IGNACIO MÉNDEZ RAMÍREZ &
PABLO GONZÁLEZ CASANOVA (ed).

1993 Matemática y ciencias sociales

Miguel A. Porrúa/CIIH/UNAM ediciones

México DF.

371 pp.

Consideraciones sobre el uso de la estadística en las ciencias sociales. Estar a la moda o pensar un poco

FERNANDO CORTÉS*
ROSA MARÍA RUBALCAVA*

INTRODUCCIÓN

"Los análisis (una serie de tabulaciones cruzadas) son relativamente poco sofisticados dada la cantidad de datos disponibles".[1]

La moda no es una buena consejera para la investigación científica; tampoco para seleccionar los instrumentos estadísticos. Sin embargo, pareciera que es uno de los criterios que se han usado y se usan en la investigación social para escoger entre el arsenal de herramientas estadísticas disponibles en las últimas décadas.

Basta una revisión somera de las principales revistas norteamericanas especializadas para corroborar que aparecen ciclos en los que una técnica particular adquiere popularidad. No es claro si este comportamiento se debe a la presencia de algunos problemas sociales relevantes cuya investigación demanda una nueva técnica o la aplicación creativa de instrumentos estadísticos ya existentes; o si obedece a que los problemas a ser investigados se delimitan a partir de los desarrollos instrumentales; o bien a que las técnicas se aplican como recetas de cocina debido a la deficiente formación metodológico-técnica que reciben los científicos sociales (Blalock H. 1989: 450); y aún habría que considerar la presión de los organismos que financian investigaciones

* Investigadores del Centro de Estudios Sociales de El Colegio de México.

[1] Párrafo del dictamen con el que una revista internacional rechazó la publicación de un artículo de un investigador mexicano en 1990.

y los comités de dictamen editorial. Cualquiera que sea la explicación a estos ciclos, se puede decir que en los últimos años la investigación social se ha caracterizado, entre otros rasgos, por la aplicación de técnicas de moda.

Una mirada superficial permite delinear la historia reciente de la investigación social según el uso de los instrumentos estadísticos. En los Estados Unidos, entre los años cincuenta y sesenta, se utilizaban profusamente métodos estadísticos para realizar clasificaciones (análisis factorial, de conglomerados, discriminante) y analizar relaciones entre variables (regresión, correlación, análisis de contingencia y de asociación). Los métodos que demandaban variables métricas[2] (factorial, conglomerados, discriminante, regresión y correlación) tuvieron una aplicación limitada en América Latina, tanto porque los recursos de cómputo eran de difícil acceso como por las debilidades en la formación de los investigadores sociales en técnicas de investigación y estadística.[3] En cambio, los análisis de asociación y de contingencia (que operan sobre variables no métricas)[4] han dominado la investigación social de América Latina hasta la actualidad.

En los años sesenta se popularizó el entonces llamado análisis causal (cuyos orígenes se remontan a un trabajo en genética poblacional escrito por Wright en 1921: 557 a 585), que se percibía como una generalización del análisis multivariado y que tuvo un fuerte impacto sobre la investigación social, a pesar de que sólo se aplicaba sobre variables métricas y variables dicotómicas. Una vez que se apreció la profundidad de la discusión epistemológica en torno a la noción de causalidad, pasó a tomar el nombre con el cual se le conoce hasta hoy: análisis de trayectorias o de senderos (*path analysis*). En este método, las complejidades de cálculo no superaban a las del análisis de regresión.

[2] Son variables métricas las que se miden en escalas de intervalo o de razón y no métricas las medidas en escalas nominal u ordinal.

[3] H. Blalock (1989: 448, 454 y 458) hace mención a la actual deficiencia en el conocimiento de instrumentos de investigación de los estudiantes en ciencias sociales (grado y posgrado) de las universidades norteamericanas. Tradicionalmente sus colegas latinoamericanos adolecen del mismo problema pero con mayor intensidad.

[4] Sólo se requería una clasificadora para realizar los cruces de variables.

La línea del desarrollo estadístico-matemático que se plantea como problema ampliar el acervo de métodos para analizar variables no métricas alcanzó resultados que atenuaron la división tajante entre variables según su nivel de medición. Por una parte, el análisis de regresión se extendió para incorporar variables explicativas y explicadas no métricas (las primeras se conocen con el nombre de variables ficticias o variables mudas, *dummy*, y se tratan como caso particular dentro del análisis de regresión; en tanto que las segundas llevan a modificaciones profundas que cristalizaron en el modelo de regresión logit o logística). Por la misma época se desarrolla una generalización del análisis multivariado (modelo loglineal) que permite modelar un conjunto de relaciones entre variables no métricas. Tanto el modelo logit como el loglineal ganaron popularidad en los años ochenta, en parte, porque sus tediosos procedimientos iterativos de ajuste pudieron realizarse con paquetes de programas estadísticos en computadoras personales.

En las ciencias sociales de América Latina se han hecho aplicaciones esporádicas, en campos temáticos reducidos, de los análisis de trayectorias, loglineal y logit.

En lugar de seguir la moda se puede tomar en cuenta una diversidad de criterios que se han propuesto para orientar la selección de "el mejor" método estadístico. Hay quienes sostienen que uno de los criterios es el nivel de medición de las variables (Siegel 1956), otros ponen el acento sobre los procedimientos que se siguieron para generar las observaciones (Campbell y Stanley 1979), también hay quienes plantean que los instrumentos de registro determinan la viabilidad de análisis estadístico: la información que se obtiene a través de métodos antropológicos (observación participante, observación directa, historia oral, filmación..) no es susceptible, se dice, de análisis cuantitativo (Magrassi, G. E., M. M. Roca y otros 1980: 14). Es cierto que este tipo

de información presenta dificultades particulares pero no son un impedimento insoslayable para aplicar la estadística.

En las páginas que siguen desarrollaremos la idea de que el uso adecuado de los instrumentos estadísticos en una investigación requiere también identificar el isomorfismo entre las estructuras lógicas de la técnica y de las respuestas provisorias (hipótesis) a las preguntas que orientan la investigación. Sostenemos que no basta considerar únicamente el nivel de medición de las variables y los procedimientos de observación, sino que, además, debe examinarse la concordancia entre las preguntas de investigación, las hipótesis de trabajo expresadas en términos de relaciones entre las observaciones o entre las variables y las técnicas que ofrecen diversas posibilidades para el análisis empírico de dichas relaciones.

En ocasiones la pregunta se limita a establecer la presencia o ausencia de relación entre variables, como: ¿hay o no relación entre la salud de los niños [SN] y el trabajo de sus madres fuera del hogar [TM]?, ¿se modifica o no esta relación según la calidad del cuidado [CC] que recibe el hijo en ausencia de la madre?, ¿hay o no relación entre la salud del pequeño y el incremento en el ingreso familiar [IIF] debido al trabajo de su madre? En otras ocasiones, interesa saber la fuerza de esas relaciones o bien cuantificar el efecto neto sobre la salud del menor, originado en el impacto positivo del mayor ingreso familiar y el negativo de la ausencia materna en el hogar, modulado por la atención alternativa al hijo.

El nivel de medición de las variables, los procedimientos que generaron los datos y las relaciones entre variables derivadas de las preguntas de investigación no necesariamente conducen a la selección de *una* técnica estadística en particular, aunque sí delimitan un subconjunto de ellas. Para establecer la relación entre TM y SN se pueden utilizar, por ejemplo, análisis de asociación, análisis de correlación o pruebas de diferencias de medias.[5] La relación TM, SN, en

[5] Se suponen variables dicotómicas.

presencia de CC, podría abordarse con análisis de asociación parcial, mediante la ecuación de covarianzas de Lazarsfeld, con correlaciones parciales o vía análisis de varianza. Para estudiar el efecto de IIF sobre SN, se podría optar entre el análisis de regresión logística, la regresión probit o el análisis discriminante. Si interesa evaluar el efecto neto de estas variables sobre SN, habría que aplicar análisis de regresión múltiple o bien análisis de trayectorias, según la estructura de los vínculos planteados entre IIF, CC, TM y SN.

La coincidencia entre el uso razonado de la estadística (que lleva en sí la correspondencia entre los planteamientos teóricos, las preguntas de investigación, los métodos y técnicas empleadas para recopilar la información y los modelos estadísticos) y la aplicación de moda sólo puede ser un hecho fortuito.

En este trabajo nos proponemos mostrar una gama de técnicas estadísticas que expondremos desde el punto de vista de la correspondencia de sus estructuras con las estructuras de las relaciones entre observaciones y entre variables, propuestas por los esquemas teóricos. Las hemos seleccionado tomando en cuenta su aplicación potencial y su frecuencia de uso en la investigación social en América Latina. Tuvimos presente, sobre todo, aplicaciones a la sociología, la sociodemografía, la antropología, los estudios urbanos y de salud pública, y la ciencia política, cuyos análisis son fundamentalmente de sección cruzada o de estática comparativa; por lo anterior, queda fuera de esta exposición el tratamiento de series de tiempo, procesos estocásticos, estudios en panel y cualquier otro tipo de análisis secuencial.

Por motivos de exposición, consideramos necesario dedicar la primera sección a la matriz de datos, que es el punto de encuentro entre las operaciones teóricas, metodológicas y estadísticas; es la carne que cubre el esqueleto lógico. En la segunda, desarrollamos la idea central de este trabajo: el isomorfismo entre el sistema de relaciones que fluye de la conceptuación y el sistema de relaciones que configura a

cada modelo estadístico. Dada la naturaleza de nuestra tarea, ponemos el énfasis en la presentación de las estructuras de algunas técnicas estadísticas. Esta parte se subdivide en técnicas para construir índices y clasificar, y técnicas para analizar relaciones entre variables. Por último, en la tercera, intentamos abogar por el uso no empirista de la estadística y perfilamos su papel en distintos momentos del proceso de investigación.

La matriz de datos

El análisis estadístico procede sólo si se dispone de un conjunto de mediciones de los atributos de las unidades de observación, unidades que suelen denominarse de diferentes maneras: individuos en sentido genérico (Bouroche, J. M. y G. Saporta 1980: 5), elementos de análisis o unidades de análisis (Galtung, J. 1966: 1), unidades (Castellanos, A. 1977: 35), objetos (Yule, U. y M. Kendall 1959: 15) o bien, casos u observaciones (SPSS, 1988: B-9). Para cada una de ellas se registran, con el fin de caracterizarlas, una serie de rasgos o propiedades que se denominan *variables*. Nosotros preferimos llamar *unidad de registro* a las unidades cuyos atributos se han registrado, con el fin de enfatizar que los datos no surgen de una percepción casi inmediata sino de operaciones que se apoyan en consideraciones teóricas y técnicas. Las unidades de registro y las variables se ordenan en la *matriz de datos*, que no es más que un arreglo rectangular con tantos renglones como unidades haya y con una columna para cada variable. Las casillas de la matriz, definidas por la intersección de renglones y columnas contienen "los valores"[6] de las variables.

La construcción de la matriz de datos es fundamental porque constituye tanto un punto de partida para la aplicación de las herramientas estadísticas, como un punto de llegada que enlaza la esfera de la conceptuación con la del registro empírico, nos parece que el precario tratamiento que han hecho de este tema, tanto la estadística como las ciencias sociales, oculta la complejidad de las operaciones que culminan en la matriz de datos y explica, en buena parte, las dificultades que enfrentan los investigadores para el mejor aprovechamiento de las técnicas en beneficio de su quehacer.

[6] Entrecomillamos este término porque en las escalas nominal u ordinal no se trata de valores en sentido estricto.

A continuación examinaremos algunos de los temas que involucra la construcción del arreglo rectangular, visto como punto de llegada, de unidades de registro y variables:

i) Las variables de la matriz de datos corresponden a los indicadores de los conceptos teóricos.[7] Hay que consignar la existencia de teorías en que algunos de sus conceptos son inobservables (Blalock, H. 1968: 5 a 27) ya que la correspondencia concepto-indicador puede involucrar varios indicadores para un concepto.

ii) La medición engloba tanto la operacionalización de conceptos teóricos como la confiabilidad y validez de los indicadores. A partir de una definición de cada concepto y de la especificación de sus dimensiones, se llega a establecer uno o más indicadores observables (Lazarsfeld, P. 1973: 35 a 45). La calidad de los indicadores se juzga por su grado de consistencia, estabilidad o precisión (confiabilidad) y por la certidumbre de que miden lo que queremos medir (validez) (Kerlinger, F. 1973: 442 a 473). La estadística auxilia la investigación social, proporcionándole conceptos y medidas (varianzas y correlaciones) que permiten evaluar la calidad de los indicadores.

iii) Cuando se tiene más de un indicador para un concepto se presenta el problema de sintetizar la información:

[7] En este trabajo reservaremos al término "variable" para referirnos a su sentido en estadística. En ciencias sociales se usa tanto en este sentido, como para designar a las categorías de mayor abstracción, a los indicadores e índices tal como se definen en la metodología de las ciencias sociales, y también se emplea en relación con un área temática (por ejemplo, cuando se dice "hay que considerar la variable poblacional").

operar sobre el subconjunto de columnas de la matriz de datos (variables) asociadas a un mismo concepto para reemplazarlas por un número menor de variables compuestas (índices). Los índices más utilizados suelen ser aquéllos que se obtienen por operaciones aritméticas elementales, como los índices sumatorios simples (en cada unidad de registro se suman los valores de las variables), o los índices que se construyen a través de cocientes y productos de variables. La estadística también proporciona técnicas que ayudan a la realización de estas operaciones teórico-metodológicas, entre las que se encuentran los análisis de componentes principales y factorial.

iv) La matriz de datos es independiente de las fuentes de información y de los métodos e instrumentos con que ésta se registre. Su forma no se modifica si el investigador obtiene su información de fuentes primarias o secundarias o si utilizó un cuestionario, una entrevista, una grabación, una filmación, un texto, o su propia observación. Lo que sí es esencial, para que sea susceptible de análisis estadístico, es que los datos sean numéricos, entendiendo por ello tanto los números que corresponden a variables métricas como los que pueden asociarse como códigos a variables no métricas.

v) Tampoco importa si la cobertura del estudio es censal o muestral. La aleatoriedad incorporada a la mayoría de las técnicas estadísticas multivariadas, se justifica no sólo por la selección de las unidades de registro a través del muestreo aleatorio, sino también por la imposibilidad de considerar todos los factores asociados a un fenómeno (Hagood, M. 1973: 65 a 78; Johnston, J. 1984: 14) o al argumentar que la aleatoriedad forma parte de los procesos sociales (King, G. 1989: 9 a 37).[8]

[8] La discusión sobre el carácter determinista o aleatorio de la realidad sigue presente en la ciencia. Véase el ríspido debate entre Ilya Prigogine y René Thom, en el encuentro organizado por la Fundación Salvador Dalí, en la Facultad de Física de la Universidad de Barcelona, en noviembre de 1985 (Wagensberg, J. 1986: 187 a 197).

vi) No es trivial, aunque parezca lo contrario, decidir cuál o cuáles son las unidades de registro pertinentes al problema que se investiga. ¿La teoría que orienta la investigación hace alusión a unidades individuales o colectivas? por ejemplo: qué es lo que interesa para un estudio: ¿la condición de "ocupado" o "desocupado" de los individuos o el número (o la proporción) de ocupados dentro de los miembros económicamente activos de un hogar? En el primer caso, las unidades de registro serán los individuos y una de las variables que los caracterice será su condición de ocupación; en el segundo, serán los hogares y una de sus variables el número de ocupados (o su proporción respecto a los económicamente activos del hogar).

vii) Las operaciones que habitualmente se aplican a la matriz de datos no se agotan en la eliminación de los indicadores que no pasaron las pruebas de confiabilidad y validez, ni en la construcción de índices sino que, en ocasiones, la investigación requiere transformar las unidades de registro. En estos casos se aprecia el doble carácter de las unidades de registro: son tales en cuanto sirvieron de base al registro empírico pero sólo constituirán *unidades de análisis* en tanto sean las relevantes para la teoría. Así, una conceptuación que centre la atención en hogares, ante la imposibilidad de acceder directamente a sus rasgos característicos, deberá llegar a la matriz de datos en dos etapas: en la primera se construirá una matriz en que las observaciones sean individuos y una de sus variables indique el hogar al que pertenecen; en la segunda etapa se construirán los hogares como nuevas unidades, unidades de análisis, generadas a partir de la matriz de datos de individuos. Los procedimientos de definición de las variables de los hogares, a partir de las de los individuos, van desde operaciones artiméticas simples como en el caso del índice de ocupación, hasta elaboraciones relacionales complejas para definir variables como el tipo de familia, a partir del parentesco de cada uno de los miembros con el jefe del hogar.

viii) Se presenta una complicación adicional cuando la investigación requiere del manejo estadístico simultáneo de unidades de registro heterogéneas, que deben combinarse en una única matriz de datos que refiera todas las variables a una misma unidad de análisis. Por caso, el problema puede demandar que se combinen las variables del hogar con las de algunos individuos seleccionados (el jefe, el cónyuge, el hijo mayor, el hijo menor) y con características de la vivienda como la zona de residencia, entre otras: a) la edad del cónyuge o la del hijo menor suelen utilizarse como indicadores del ciclo doméstico que, a su vez, se considera como uno de los factores explicativos de la participación femenina en el mercado de trabajo; b) algunas variables de la vivienda y de la zona de residencia pueden usarse en la construción de estratos sociales que, junto con el ciclo doméstico y otras variables, complementan el abanico de factores explicativos de la tasa de participación de la mujer. En los casos en que la estratificación se basa en pocas variables se pueden aplicar procedimientos estadísticos relativamente simples: representación gráfica, comparación de promedios y descomposición de la varianza; pero cuando son muchas, estos métodos no son eficientes y se requiere de algunos más elaborados: análisis de conglomerados, análisis clasificatorio múltiple (*multiple classification analysis*) y análisis discriminante, por citar algunos.

Para finalizar esta sección, debemos hacer notar que la matriz de datos, en la mayoría de las investigaciones (si no es que en todas), experimenta sucesivas transformaciones, las que difícilmente se comprenden con la imagen de un proceso de investigación que avanza, en forma continua, en dirección a su término y cuyo producto se integra "acumulándose en el cuerpo de conocimiento disponible" (Bunge, M. 1979: 19 a 37). Las transformaciones de la matriz de datos aparecen como "casi naturales" si la investigación se conceptúa como un proceso que combina fases de continuidad con rupturas y reordenaciones, es decir,

como un proceso caótico (piaget, J. y R. García,1982: 192 a 194; Prigogine, I. e I. Stengers 1983: 166 a 187; Lazlo, I. 1990: 137 a 149; Balandier, G. 1989: 226 a 237).

Técnicas estadísticas

La matriz de datos, si bien es el punto de llegada de las operaciones teórico-metodológicas, es también el punto de partida del análisis estadístico. En ocasiones, desde el momento en que se selecciona un procedimiento estadístico particular es necesario, para satisfacer sus supuestos, introducir cambios en la matriz de datos. A su vez, las preguntas llevan a seleccionar subconjuntos de variables o de unidades de registro: casi nunca se utiliza la matriz de datos en su totalidad. Más aún, los resultados que arrojan los primeros análisis (supongamos que sea un simple análisis de frecuencias) conllevan, la mayoría de las veces, recodificaciones y redefiniciones de las variables, lo que implica volver al plano de la teoría a la vez que inducen nuevas preguntas de investigación que pueden requerir instrumentos estadísticos diferentes.

Dedicaremos esta sección a presentar la relación entre la estructura del sistema de hipótesis teóricas y la estructura de los instrumentos básicos que proporciona la estadística social. Dados los propósitos de este trabajo se acentuará la exposición de la estructura y características de las técnicas, pero el lector debe retener que, en la práctica, la referencia al campo teórico es permanente.

Para realizar la exposición dividiremos los modelos de análisis estadístico en dos grandes grupos. El criterio de clasificación se inclina en favor del tipo de problema sustantivo que se quiere resolver y responde al uso más frecuente que se hace de las distintas técnicas en las ciencias sociales. En el primer grupo incluimos las que se utilizan con la finalidad principal de construir índices o de clasificar (formar

grupos, estratos, zonas, regiones). En el segundo, están los instrumentos que permiten analizar relaciones conceptuales en la forma de relaciones entre variables. Aunque sea trivial, vale la pena destacar que ésta es sólo una de las varias maneras en que se pueden agrupar las técnicas estadísticas y que responde al tipo de aplicación que habitualmente se hace de ellas, pero de aquí no debe derivarse que no se puedan usar para propósitos distintos; por ejemplo, si bien el análisis factorial se utiliza preferentemente para construir índices, también puede emplearse para contrastar hipótesis (Kim J. O. y Ch. W. Mueller 1978: cap. v; Long, J. S. 1983).

Técnicas para construir índices y clasificar

Análisis de conglomerados (o de cúmulos)

Tiene por propósito agrupar a las unidades de registro cuyas características son las "más parecidas".

La idea básica que subyace a la formación de grupos es que éstos deben ser internamente homogéneos y, a la vez, lo más diferenciados posible entre sí. El caso más simple es el de la formación de grupos a partir de una sola variable, para ello se utilizan procedimientos gráficos, análisis de la distribución de frecuencias y técnicas de descomposición de la varianza. En el método gráfico se toman como referencia los máximos y mínimos de la distribución de frecuencias y se delimitan los valores de variable que marcarán las fronteras entre los grupos. Este recurso se puede complementar analizando el comportamiento de los promedios grupales y de las inter e intravarianzas. También es común examinar los valores de la variable y tomar como puntos de corte aquéllos en que se aprecian discontinuidades.

Sin embargo, cuando el problema involucra a más de una variable, estos criterios pierden operatividad. Las decisiones de agrupación no tienen por qué ser las mismas para

todas las variables: una unidad de registro según una de sus características debería pertenecer a un grupo y, según otra, a uno distinto.

El análisis de conglomerados resuelve este problema considerando simultáneamente *todas las variables*.

El problema consiste en comparar los renglones de la matriz de datos, en las variables seleccionadas, para definir los grupos o conglomerados y decidir cuáles son los más parecidos. Para iniciar un agregado se razona del siguiente modo: dado un renglón se busca entre los restantes al que más se le parezca, se tiene así un grupo de dos elementos; si no hay ninguno similar se comienza la formación de un segundo grupo. El tercer caso se asigna a uno de los grupos ya creados, excepto si es muy distinto. Este procedimiento se repite hasta que todas las unidades de registro hayan sido asignadas a un grupo.

Lo importante para el procedimiento descrito es tener un índice que permita medir el "parecido" de las unidades de registro. El análisis de conglomerado pone varias opciones a disposición del investigador: la distancia euclidiana entre las unidades de registro, la distancia euclidiana al cuadrado, la diferencia entre los valores absolutos, etcétera. (Tryon, R. y D. Bailey 1970: 135 a 181).

Este método se inicia con tantos conglomerados como casos haya y concluye con todos los casos formando un solo grupo. Es decisión del investigador seleccionar cuántos grupos quiere distinguir. Para ello se apoya bien sea en la evolución del valor de la medida resumen utilizada, en gráficas que producen el mismo método o en el análisis de promedios y de varianzas (de las variables) de los conglomerados construidos. Además, puede tomarse en cuenta información externa para afinar la conformación de los grupos, como sería la contigüidad geográfica para una regionalización.

Dependiendo del tamaño de la matriz de datos y de los recursos de cómputo, una aplicación particular puede de-