

ANÁLISIS EXPLORATORIO Y ANÁLISIS CONFIRMATORIO DE DATOS

Javier Parra Olivares*

Resumen

Hoy día, en el marco de la crisis de los grandes paradigmas teóricos, las universidades y centros de investigación siguen privilegiando, con marcado énfasis, los análisis de datos que conducen a la Confirmación de teorías o hipótesis, en contraste con un reducido interés en las herramientas estadísticas que orientan a la Exploración de datos sobre problemas de la realidad social. Ésto es común aún en campos o áreas relativamente nuevas o poco trabajadas por la ciencia social (Parra, 1990). A menudo, se olvida que, en la ciencia, muchas veces es más importante encontrar la pregunta que hallar la respuesta, lo cual apunta a la posible formulación de enfoques diferentes a los actuales y al enriquecimiento del universo intelectual de las diversas disciplinas de lo social.

Como consecuencia de lo anterior, se requiere la enseñanza y empleo de ambos enfoques de análisis de datos de investigación: tanto el exploratorio como el confirmatorio, como medios para el avance de las ciencias sociales. Este trabajo intenta realizar un aporte en este sentido, presentando algunas de las cuestiones claves que plantea la lógica o la práctica de la investigación científica y derivando de aquellas las diferencias de la Exploración y de la Confirmación, en especial en lo atinente a la actitud ante el conocimiento.

Palabras clave: *Análisis exploratorio, análisis confirmatorio, paradigmas, actitud científica, estadística, ciencias sociales.*

Recibido: 27-10-00 Aceptado: 10-11-01

* Escuela de Sociología. Universidad del Zulia. Maracaibo, Venezuela. E-mail: animarte@telcel.net.ve

Exploratory and Confirmatory Analysis of Data

Abstract

Today, within the framework of the crisis of the great theoretical paradigm, universities and research centers in Latin America still use, with high frequency, confirmatory data analysis in relation to theories and hypothesis, in contrast with the little interest seen in statistical tools oriented towards exploratory data analysis of social realities. This is common even in new and little worked fields or areas of social science. We often forget that finding the question is often more important than finding the answer in science. This is very useful for the probable formulation of different approaches in order to increase the intellectual universe of social science. For this reason, we need to teach and use both approaches to data analysis: exploratory and confirmatory, as ways to advance social science. This paper attempts to contribute to this theme, presenting some of the key areas that propose the logic and practice of deducing the differences between confirmatory and exploratory data analysis, with special emphasis on attitudes towards knowledge.

Key words: *Exploratory analysis, confirmatory analysis, paradigms, scientific attitude, statistics, social sciences.*

El Problema

Durante mucho tiempo, en las ciencias sociales, el análisis de datos ha tenido diversidad de papeles que podrían ser resumidos o clasificados en dos tipos: el Confirmatorio y el Exploratorio. El **paradigma de investigación científica** más familiar para la mayoría de la gente podría ser resumido como se observa en el siguiente cuadro:

Sin embargo, no se puede obviar que, si se pensara a la producción científica sólo como resultado de este esquema confirmatorio, se estaría dejando de lado a buena parte de la práctica y el esfuerzo cotidiano de la investigación.

En este sentido, cabría preguntarse (y responderse): Como se observa en el siguiente cuadro:

Se podría decir, en consecuencia, que para poder implementar el paradigma confirmatorio se requiere, sin duda, realizar una gran cantidad de trabajo exploratorio. Por lo tanto, ninguno de los dos tipos de análisis son suficientes por sí solos para contribuir al avance de la ciencia.

Pregunta
(Problema de Investigación)



Diseño



Recolección de Datos



Análisis



Respuesta



Si se retoma el esquema del paradigma anterior, agregándole un paso en la primera etapa, se podría visualizar mejor lo que realmente se da en la investigación:

Por ejemplo: si tenemos la idea de que existe un cierto **valor A** en las personas que contribuye sólidamente a que éstas mantengan una determinada conducta **B** ante su entorno y queremos comprobarlo, entonces aún no hemos formulado una pregunta (o problema) en el sentido subyacente en los esquemas paradigmáticos arriba mencionados. Lo que tenemos es una idea de una pregunta pero no el tipo de pregunta que puede tener una respuesta con soporte estadístico (o fundamentada en datos) observese en le siguiente cuadro.

La clase de pregunta que si tiene una respuesta debe estar mejor delimitada y su selección debe depender de cuestiones prácticas y no sólo de deseos. Por ejemplo, desearíamos responder una pregunta como esta: “Entre las personas que poseen el **valor A**, ¿cuál porcentaje presenta la conducta B ante su entorno y qué porcentaje no tiene esa conducta?”

La formulación de la pregunta en sí involucra a) de hecho, qué puede ser preguntado, b) cuáles diseños son factibles y c) cuán probable es que un diseño

Pregunta	Respuesta
1) ¿Cómo se generan las preguntas?	La mayoría de las preguntas se producen a partir de ideas “cuasi-teóricas” y por la exploración de datos pasados.
2) ¿Cómo se orientan los diseños?	Con frecuencia se guían por la mejor información (cualitativa o cuantitativa) que esté disponible proveniente, también, de la exploración de datos pasados.
3) ¿Cómo se vigila la recolección de datos?	Con base en la exploración de los datos tal como se presentan con la finalidad de encontrar comportamientos no esperados.
4) ¿Cómo se supervisa el análisis o qué nos indica cuáles datos debemos descartar?	Generalmente se hace por la exploración de los datos antes, durante y después del análisis en busca de rastros, ideas y algunas conclusiones a un tanto por ciento de confianza.

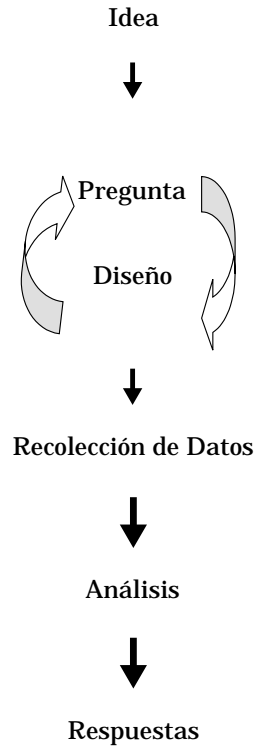
dado genere una respuesta útil. Por tanto, de entrada la exploración perspicaz y extensiva (de datos pasados) pueden (y deben) estar presentes en este proceso de formular la pregunta de investigación.

En consecuencia, **la ciencia ni comienza con una pregunta ordenada o clara, ni finaliza con una respuesta ordenada y clara.**

Por tanto, necesitamos como analistas de datos o como estadísticos recordar que a menudo es más importante encontrar la pregunta que hallar la respuesta.

Análisis Exploratorio de Datos

Es un modo de análisis de los datos que utiliza el resumen numérico y visual para explorar datos **en busca de patrones no anticipados**. Autores clásicos como John Tukey, Frederick Hartwig y Brian Dearing lo catalogan como



un “estado mental” ante el conocimiento. Es así como el primero lo define como una actitud, una flexibilidad y “algunas hojas con gráficos” (o transparencias, o ambos). Esto último como un reconocimiento de que el ojo que mira al horizonte es el mejor instrumento que tenemos para observar, de manera completa, lo no anticipado (Tukey, 1980). Por su parte, Hartwig y Dearing argumentan que el investigador debe aprender todo lo posible acerca de una variable o conjunto de variables antes de utilizar los datos para probar hipótesis o teorías acerca de las relaciones sociales (Hartwig y Dearing, 1979). Más recientemente, Eugene Horber y Dominique Ladiray plantean que el “razonamiento” exploratorio es un esquema de análisis que enriquece las posibilidades del investigador para hallar nuevas respuestas a los problemas que se planteen (Horber y Ladiray, 1995).

En general, el análisis exploratorio de datos se caracteriza por el uso de herramientas o técnicas con mucha carga visual o gráfica, con énfasis en revelar información vital sobre la data examinada. El arsenal correspondiente está compuesto, entre otros, por instrumentos como:

Diagrama de caja y bigotes (Box-and-whisker): un resumen visual de la distribución (comportamiento) de una variable que provee detalles acerca de si uno o ambos extremos de la distribución contienen valores inusualmente grandes o pequeños (Ver Gráfico No. 1).

Diagrama de tallo y hojas (Stem-and-leaf): es una muestra visual de la distribución de una variable. Se asemeja a un histograma y "... se usa mucho para mostrar tanto el orden de rangos como la forma de un conjunto de datos en forma simultánea". (Anderson y otros, 1999: 40). Cada caso es representado por uno o más dígitos colocados a la derecha de una línea vertical y en la fila correspondiente al primer dígito del valor observado (Ver Gráfico No. 2).

Diagrama de dispersión (Scatter plot): gráfico que muestra la relación entre variables. Es útil para examinar la dirección, fuerza y forma de la relación (Ver Gráfico No. 3).

Gráfico N° 1
Diagrama de Caja y Bigotes

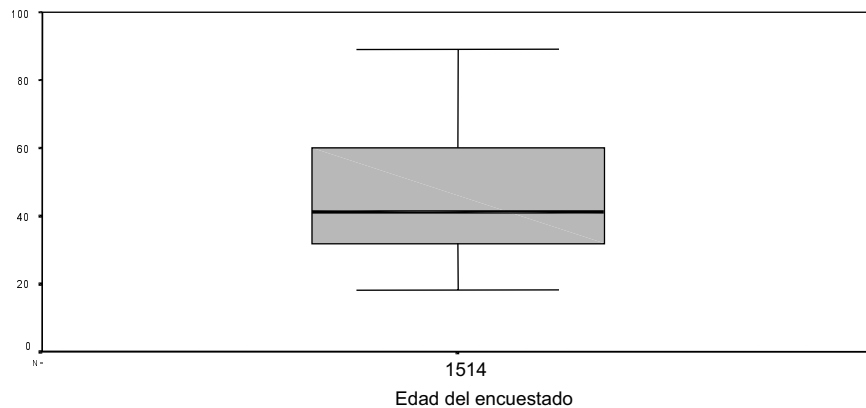


Gráfico N° 2

Diagrama de Tallo y Hojas

Número de años de escolarización Stem-and-Leaf Plot

Frecuencia Tallo y hojas

55,00 Extremos (=,0)

68,00 8. 0000000

56,00 9. 0000000

73,00 10. 00000000

85,00 11. 00000000

461,00 12. 000

130,00 13. 00000000000000

175,00 14. 000000000000000000

73,00 15. 00000000

194,00 16. 00000000000000000000

43,00 17. 00000

45,00 18. 00000

22,00 19. 00

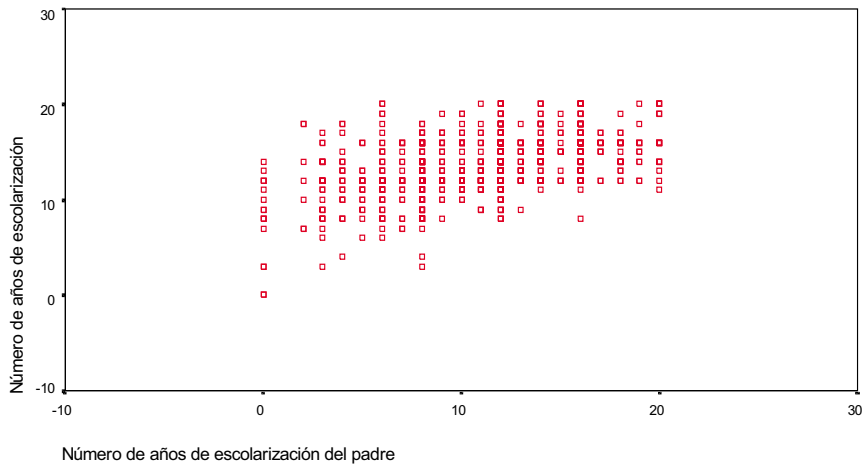
30,00 Extremos (=20)

Ancho de las hojas: 1

Cada hoja: 10 caso(s)

Gráfico N° 3

Diagrama de Dispersión



El uso de las herramientas para el análisis exploratorio de los datos requiere una mente muy abierta para la búsqueda de información y una actitud muy dispuesta y paciente para el “rastreo” del comportamiento de las variables.

Análisis Confirmatorio de Datos

Es un modo de análisis de datos que utiliza estadísticos numéricos de resumen generados a partir del empleo de un modelo, definido a priori, para confirmar o no una hipótesis. Se caracteriza por el empleo de indicadores como la media, la varianza y los coeficientes de correlación y regresión, así como las pruebas de hipótesis. Es, sin duda, el modelo de análisis de datos más enseñado y, por ende, más empleado en las ciencias sociales, por lo que no se requiere presentar mayores detalles del mismo en este artículo.

No obstante, es importante resaltar que, en muchos casos, la actitud ante los datos que se deriva de este esquema de análisis está muy orientada a registrar sólo ciertos indicadores de resumen y a emplear, de manera mecánica, determinadas técnicas de comprobación de hipótesis, sin aprovechar, lo más posible, la riqueza de la información contenida en la data.

Así mismo, puede decirse que este es el modelo de análisis de datos cuya rutina es más sencilla de enseñar y de estandarizar mediante el uso de computadores, como se puede constatar por el desarrollo de programas informáticos como el SPSS, SAS, SPAD y STATGRAPHICS, entre otros. Posiblemente esto ha contribuido, aún más, a reforzar la actitud mencionada en el párrafo anterior acerca del empleo mecánico, casi sin reflexión, del modo de análisis confirmatorio de datos.

A manera de comparación

La mayoría de las veces los análisis estadísticos están diseñados para responder preguntas como: ¿Confirman estos datos la hipótesis de que la variable **X** está relacionada con la variable **Y** en un contexto definido?. En este esquema confirmatorio de análisis, se ajusta un modelo de relación (a menudo lineal) a los datos, se obtienen estadísticos de resumen de los datos (tales como las medias y las varianzas explicadas) y estos últimos se prueban contra la probabilidad de que valores tan altos como los obtenidos podrían haber ocurrido por azar. Esta manera de analizar los datos no sólo pone demasiada confianza en los estadísticos de resumen,

sino que también carece de “apertura”, en la medida en que sólo se han considerado dos alternativas para el análisis. Es decir, los datos no se han explorado para observar qué otros patrones de comportamiento pueden existir en ellos.

Un enfoque alternativo consiste en realizar una pregunta diferente: “¿Qué pueden decir estos datos acerca de la relación entre las variables X e Y en el contexto definido? En contraste con el modo confirmatorio, este esquema exploratorio de análisis está abierto a un rango más amplio de explicaciones alternativas, que incluyen, entre otras, la aleatorización y algunos modelos teóricos ya existentes.

El investigador o analista, en consecuencia, debe estar abierto a posibilidades que no espera encontrar, particularmente en el caso de teorías o marcos teóricos muy débiles que no especifican los modelos o tipos de relación entre las variables, sino que sólo plantean que están relacionadas, situación muy común en estos tiempos de caída de paradigmas conceptuales.

Reflexiones finales

- a. Es recomendable enseñar en las universidades tanto el Análisis Exploratorio de datos como el Confirmatorio, incluso en ese orden.
- b. La actitud ante la investigación sería más abierta si se enseñara también el análisis de los datos con la finalidad de buscar comportamientos inusuales de los mismos y no sólo para juzgar si la data se ajusta a lo planteado por ciertas hipótesis o modelos.
- c. En los tiempos actuales, cuando los grandes paradigmas teórico-explicativos han entrado en crisis, la mirada de los científicos sociales debería girar hacia el uso de modelos analíticos que permitan explorar datos, en aras de plantearse nuevas y diferentes relaciones entre las variables, enriqueciendo así el ambiente intelectual y la producción de conocimientos.
- d. Lo que se afirma aquí no es que el análisis confirmatorio de datos no tiene lugar en las Ciencias sociales, sino que el mismo no debe ser el único y exclusivo modelo a enseñar, máxime cuando, en muchos casos hoy día, ni siquiera se cuenta con teorías sólidas para ser confirmadas.

Referencias Bibliográficas

- ANDERSON, David; SWEENEY, Dennis y WILLIAMS, Thomas (1999) **Estadística para Administración y Economía**. México: International Thomson Editores.
- HARTWIG, Frederick y DEARING, Brian (1979) **Exploratory Data Analysis**, California: Sage Publications.
- HORBER, Eugene y LADIRAY, Dominique (1995) **Análisis Exploratorio de los datos**, II Seminario de Capacitación de docentes en "Producción y tratamiento de datos de Investigación en ciencias humanas", Caracas, Venezuela.
- PARRA, Javier (1999) "Contenidos de Estadística en Universidades de Latinoamérica". (monografía). Centro de Estadística e Investigación de Operaciones de La Universidad del Zulia, Maracaibo, Venezuela.
- TUKEY, John (1980) "We need both Exploratory and Confirmatory" **The American Statistician**, Vol. 34, No. 1.