sólo a dos conceptos (lo que para nuestros propósitos no resta generalidad al tratamiento) y que ya se llegó a sus referentes empíricos. Esto quiere decir que la estructura de la situación que nos interesa entraña, por una parte, una relación teórica entre dos conceptos y, por otra, dos índices (no relacionados), uno para cada concepto. Examinaremos a continuación cómo se puede usar la conexión teórica para vincular estadísticamente los índices (datos); vale decir, cómo el pensamiento teórico puede orientar la utilización de la estadística de relaciones (en particular del análisis de asociación y de regresión).

En nuestra concepción, el análisis estadístico requiere, para determinar el grado y la forma de la asociación entre los datos, que se establezca la proposición[9] correspondiente a la oración expresada en lenguaje natural a través de la cual se manifiesta el pensamiento teórico. El buen uso del instrumento implica cambio de lenguaje.

El proceso de formalización consiste en pasar el discurso teórico al plano del lenguaje formal, es decir, en traducir a proposiciones el contenido de los planteos discursivos. En cambio, el de matematización implica el pasaje del dominio de las oraciones al del lenguaje matemático, cualquiera que éste sea, aunque sabemos que los más populares son el del álgebra, la geometría plana y el del cálculo, ya sea diferencial o integral.

[8] Para determinar el referente empírico de una relación entre conceptos se agrega a las operaciones de objetivación y operacionalización la necesidad de establecer el vínculo empírico entre las variables.

[9] Respecto a la noción de proposición se puede consultar el viejo texto de Cohen Morris y Ernest Nagel, *Introducción a la lógica y al método científico 1,* Buenos Aires, Amorrortu Editores, cuarta edición, 1976, Caps. II y III. La primera edición en idioma inglés es de 1934. Un libro reciente donde se encuentra una excelente sistematización del concepto de proposición es el de Mario Bunge, *Epistemología,* Barcelona, Caracas y México, Editorial Ariel, 1980, Cap. 4.

II. *Teoría, formalización y análisis de asociación*

Si la proposición teórica [10] está claramente establecida y se vincula a conceptos no cuantitativos [11] (es decir, si ambos son conceptos individuales, de clase, relacionales no comparativos, relacionales comparativos o una combinación de ellos) entonces la relación entre los índices o variables, en términos de la estadística, puede tratarse con la técnica de asociación.

Tomemos como ejemplo una investigación realizada por José Nun [12] sobre la reinserción laboral de los trabajadores de la industria automotriz argentina. Nuestra lectura [13] nos lleva a sostener que uno de los planteamientos del trabajo establece que en la resinserción laboral juega un papel importante la relación entre los procesos de segmentación del mercado de trabajo y de heterogeneidad/homogeneidad de la fuerza de trabajo. Como ambos conceptos no son directamente observables les hace corresponder el tipo de ocupación actual y el nivel de calificación del trabajador respectivamente. Su planteamiento sostiene que los trabajadores se reinsertarán en el mercado de trabajo primario o secundario en función del nivel de calificación que tengan. El mercado primario está definido por las ocupaciones en los planteles industriales grandes (que denotaremos por A) en tanto que el secundario lo está por las actividades desarrolladas en las industrias pequeñas y en los servicios (simbólicamente A'). La calificación de la fuerza de trabajo se dicotomiza en calificada (B) y no calificada (B'). La proposición se puede expresar de la siguiente manera:

Si B, entonces A y si B' entonces A'.

Esto quiere decir que al cruzar el nivel de ocupación con el tipo de ocupación actual debería generarse una tabla como la siguiente:

	Nivel de calificación	
	Calificado (B)	*No calificado (B')*
Plantas Ind. grandes	XXXXX	O
Tipo de Oc. actual		
Ind. peq. y servicios	O	XXXXX

[10] Entendemos por proposición teórica aquella que involucra dos o más conceptos inobservables o susceptibles de observarse.

[11] Esta clasificación está tomada de Mario Bunge, *La lógica de la investigación científica, op. cit.*, págs. 78 y 79.

[12] Nun, José, "La Industria Automotriz Argentina: estudio de un caso de superpoblación flotante", *Revista Mexicana de Sociología 1/78*, México, págs. 55 a 106.

[13] Una revisión detallada y discutida con el autor de este artículo se encuentra en nuestro trabajo por publicarse intitulado *Análisis cuantitativo de variables cualitativas: introducción al análisis de asociación.*

La distribución de los datos que se ha marcado en la tabla (donde las series de X representan las casillas donde debiera localizarse la información) sería la observada en caso de que la proposición gozase de sustento empírico.

Para analizar esta tabla se abren dos caminos estadísticos alternativos. El que define asociación por negación de independencia estadística, que se encuentra en la base de los análisis que toman pie en ji-cuadrado [14] y que origina una serie de coeficientes que son función de dicha estadística. Y aquel que intenta una definición de asociación por lo que es,[15] en términos de proposiciones, y no por negación. Según esta última vertiente, se tienen tantos coeficientes de asociación como definiciones de la misma, lo que lleva a explicar el por qué del abanico de valores que pueden asumir dichos coeficientes, y a la vez individualiza al que responde a la estructura lógica contenida en la proposición.

La primera de estas vertientes inicia sus desarrollos vinculando la fuerza con que se relacionan las variables a la diferencia entre porcentajes, continúa con el análisis de las discrepancias entre frecuencias observadas y esperadas y culmina con un conjunto de coeficientes que se definen en función de ji-cuadrada.

La segunda línea se inaugura, según nuestro parecer, con el trabajo de Goodman y Kruskal, y en ella nos parece un hito importante la contribución de Hildebrand, Laing y Rosenthal, quienes dan a la noción de asociación un contenido específico y cambiante según lo requiera el tipo de proposición que interesa someter a contrastación empírica.

Es interesante notar que la estrategia de análisis ji-cuadrada mide la asociación sobre la base de un criterio estadístico, por lo demás perfectamente explícito, que no presenta conexiones claras con la forma como se define teóricamente la relación entre las variables. Por el contrario, la vía que se descuelga por el lado de la definición proposicional de asociación, establece un vínculo entre teoría, técnica estadística adecuada e información que hace imposible contrastar la proposición con la distribución de las observaciones, si no es sobre la base de la integración de esos niveles. Todos sabemos que una de las dificultades que enfrenta el investigador al analizar tablas de contingencias es la de decidir qué coeficiente utilizar, especialmente cuando se dispone de computadoras en que los programas presentan como opción más de una decena de coeficentes. Pareciera que la raíz de este problema se encuentra en la carencia de definición del sentido lógico que se le asigna a la idea de asociación en la aplicación

[14] Los coeficientes de asociación que son función de ji-cuadrado la definen a partir de la diferencia entre las frecuencias observadas y las que se esperaría si los atributos fueran estadísticamente independientes.

[15] La definición precisa de asociación está determinada por la proposición teórica, de manera que a proposiciones diferentes corresponderán coeficientes distintos.

específica y la coordinación de esta definición, con aquella que subyace a los diferentes coeficientes estadísticos de asociación.

Aún cuando los coeficientes biserial, biserial punto y tetracórico también permiten trabajar con variables cualitativas, hemos decidido no incluirlos por cuanto se inscriben dentro del campo del análisis de correlación.

La claridad lógica respecto a lo que se entenderá por asociación no es condición necesaria y suficiente para resolver el problema de seleccionar el coeficiente más adecuado. De todas maneras, el investigador, sin decirlo, puede elegir el índice que arroje el valor más alto, y plantear *a posteriori* la proposición que le corresponde, pero si entrega las bases para hacer un uso no empirista de la técnica.

En la actualidad, se distinguen con toda claridad dos vertientes para analizar tablas de contingencia. Una es la línea inaugurada por Goodman y Kruskall, que ha llevado a la construcción del coeficiente delta-ro.[16] Y la otra que toma pie en ji-cuadrada y que ha desembocado en el modelo logarítmico lineal (log-lin).[17] Sería interesante investigar acerca de la posible convergencia entre ambas, porque de ese modo se acoplaría la potencialidad analítica de la técnica log-lin con las preguntas que surgen desde el ámbito del pensamiento teórico.

Otro tema que surge al aplicar estadística de relaciones es el de su uso descriptivo o inferencial.[18] En efecto, si suponemos que la proposición es no probabilística, la única manera como podría entrar la aleatoriedad (que justificaría las desviaciones con respecto a la aseveración teórica formalizada) sería a través del proceso de selección de las observaciones (muestreo aleatorio); esto quiere decir que, si consideramos que la proposición es exacta y los datos son poblacionales, entonces no cabe utilizar inferencia estadística.

Una manera diferente de considerar el mismo tema parte del reconocimiento de que usualmente el pensamiento teórico se expresa a través de proposiciones no probabilísticas; pero que, al ponerlo en correspondencia con la información, se introduce la aleatoriedad porque: i) se come-

[16] Goodman, Leo y Kruskall, "Measures of Association for Cross-clasifications", en *Journal of American Statistical Association,* número 49, 1954. El desarrollo meticuloso del coeficiente delta-ro se encuentra en Hildebrand D., J. Laing y H. Rosenthal, *Analysis of Ordinal Data,* California, Sage Publications, 1977.

[17] Una buena introducción al método logarítmico lineal se encuentra en Everitt, B. S. *The Analysis of Contingency Tables,* London, Chapman and Hall, 1977. Dos excelentes tratados son el de Bishop, Yvonne, Stpehen Fienberg and Paul Holland, *Discrete Multivariate Analysis: theory and practice,* Cambridge Massachusetts and London England, The MIT Press, 1975; y el de Haberman Shelby, *Analysis of Qualitative Data,* New York, San Francisco y London, Academic Press, 1978, volúmenes I y II.

[18] Sobre este tema hay que revisar el trabajo ya clásico de Hagood, Margaret, "The Notion of a Hipothetical Universe", en Morrison and Henkel, *The Significance Test Controversy,* Chicago, Aldine Publishing, 1970.

ten errores de medición; ii) falta considerar factores explicativos incorporándose sólo los relevantes, y/o iii) el fenómeno en sí es aleatorio.

El segundo argumento que racionaliza la incorporación de la aleatoriedad pareciera suponer que todo fenómeno estaría determinado por un conjunto n de factores, con n tendiendo a infinito, y que sería cuestión de conocerlos, así como la *forma* en que se relacionan, para explicar perfectamente el fenómeno.[19] En cambio, el tercer argumento es consistente con una proposición teórica conceptualizada aleatoriamente según la cual la explicación descansaría en un proceso, por ejemplo, tipo markoviano, o bien que se aceptara la aleatoriedad como una concesión a la estadística. En este último caso se parte reconociendo que hay una laguna entre la proposición, tal como ha sido formulada por el desarrollo teórico, y la información. Si estamos interesados en contrastar empíricamente nuestra proposición, podemos recurrir al análisis estadístico. Pero este último basa toda su construcción moderna sobre el principio de determinación aleatoria, de manera que no importa cuál de los principios[20] se haya usado en las elucubraciones teóricas, de todos modos se establece un modelo aleatorio como una manera de vincular teoría con información. Sobre la base de la estimación inferencial realizada dentro del dominio de la estadística, se pasa posteriormente al ámbito de las proposiciones.

De lo anterior se derivan algunas consecuencias inmediatas: i) la inferencia tiene cabida independientemente de la técnica de recolección de información aplicada; ii) no es posible pedir a la estadística que incorpore el principio de determinación utilizado en el desarrollo de la teoría, cualquiera que éste sea (estructuralista, dialéctica, estructural funcionalista, teleonómica, etcétera); iii) es necesario que se articulen ambos principios de determinación, de manera que las conclusiones estadísticas se puedan interpretar teóricamente, y iv) la construcción de proposiciones a partir del análisis estadístico de la información impone el principio de determinación aleatorio en la explicación conceptual, lo que puede entrar en abierta contradicción con los principios metateóricos en los cuales se asienta la teoría.

En resumen, el análisis de asociación resulta útil para vincular teoría con información, cuando el planteamiento conceptual involucra la conexión de dos o más conceptos no cuantitativos. Disponemos de dos vertientes; la que define asociación como no independencia estadística y aquélla que busca precisar, en el campo de la lógica, lo que se entenderá por asociación en cada aplicación particular. La segunda provee de bases conceptuales para hacer un uso no empirista de la técnica. Por otra parte, la misma técnica se puede utilizar desde un punto de vista descriptivo o inferencial. El primero tiene lugar cuando el desarrollo teórico lleva a

[19] Véase, por ejemplo, Johnston, J., *Econometrics Methods,* New York, McGraw Hill, Second Edition, 1979, pág. 10.

[20] Bunge, Mario, *El principio de causalidad en la ciencia moderna,* Buenos Aires, EUDEBA, 1961, págs. 29 a 33.

proposiciones no probabilísticas y la información es censal. El segundo puede tener los siguientes orígenes: i) la proposición teórica es no probabilística pero la información es muestral; ii) la proposición teórica es no probabilística y la información es censal. En este caso, se plantea un modelo estadístico homólogo al modelo teórico que difiere de aquél en el principio de determinación que le da sustento; iii) la proposición teórica es en sí probabilística y la información es muestral, en cuyo caso coexisten dos fuentes de aleatoriedad: una introducida por el principio de determinación y otra que ingresa por la vía del mecanismo de selección de la información, y iv) la proposición teórica es probabilística y la información es censal, por lo que la aleatoriedad sólo puede justificarse teóricamente.

III. *Teoría, matematización y análisis de regresión*

El análisis de regresión puede verse también como una técnica que ayuda a establecer el vínculo entre teoría e información. Requiere, en primer término, que el pensamiento conceptual se exprese en el lenguaje matemático de las funciones (lineales o susceptibles de linealizarse a través de transformaciones) y en segundo término, que se articulen los principios de determinación en el caso en que la teoría se base en algún principio de determinación no aleatorio.

Tenemos, entonces, que la regresión exige, en primer término, que se traduzca el contenido teórico, expresado en lenguaje natural, al lenguaje matemático de las funciones, el que nos entrega una sintaxis poderosa que permite extraer una serie de conclusiones de naturaleza teórica difícilmente deducibles sobre la base del lenguaje natural.[21] No insistiremos acerca de las bondades y de las limitaciones que se derivan de la matematización de la teoría, ni sobre la parte que en ella ponen la teoría y la matemática; pero hay que señalar que expresar el pensamiento teórico en términos de funciones no necesariamente conduce a aplicar análisis de regresión. Es posible usar la matematización para sacar las consecuencias entrañadas en los conceptos y sus relaciones, sin establecer ningún vínculo con la información estadística. Las matemáticas también son útiles para apoyar la elaboración de ensayos teóricos.[22]

[21] Una buena discusión de las bondades y limitaciones de la matematización se encuentra en Bunge, Mario, *La lógica de la investigación científica, op. cit.*, págs. 503 a 516.

[22] Respecto a este uso de las matemáticas en ciencias sociales hay una serie de escritos. Sugerimos que de entre ellos se consulte dos trabajos realizados a partir de ópticas teóricas distintas. Uno es el de Przeworski, Adam A. y Michael Wallerstein, "The Structure of Class Conflict in Democratic Capitalist Societes", en *The American Political Science Review*, Vol. 76, Núm. 2, junio de 1982, y el otro es el de, Rapoport Anatole, *Fights, Games and Debates*, Ann Arbor, The University of Michigan Press, 1960.

Si la relación teórica (que supondremos, por razones de simplicidad, involucra sólo dos conceptos) ha sido matematizada en el lenguaje funcional (con el requisito de linealidad que ya hemos señalado), los conceptos son cuantitativos [23] e interesa conectar el desarrollo del pensamiento con los referentes empíricos, entonces la relación entre los indicadores o índices puede ser tratada con el análisis de regresión.

Tomemos como ejemplo una de las matematizaciones que nos ofrece R. Boudon [24] para tratar la hipótesis "la mayoría de los votos de izquierda son de origen obrero".[25] Si se supone que Yi y Xi simbolizan las proporciones de votos a favor de la izquierda y de obreros en el distrito electoral i, y p y q [26] denotan las propensiones de los obreros y de los no obreros a votar por la izquierda, entonces la relación entre ambos conceptos se puede expresar a través de la función lineal:

$$Yi = pXi + q(1 - Xi)$$

que nos permite identificar las dos fuentes posibles de votos a favor de la izquierda y que nos da pie, previa evaluación estadística de ella a través del análisis de regresión, formarnos una idea respecto a los tamaños relativos de p y q, lo que nos permitirá verificar si se puede sostener empíricamente que la mayoría de los votos de izquierda tienen origen obrero.

La ecuación expresa en el lenguaje del álgebra (y si se quiere, también en el de la geometría analítica) el contenido de la hipótesis teórica, aunque se ha agregado un mayor nivel de precisión al especificar la naturaleza de la relación. A partir de ella se pueden obtener algunas conclusiones que, si bien son estrictamente matemáticas, admiten interpretaciones sustantivas. Por ejemplo, reordenando términos llegamos a:

$$Yi = q + (p - q)Xi$$

Por lo tanto, la ordenada al origen es la propensión de los no obreros a votar por la izquierda y la pendiente es la diferencia de las preferencias por la izquierda de los obreros y de los no obreros.

Dado que nuestro interés consiste en conectar el modelo matemático (teórico) con la información, debemos ahora poner el foco en el otro lado de la relación. En efecto, es necesario que se defina con claridad el indicador que se usará para medir el "voto por la izquierda", lo que implica, en primer lugar, precisar el concepto de izquierda que se usará y

[23] Bunge, Mario, *La lógica de la investigación científica, op. cit.*, págs. 78 y 79.

[24] Boudon, R., *L'Analyse Mathematique des Faits Sociaux*, París, Plon, 1967, págs. 183 a 186.

[25] Obviamente es posible utilizar una justificación con más contenido teórico. Podríamos fundamentar la hipótesis en un argumento que ligue la inserción de la fuerza de trabajo en relaciones sociales de producción y el comportamiento político, pasando por la discusión de la conciencia de clase. Sin embargo, para nuestros propósitos basta con la forma como ha sido planteada.

[26] Nótese que esto implica que se ha supuesto que pi = p y qi = q. Esto quiere decir que aceptamos que las propensiones de los obreros y de los no obreros a votar por la izquierda son las mismas para todos los distritos electorales.

qué candidatos o partidos se considerarán como tales; también hay que decidir si la proporción se calculará en relación con el total de votos válidamente emitidos, al total de votantes o al total de la población que cumple con los requisitos de ser ciudadano. Algo similar pasa con el caso de la medición del concepto obrero, cuestión que puede ser bastante escabrosa dentro de algunos discursos teóricos. La situación se complica aún más en aquellos casos en que la información es censal.

El análisis de regresión nos puede ser útil en una situación en que, por una parte, tenemos una teoría expresada en lenguaje matemático y, por otra, índices, indicadores o escalas no relacionadas, resultantes de los procesos de objetivación y de operacionalización de los conceptos. En otros términos, la regresión puede servirnos en aquellas situaciones en que nos interese relacionar valores numéricos de variables, apoyándonos en las conexiones teóricas (planteadas en lenguaje funcional).

La regresión, así como la asociación, se puede usar como técnica descriptiva o inferencial. Cabe una utilización descriptiva cuando el principio de determinación teórico es no probabilístico, los datos son censales y la hipótesis teórica se vincula directamente con la información, es decir sin que medie un modelo estadístico probabilístico. La aleatoriedad se puede introducir suponiendo que: i) el principio de determinación teórico es no estadístico, pero las variables provienen de una muestra aleatoria, o ii) el principio de determinación es aleatorio, lo que introduce una justificación neta para la incorporación de argumentos estadísticos y da lugar para la aplicación de la estadística inferencial aunque los datos sean censales.

Tenemos, entonces, que el uso descriptivo de la regresión supone que una vez que hemos matematizado la teoría, hay que estimar, en un modelo estadístico, que se diferencia del teórico en que los conceptos se han sustituido por variables.[27] Se escribe un modelo análogo a aquél que expresa el planteo teórico, pero entre ambos median los procesos de objetivación, operacionalización y construcción de escalas o índices. Esta óptica sitúa al modelo estadístico en el plano de la observación y de la medición.

Desde el punto de vista inferencial hay que repensar el problema de la relación entre teoría e información poniendo el énfasis sobre la articulación de los principios de determinación. En efecto, pongámonos en el caso, por cierto el más frecuente, en que hemos recurrido a funciones matemáticas para expresar un pensamiento no probabilístico y que, además, no sólo interesa extraer sus consecuencias teóricas, sino que también nos preocupa establecer la correspondencia que guarda con los datos. Ya examinamos uno de los caminos posibles que nos brinda el análisis de regresión (su uso descriptivo). La otra vía consiste en admitir el supuesto de que el concepto simbolizado por Y[28] (localizado en el dominio de la

[27] Recuérdese que una variable es un indicador que carece de vinculación con el concepto.

[28] En el plano de la matematización se usa la noción de variable para representar el contenido de un concepto.

teoría sustantiva) se puede representar por una variable aleatoria Y. Esto implica que al modelo teórico le hacemos corresponder un modelo estadístico análogo. El paso de la función matemática a la ecuación estadística lleva consigo no sólo la diferencia entre dos dominios conceptuales, sino también entre dos principios de determinación: el que subyace a la construcción teórica y el estadístico.

Los modelos matemático y estadístico difieren en que en el primero las variables, usualmente, son no aleatorias y representan conceptos que no necesariamente deben conectarse con los observables, en tanto que en el segundo, las variables son aleatorias y deben estar vinculadas, por un lado, con las variables matemáticas y por otro, con la información o los datos. Se corresponden en la medida en que debe ser posible pasar de las estimaciones de los parámetros estadísticos a los parámetros de la ecuación teórica.

Supongamos que nuestro problema consiste en estimar las propensiones a votar por la izquierda en el modelo teórico:

$$Yi = q + (p - q)Xi$$

de modo que podamos dar sustento empírico a la aseveración de que la mayoría de los obreros vota por la izquierda. Las definiciones de Xi, Yi, p y q se han explicitado en páginas anteriores.

Para resolver el problema de estimación (desde el punto de vista inferencial) debemos suponer, en primer lugar, que Yi es una variable aleatoria que incorpora una parte sistemática (la señalada por la teoría: $q + [p - q]Xi$) y otra estocástica (que simbolizaremos por Ui). El modelo estadístico correspondiente al matemático sería:

$$Yi = a + bXi + Ui \qquad (i = 1, 2, \ldots\ldots, n)$$

La simple operación de agregar un término de error aleatorio (Ui) entraña un cambio en el principio de determinación (supuesto que el teórico no es estadístico) pasándolo al dominio probabilístico.

El problema de estimación, propio del campo de la inferencia estadística, se realiza sobre la base de un conjunto elemental de supuestos.

i) Los promedios de las variables aleatorias Yi se localizan sobre la recta:

$$\mathrm{E}[Yi] = a + bXi \qquad (i = 1, 2, \ldots\ldots, n)$$

o, equivalentemente que $\mathrm{E}[Ui] = 0$.

ii) Las varianzas de las variables aleatorias Yi (o de los términos de error Ui) son iguales entre sí:

$$\mathrm{E}[Yi - \mathrm{E}Yi]^2 = V, \text{ donde } i = 1, 2, \ldots\ldots n \text{ y}$$

V es una constante.

Esta expresión implica que:

$$\mathrm{E}[Ui]^2 = V, \text{ para todo } i.$$

iii) Las covarianzas entre las variables aleatorias Yi son iguales a cero.

$$E[(Yi - EYi)\ (Yj - EYj)] = 0, \text{ para todo } i$$

distinto de j.

Se puede demostrar que este supuesto conduce a que:

$$E[UiUj] = 0. \text{ Suponiendo } i \text{ diferente de } j.$$

Es decir, a que la covarianza entre los términos de error sea también igual a cero.

A este conjunto de supuestos se agrega, por consideraciones de simplicidad en el tratamiento estadístico, que la variable explicativa sea conceptualizada como fija o predeterminada. Esta suposición, así como las anteriormente especificadas para el modelo estadístico, pueden someterse a contrastación empírica. Una de las preocupaciones centrales de la estadística y de la econometría es el estudio de las complejidades que surgen en el análisis de la información cuando estos supuestos no se satisfacen.

El problema de estimación consiste en asociar valores numéricos a los parámetros del modelo estadístico bajo el supuesto de que éste ha generado los pares ordenados $(X1, Y1), (X2, Y2), \ldots\ldots, (Xn, Yn)$. Con este conjunto de supuestos se tiene la estructura típica de los problemas de inferencia estadística.

Los parámetros estadísticos son a, b, las varianzas y covarianzas de las variables aleatorias Yi y los de la matematización de la teoría son p y q. Como se puede apreciar, los parámetros de ambos modelos son diferentes, aunque están relacionados matemáticamente a través de:

$$a = q \text{ y, } b = p - q$$

son estas ecuaciones las que nos permiten ligar las estimaciones estadísticas con las propensiones de los sectores obreros y no obreros a votar por la izquierda.

Hemos localizado el análisis de regresión, tanto en su vertiente descriptiva como inferencial, según una perspectiva que lo define en relación con un modelo teóricamente construido. Sin embargo, no es poco usual que se recurra al análisis estadístico para construir el modelo teórico. Esta inversión en el papel de la estadística en la investigación, en general, y en la investigación social, en particular, se encuentra privilegiada en los propios libros de texto dedicados a esta materia. Al exponer el modelo de regresión, casi sin excepciones, señalan la conveniencia de empezar el análisis dibujando un diagrama de dispersión y en función de la distribución de la nube de puntos elegir el modelo que "mejor ajuste". Dejando a un lado el hecho de que este método es práctico sólo cuando se trata de dos variables, se complica bastante cuando son tres y no se puede aplicar cuando son cuatro o más, la estrategia que se sugiere lleva a confundir el modelo teórico con el modelo estadístico. Las consecuencias de superponer ambas construcciones son inmediatas:

i) pueden aparecer contradicciones entre el principio de determinación aleatorio que se usó para fundamentar las estimaciones y el principio de determinación que subyace a las conceptualizaciones generales que dan sustento al modelo, y ii) se confunden los índices, indicadores o escalas, es decir, las variables, con los conceptos teóricos.

IV. *Conclusiones*

La preocupación central de este trabajo fue la de explorar la ayuda que puede prestar el instrumental estadístico que sirve para estudiar relaciones entre variables, en la vinculación entre teoría e información. La manera de enfocar el problema consistió en suponer que la estadística es un poderoso auxiliar en el proceso de investigación, siempre que haya un mínimo de claridad conceptual que permita formalizar o matematizar el contenido teórico.

Dada una relación entre dos o más conceptos, recurrimos a la objetivación, operacionalización y construcción de escalas o de índices para determinar sus correspondientes variables (en la terminología de la estadística). Tenemos, por una parte, relaciones entre conceptos, y por otra, mediciones correspondientes a cada uno de ellos, por separado. *La estadística nos provee de un instrumento para establecer las relaciones entre las manifestaciones empíricas de los vínculos entre los conceptos.* Al estudiar estadísticamente el nivel y la forma de la asociación entre las variables, analizamos también la conexión entre los conceptos.

A partir de la postura que sostiene que el análisis de la información debe estar teóricamente orientado, no necesariamente se deduce una visión lineal del proceso de investigación, el que supuestamente partiría de la teoría y culminaría con los datos. Sólo planteamos que en la constrastación o búsqueda de soporte empírico para las relaciones teóricas (obviamente, una vez que éstas se han constituido) es necesario distinguir el dominio de la teoría del de la estadística, así como tener claridad respecto a las articulaciones que los ligan. Esto no impide que se recurra a la estadística (aunque no necesariamente al análisis de asociación y de regresión según un enfoque inferencial) en otros momentos de la espiral investigativa que se mueve entre el objeto teórico y el objeto empírico.

Por otra parte, el intento de construir modelos teóricos sobre la base de la información y de la inferencia estadística que sirve para tratar relaciones entre variables (de las cuales hemos destacado la asociación y la regresión), yuxtapone los modelos teórico y estadístico, lo que origina una serie de dificultades que nacen, al parecer, de la confusión de los conceptos con sus mediciones (variables) y de la mezcla de principios de determinación.