CENTRO NACIONAL
DE EVALUACIÓN PARA
LA EDUCACIÓN SUPERIOR, A.C.

CENEVAL®

Análisis de clases latentes
Una técnica para detectar heterogeneidad en poblaciones

Lucía Monroy Cazorla
Rafael S. Vidal Uribe
Antonio Saade Hazin

Cuaderno técnico 2

Primera edición

CENTRO NACIONAL
DE EVALUACIÓN PARA
LA EDUCACIÓN SUPERIOR, A.C.

CENEVAL®

Análisis de clases latentes

Una técnica para detectar heterogeneidad en poblaciones

Cuaderno técnico 2

Análisis de clases latentes
Una técnica para detectar heterogeneidad en poblaciones
Cuaderno técnico 2

Lucía Monroy Cazorla
Rafael S.Vidal Uribe
Antonio Saade Hazin

Revisión técnica:
Arturo Bouzas Riaño

Análisis de clases latentes
Una técnica para detectar heterogeneidad en poblaciones
Cuaderno técnico 2

D.R. © 2009, Centro Nacional de Evaluación
para la Educación Superior, A.C. (Ceneval)
Av. Camino al Desierto de los Leones 19,
Col. San Ángel, Deleg. Álvaro Obregón,
C.P. 01000, México, D.F.
www.ceneval.edu.mx

Diseño: Mónica Cortés Genis
Formación: Alvaro Edel Reynoso Castañeda

Primera edición, septiembre de 2010

Impreso en México • Printed in México

Directorio

Dirección General
Rafael Vidal Uribe

Dirección General Adjunta de los EGEL
Jorge Hernández Uralde

Dirección General Adjunta de los EXANI
José O. Medel Bello

Dirección General Adjunta de Programas Especiales
Rocío Llarena de Thierry

Dirección General Adjunta Técnica y de Investigación
Lucía Monroy Cazorla

Dirección General Adjunta de Operación
Francisco Javier Apreza García Méndez

Dirección General Adjunta de Difusión
Javier Díaz de la Serna Braojos

Dirección General Adjunta de Administración
Francisco Javier Anaya Torres

Dirección de Procesos Ópticos y Calificación
María del Socorro Martínez de Luna

Dirección de Tecnologías de la Información
y las Comunicaciones
Francisco Manuel Otero Flores

Índice de figuras

Índice de tablas

Prefacio

El Centro Nacional de Evaluación para la Educación Superior (Ceneval) es una institución de carácter eminentemente técnico. A lo largo de tres lustros su actividad esencial ha sido promover la calidad de la educación mediante evaluaciones válidas, confiables y pertinentes de los aprendizajes.

Primordialmente, evalúa los conocimientos y habilidades adquiridos por los individuos en los procesos de enseñanza-aprendizaje, formales o no formales, de los sistemas educativos. Así contribuye a la toma de decisiones fundamentadas. De hecho, con sus servicios de evaluación atiende instituciones de educación media superior y superior, autoridades educativas, organizaciones profesionales y otras instancias públicas y privadas y, desde luego, al destinatario final –y el más importante– de sus pruebas: el propio sustentante.

Con la serie *Cuadernos técnicos* el Centro promueve también el uso de herramientas de análisis en círculos cada vez más amplios. El propósito de estos títulos es contribuir a elevar la calidad de la educación mexicana y fomentar una auténtica cultura de la evaluación.

En el ámbito educativo, detectar la variabilidad de los individuos que conforman una muestra permite elaborar estrategias diferenciales para la mejora educativa, la asignación de recursos, la planeación del desarrollo, etcétera. *Análisis de clases latentes: una técnica para detectar heterogeneidad en poblaciones,* cuaderno técnico número 2, busca adentrar al lector en el conocimiento de las clases latentes, herramientas estadísticas que permiten precisamente modelar las relaciones entre las variables observadas.

Las clases latentes de datos categóricos son un insumo medular para una gama de estudios sociales. Aquí se explican las diferencias conceptuales entre las variables latentes y las manifiestas, así como las métricas con que pueden ser medidas. Se detalla, asimismo, un ejemplo del uso de la técnica para tipificar a los candidatos a ingresar al Sistema de Educación Superior Tecnológica en cuanto a su calidad académica.

En el ámbito de las ciencias sociales se ha generado un cuerpo amplio de datos empíricos que presentan evidencias acerca de la heterogeneidad de las poblaciones bajo estudio (Bauer y Curran, 2004; Muthén y Muthén, 2000). En virtud de que éstas suelen no ser homogéneas, las técnicas estadísticas que permiten la clasificación de los sujetos han tenido una importancia medular durante muchos años.

Existen múltiples estudios para tipificar a las poblaciones; por ejemplo, se han presentado en la literatura de investigación diferentes tipos de consumidores de alcohol (Chassin , Pitts y Prost, 2002; Bucholz *et al.,* 2006), de niños con patrones agresivos (Tremblay *et al.,* 2004), de votantes en elecciones (Pérez y Fajardo, 2001), de consumidores (Rondán, Sánchez y Villarejo, 1999) y, en educación, algunos estudios han abordado la variabilidad presente en los planteles educativos y en los alumnos (Cervini, 2002; Howley y Howley, 2004; Willms, D., 2006).

En el medio educativo, detectar la variabilidad de los individuos que conforman una muestra permite generar estrategias diferenciales para la mejora educativa, la asignación de recursos, la planeación del desarrollo, etcétera. Asimismo, detectar los segmentos de una población en el área educativa ayuda a contextualizar las medidas arrojadas por instrumentos de evaluación, tales como las pruebas de logro académico, el alcance de metas, entre otros. Finalmente permite una mejor clasificación de los niveles de logro en las pruebas de rendimiento.

Cuando se habla de la heterogeneidad de una población se hace referencia no a las características únicas y distintivas de un individuo u objeto, sino a la presencia de grupos en una muestra. Cada unidad observada consiste en un conjunto de valores; por ejemplo, si observamos los rasgos físicos de los humanos, cada individuo tiene un valor en estatura, peso, color de piel, de cabello, de ojos, etcétera. Si el ejemplo fuera en el ámbito educativo, la medición de un individuo podría consistir en el puntaje promedio que obtuvo en diferentes asignaturas del ciclo escolar.

El objetivo de las técnicas clasificatorias utilizadas en las ciencias sociales es detectar patrones similares de respuesta para agrupar a los sujetos de una muestra, de tal forma que los individuos de un grupo se parezcan entre sí, pero que sean diferentes a los objetos clasificados en otro segmento, clase o grupo. En el primer ejemplo del párrafo anterior, el objetivo de la clasificación podría ser buscar las características distintivas de diferentes grupos étnicos; así, agrupados como caucásicos estarían los individuos de tez blanca, cabello rubio y ojos claros, mientras que en el grupo de los asiáticos estarían individuos con tez amarilla, cabello oscuro y ojos café oscuro. En nuestro segundo ejemplo se buscarían grupos de estudiantes por nivel de competencia. Así, en un grupo se ubicarían los que obtuvieron promedios superiores a 9 en todas las materias; en otro, los estudiantes de bajos promedios. En ambos casos observamos que los valores de las variables medidas (color de tez, de ojos y de cabello) tienden a estar correlacionados entre sí. Debe notarse que la asociación de las variables manifiestas en ningún momento implica causalidad.

En las técnicas clasificatorias se modela la estructura de la asociación de las variables, sin proponer relaciones de causalidad o de contribución sobre una variable dependiente.

Tradicionalmente, la clasificación de los objetos bajo estudio se llevaba a cabo utilizando la técnica multivariada denominada *análisis de cúmulos o conglomerados* (análisis de *cluster*). Con esta herramienta se conforman conglomerados de objetos (escuelas, alumnos, programas de estudio…) respecto a algunos parámetros de selección predeterminados. Un problema importante de esta metodología es la ausencia de un criterio objetivo que guíe a los investigadores en la determinación del número óptimo de grupos (Kaufman y Rousseeuw, 1990). Por ello, en fechas recientes, muchos investigadores han seleccionado *al análisis de clases latentes* como una metodología alternativa para la clasificación de datos, en virtud de que provee criterios menos arbitrarios para determinar el número de grupos presentes en la población (Vermunt).

El análisis de clases latentes (ACL) es una herramienta estadística que permite modelar las relaciones entre las variables observadas, suponiendo que la estructura de relaciones subyacentes es explicada por una variable latente categórica (no observada). Esta metodología clasificatoria se basa en la estimación de probabilidades condicionales, lo que permite analizar variables medidas en diferentes métricas, especialmente datos categóricos (Magidson y Vermunt, 2001, 2004).

Al igual que el análisis factorial, el ACL permite tanto explorar las relaciones entre las variables como probar hipótesis acerca de las estructuras. Como técnica exploratoria con esta herramienta estadística es posible reducir datos en una sola variable latente que identifica la membrecía de las clases. Como análisis confirmatorio, esta técnica puede confirmar la heterogeneidad de la población bajo estudio, permitiendo al investigador probar sus hipótesis acerca de la estructura de las relaciones entre las variables manifiestas.

Este cuaderno técnico introduce al lector al análisis de clases latentes de datos categóricos, los cuales son un insumo medular para una extensa gama de estudios sociales. En el capítulo I se explican las diferencias conceptuales entre variables latentes y variables manifiestas, así como las métricas en que pueden ser medidas.

En los capítulos II y III se presentan las herramientas estadísticas que permiten estudiar la variabilidad de los individuos de una población, dando énfasis al modelo de análisis de clases latentes como una herramienta para tipificar los objetos bajo estudio.

En el capítulo IV se detalla un ejemplo del uso de la técnica para tipificar a los candidatos para ingresar al Sistema de Educación Superior Tecnológica en cuanto a su calidad académica (variable latente definida para el ejemplo).

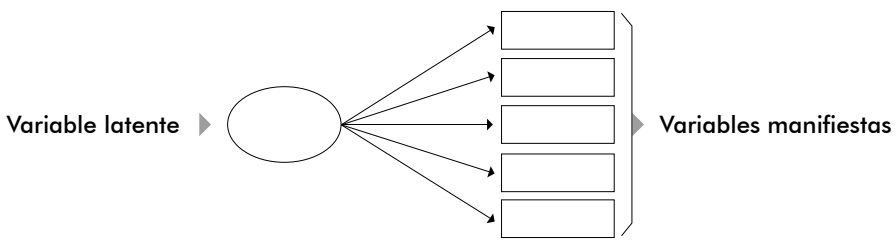
Finalmente, en el anexo se presenta una información complementaria respecto a las variables categóricas y algunas de las operaciones que se pueden realizar con este tipo de datos.

En las ciencias sociales, buena parte de las variables de interés en la disciplina no pueden ser observadas de manera directa. Por ejemplo, no podemos obtener una medición directa de la felicidad, el racismo, la inteligencia, la calidad de un docente o del capital cultural de una familia. A estas variables que no pueden ser medidas directamente se les denomina *variables latentes* y su valor depende de las variables observadas o manifiestas. Las *variables manifiestas* pueden ser medidas a través de instrumentos como las preguntas de una encuesta, los reactivos de un examen o las observaciones directas que se realicen del comportamiento de los individuos (como los registros que reportan el tiempo que duró una conducta o la frecuencia con la que ésta se presentó en un lapso determinado).

En el ámbito educativo, un ejemplo de variable latente es la habilidad verbal de un estudiante. Esta variable no puede ser medida directamente, por lo que se toma su valor dependiendo de la ejecución que tenga el alumno en una prueba con reactivos que exploren sinónimos, antónimos y analogías.

En la representación gráfica de los modelos con variables latentes, las variables observadas se representan con rectángulos; las variables latentes, con óvalos, y la dirección de la relación, con flechas rectas (figura 1).

Figura 1. Ejemplo de un modelo latente



Los análisis en los que se incorporan variables latentes tienen como finalidad detectar si las relaciones entre las variables manifiestas (dependencias) pueden ser explicadas por una o más variables latentes. De esta forma, los análisis de este tipo reducen el número de variables de un estudio y definen las relaciones entre las variables observadas.

Una variable latente (figura 1) explica las relaciones que mantienen las variables observadas, de forma que representa la fuente o causa "verdadera" de la asociación. Si esta variable puede ser caracterizada, entonces al controlarla se desvanecerá la asociación entre las variables observadas. En el ejemplo anterior, "la habilidad verbal" explica las relaciones entre los reactivos de la prueba. Los estudiantes de mayor habilidad tendrán probabilidades altas de contestar más reactivos del examen, mientras que los estudiantes con poca habilidad tenderán a fallar en más reactivos.

Métrica de las variables

Los métodos estadísticos que brindan la posibilidad de analizar datos para comprender la variabilidad de un fenómeno se han venido perfeccionando a través de los años. De hecho, estos métodos orientados a responder una gama extensa de preguntas de investigación tratan, generalmente, con variables que fueron medidas en diferentes métricas; es decir, pueden ser variables categóricas (si a las observaciones se les dio una etiqueta que las agrupa como "frecuente" o "infrecuente") o numéricas (cuando cada observación tomó un valor numérico que representa diferentes magnitudes del atributo que se mide).

Las variables numéricas permiten utilizar un número amplio de herramientas estadísticas para describir la variabilidad que capturan. Generalmente, con estas variables se pueden calcular medidas de tendencia central y de dispersión que describen los datos. Por su lado, las variables categóricas permiten describir

puntualmente las características que tienen los individuos que están agrupados en una categoría, y reportar el número o proporción de observaciones que se encuentran en cada segmento o grupo. Por ejemplo, se pueden identificar las habilidades y conocimientos que poseen los estudiantes que obtuvieron en una prueba un nivel de desempeño satisfactorio o sobresaliente y reportar la proporción de estudiantes que están en cada uno de estos niveles.

La métrica de las variables determina las preguntas de investigación que pueden ser planteadas y los métodos estadísticos que deben usarse para darles una respuesta.

Dado que las ciencias sociales estudian una gran diversidad de constructos que no pueden ser medidos de manera directa, el uso de los modelos latentes se ha ido diseminando. En estos modelos, el valor de la variable latente (no observada) se deriva de la información de las variables manifiestas, capturada generalmente por instrumentos de medición: cuestionarios, encuestas, exámenes o registros observacionales.

Los modelos latentes asumen que las respuestas en las variables manifiestas son el resultado de la posición en la variable latente y que *las relaciones entre los datos observados desaparecen una vez que se introduce la variable latente en el modelo.*

Actualmente, existen diversos modelos que incorporan variables latentes. Todos parten de los mismos supuestos y se diferencian por la escala de medición de las variables modeladas. Su selección debe basarse tanto en la métrica de las variables observadas como en el nivel de medición (continuo o categórico) que se asuma en las variables latentes. En la tabla 1 se muestra una clasificación de los modelos básicos que incorporan variables latentes. De estos modelos existen ampliaciones que incorporan medidas repetidas o modelos con variables observables mixtas (continuas y categóricas).

Una clara exposición sobre estos modelos puede encontrarse en el libro de Bartholomew *et al.* (2002, caps. 6-9). En el siguiente capítulo sólo se expondrá

el modelo de clases latentes y, en los subsecuentes, las extensiones de este modelo que fueron utilizadas en los tres estudios sobre el comportamiento de riesgo juvenil.

Tabla 1. Clasificación de modelos con variables latentes de acuerdo con los niveles de medición

		Variables manifiestas	
		Continuas	**Categóricas**
Variables latentes	**Continuas**	Análisis factorial (*Factor analysis*)	Análisis de rasgo latente (*Latent trait analysis*)
	Categóricas	Análisis de perfil latente (*Latent profile analysis*)	Análisis de clases latentes (*Latent class analysis*)

Modelos latentes

En gran parte de los trabajos de las ciencias sociales la población bajo estudio no es homogénea y los sujetos podrían agruparse de acuerdo con varias características que comparten entre sí. En estas disciplinas una herramienta clasificatoria tradicional es la de los modelos de cúmulos, como el análisis de *cluster* con K-medias. Estos modelos basan la clasificación de las observaciones (por ejemplo, de individuos) en algoritmos que consideran, como medida de proximidad, la distancia de un objeto a otro (generalmente estimando distancias euclidianas).[1] En estos análisis de agrupamiento no existen estimadores estadísticos que permitan evaluar modelos con un número diferente de grupos, por lo que el investigador tiene que determinar en cuántos segmentos puede dividirse la población (Kaufman y Rousseeuw, 1990; Everitt, Sabine y Morven, 2001).

Además de los modelos que estiman distancias entre los objetos, para detectar la heterogeneidad de una población suelen utilizarse modelos que incorporan variables latentes discretas que clasifican a los objetos de acuerdo con la probabilidad de presentar un patrón de respuesta determinado, por lo que en cada clase o segmento de la población se incorporan a los objetos con alta probabilidad de tener un patrón o vector de respuesta similar. Estos modelos probabilísticos tienen la ventaja de estimar criterios menos arbitrarios para identificar el número de grupos presentes en una muestra. Es decir, estiman parámetros que permiten seleccionar el modelo que mejor ajusta a los datos (Wolfe, 1970; McLachan y Peel, 2000; Vermunt y Magidson, 2002).

Diversos modelos de agrupamiento latente se emplean para dar respuesta a diferentes preguntas de investigación. De manera general, estas herramientas analíticas podrían agruparse en tres vertientes:

[1] La distancia euclidiana entre dos puntos es la longitud del camino que los conecta.

1. *Análisis de clases latentes (latent class analysis).* Estudian la variabilidad del comportamiento de los individuos de una población; son útiles para detectar tipologías.
2. *Modelos factoriales discretos o análisis factorial con clases latentes (latent class factor models).* Sirven para reducir el número de variables en un estudio, o bien para analizar las relaciones estructurales de un conjunto de variables que se asume provienen del mismo dominio (McLachlan y Peel, 2000; Magidson y Vermunt, 2001).
3. *Modelos de regresión latente o análisis de regresión con clases latentes (latent regression analysis).* Permiten detectar la contribución diferencial de las variables independientes, condicionadas a su pertenencia a una clase o segmento de la población. En estos análisis se pueden detectar efectos dependientes o independientes de las clases (Wedel y De Sarbo, 1994; Vermunt y Van Dijk, 2001).

Los modelos denominados análisis de clases latentes se explicarán con detalle más adelante, en virtud de que son una herramienta de gran utilidad para detectar el número de segmentos o grupos presentes en las poblaciones que sustentan los exámenes de logro educativo.

Análisis de clases latentes

El análisis de clases latentes (ACL) fue reportado por primera vez por Lazarsfeld (1950) como herramienta para construir una tipología en el análisis de un conjunto de variables dicotómicas. Años después, Lazarsfeld y Henry (1968) continuaron utilizando un modelo latente en un estudio sobre actitudes para determinar la presencia de diferentes grupos entre los sujetos observados.

Leo Goodman (1974) logró que los modelos de clases latentes pudieran aplicarse en una mayor diversidad de estudios, desarrollando un algoritmo para obtener las estimaciones por máxima verosimilitud. Él propuso la extensión del modelo para variables manifiestas politómicas y realizó importantes mejoras para la identificación de los modelos.

El ACL permite detectar la heterogeneidad de una población identificando el menor número posible de grupos presentes en el universo que se estudia. En el análisis de clases latentes se propone que una variable discreta no-observada (latente) describe las relaciones entre las variables manifiestas.

Diversos investigadores (Agresti, 2002; Bartholomew *et al.,* 2002; Hagenaars, 1990; McCutcheon, 1987; Vermunt, 2003 y 2004) han resaltado algunas de las bondades de los modelos de clases latentes:

- Reducen la complejidad de los datos identificando un número pequeño de variables (clases latentes) que son suficientes para explicar las relaciones entre las variables manifiestas.
- Explican las relaciones "verdaderas" entre variables observadas. Se dice que las variables no-observadas (latentes) explican relaciones "verdaderas", ya que al incorporarlas en los modelos controlan diversas fuentes de error tales como casos ausentes, variables omitidas, correlaciones entre las observaciones, etcétera.
- Permiten estimar la probabilidad que tiene cada uno de los participantes de pertenecer a una de las clases latentes.
- Analizan datos categóricos en las escalas en que fueron medidos, sin requerir transformaciones para lograr normalidad multivariada.

El ACL puede servir como una herramienta exploratoria que evalúa el ajuste de modelos con diferente número de clases. Se utiliza también para confirmar hipótesis respecto a la estructura latente de un conjunto de variables, hipótesis

relativas al tamaño de los grupos o clases, o bien, hipótesis sobre relaciones específicas entre las variables manifiestas. Para un mayor detalle sobre las posibles relaciones que se pueden asumir entre las variables manifiestas y la variable latente consúltese el artículo "Análisis de clases latentes", de Goodman (2002).

De acuerdo con Bartholomew *et al.* (2000, p. 236-237) en el análisis de clases latentes se asume que:

- Cada participante de una muestra aleatoria pertenece sólo a una de las clases latentes detectadas.
- La probabilidad de dar una respuesta a un ítem particular es la misma para todos los individuos que comparten la membrecía de una clase, pero diferente a la de los individuos que pertenecen a una clase diferente.
- Dada la pertenencia de un individuo a una clase latente, sus respuestas a cada uno de los ítems son condicionalmente independientes.

Hoy, gracias al desarrollo de importantes algoritmos que pueden llevar a cabo diversos paquetes estadísticos comerciales, el ACL se ha extendido en virtud de que permite tener criterios menos arbitrarios para la selección del número de grupos en la población y porque permite incorporar en el análisis variables con diferentes escalas de medición. Por ello, esta herramienta clasificatoria se ha convertido en un método versátil y preciso que permite dar respuesta a importantes preguntas sobre *variabilidad*.

Análisis de clases latentes (básico o estándar)

El ACL básico es usado generalmente como un método analítico que permite identificar el menor número posible de clases latentes que son suficientes para explicar las relaciones entre las variables observadas o manifiestas que pueden

reportar datos dicotómicos, politómicos, nominales, ordinales, o bien combinaciones de variables en diferentes escalas de medición.

Un supuesto muy importante que subyace en el análisis de clases latentes básico es el de independencia local (Lazarsfeld y Henry, 1968). Si al incluir una variable latente en el análisis, las correlaciones entre las variables observadas son cercanas a cero, se dice que las variables manifiestas son independientes localmente. Esta condición es un método relevante para determinar si las relaciones de un conjunto de variables observadas son espurias, es decir, si desaparecerían al incorporar una variable no-observada (McCutcheon, 1987).

Notación del modelo

Antes de formalizar el análisis de clases latentes explicaremos la notación empleada con más frecuencia en el área para plantear los modelos. El símbolo π se utiliza para referir probabilidad. La letra X denota en general una variable latente; el sufijo t, las categorías de la variable latente, y T, el total de clases o categorías de la variable latente.

De esta forma $\pi(X_t)$ representa la probabilidad de que un individuo seleccionado aleatoriamente pertenezca a la clase latente t (t= 1, 2,...,T). Esta probabilidad es equivalente a decir que $\pi(X_t)$ es la proporción de individuos que pertenecen a la clase latente t.

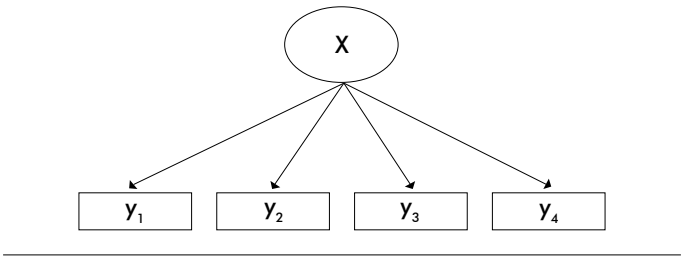
Cada una de las variables observadas puede ser denotada con una letra del alfabeto y un sufijo que indicará su nivel o categoría de respuesta. Cuando se utiliza una línea vertical entre las variables (|) se indica una probabilidad condicional. En $\pi(y_i|X_t)$ se denota la probabilidad de tener el valor i en la variable y, dada la pertenencia a la clase t de la variable X.

Se denota $\pi(y_1\ y_2\ y_3\ y_4|X_t)$ a la probabilidad conjunta de una serie de valores de respuesta, dada su pertenencia en la clase t de la variable latente X.

En la figura 2 se muestra la representación gráfica del modelo latente básico.

Como puede observarse, asumimos que las variables manifiestas o indicadores (y_1 y_2 y_3 y_4) son mutuamente independientes, dada su membrecía en una clase latente (X). Nótese que en este modelo la variable latente no está influida por ninguna variable. A las variables que no reciben la influencia de otras variables se les suele dar el nombre de variables exógenas, mientras que a las variables que son impactadas por otras variables se les llama endógenas.

Figura 2. Representación gráfica de un modelo de clases latentes básico



En el ACL la variable X del modelo gráfico es una variable latente categórica nominal. Asumiendo que cada variable manifiesta es una variable categórica que puede tener el valor de 0 o 1, podemos tener diferentes patrones de respuesta a los que comúnmente se les denomina vector. Así, de las variables $y_1...y_4$, podríamos tener un vector de respuesta como 0001, el cual refleja que un sustentante tuvo valores de 0 en y_1, y_2, y_3, y de 1 en la variable y_4. Este vector de respuesta es una función de dos probabilidades:

- La probabilidad de que el individuo pertenezca a una clase de la variable latente.
- La probabilidad de que en cada variable manifiesta obtenga el valor 0 o 1, dado la pertenencia a la clase latente.

El supuesto de la *independencia local,* en el que las variables están asociadas a través de la variable latente que las explica, es muy importante: permite estimar la probabilidad conjunta del vector de respuestas, dada la pertenencia a la clase latente, como el producto de las probabilidades de cada respuesta.

Formalmente, el modelo de clases latentes para variables categóricas se puede escribir de la siguiente manera:

$$\pi(\mathbf{Y}_i = y) = \sum_{t=1}^{T} \pi(X_i = t) \prod_{j=1}^{J} \pi(Y_{ij} = y_j \mid X_i = t)$$

Donde:

$\mathbf{Y}_i$ indica el vector de respuesta del caso *i*.

Y_{ij} indica la respuesta del caso *i* en la variable *j*; con *J*, el número de variables en el modelo.

X_i representa a la variable latente; *t* indica una clase latente particular; con *T*, el número de clases latentes.

Las letras en negritas representan un vector de respuesta, mientras que las letras minúsculas representan las realizaciones de una variable.

Probabilidad de las clases latentes

La probabilidad de las clases latentes $\pi(X_t)$ describe la distribución de los niveles detectados en una variable no observada, a través de los cuales las variables observadas son independientes. A estos subgrupos suele identificárseles como clases, segmentos o grupos.

Hay dos aspectos muy importantes en las probabilidades de las clases latentes:

- *El número de clases*. Número total de grupos o clases (*T*) de la variable latente (*X*) que se modelaron para explicar las relaciones entre las variables observadas.

En los modelos de clases latentes, el número menor de clases que se modelan es dos, ya que el modelo de una clase equivale a un modelo sin clases (en el que no hay heterogeneidad entre los participantes del estudio). Generalmente, en el ACL se inicia la evaluación con el modelo de una clase *y*, posteriormente, se va incrementado una clase *T*+1 hasta encontrar el modelo que se ajuste mejor a los datos.

- *El tamaño de las clases*. El tamaño relativo de cada una de las clases latentes indica cómo se distribuye la población entre el número total de clases (*T*). Así, podemos identificar grupos normativos, minoritarios o equivalentes en la población.

  La suma de las probabilidades de las clases debe ser igual a 1. El tamaño de las clases latentes es un parámetro muy importante para comparar diferentes poblaciones.

Probabilidades condicionales

Estos parámetros de los modelos de clases latentes representan la probabilidad de que un individuo obtenga un valor determinado en una variable, dada su pertenencia a una clase latente. Las probabilidades condicionales como $\pi(y_1|X_1)$ indican la probabilidad de tener el valor 1 en la variable *y*, dado que se pertenece a la clase latente 1.

Estas probabilidades condicionales reflejan las características de los miembros de una clase latente, por lo que de acuerdo con ellas se asigna el nombre a la clase. Por ejemplo, si en cuatro ítems que exploran el uso de sustancias adictivas, las probabilidades condicionales del valor que representa la categoría de "no-consumo" son altas; entonces, podríamos denominar a este grupo como "No consumidor".

En cada clase latente, las probabilidades condicionales de las variables observadas deben sumar 1, por lo que cada observación tiene una probabilidad específica de estar en un nivel de la variable observada.

Estimación de los modelos

La estimación de los modelos de clases latentes dependen de la escala de medición de las variables observadas, ya que se asumen diferentes distribuciones para las variables nominales, ordinales y continuas. Los modelos pueden incluir un conjunto de variables medidas en diferentes escalas.

Las variables observadas nominales que se asumen provienen de una distribución multinomial, se modelan mediante una regresión logística multinomial; las variables ordinales, mediante regresiones logísticas ordinales, y las variables dicotómicas, mediante regresiones logísticas binarias. Las variables continuas se estiman mediante regresiones lineales estándares (Vermunt y Magidson, 2005).

Evaluación del ajuste del modelo

Una vez que conocemos cómo modelar clases latentes, otra herramienta indispensable es la que nos permite valorar cuál de los modelos propuestos se ajusta mejor a los datos que estamos trabajando. En el ACL el estadístico más usado para evaluar el ajuste de los modelos de clases latentes es el *criterio de información bayesiana (bayesian information criterion,* BIC). Este y otros estadísticos similares ponderan, según el número de parámetros, la bondad del ajuste de un modelo medido por el valor de máxima verosimilitud obtenido. Este estadístico es especialmente útil cuando en la población que se estudia hay datos esparcidos o casos escasos. Como regla para la selección del modelo que mejor se ajusta a los datos, se debe identificar el modelo que obtenga el menor valor de BIC. (Gill, 2002).

$$BIC = -2\,\ell\,(\,\hat{\theta}\,|x) + p\,log(n)$$

Donde:

$\ell(\hat{\theta}|x)$ es el valor maximizado del logaritmo de la función de verosimilitud

p es el número de variables explicativas del modelo (incluyendo la constante)

n es el tamaño de la muestra.

Análisis de clases latentes con covariables

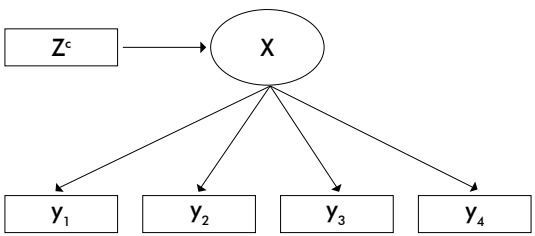
El modelo de análisis de clases latentes básico analiza las relaciones de variables manifiestas (generalmente categóricas), proponiendo que sus relaciones son explicadas por una variable latente que tiene más de una clase.

Los modelos de clases latentes han incorporado variables con diversas escalas de medición (continua, ordinal o nominal). Además, los modelos también suelen incorporar variables independientes que afectan la pertenencia a las clases latentes. A estas variables exógenas se les reconoce como *covariables o variables de agrupamiento* (Hagenaars, 1993; McCutcheon, 1987; Vermunt y Magidson, 2003).

Los modelos que incorporan covariables son las extensiones más importantes porque permiten modelar variables explicativas que afectan las respuestas (Wedel y De Sarbo, 1994) o la membrecía de las clases (Dayton, 1999).

En los modelos de clases latentes las covariables también pueden ser variables nominales, ordinales o continuas (figura 3).

Figura 3. Modelo de clases latentes con una covariable



En la siguiente ecuación se modeló una covariable (z^c) que afecta la membrecía de los individuos en la clase latente. Cabe resaltar que en este modelo las variables observadas no reciben influencia de esta variable.

$$\pi(\mathbf{Y}_i = y \mid z^c) = \sum_{t=1}^{T} \pi(X_i = t \mid z^c) \prod_{j=1}^{J} \pi(Y_{ij} = y_j \mid X_i = t)$$

Donde:

Y_{ij} indica la respuesta del caso i en la variable j, siendo J el número de variables en el modelo.

X_i: representa a la variable latente; t, una clase latente particular, siendo T el número de clases latentes.

Z_i indica una variable independiente que afecta la pertenencia a las clases latentes.

Las letras en negritas representan un vector, mientras que las letras minúsculas representan las realizaciones de una variable.

En las extensiones al modelo del ACL-básico también se pueden modelar variables independientes que afectan las relaciones entre las variables manifiestas y la variable latente. A estas variables se les suele identificar como variables *predictoras*. En este ejemplo se les denota como Z^p. Supongamos un caso en que la variable predictora Z^p afecta a las variables manifiestas, mientras que otra variable denominada Z^c afecta la pertenecía a la clase latente (figura 4).

$$\pi(\mathbf{Y}_i = y \mid z^c, z^p) = \sum_{t=1}^{T} \pi(X_i = t \mid z^c) \prod_{j=1}^{J} \pi(Y_{ij} = y_j \mid X_i = t \mid z^p)$$

Donde:

Y_{ij} indica la respuesta del caso i en la variable j, siendo J el número de variables.

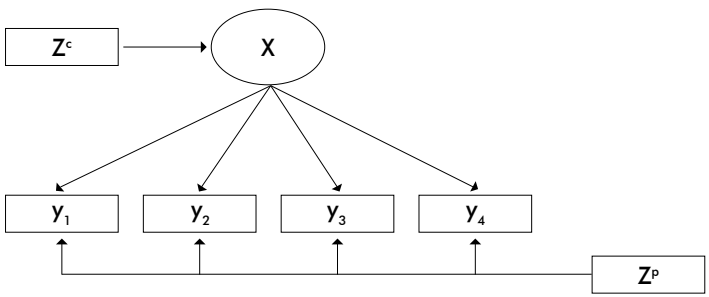
X_t representa a la variable latente; t, a una clase latente particular, siendo T el número de clases latentes.

Z^c indica una variable independiente que afecta la pertenencia a las clases latentes.

Z^p indica una variable independiente predictora que afecta a las variables observadas.

Las letras en "negritas" representan un vector; las letras minúsculas, las realizaciones de una variable.

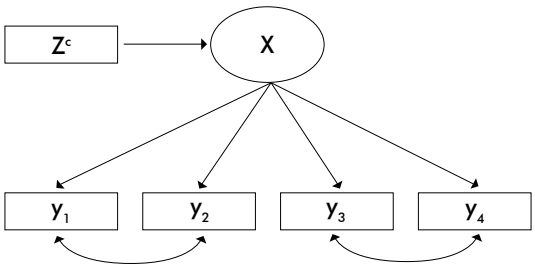
Figura 4. Modelo de clases latentes con una covariable que afecta las variables observadas (Z^p) y otra que afecta a la variable latente (Z^c)



Análisis de clases latentes con covariables y dependencias

Otra innovación al modelo básico de clases latentes fue la posibilidad de *relajar el supuesto de independencia local*. Una forma de mejorar los modelos que presentan un pobre ajuste a los datos es permitir asociaciones entre las variables manifiestas, así como dependencias entre las covariables y los indicadores (Vermunt y Magidson, 2003).

Figura 5. Modelo de clases latentes con una variable agrupadora y dependencias



En diversos modelos, al incorporar dependencias entre las variables manifiestas, los errores en la clasificación pueden disminuir de manera importante. En el ejemplo anterior, al tiempo de incorporar el efecto de la covariable (Z^c) sobre la variable latente (X), podemos modelar dependencias entre las variables y_1-y_2, así como entre y_3-y_4.

$$\pi(y_1 y_2 y_3 y_4 \mid z^c) = \sum_{t=1}^{T} \pi(x_t \mid z^c)\, \pi(y_1 y_2 \mid x_t)\, \pi(y_3 y_4 \mid x_t)$$

Donde:

$\pi(x_t \mid z^c)$ denota que la pertenencia a la clase latente t está condicionada al covariado z_q.

Heterogeneidad de los estudiantes que solicitan ingresar a un plantel del Sistema de Institutos Tecnológicos de Educación Superior

Con la finalidad de ejemplificar el ACL, en esta sección se presenta un análisis de los datos capturados en el cuestionario de contexto que acompaña el Examen Nacional de Ingreso a la Educación Superior (EXANI-II) presentado en 2008 por los candidatos para ingresar a un plantel del Sistema de Institutos Tecnológicos de Educación Superior.

Objetivo

El estudio se llevó a cabo para identificar la heterogeneidad de la calidad de la trayectoria académica de la población. Con él se dio respuesta a las siguientes preguntas:

1. ¿Los alumnos que solicitan su ingreso en los diferentes planteles de educación superior tecnológica tienen similar calidad académica?
2. ¿Cuántos grupos se pueden distinguir en la población?
3. ¿Cuáles son los patrones de respuesta que permiten distinguir a los estudiantes?

Participantes

En este estudio se consideraron los datos del cuestionario de contexto de 109,810 personas que se registraron para presentar el EXANI-II, a fin de competir por un lugar en los planteles del Sistema de Institutos Tecnológicos de Educación Superior. Solo se consideró la información de los sustentantes que contestaron todas las preguntas que se seleccionaron para llevar a cabo la

ejemplificación del ACL. En esta población solicitante, 65% eran hombres. La distribución de los candidatos por género y nivel socioeconómico se aprecia en la tabla 2.

Tabla 2. Distribución de los sustentantes por género y nivel socioeconómico

Nivel económico	Hombre		Mujer		Total	
	Núm.	%	Núm.	%	Núm.	%
Bajo	13 675	19	10 013	26	23 688	22
Medio bajo	15 747	22	9 427	24	25 174	23
Medio	16 767	24	8 909	23	25 676	23
Medio alto	13 530	19	6 190	16	19 720	18
Alto	11 312	16	4 240	11	15 552	14
Total	71 031	100	38 779	100	109 810	100

Descripción de las variables

En el cuestionario de contexto que aplica el Ceneval se incorporan exclusivamente preguntas de opción múltiple, de auto-reporte y comprenden una gama extensa de variables personales, académicas, familiares y económicas de los sustentantes. Para este ejemplo se seleccionaron, exclusivamente, las variables que nos permitieron construir un indicador acerca de la calidad académica de los sustentantes: uno sobre el nivel económico familiar y otro sobre el capital cultural de la familia.

- ***Calidad académica***

  Se consideró como una variable latente categórica nominal estimada a partir de las siguientes variables manifiestas:

  - Promedio general del bachillerato (PG). Las opciones de respuesta categorizaban los puntajes obtenidos.
  - Habilidad para el uso de un procesador de textos en la computadora (P_TX). Esta variable captura la percepción de habilidad de los sustentantes para elaborar documentos usando un procesador de textos de la computadora. La categoría de *No lo sé hacer* se unió a la de *Poco hábil,* en virtud de la baja frecuencia de respuesta.
  - Exámenes extraordinarios (E_EXT). Se consideró cómo tener el atributo cuando el sustentante hubiera presentado al menos un examen extraordinario durante el bachillerato.
  - Lectura de textos académicos en inglés (Ing). El sustentante reportó si era capaz de realizar esta actividad.

- ***Nivel económico familiar***

  Esta variable se construyó dividiendo a la población en quintiles de acuerdo con su puntuación en una escala construida utilizando un modelo de crédito parcial de teoría de respuesta al ítem (TRI), la cual conformó con los siguientes reactivos: bienes y servicios en la vivienda (televisión, teléfono, reproductor de DVD, internet y horno de microondas); recubrimiento asfáltico en la calle donde está ubicada su vivienda y pertenencia de bienes de uso personal, como reproductor de mp3 y teléfono celular. Para este ejemplo, los quintiles se agruparon en tres categorías: el primer y segundo quintiles en la categoría *Bajo,* el tercer quintil en *Medio* y el cuarto y quinto quintiles en *Alto.*

- *Capital cultural de la familia*

  Esta variable se construyó dividiendo la población en quintiles de acuerdo con su puntuación en una escala construida utilizando un modelo de crédito parcial de teoría de respuesta al ítem (TRI), la cual se conformó con los siguientes reactivos: escolaridad del padre y de la madre, número de libros en casa, frecuencia con la que asiste a espectáculos artísticos y al cine, y número de estados de la República que ha visitado como turista durante el último año. Para este ejemplo, los quintiles se agruparon en tres categorías: primer y segundo quintiles en la categoría *Bajo;* tercer quintil en la de *Medio,* y cuarto y quinto quintiles en la de *Alto.*

En la tabla 3 se puede observar la métrica de cada una de las variables incorporadas en el modelo y sus categorías.

Tabla 3. Variables manifiestas que definen la calidad de la trayectoria académica

Variable	Categorías	Métrica declarada para estimación
Promedio general en el bachillerato (PG)	6.0-6.9 7.0-7.9 8.0-8.9 >=9	Ordinal
Habilidad en el uso de procesadores de texto (P_Tx)	Poco hábil Hábil Muy hábil	Nominal
Exámenes extraordinarios (E_Ex)	No Sí	Nominal
Lectura de textos académicos en inglés (Ing)	No Sí	Nominal
Nivel económico familiar (Econ)	Bajo Medio Alto	Ordinal
Capital cultural de la familia (Cult)	Bajo Medio Alto	Ordinal

Procedimiento

El cuestionario de contexto fue llenado por los candidatos al registrarse para sustentar el examen. El registro se llevó a cabo por medio de Internet, con un sistema diseñado por el Ceneval. El cuestionario cuenta con 186 preguntas y el tiempo promedio que tardaron los estudiantes para contestarlo fue de 25 minutos.

Análisis de datos

La heterogeneidad de los sustentantes que presentaron el EXANI-II fue explorada a través de un *análisis de clases latentes*. Este método nos permitió detectar el número menor de grupos presentes en la población y mostró estimados sobre el tamaño de los grupos y sobre las probabilidades conjuntas de los vectores de respuesta condicionados para la clase latente de pertenencia (McCutcheon, 1987, 2002). De esta forma, los sustentantes que fueron agrupados en la misma clase son similares respecto a las variables incorporadas en los análisis.

El ajuste del modelo fue evaluado con el criterio de información bayesiana (BIC), que permite identificar el modelo con el menor número de clases que se ajusta mejor a los datos. El menor valor del BIC fue considerado como indicador del mejor modelo (Vermunt y Magidson, 2002; Vermunt y Magidson, 2003). Los modelos fueron evaluados con diferente número de clases, empezando con el modelo de 1-clase para, posteriormente, ir agregando una clase hasta que el indicador BIC mostrara un incremento en su valor, o bien, hasta que se obtuvieran grupos de sustentantes que incorporaran no menos del 1% de la población.

Para confirmar que el modelo seleccionado fuera el más adecuado a los datos, se consideraron además el error de clasificación, la reducción del error

del modelo y un análisis de los residuos de cada modelo. Los detalles de los modelos de análisis de clases latentes se presentaron en un capítulo anterior de este cuaderno técnico. En los siguientes párrafos se desglosará el procedimiento con el que se llevó a cabo el *análisis de clases latentes*.

Paso I. Inspección de las variables y formulación de los modelos por evaluar

Iniciemos con el análisis de la estructura que tienen los datos. La estructura de la base de datos se ejemplifica mostrando los resultados de los primeros diez sustentantes (tabla 4).

Tabla 4. Estructura de la base de datos

Casos	PG	P_TX	E_Ext	Ingl	Econ	Cult
1	2	2	1	0	1	1
2	2	1	1	0	1	1
3	3	1	0	1	1	2
4	2	2	1	0	1	1
5	3	1	0	0	2	1
6	2	2	1	0	1	1
7	2	2	1	1	1	1
8	3	1	0	0	1	1
9	2	1	1	1	3	2
10	2	2	0	0	1	1

PG=promedio general del bachillerato; P_Tx=habilidad para usar procesador de textos; E_Ext=presentación de exámenes extraordinarios; Ing=lectura de textos académicos en inglés; Econ=nivel económico familiar; Cult=capital cultural de la familia.

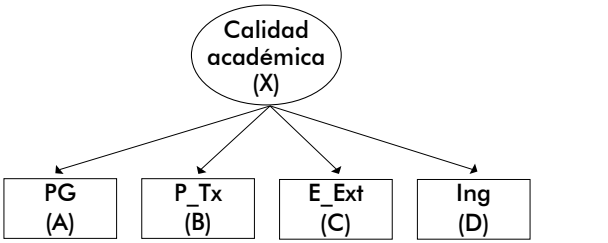
En este ejemplo se consideró que las cuatro variables reportadas por los sustentantes eran indicadores del fenómeno al que denominamos "calidad académica" y que esta variable latente categórica explicaba las relaciones de las variables manifiestas. Por supuesto que en este estudio se asumió que algunas de las variables están asociadas; sin embargo, no se plantearon relaciones de causalidad entre ellas. En la tabla 5 se presentan las correlaciones (policóricas) entre las variables manifiestas que definen el constructo latente del modelo.

Tabla 5. Correlación entre las variables manifiestas que definen calidad académica

Variables	PG	P_Tx	E_Ext	Ingl
Promedio general (PG)	1.000			
Habilidad en el uso de procesador textos (P_Tx)	0.117	1.000		
Exámenes extraordinarios (E_Ext)	-0.639	-0.057	1.000	
Leer textos académicos en inglés (Ing)	0.081	0.216	-0.058	1.000

Para la estimación del modelo de clases latentes se incorporaron las variables manifiestas asumiendo independencia local entre ellas. La representación gráfica del modelo se puede observar en la figura 6.

Figura 6. Modelo básico



PG=promedio general del bachillerato; P_Tx=habilidad para usar procesador de textos; E_Ext=presentación de exámenes extraordinarios; Ing=lectura de textos académicos en inglés. En los paréntesis se muestran las etiquetas que se dieron a las variables para modelarlas formalmente en la ecuación.

Formalmente, el modelo que analizaremos podría escribirse:

$$\pi(A_p B_q C_r D_s|X_t)=\sum_{t=1}^{T}\pi(X_t=t)\prod_{p=1}^{P}\pi(A_p|X_t)\prod_{q=1}^{Q}\pi(B_q|X_t)\prod_{r=1}^{R}\pi(C_r|X_t)\prod_{s=1}^{S}\pi(D_s|X_t)$$

Donde:

A, B, C y D son las variables manifiestas ordinales con *J* categorías. En este ejemplo se asignaron las letras p, q, r y s para denotar las categorías de las variables manifiestas. Así, (p= 1,2,3), (q= 1,2,3,4), (r= 1,2), (s= 1,2) y (o=1,2). La variable *latente* está representada con la letra *X* con realización *t* (t=1,2, ...*T*).

Paso II. Evaluación y selección de modelos

Una vez clarificado el modelo de partida que nos ayudará a dar respuesta a las preguntas de investigación, se estimaron los estadísticos que permitieron guiar la selección del modelo que mejor ajustaba a los datos. Recuérdese que

en el modelo presentado anteriormente (figura 6) no se modelaron covariables o dependencias entre las variables observadas, por lo que lo denominaremos *modelo básico*.

Como primer paso se estimó el modelo con una sola clase latente (T=1), el cual fue considerado como *modelo de contraste*. Después se evaluaron modelos en los que se incrementó el número de clases latentes, de uno en uno. En la tabla 6 se presentan los resultados obtenidos con los modelos con 1, 2, 3, 4 y 5 clases latentes sin restricciones. Considerando el estadístico para valorar la bondad de ajuste de cada modelo, el modelo de 4-clases latentes fue seleccionado como el mejor. Este modelo tuvo el menor valor en BIC y su error de clasificación fue de 0.27

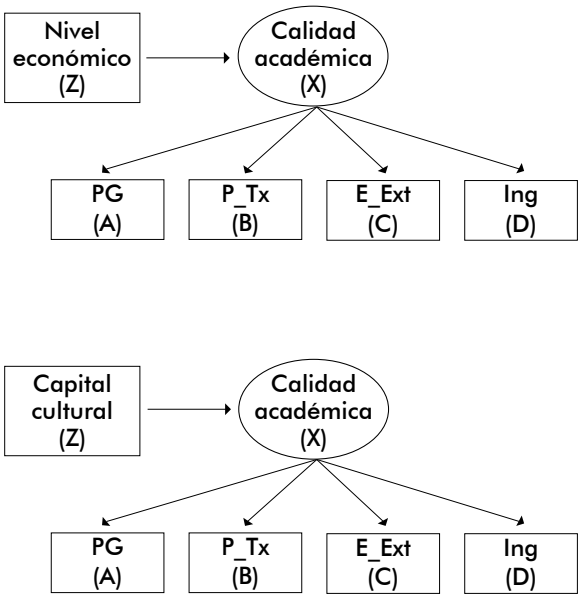
Tabla 6. Resultados de la evaluación de modelos básicos con diferente número de clases latentes

Modelos	BIC	Npar	p-value	Class.Err.
1-clase	789892.31	7	4.3e-7191	0.00
2-clases	760011.93	12	1.1e-713	0.08
3-clases	758283.71	17	2.6e-335	0.15
4-clases	**756879.96**	**22**	**1.5e-32**	**0.27**
5-clases	756918.40	27	3.4e-31	0.29

BIC=criterio bayesiano de información; Npar=número de parámetros; p-value=significancia del modelo, y Class.Err.=clasificación del error. En negritas se marcó el modelo con un valor menor al del BIC.

Posteriormente, se evaluaron dos modelos en los que se incorporaron covariables. El primer modelo incorporó, como covariable que impacta a la variable latente, al nivel económico familiar, y el segundo modelo, al capital cultural (figura 7). Como las variables *nivel económico* y *capital cultural* no fueron estimadas en el modelo, se consideran como variables manifiestas en virtud de que sus valores ya no tienen que ser estimados.

Figura 7. Modelos con covariables



PG=promedio general del bachillerato; P_Tx=habilidad para usar procesador de textos; E_Ext=presentación de exámenes extraordinarios; Ing= lectura de textos académicos en inglés. En los paréntesis se muestran las etiquetas que se dieron a las variables para modelarlas formalmente en la ecuación.

Formalmente, los modelos de análisis de clases latentes con un covariado podrían escribirse así:

$$\pi(A_p B_q C_r D_s | X_t Z_1) = \sum_{t=1}^{T} \pi(X_t | Z_1) \prod_{p=1}^{P} \pi(A_p | X_t) \prod_{q=1}^{Q} \pi(B_q | X_t) \prod_{r=1}^{R} \pi(C_r | X_t) \prod_{s=1}^{S} \pi(D_s | X_t)$$

Donde:

A, B, C y *D* son las variables manifiestas ordinales con *J* categorías. En este ejemplo se asignaron las letras *p, q, r* y *s* para denotar las categorías de las variables manifiestas. Así, (*p*= 1,2,3,4), (*q*= 1,2,3), (*r*= 1,2) y (*s*= 1,2). La variable *latente* está representada por la letra *X* con realización t (*t*=1, 2, ... *T*) y la covariable con la letra *Z*.

Debido a que se seleccionó el modelo de 4-clases como el mejor, se quiso evaluar si mejoraba al controlar el *nivel económico familiar* o el *capital cultural de la familia*. En la tabla 7 se muestran los resultados del estadístico de ajuste para cada modelo.

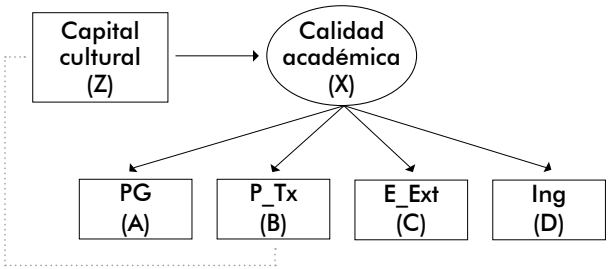
Tabla 7. Resultados de la evaluación de los modelos básico y con covariables

Modelos	BIC	Npar	p-value	Class.Err.
1-clase	789892 .31	7	4.3e-7191	0.00
2-clases	760011 .93	12	1.1e-713	0.08
3-clases	758283 .71	17	2.6e-335	0.15
4-clases	756879 .96	22	1.5e-32	0.27
5-clases	756918 .40	27	3.4e-31	0.29
4-clases_económico	749386 .011	25	1.1e-196	0.26
4-clases_cultural	**746198.049**	**25**	**5.8e-165**	**0.25**

BIC=criterio bayesiano de información; Npar=número de parámetros; p-value=significancia del modelo, y Class.Err.=clasificación del error. En negritas, el modelo con un valor menor al del BIC.

Después de analizar los residuos de los modelos se decidió modelar la dependencia entre las variables *capital cultural* (variable *Z*) y *nivel de habilidad en el uso del procesador de textos* (variable *B*) (figura 8).

Figura 8. Modelo con una covariable y dependencia



PG=promedio general del bachillerato; P_Tx=habilidad para usar procesador de textos; E_Ext=presentación de exámenes extraordinarios; Ing=lectura de textos académicos en inglés. La dependencia de la relación entre capital cultural y habilidad para usar procesador de textos se muestra con una línea punteada. En los paréntesis se muestran las etiquetas que se dieron a las variables para modelarlas formalmente en la ecuación.

Al analizar la tabla de resultados se detectó que el modelo con una covariable y dependencias (al que se denominó 4-*Cl-cul/p_tx*) tenía mejor ajuste. En la tabla 8 se muestran los resultados de todos los modelos analizados.

Tabla 8. Resultados de los modelos básicos y de los modelos con covariables y dependencias

Modelos	BIC	Npar	p-value	Class.Err.
1-clase	789892.031	7	4.3e-7191	0.00
2-clases	760011.093	12	1.1e-713	0.08
3-clases	758283.071	17	2.6e-335	0.15
4-clases	756879.096	22	1.5e-32	0.27
5-clases	756918.040	27	3.4e-31	0.29
4-cla-econ	749386.011	25	1.1e-196	0.26
4-cla-cult	746198.049	25	5.8e-165	0.25
4cla-cul/p_tx	**746015.075**	**26**	**7.4e-128**	**0.28**

BIC=criterio bayesiano de información; Npar=número de parámetros; p-value=significancia del modelo, y Class.Err.=clasificación del error. En negritas, el modelo con un valor menor al del BIC.

Asimismo, se consideró conveniente analizar los residuos de los modelos. Este análisis confirmó que el modelo "4-*cla-cul/p_tx*" era el mejor ya que tenía menores residuos. En la tabla 9 se muestran los residuos del modelo sin clases; el modelo de 4-clases; el modelo de 4-clases y nivel económico; el modelo de 4-clases y capital cultural, y el modelo de 4-clases, capital cultural y dependencias entre capital cultural y procesador de textos.

Tabla 9. Residuos de los modelos

Modelo sin clases

	PG	P_tx	Ex_extr	Ing
PG	.			
P_tx	185.119	.		
Ex_extr	**8839.526**	93.1472	.	
Tx_Ing	131.8557	1317.3246	151.0206	.

Modelo de 4-clases

	PG	P_tx3	Ex_extr	Ing
PG	.			
P_tx	**7.557**	.		
Ex_extr	1.5572	0.0145	.	
Ing	2.9662	0.4805	0.0985	.

Modelo de 4-clases y nivel económico (como covariado)

	PG	P_tx	Ex_extr	Ing
PG	.			
P_tx	10.8848	.		
Ex_extr	14.5931	1.1847	.	
Ing	6.0599	2.2151	0.09	.
Covariates	PG	P_tx	Ex_ext	Ing
Nivel económico	24.4218	**31.2356**	0.8829	9.5395

Modelo de 4-clases y capital cultural (como covariado)

	PG	P_tx	Ex_extr	Ing
PG	.			
P_tx3	11.5473	.		
Ex_extra2	11.9591	2.2362	.	
Tx_Ing2	4.5028	0.826	0.0056	.
Covariates	PG	P_tx	Ex_ext	Ing
Capital cultural	6.3126	**33.5607**	1.2991	12.4493

Modelo de 4-clases, capital cultural (como covariado) y dependencias entre capital cultural y procesador de textos

	PG	P_tx	Ex_extr	Ing
PG	.			
P_tx	13.0073	.		
Ex_extr	8.2832	2.4784	.	
Ing	5.3433	0.0667	1.8169	.
Covariates	PG	P_tx	Ex_ext	Ing
Capital cultural	7.0481	**15.6309**	1.4536	8.649

PG=promedio general del bachillerato; P_Tx=habilidad para usar procesador de textos; E_Ext=presentación de exámenes extraordinarios; Ing=lectura de textos académicos en inglés. Se resalta con negritas el residuo más grande.

Paso III. Resultados

Análisis de clases latentes

Para determinar si los candidatos a ingresar a los planteles del Sistema de Educación Superior Tecnológica conformaban una población heterogénea en cuanto a su calidad académica, se llevó a cabo un análisis de clases latentes. En él se probaron modelos que permitían detectar la ausencia de clases y modelos que proponían diferente número de grupos en la población.

La estructura de relaciones entre las variables manifiestas indicó la presencia de *cuatro clases* con diferentes atributos académicos. El modelo final incorporó como covariado afectando al *capital cultural de la familia* y a dependencias entre dos indicadores: *capital cultural y procesador de textos.*

En la tabla 10 se muestran los resultados obtenidos por el modelo seleccionado. En la parte superior se señala el tamaño y el nombre que se asignó a los cuatro grupos (clases) del modelo. En secciones posteriores se muestra la probabilidad condicional de obtener un valor en las variables manifiestas dada la membrecía de la clase. Así, se puede observar que los candidatos que pertenecen al grupo denominado *calidad académica elemental* tienen alta probabilidad de haber presentado exámenes extraordinarios durante el bachillerato.

Tabla 10. Resultados del modelo de 4-clases con capital cultural que afectan a las clases y dependencias entre el capital cultural y la habilidad para usar el procesador de textos

	Grupo 1 Elemental	Grupo 2 Satisfactoria	Grupo 3 Deficiente	Grupo 4 Sobresaliente
Tamaño del grupo	0.31	0.25	0.24	0.20
Promedio general				
6-6.9	0.04	0.00	0.08	0.00
7-7.9	**0.50**	0.09	**0.60**	0.08
8.8.9	0.42	**0.56**	0.31	**0.54**
>=9	0.04	0.35	0.02	0.39
Exámenes extraordinarios				
No	0.05	**0.92**	0.18	**0.98**
Sí	**0.95**	0.08	**0.82**	0.02
Uso de procesador de textos				
Poco hábil	0.12	0.30	**0.42**	0.06
Hábil	0.39	**0.45**	**0.42**	0.33
Muy hábil	**0.49**	0.24	0.15	**0.61**
Lectura de textos académicos en inglés				
No	0.32	**0.60**	**0.65**	0.18
Sí	**0.68**	0.40	0.35	**0.82**
Covariable Capital cultural				
Bajo	0.23	**0.63**	**0.57**	0.22
Medio	0.23	0.20	0.22	0.21
Alto	**0.55**	0.17	0.21	**0.57**

Se presentan el tamaño de los grupos, la probabilidad condicional de responder a una categoría, dada la clase latente a la que pertenece y las probabilidades de estar en una de las categorías de las covariables.

En los siguientes párrafos se describen las características de los grupos detectados en la población.

1. El grupo más grande incorpora a 31% de la población y presenta un patrón al que denominamos *calidad académica elemental*. Este grupo de sustentantes obtuvo alta probabilidad de tener un promedio general bajo (entre 7 y 7.9) y de haber presentado exámenes extraordinarios durante el bachillerato. No obstante, posee los otros dos recursos que le permitirán afrontar algunas demandas de la educación superior: la habilidad para elaborar documentos en la computadora con un procesador de textos y la de leer textos académicos escritos en inglés.

   Por lo que se refiere al capital cultural de la familia, este grupo tiene altas probabilidades de estar en el grupo más alto de la población.

2. El siguiente grupo, al que identificamos como de *calidad académica satisfactoria,* muestra a 25% *de los participantes*. Estos sustentantes tuvieron un promedio general alto y no presentaron exámenes extraordinarios durante el bachillerato. Sin embargo, no cuentan con la habilidad de leer textos académicos escritos en inglés y los niveles de habilidad para usar un procesador de textos se distribuyen en todas las categorías.

   Por lo que se refiere al capital cultural de la familia, en esta clase se agruparon estudiantes provenientes de un bajo nivel.

3. El tercer grupo representa 24% *de la población* con un patrón de *calidad académica deficiente*. En él, los sustentantes tuvieron bajos promedios en bachillerato (entre 7 y 7.9) y presentaron exámenes extraordinarios. Asimismo, no tienen niveles altos de habilidad para usar un procesador de textos y no pueden leer textos académicos en inglés. Este grupo de candidatos está en los niveles bajos de capital cultural familiar.

4. La cuarta clase aglutina a 20% *de los participantes*. Se caracteriza por tener un patrón de *calidad académica excelente,* con altos promedios durante el bachillerato sin la presentación de exámenes extraordinarios, con niveles altos de habilidad para el uso de procesadores de texto y la capacidad para leer textos académicos en inglés. Estos candidatos provienen de entornos con un alto capital cultural familiar.

Perfil descriptivo

En esta sección se presentan algunos datos descriptivos que permiten conformar un panorama general de los insumos estudiantiles que reciben los planteles que conforman el Sistema Nacional de Institutos Tecnológicos de Educación Superior.

La Dirección General de Educación Superior Tecnológica (DGEST) tiene bajo su responsabilidad el Sistema Nacional de Institutos Tecnológicos (SNIT), entidad académica que forma ingenieros y profesionales de las áreas económico-administrativas. Cuenta con una matrícula cuya proporción es de 80% en las áreas de ingeniería y tecnología, y de 20% en las áreas económico-administrativas.

El SNIT está conformado por 218 instituciones: 212 institutos y seis Centros especializados, de los cuales 110 son *instituciones federales* con presencia en los 31 estados de la República, y 108 *estatales,* ubicados en 22 entidades federativas. Por su vocación, 185 son tecnológicos industriales; 20, agropecuarios; seis, del mar, y uno, forestal. Los seis centros especializados son el Centro Nacional de Investigación y Desarrollo Tecnológico (Cenidet), el Centro Interdisciplinario de Investigación y Docencia en Educación Técnica (CIIDET) y los cuatro Centros Regionales de Optimización y Desarrollo de Equipo (CRODE).

Con esta sección descriptiva se proporciona un ejemplo, con datos reales, sobre la distribución de las clases o grupos en los planteles de un sistema

educativo. La información permite reconocer la variabilidad en el insumo estudiantil de los planteles federales y descentralizados, así como en las entidades federativas.

La distribución de los grupos se presenta de acuerdo con el régimen de los planteles donde los solicitantes presentaron el examen de ingreso. Primero se muestra la distribución de los grupos de sustentantes con diferente calidad académica que solicitaron ingresar a planteles federales (figura 9); después, la distribución para los planteles descentralizados o estatales (figura 10).

Figura 9. Distribución de los grupos en los planteles federales del SNIT

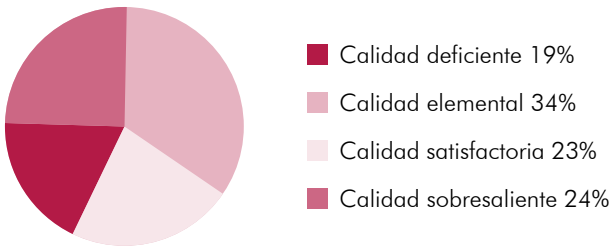
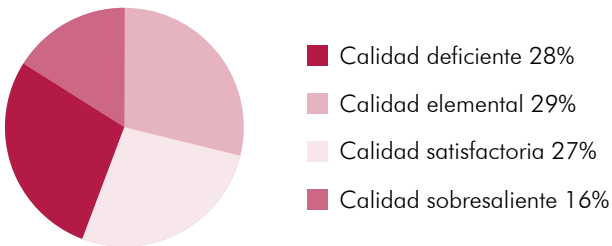
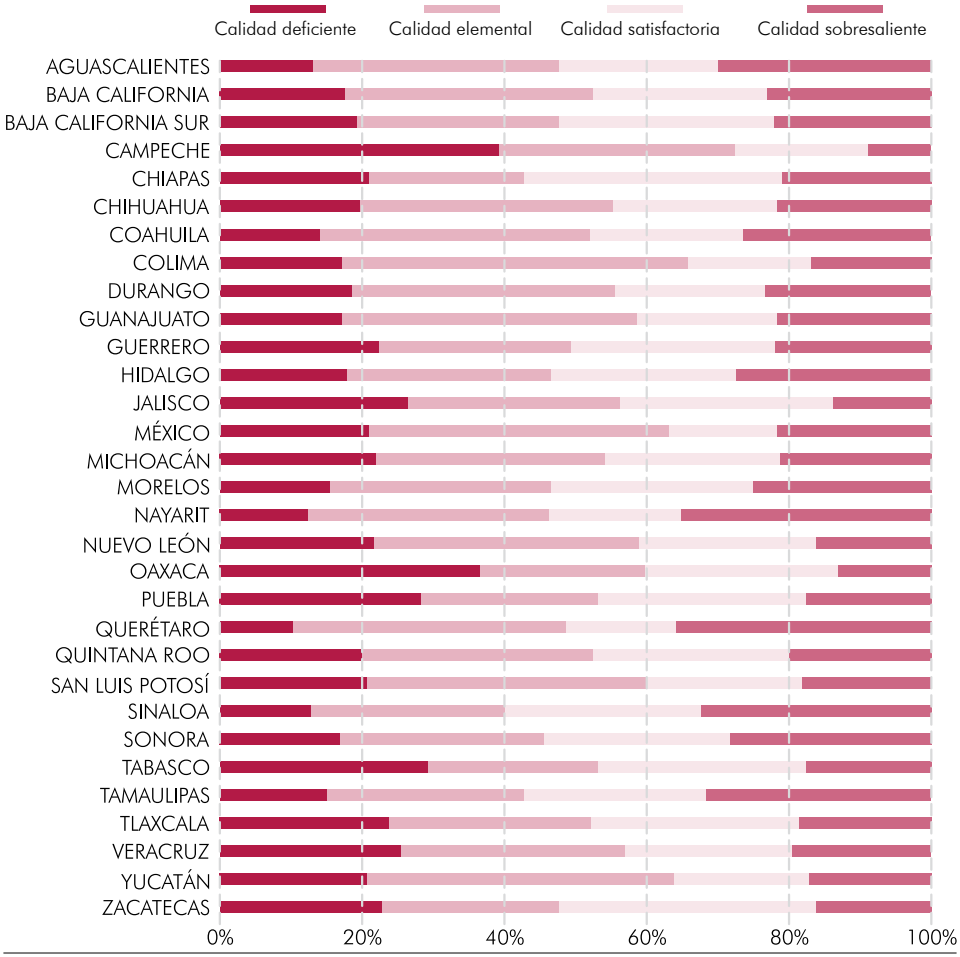


Figura 10. Distribución de los grupos en los planteles descentralizados del SNIT



Por último se muestra la distribución porcentual de los grupos de estudiantes con diferente calidad académica en cada entidad federativa, considerando tanto los planteles federales como los descentralizados (figura 11).

Figura 11. Distribución de los grupos por entidad federativa



Variables categóricas

Las variables categóricas son datos recolectados por los instrumentos de medición que se utilizan en el ámbito educativo y pueden ser analizados con los métodos cuantitativos que reconocen diferentes métricas en las variables manifiestas, como el análisis de clases latentes.

En las variables categóricas, también llamadas variables discretas (si son ordinales o dicotómicas), los valores se distribuyen en un número pequeño de valores o categorías. Una explicación sobre escalas de medición y las distribuciones asociadas al tipo de variable puede encontrarse en el libro de Alan Agresti (2002).

Cuando las respuestas o categorías de una variable discreta reflejan un orden se les denomina *variables ordinales*. Nivel de habilidad en cómputo (*poco hábil, hábil* y *muy hábil*) es un ejemplo de una variable en la que las respuestas están ordenadas; en este caso, de menor a mayor nivel de habilidad. Los análisis estadísticos para este tipo de variables deben considerar el orden de las categorías, por lo que una alteración en el orden podría generar que variaran los resultados.

A las variables categóricas que no tienen un orden se les denominan *variables nominales*. Como ejemplo podríamos considerar el reactivo de un cuestionario que explora el lugar donde se compran cigarros (tienda de autoservicio, miscelánea de barrio, tienda de vinos y licores, restaurantes o farmacias).

Las variables nominales también pueden subdividirse en *nominales dicotómicas* (cuando tienen sólo dos categorías) o *nominales politómicas* (cuando tienen más de dos). Un ejemplo de variable dicotómica nominal es el género del sustentante de un examen (hombre/mujer), mientras que un ejemplo de variable nominal politómica sería el tipo de tabaco que se prefiere consumir (claro/oscuro/suave/mezcla).

Es importante resaltar que las técnicas estadísticas desarrolladas para analizar datos nominales también pueden ser utilizadas para variables ordinales, aunque la información sobre el ordenamiento de las categorías se perdería. Sin embargo, las técnicas desarrolladas para variables ordinales no podrían ser aplicadas a variables nominales.

Tablas de contingencia

En esta sección se repasa la notación común que se usa en los análisis de tablas cruzadas o de contingencia. Una amplia explicación sobre este tema se puede encontrar en los primeros capítulos del libro de Wickens (1989), así como en el libro introductorio publicado por Agresti y Franklin (2007).

En una *tabla de contingencia* se presentan los datos de la frecuencia cruzada de al menos dos variables categóricas. Los renglones muestran las categorías de una variable y las columnas, las de la otra variable. Cada entrada en la tabla presenta la frecuencia de casos de la muestra que están en cierto valor, considerando las dos variables. A estas entradas se les conoce con el nombre de *celdas*. El elemento *ij* en una tabla denota a un individuo con respuesta *i* en el renglón, y respuesta *j* en la columna.

En la tabla 11 se presenta una tabla de dos variables; una con los datos del género de los sustentantes, y otra con información relativa a si los estudiantes trabajaron mientras estudiaban el bachillerato. Ambas variables tienen dos categorías cada una. Por tal razón, se le reconoce como tabla de contingencia de 2X2. Por ejemplo, el valor de la primera celda nos indica que, en la población bajo estudio, hay 25,070 hombres que sí trabajaban mientras estudiaron el bachillerato.

Tabla 11. Número de sustentantes, por género, que trabajaban durante el bachillerato

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	25 070	7 937	33 007
	No	45 644	30 751	76 395
	Total	70 714	38 688	109 402

Ji-cuadrada de Pearson

Alguno de los términos más utilizados cuando se trabaja con tablas de contingencia son las *frecuencias observadas* y las *frecuencias esperadas*.

Las primeras se refieren a los datos recabados en el estudio, mientras que el valor de las frecuencias esperadas puede estimarse para cada una de las celdas de la tabla. Para el caso en el que las variables X, Y son independientes, las *frecuencias esperadas* se pueden estimar de la siguiente manera:

$$\textit{frecuencias esperadas} = \frac{(\textit{total fila}) \times (\textit{total columna})}{\textit{total observaciones}}$$

Para comparar qué tanto las frecuencias observadas se alejan de las frecuencias esperadas, y tener así un indicador de su independencia, generalmente se utiliza la ji-cuadrada de Pearson (x^2) con los grados de libertad (I-1) (J-1), con I=número de categorías de X y J= número de categorías de Y.

$$x^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(observadas - esperadas)^2}{esperadas}$$

Las frecuencias de cada celda de las tablas de contingencia pueden ser transformadas fácilmente a proporciones, que se obtienen dividiendo el valor de una celda entre el número total de observaciones; se les reconoce como *proporciones conjuntas*.

En la mayoría de las tablas de contingencia las variables tienen una relación "causal", en la que una de las variables (*X*) se convierte en una variable *explicativa,* y otra, la variable (*Y*), en una de *respuesta*. En este caso, lo conveniente es estimar *proporciones condicionales*.

Cuando las variables mantienen relaciones simétricas, no tendría sentido denominar a una explicativa y a otra de respuesta. Para describir las asociaciones en estas tablas, recurriremos a distribuciones conjuntas, distribuciones condicionales de *Y* dado *X* o distribución condicional *X* dado *Y*.

Distribuciones de probabilidad

Una tabla de contingencia puede tener varias distribuciones de probabilidad.

- *Distribución conjunta*. Es la probabilidad de que un sujeto tomado aleatoriamente de la muestra obtenga un valor localizado en la fila *i* y columna *j* de una tabla. Su notación es π_{ij}.

En la tabla 12 se muestran ejemplos de probabilidades conjuntas en las celdas sombreadas.

Tabla 12. Distribución conjunta

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	π_{11}	π_{12}	
	No	π_{21}	π_{22}	
				1

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	.23	.07	
	No	.42	.28	
				1

Las sumas de la probabilidad conjunta de todas las celdas de la tabla son iguales a la unidad, lo mismo que la suma de las probabilidades marginales de las filas o las columnas.

$$\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$$

- *Distribuciones marginales*. Estas distribuciones son la suma de las probabilidades conjuntas a través de todas las celdas de la fila o renglón. Se denota a las probabilidades marginales del renglón como π_{i+}.

$$\pi_{i+} = \sum_j \pi_{ij}$$

Por ejemplo, la sumatoria del "margen" de la fila 1 de la tabla 13 sería: $\pi_{i+} = \pi_{11} + \pi_{12}$

Tabla 13. Distribuciones de probabilidad marginal asociada a trabajar durante el bachillerato

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	π_{11}	π_{12}	π_{1+}
	No			
				1

		Género		
		Hombre	Mujer	Total
Trabajaban durante el bachillerato	Sí	.23	.07	**.30**
	No			

Asimismo, son distribuciones marginales la suma de las celdas de la columna, denotadas generalmente como π_{+j}.

$$\pi_{+j} = \sum \pi_{ij}$$

Por ejemplo, la sumatoria del "margen" de la columna 1 de la tabla 14 sería $\pi_{+1} = \pi_{11} + \pi_{21}$

Tabla 14. Distribución de probabilidad marginal de los hombres

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	π_{11}		
	No	π_{21}		
		π_{1+}		

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	.23		
	No	.42		
		.65		

- *Distribuciones condicionales*. Son aquellas en las que el valor de una de las variables depende del valor de otra. Para escribir una probabilidad condicional se usa una línea vertical entre las variables: por ejemplo, cuando se dice que $\pi(X|Y)$. significa que la probabilidad de que ocurra X está condicionada por (depende de) la ocurrencia de Y y *la relación que exista entre ambos eventos*. En la tabla 15 se muestra la notación de probabilidades condicionales.

  La falta de independencia entre las variables que definen una tabla de contingencia, es un indicador del grado de asociación entre ellas.

Tabla 15. Distribución condicional

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	$\pi_{1\|1}$	$\pi_{1\|2}$	
	No	$\pi_{2\|1}$	$\pi_{2\|2}$	

		Género		Total
		Hombre	Mujer	
Trabajaban durante el bachillerato	Sí	.35	.20	
	No	.65	.80	

Agresti, A. & Franklin, C. (2007). *Statistics: The Art and Science of Learning from Data,* New Jersey: Pearson Prentice Hall.

Agresti, A. (2002). *Categorical Data Analysis,* 2nd edition, New York: Wiley-Interscience.

Agresti, A. (2007). *Introduction to Categorical Data Analysis,* 2nd edition, New Jersey: John Wiley & Sons, Inc.

Bartholomew, D. J., Steele, F., Moustaki, I. & Galbraith, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists.* New York: Chapman & Hall.

Bauer, D.J. & Curran, P.J. (2004). "The integration of continuous and discrete Latent Variable Models: Potential Problems and Promising Opportunities", *Psychological Methods,* 9 (1), 3-29.

Bucholz, K. K., Andrew, C. H., Reich, T. Hesselbrock, V. M., Krarner, J. R., Nurnberger, J. I. & Schuckit, M. A. (2006). "Can we subtype alcoholism? A latent class analysisi of data from relatives of alcoholics in a multicenter family study of alcoholism", *Alcoholism Clinical and Experimental Research,* 20(8), 1462-1471.

Cervini, R. (2002). "Desigualdades en el logro académico y reproducción cultural en Argentina", *Revista Mexicana de Investigación Educativa,* 16 (7), 445-500.

Chassin, L., Pitts, S.C. & Prost, J. (2002). "Binge drinking trajectories from adolescence to emerging adulthood in a high-risk sample: predictors and substance abuse outcomes", *Journal of Consulting and Clinical Psychology,* 70(1), 67-78.

Croon, M.A. (1990). "Latent class analysis with ordered latent classes", *British Journal of Mathematical and Statistical Psychology,* 43, 171-192.

Dayton, C.M. (1999). *Latent class scaling analysis,* Thousand Oaks: Sage Publications.

Everitt, B.S., Sabine, L. & Morven, L. (2001). *Cluster Analysis.* 4th edition, New York: Oxford University Press.

Gill, J. (2002). Bayesian methods: A social and behavioral sciences approach. Boca Raton, FL: CRC Press

Goodman, L. (2002). "Latent class analysis: the empirical study of latent types, latent variables and latent structures", en J. Hagenaars & A. McCutcheon (eds.). *Applied Latent Class Models* (pp. 3-55), New York: Cambridge University Press.

Goodman, L.A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models", *Biometrika,* 61, 215-231.

Hagenaars, J.A. (1988). "Latent structure models with direct effects between indicators: local dependence models". *Sociological Methods and Research,* 16, 379-405.

Hagenaars, J.A. (1990). *Categorical longitudinal data: loglinear analysis of panel, trend and cohort data,* Newbury Park, CA: Sage Publications.

Hagenaars, J.A. & McCutcheon, A.L. (2002). *Applied latent class analysis.* Cambridge, UK: Cambridge University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models: similarities and differences,* Thousand Oaks, CA: Sage Publications.

Howley, C.B. & Howley, A.A. (2004). "School size and the influence of socioeconomic status on student achievement: confronting the threat of size bias in national data sets". *Education Policy Analysis Archives,* 12(52).

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding groups in data: an introduction to cluster analysis,* New York: John Wiley and Sons, Inc.

Lazarsfeld, P.F. (1950). "The logical and mathematical foundation of latent structure analysis and the interpretation and mathematical foundation of latent structure analysis", en Stouffer, S.A. *et al.* (eds.), *Measurement and prediction* (pp. 362–472), Princeton, NJ: Princeton University Press.

Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis,* Boston, MA: Houghton Mill.

Magidson, J. & Vermunt, J.K, (2004). "Latent class models", en D. Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences,* chapter 10, 175-198, Thousand Oakes: Sage Publications.

Magidson, J. & Vermunt, J.K. (2001). "Latent class factor and cluster models, bi-plots and related graphical displays", *Sociological Methodology,* 31, 223–264.

Magidson, J. & Vermunt, J.K. (2002). "Latent Class Models for clustering: a comparison with K-means", *Canadian Journal of Marketing Research,* 20, 37-44.

McCutcheon, A.L. (1987). *Latent Class Analysis,* Newbury Park, CA:Sage Publications.

McCutcheon, A.L. (2002). "Basic concepts and procedures in single and multiple-group latent class analysis", en Hagenaars, J. & McCutcheon, A. (eds.). *Applied Latent Class Models* (pp. 89-85), New York: Cambridge University Press.

McLachan, G. & Peel, D. (2000). *Finite mixture models,* New York: Wiley Series in Probability and Statistics, John Wiley and Sons, Inc.

Muthén, B. & Muthén, L. (2000). "Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes", *Alcoholism: Clinical and Experimental Research,* 24, 882-891.

Pérez, J. y Fajardo, M.A. (2001). Determinación de la lealtad de voto mediante un modelo de clases latentes, *Estadística española,* 43 (147), 89-103.

Rondán, F.J., Sánchez, M.J. y Villarejo, A.F. (1999). "Análisis de clases latentes en la relación entre calidad de servicio, satisfacción y confianza con la intención de recompra", *Memorias del IX Congreso Hispano-Francés, Logroño (La Rioja),* 16 ,17 y 18 de junio, 1999 (pp. 2025-2036), España: Universidad de la Rioja.

Rose, L. Chassin, Presson, C. & Sherman, S. (2000), *Multivariate applications in substance use research* (pp. 259-258), Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Tremblay, R. E., Nagin, D. S., Seguin, J. R., Zoccolillo, M., Zelazo, P. D., Boivin, M., Perusse, D. & Japel, C. (2004). Physical aggression during early childhood: trajectories and predictors. *Pediatrics,* 114(1), 43-50

Vermunt, J.K. (en prensa). "Latent class models", en Baker, E., McGaw, B. & Peterson, P. (eds.), *International Encyclopedia of Education,* Oxford, UK: Elsevier.

Vermunt, J.K. & Magidson, J. (2002). "Latent class cluster analysis", en J. Hagenaars & A. McCutcheon (eds.). *Applied Latent Class Models* (pp. 89-106), New York: Cambridge University Press.

Vermunt, J.K. & Magidson, J. (2003). *Addendum to Latent GOLD User's Guide: Upgrade for Version* 3, Boston, MA: Statistical Innovations Inc.

Vermunt, J.K. & Magidson, J. (2005). *Technical Guide for Latent GOLD 4.0: Basic and Advanced*, Boston, MA: Statistical Innovations, Inc.

Vermunt, J.K. & Van Dijk, L.A. (2001). "A Nonparametric Random-coefficients Approach: the Latest Class Regression Model", *Multilevel Modelling Newsletter,* 13(2), 6-13.

Wedel, M. & De Sarbo, W. (1994). "A review of recent developments in latent class regression models", en Bagozzi, R.P. (ed.). *Advanced methods of marketing research,* 352-388, Cambridge, MA: Blackwell Publishers.

Willms, D. (2006). *Las brechas de aprendizaje: diez preguntas de la política educativa a seguir en relación con el desempeño y la equidad en las escuelas y los sistemas educativos,* UIS working papers, Montreal, Canada: Instituto de Estadística de la UNESCO.

Wolfe, J.H. (1970). "Pattern clustering by mulltivariate cluster analysis". *Multivariate Behavioral Research,* 5, 329-350.

Paquetes estadísticos:

Muthén, Bengt & Muthén, Linda (2006). *Mplus* (versión 4.0) [software de cómputo], Los Ángeles, CA: Muthén & Muthén.

SAS Institute, Inc. (2003). *PROC LCA* (Versión 1.1.5) [software de cómputo], Cary, NC: SAS Institute, Inc.

Statistical Innovations, Inc. (2005). *Latent GOLD* (Versión 4.5) [software de cómputo], Belmont, MA: Statistical Innovations, Inc.

El Centro Nacional de Evaluación para la Educación Superior es una asociación civil sin fines de lucro constituida formalmente el 28 de abril de 1994, como consta en la escritura pública número 87036 pasada ante la fe del notario 49 del Distrito Federal. Sus órganos de gobierno son la Asamblea General, el Consejo Directivo y la Dirección General. Su máxima autoridad es la Asamblea General, cuya integración se presenta a continuación, según el sector al que pertenecen los asociados, así como los porcentajes que les corresponden en la toma de decisiones:

Asociaciones e instituciones educativas (40%): Asociación Nacional de Universidades e Instituciones de Educación Superior, A.C. (ANUIES); Federación de Instituciones Mexicanas Particulares de Educación Superior, A.C. (FIMPES); Instituto Politécnico Nacional (IPN); Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM); Universidad Autónoma del Estado de México (UAEM); Universidad Autónoma de San Luis Potosí (UASLP); Universidad Autónoma de Yucatán (UADY); Universidad Nacional Autónoma de México (UNAM); Universidad Popular Autónoma del Estado de Puebla (UPAEP); Universidad Tecnológica de México (UNITEC).

Asociaciones y colegios de profesionales (20%): Barra Mexicana Colegio de Abogados, A.C.; Colegio Nacional de Actuarios, A.C.; Colegio Nacional de Psicólogos, A.C.; Federación de Colegios y Asociaciones de Médicos Veterinarios y Zootecnistas de México, A.C.; Instituto Mexicano de Contadores Públicos, A.C.

Organizaciones productivas y sociales (20%): Academia de Ingeniería, A.C.; Academia Mexicana de Ciencias, A.C.; Academia Nacional de Medicina, A.C.; Fundación ICA, A.C.

Autoridades educativas gubernamentales (20%): Secretaría de Educación Pública.

- Ceneval, A.C.®, EXANI-I®, EXANI-II® son marcas registradas ante la Secretaría de Comercio y Fomento Industrial con el número 478968 del 29 de julio de 1994. EGEL®, con el número 628837 del 1 de julio de 1999, y EXANI-III®, con el número 628839 del 1 de julio de 1999.
- Inscrito en el Registro Nacional de Instituciones Científicas y Tecnológicas del Consejo Nacional de Ciencia y Tecnología con el número 506 desde el 10 de marzo de 1995.
- Organismo Certificador acreditado por el Consejo de Normalización y Certificación de Competencia Laboral (CONOCER) (1998).
- Miembro de la International Association for Educational Assessment.
- Miembro de la European Association of Institutional Research.
- Miembro del Consortium for North American Higher Education Collaboration.
- Miembro del Institutional Management for Higher Education de la OCDE.

La publicación de esta obra la realizó
el Centro Nacional de Evaluación
para la Educación Superior, A.C.
Se terminó de imprimir el 29 de octubre de 2010
en los talleres de Winkilis, Bugambilias 131,
Col. El Rosario, México, D.F., C.P. 09930,
con un tiraje de 500 ejemplares

El Ceneval promueve la calidad de la educación mediante evaluaciones válidas, confiables y pertinentes de los aprendizajes. Así contribuye a la toma de decisiones fundamentadas. El trabajo de centenares de personas dentro y fuera de la propia institución hace posible la renovación constante de las pruebas y la capacitación y actualización de todos los involucrados en el quehacer evaluativo: autoridades educativas, especialistas, académicos, estudiantes y sociedad en general.

La serie de cuadernos técnicos del Ceneval ofrece al estudioso y al practicante de la evaluación educativa algunas técnicas de uso común que le permitirán mejorar el análisis de su objeto de trabajo. Orientados hacia el público conocedor y también hacia el especializado, la intención de estos títulos es promover el uso de herramientas de análisis en círculos cada vez más amplios. Al ser provechosa su lectura se cumplirá uno de los propósitos del Centro: impulsar la cultura de la evaluación en nuestro país para saber quiénes somos y cómo podremos ser mejores.