Modelos de clases latentes aplicados a las encuestas de percepción ciudadana: estudio de caso

*Claudio R. Castro López**
*Alma Janett Tenorio Aguirre***

Introducción

En el ámbito de las ciencias sociales existe una gran cantidad de situaciones o fenómenos que no pueden ser directamente observados o cuantificados, ya sea porque se trata de un concepto abstracto o una característica subyacente (la calidad de vida, el liderazgo de un gobernante, los resultados de una gestión gubernamental, etcétera. Los conceptos de esta naturaleza pueden agruparse bajo la denominación genérica de variables latentes, es decir, subyacen en el fenómeno bajo estudio, pero no son directamente observables. Su estudio se lleva a cabo mediante variables observadas (manifiestas), que se consideran indicadoras de estas variables. La idea principal es que las variables indicadoras sirvan para definir o medir la variable latente (Vermunt & Magidson, 2000). El estudio de las llamadas variables latentes, ha dado lugar al surgimiento de los conocidos modelos de variables latentes.

* Doctor en Estadística Multivariante Aplicada de la Universidad Veracruzana. Líneas de investigación: educación estadística, técnicas estadísticas aplicadas al marketing, análisis multivariante en datos cualitativos, modelación de tablas de contingencia multivariantes. Correo electrónico: ccastro@uv.mx

** Licenciada en Estadística y candidata a maestra en Gestión de la Calidad de la Universidad Veracruzana. Líneas de investigación: análisis de datos y elaboración de reportes, muestreo, métodos multivariados, nuevas tecnologías para el desarrollo de encuestas y control de calidad. Correo electrónico: almtenorio@uv.mx

Para el estudio de estos tipos de fenómenos existe una serie de técnicas y modelos estadísticos que aunque han demostrado su utilidad en el análisis, son poco conocidos por los investigadores sociales. El objeto principal del presente artículo es plantear los aspectos metodológicos más destacados de estas técnicas y modelos, así como promover su uso mostrando una aplicación en estudios de percepción ciudadana.

El análisis de clases latentes es una técnica estadística que consiste en clasificar a los individuos de una población en segmentos o clases de naturaleza exhaustiva y excluyente; es una técnica óptima basada en criterios relacionados con aspectos internos de los individuos, como actitudes, percepciones, preferencias y, en general, cualquier otro aspecto de naturaleza subjetiva.

Autores como Lazarsfeld y Henry (1968) o Goodman (1974), entre algunos otros, aportan las ideas iniciales de este tipo de modelos. Trabajos como los de Lindsay, Clogg y Greco (1991); Uebersax (1993); Magidson y Vermunt (2001) o Sepúlveda (2004), dan cuenta del gran desarrollo que han alcanzado algunos métodos y modelos relacionados con el análisis de clases latentes.

Una de las principales ventajas de esta técnica de segmentación frente a las técnicas tradicionales es su carácter confirmatorio. Al igual que otras técnicas estadísticas empleadas para segmentar, como el análisis factorial o el análisis cluster, el análisis de clases latentes es un método exploratorio de poblaciones o muestras, sin embargo, por encima de su naturaleza exploratoria, el análisis de clases latentes permite realizar todo tipo de investigaciones confirmatorias sobre la naturaleza del concepto latente (como la propia existencia del concepto, la adecuación de los indicadores empleados para su estudio, la óptima distribución de la población en los segmentos identificados, el tamaño de cada segmento, el comportamiento de los individuos ubicados en cada segmento, etcétera).

El análisis de información contenida en una encuesta puede ser de gran utilidad cuando se realiza de forma correcta; sin embargo, existe una gran variedad de estudios de opinión que consideran fenómenos que no son observados de manera directa; tales fenómenos podrían ser estudiados en una denominación de conceptos latentes. Aquí se

presenta un uso de las clases latentes aplicadas a una encuesta de percepción ciudadana en el estado de Veracruz.

Modelo básico del análisis de clases latentes

Un modelo de variables latentes se define simplemente como un modelo estadístico que especifica la distribución conjunta de un grupo de variables aleatorias en el cual alguna de estas variables —variable latente— no es observable. Las relaciones de dependencia entre las variables categóricas de una tabla de contingencia en muchos casos están provocadas por la existencia de una asociación entre cada una de ellas y otra variable no observable directamente, llamada *variable latente.*

El análisis de clases latentes es una técnica estadística que considera la obtención de una variable latente con C categorías, las cuales representan un grupo; es decir, ésta permite estudiar la existencia de variables latentes a partir de un conjunto de variables explicativas observadas y así definir una clasificación. Esta técnica surge por la necesidad de explicar la relación existente entre un conjunto de p variables observadas directamente, $\mathbf{X}'=\left(X_1, X_2, \ldots, X_p\right)$, medidas sobre una muestra de n individuos. Pioneros como Galton y Spearman hicieron una valiosa aportación, ya que fueron quienes plantearon que la relación existente podría ser definida mediante las variables latentes denotadas por Y, las cuales se expresan mediante el vector $\mathbf{Y}'=\left(Y_1, \cdots, Y_q\right)$, con $q < p$, y que, por tanto, no era necesario obtenerlas físicamente, razón por la cual fueron denominadas variables latentes y cada categoría de Y es denominada una clase latente.

Los modelos de clases latentes pueden clasificarse de acuerdo con la escala de las variables observadas y las variables latentes; según Bartholomew y Knott (1999), existe una doble clasificación en variables métricas y categóricas. Las variables métricas son aquellas que toman valores en el conjunto de los números reales y éstas pueden ser tanto discretas como continuas; las variables categóricas son las formadas por un conjunto de categorías, no necesitan ser nominales, también pueden considerarse variables ordinales o de intervalo dis-

cretizadas. Según lo anterior, en la Tabla 1 se plantea un esquema de clasificación para los análisis de clases latentes.

Tabla 1. Esquema de clasificación

		Variables manifiestas X	
		Métricas	*Categóricas*
Variables latentes *Y*	Métricas	*Análisis factorial	*Análisis de rasgos latentes *Análisis factorial de datos categóricos
	Categóricas	*Análisis de perfiles latentes	* Análisis de clases latentes

Algunos aspectos que caracterizan los modelos de clases latentes orientados hacia la identificación de segmentos son:

- *Clasificación de personas en distintos segmentos (categorías de la variable latente resultante):* se basa en probabilidades de pertenencia estimadas directamente a partir del modelo, mientras que en algunos algoritmos tradicionalmente utilizados la clasificación está basada en la proximidad de un individuo a otro, conforme a alguna medida de distancia específica y algún algoritmo de asociación entre los individuos.
- *La escala de medición de las variables:* de entrada al modelo puede ser continua, dicotómica, nominal u ordinal, conteos o combinaciones de ellas, mientras que algunas de las herramientas estadísticas tradicionales no permiten el uso de variables dicotómicas o nominales, y algunas veces incluso ordinales.

Existen dos supuestos a considerar en un análisis de clases latentes, uno de los supuestos básicos en el modelo de clases latentes es el de independencia local o condicional: las variables indicadoras son estadísticamente independientes dentro de cada clase latente y, por tanto, la variable latente es suficiente para explicar las relaciones exis-

tentes entre estas variables. Este supuesto implica que las variables latentes causan la relación existente entre las variables observadas, por consiguiente no existe una relación directa entre las variables observadas; es decir, éstas están correlacionadas entre sí, pero esta correlación desaparece si las variables latentes permanecen constantes.

Otro de los supuestos básicos en el modelo es el de homogeneidad interna de las variables latentes: cada uno de los miembros de una clase latente tiene una distribución de probabilidad igual respecto a la de la variable latente, y ésta será diferente a la de los individuos pertenecientes a cada clase, por lo que cada individuo de diferente clase tendrá características diferentes, es decir, este supuesto se utiliza para diferenciar a los individuos pertenecientes a diferentes clases y poder diferenciar tanto la variable latente como las clases latentes.

Si se considera una variable latente Y con C categorías o clases latentes y *p* variables observadas indicadoras ($x_1, x_2, ..., x_p$) de la variable latente, éstas variables conforman el modelo de clases latentes, el cual está definido por:

$$p(\mathbf{X}=\mathbf{x}) = \sum_{c=1}^{C} p(\mathrm{Y}=c, \mathbf{X}=\mathbf{x}) \tag{1}$$

donde $X = x_1, x_2, ..., x_p$ es el vector de variables manifiestas; $X = x_1, x_2, ..., x_3$ es un patrón de respuesta cualesquiera; $\mathrm{p}(\mathbf{X}=\mathbf{x})$ es la probabilidad conjunta de que las variables manifiestas sean iguales a un cierto patrón de respuesta, y

$$\sum_{c=1}^{C} p(\mathrm{Y}=c, \mathbf{X}=\mathbf{x})$$

es la probabilidad conjunta de tener un patrón de respuesta x, el cual pertenece a la clase latente *c.* El modelo (1) se puede expresar por:

$$\begin{aligned} p(\mathbf{X}=\mathbf{x}) &= \sum_{c=1}^{C} p(Y=c, X=x) \\ &= \sum_{c=1}^{C} p(Y=c)p(X=x/Y=c) \end{aligned} \tag{2}$$

donde $p(Y = c)$, es la probabilidad de pertenecer a la clase latente c, conocida como probabilidad *a priori*, y $p(\mathbf{X} = \mathbf{x}/Y = c)$ es la probabilidad condicional de obtener un determinado patrón de respuesta para un individuo de la clase latente c. Así, (2) se puede expresar por:

$$\begin{aligned} p(\mathbf{X} = \mathbf{x}) &= \sum_{c=1}^{C} p(Y = c, \mathbf{X} = \mathbf{x}) \\ &= \sum_{c=1}^{C} (Y = c) p(\mathbf{X} = \mathbf{x}/Y = c) \\ &= \sum_{c=1}^{C} p(Y = c) \prod_{p=1}^{p} p(\mathbf{x}_p = x_p/Y = c) \end{aligned} \tag{3}$$

donde $p\, x_p = x_p/Y=c$ es la probabilidad de obtener un determinado valor en la variable x_p, para un individuo de la clase latente c.

Individuos con patrón de respuesta son clasificados dentro de la clase latente c, utilizando un asignamiento modal, es decir, los individuos se asignan a la clase latente para la cual su probabilidad *a posteriori*, $p(Y = c/\mathbf{X} = \mathbf{x})$, es mayor.

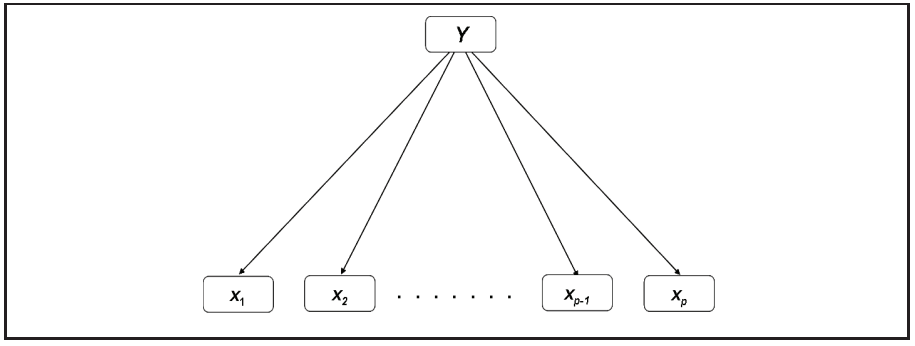
Para estimar las probabilidades *a posteriori*, se utiliza el teorema de Bayes:

$$p(Y = c/\mathbf{X} = \mathbf{x}) = \frac{p(Y = c, \mathbf{X} = \mathbf{x})}{p(\mathbf{X} = \mathbf{x})} \tag{4}$$

cuyo numerador y denominador están definidos en las fórmulas anteriores.

El modelo de clases latentes puede ser descrito según la perspectiva de la teoría de grafos en la Figura 1, la cual describe que las variables observadas $X_1, X_2, \ldots X_{p-1}, X_p$, no se encuentran directamente relacionadas entre sí, pero cada una de éstas puede ser afectada por la variable Y, y cuando la categoría o nivel de la variable Y deja de ser fija, estos efectos entre Y, y las p variables pueden producir la aparente relación entre estas últimas.

Figura 1. Representación gráfica de un modelo de clases latentes formado por una variable latente y p variables manifiestas



Modelo de clases latentes aplicado al análisis de la percepción ciudadana del estado de Veracruz

Los estudios de opinión constituyen una herramienta fundamental para obtener información sobre diversos fenómenos en las sociedades contemporáneas. En el marco de las diversas actividades de gobierno, los estudios de opinión proveen información de utilidad para las evaluaciones de desempeño de los programas y acciones, así como para conocer la percepción que los ciudadanos tienen sobre diversos rubros de la acción política, de los gobernantes y de su gobierno.

En este sentido, se constituyen en insumos importantes para el análisis y la toma de decisiones. Desde luego que esto último se logrará, en la medida que los datos se procesen y analicen apropiadamente, que se obtengan resultados que después se conviertan en información útil, y que, finalmente, la información se convierta en conocimiento.

Actualmente, en los ámbitos público y privado se confía en la investigación mediante el análisis de encuestas de opinión con la finalidad de conocer cuáles son las necesidades de las personas y sus opiniones respecto a distintos temas de interés. En este sentido, si tal análisis se realiza correctamente, puede proporcionar información confiable; sin embargo, es posible que exista información que involucra un conjunto de variables que tratan de describir un fenómeno,

y que sea necesario analizarla con estudios más complejos como el análisis de clases latentes.

Los datos que se usan en el análisis estadístico de este artículo se recopilaron de encuestas realizadas por Percigove (Sistema de Percepción Ciudadana sobre las Acciones del Gobierno de Veracruz, de la Red Universitaria de Estudios de Opinión: Universidad Veracruzana) durante 2005. Se trabajó con una base de datos en la que se concentró la opinión se 5,127 ciudadanos del estado de Veracruz. La encuesta se realizó con una metodología de muestreo estratificado con un error de 2% y un nivel de confianza de 95 por ciento.

El diseño estadístico de clases latentes permite construir una variable nominal no observada (latente) con C categorías, las cuales representan a cada uno de los segmentos identificados en la población bajo estudio.

Uno de los objetivos de la encuesta fue clasificar a los individuos de acuerdo con su opinión respecto a si el gobernador del estado posee o no determinadas características. En el cuestionario se incluyó un grupo de variables empleadas para describir algunas características del gobernador en ese momento. La Tabla 2 muestra este conjunto de cuestionamientos.

En cada una de las variables se tomaron en cuenta cuatro categorías: "Totalmente en desacuerdo", "En desacuerdo", "De acuerdo" y "Totalmente de acuerdo".

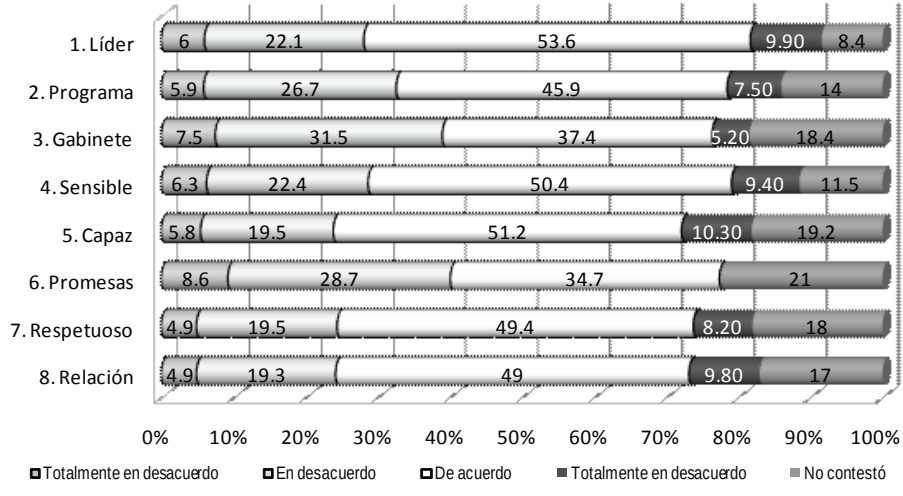
En la Figura 2 se presentan los resultados exploratorios de estas variables con el fin de describirlas. Se puede destacar que 31.5% de los ciudadanos mencionó estar totalmente en desacuerdo con que el gobernador *tiene un gabinete que conoce sus respectivas áreas de atención.* De igual forma se destaca que las afirmaciones en las que están mayormente de acuerdo o totalmente en desacuerdo son: *Es un líder político que conoce los problemas del estado de Veracruz; Es una persona honesta y capaz; y Es un veracruzano sensible a escuchar las necesidades ciudadanas.*

Para realizar el análisis de clases latentes se utilizó el *software* Latent GOLD® 4.0; y como primer paso, se busca la existencia de independencia entre las variables o si, por el contrario, puede explicarse su comportamiento temporal a través de una variable latente.

Tabla 2. Grupo de variables observadas

1.	Es un líder político que conoce los problemas del estado de Veracruz.
2.	Tiene un programa de gobierno congruente con las necesidades del estado.
3.	Tiene un gabinete que conoce sus respectivas áreas de atención.
4.	Es un veracruzano sensible a escuchar las necesidades ciudadanas.
5.	Es una persona honesta y capaz.
6.	Será un gobernador que cumpla sus promesas de campaña.
7.	Será un gobernador respetuoso del Congreso del Estado.
8.	Será un gobernador que mantendrá una buena relación con el presidente de México.

Figura 2. Resultados exploratorios sobre el gobernador del estado



La Tabla 3 muestra que se rechaza el supuesto de independencia dado el valor (p=0.0000); por tanto, se puede admitir que existe una asociación entre las variables observadas a través de una variable latente.

Al considerar que no existe modelo perfecto y que siempre son preferibles los modelos con menos variables, puesto que además de ser más sencillos son más estables y menos sometidos a sesgo, en este caso para elegir el modelo adecuado se consideró el criterio conocido como AWE (peso promedio de evidencia) (Banfield y Raftery, 1993).

La Tabla 4 presenta las proporciones (probabilidades) estimadas a partir del modelo seleccionado, encontrándose que el segmento 1 es el más grande y representa 43.91% de la población.

Tabla 3. Prueba de independencia

	X^2	Prob.
Independencia	51034000000	0.000

Tabla 4. Tabla de probabilidades

	Segmento 1	*Segmento 2*	*Segmento 3*	*Segmento 4*	*Segmento 5*
Tamaño del segmento	0.4391	0.2787	0.1894	0.0629	0.0298
1) Es un líder político que conoce los problemas del estado de Veracruz					
Totalmente en desacuerdo	0.0013	0.0209	0.1819	0	0.8432
En desacuerdo	0.0891	0.3181	0.5761	0.0012	0.1532
De acuerdo	0.7895	0.6391	0.2403	0.2336	0.0037
Totalmente de acuerdo	0.12	0.022	0.0017	0.7652	0
2) Tiene un programa de gobierno congruente con las necesidades del estado					
Totalmente en desacuerdo	0.0006	0.0375	0.1706	0	0.8026
En desacuerdo	0.0853	0.5052	0.6594	0.0007	0.1943
De acuerdo	0.8325	0.4533	0.1696	0.2276	0.0031
Totalmente de acuerdo	0.0816	0.0041	0.0004	0.7717	0
3) Tiene un gabinete que conoce sus respectivas áreas de atención					
Totalmente en desacuerdo	0.0043	0.0667	0.2098	0	0.8986
En desacuerdo	0.213	0.5834	0.6524	0.0095	0.1007
De acuerdo	0.7197	0.3446	0.137	0.4462	0.0008
Totalmente de acuerdo	0.063	0.0053	0.0007	0.5444	0

4) Es un veracruzano sensible a escuchar las necesidades ciudadanas					
Totalmente en desacuerdo	0.0007	0.0218	0.2016	0	0.8671
En desacuerdo	0.0769	0.3616	0.6113	0.0006	0.1309
De acuerdo	0.8102	0.6034	0.1863	0.1864	0.002
Totalmente de acuerdo	0.1121	0.0132	0.0007	0.8131	0
5) Es una persona honesta y capaz					
Totalmente en desacuerdo	0.0007	0.0193	0.1806	0	0.8654
En desacuerdo	0.0645	0.3058	0.5728	0.0007	0.1319
De acuerdo	0.7982	0.6555	0.2452	0.2044	0.0027
Totalmente de acuerdo	0.1365	0.0194	0.0015	0.7949	0
6) Será un gobernador que cumpla sus promesas de campaña					
Totalmente en desacuerdo	0.0009	0.0898	0.2462	0	0.9683
En desacuerdo	0.1179	0.6517	0.6581	0.0012	0.0316
De acuerdo	0.7981	0.2569	0.0955	0.2444	0.0001
Totalmente de acuerdo	0.0831	0.0016	0.0002	0.7544	0
7) Será un gobernador respetuoso del Congreso del Estado					
Totalmente en desacuerdo	0.0003	0.0104	0.1401	0	0.9386
En desacuerdo	0.0612	0.3163	0.6516	0.0004	0.0612
De acuerdo	0.8422	0.6619	0.2077	0.1868	0.0003
Totalmente de acuerdo	0.0963	0.0115	0.0005	0.8128	0
8) Será un gobernador que mantendrá una buena relación con el presidente de México					
Totalmente en desacuerdo	0.0021	0.0203	0.1536	0	0.7675
En desacuerdo	0.101	0.2902	0.5409	0.001	0.2223
De acuerdo	0.7716	0.6578	0.3019	0.1938	0.0102
Totalmente de acuerdo	0.1254	0.0317	0.0036	0.8052	0

Según los resultados obtenidos, se pueden formar los siguientes segmentos:

Segmento 1. (43.9% de los entrevistados): veracruzano que opina estar de acuerdo *en todas las características personales del gobernador* y tiene una expectativa de gobierno favorable.

A continuación se muestra, como ejemplo, la descripción de las proporciones con las que se llega a la descripción del grupo.

- El 78.95% está de acuerdo con que el gobernador es un líder político.
- El 83.25% está de acuerdo con que tiene un programa congruente con las necesidades del estado.
- El 71.97% está de acuerdo con que tiene un gabinete de gobierno que conoce sus respectivas áreas de atención.
- Un 81.02% está de acuerdo con que es un veracruzano sensible a escuchar las necesidades ciudadanas.
- Un 79.82% está de acuerdo con que es una persona capaz.
- El 79.81% está de acuerdo con que será un gobernador que cumpla sus promesas de campaña.
- El 84.22% está de acuerdo con que será un gobernador respetuoso del Congreso del Estado.
- Un 77.16% menciona estar de acuerdo con que será un gobernador que mantendrá una buena relación con el presidente de México.

Segmento 2. (27.9% de los entrevistados): entrevistado que opina estar *de acuerdo* en las características, *Es un líder político que conoce los problemas del estado de Veracruz; Es un veracruzano sensible a escuchar las necesidades ciudadanas; Es una persona honesta y capaz; Será un gobernador respetuoso del Congreso del Estado,* y *Será un gobernador que mantendrá una buena relación con el presidente de México*. Pero está *en desacuerdo* en las siguientes características: *Tiene un programa de gobierno congruente con las necesidades del estado; Tiene un gabinete que conoce sus respectivas áreas de atención,* y *Será un gobernador que cumpla sus promesas de campaña*.

Segmento 3. (18.9% de los entrevistados): informante que opina estar *en desacuerdo* en todas las características que se le han planteado.

Segmento 4. (6.3% de los entrevistados): veracruzano que opina estar *totalmente en desacuerdo* en todas las características que se le han planteado.

Segmento 5. (3.0% de los entrevistados): veracruzano que opina estar *totalmente en desacuerdo* en todas las características que se le han planteado.

Conclusiones

La presente nota hace referencia de manera general a los llamados modelos de clases latentes, así como a la utilidad de esta herramienta en los estudios de tipo social; se trata de una herramienta estadística que comúnmente no se aplica. El modelo de análisis de clases latentes presenta una solución única, además de ser una herramienta de carácter confirmatorio sobre la naturaleza del concepto latente. Es una técnica de gran utilidad para estudios cuya medida resulta complicada debido a que no son observados directamente.

El modelo de clases latentes está acompañado de un conjunto de ecuaciones, las cuales resumen la relación existente entre las variables latentes. El modelo supone que una población se encuentra dividida en cierto número de clases latentes, tantas como categorías tenga la variable latente. Por tanto, cada individuo de la población estudiada pertenece únicamente a una clase latente.

El estudio de caso es un claro ejemplo del uso de los modelos de clases latentes, en el análisis de información contenida en una encuesta. A la población anteriormente mencionada se le aplicó un modelo de clases latentes confirmando la óptima distribución de la población en segmentos. Además proporcionó el tamaño de cada segmento, así como el comportamiento de los individuos de cada uno de éstos.

Bibliografía

Banfield, J. D. y A. E. Raftery (1993), "Model-based Gaussian and non-Gaussian clustering", *Biometrics*, núm. 49, pp. 803-821.

Bartholomew, D. J. y M. Knott (1999), *Latent Variable Models and Factor Analysis*, 2a. ed., Londres, Oxford University Press.

Goodman, L. A. (1974a), "The analysis of system of qualitative variables when some of the variables are unobservable, Part I- a modified latent structure approach", *American Journal of Sociology*, núm. 79, pp. 1179-259.

—— (1974b), "Exploratory latent structure analysis using both identifiable and unidentifiable models", *Biometrika*, núm. 61, pp. 215-231.

Lazarsfeld, P. F. y N. W. Henry (1968), "Latent Structure Analysis", Boston, Houghton Miffin.

Lindsay, B., C. C. Clogg y J. Greco (1991), "Semiparametric estimation in the Rash model and related exponential response models, including a simple latent class model for item analysis", *Journal of the American Statistical Association*, núm. 86, pp. 96-107.

Uebersax, J. S. (1993), "Statistical modeling of expert ratings on medical treatment appropriateness", *Journal of the American Statistical Association*, núm. 88, pp. 421-427.

Fecha de recepción: 21 de septiembre de 2010.
Fecha de aceptación: 1 de noviembre de 2010.
Fecha de publicación: 17 de diciembre de 2010.