Variables Latentes: Introducción y conceptos básicos

David Torres Irribarra

Escuela de Psicología
Pontificia Universidad Católica de Chile

30 de enero de 2018

CONTEXTO

Medición, modelos, variables latentes
y modelos de variables latentes como modelos de medición.

Medición

Hay muchas—y muy distintas—teorías de medición.

> *El proceso de obtener experimentalmente uno o más valores de una cantidad que pueden ser razonablemente atribuidos a una cantidad*
>
> — JCGM, 2012

> *La estimación numérica de la razón de una magnitud de un atributo cuantitativo a una unidad del mismo atributo*
>
> — Michell, 1997

> *Una medición es cualquier operación precisamente especificada que produce un número*
>
> — Dingle, 1950

> *Medición es la asignación de numerales para representar propiedades*
>
> — Campbell, 1920

Medición en Ciencias Sociales

En general en ciencias sociales se utiliza—para bien o para mal—la definición de Stevens (1946) como punto de partida:

> *Podemos decir que medición, en el sentido más amplio, es definida como la asignación de numerales a objetos o eventos de acuerdo a reglas*
>
> — Stevens, 1946

La idea clave aquí es que "medición es asignación".

Modelos

La idea de *modelo* es muy amplia, pero podemos entenderla en general de la siguiente forma:

> *Para un observador B, un objeto A* es un modelo de un objeto A en la medida que B pueda usar A* para responder preguntas que le interesen respecto a A*
>
> — Minsky, 1965

Las idea de modelos y modelamiento juegan un rol central en la investigación científica en general y en psicometría en particular.

Modelos

Algunos aforismo

Es inusual que un modelo haga predicciones perfectas o que describa los datos perfectamente.

> *Todos los modelos son incorrectos, pero algunos son útiles*
>
> — Box, 1987

> *Todos los modelos son correctos, pero la mayoría son inútiles*
>
> — Tarpey, 2009

> Yo simplemente diría: *Algunos modelos son útiles.*
>
> —

Nos interesa usar modelos cuyas predicciones son suficientemente buenas para ser útiles en contextos de aplicación específicos.

Variables latentes (VL)

Conceptualmente, una VL es simplemente **una variable** (aleatoria) **no observada que es modelada a través de** (la variabilidad en) **variables observadas**.

La idea se remonta a Spearman (1904) quien postuló que la variación común en variables observadas podría ser utilizada para hacer inferencia respecto a una causa común no observada.

Las variables latentes pueden ser interpretadas y aplicadas de múltiples maneras en diversas disciplinas.

Variables latentes (VL)

> *Latent variables are **random variables** whose realized values are hidden. Their properties must thus be inferred indirectly using a statistical model connecting the latent (unobserved) variables to observed variables.*
>
> — Skrondal & Rabe-Hesketh, 2007

Es clave recordar que **las variables latentes son variables aleatorias**.

Modelos de Variables Latentes (MVL)

Los MVL son **un tipo de modelos estadísticos** que nos permiten hacer inferencias respecto a *una variable aleatoria no observada* a partir de las relaciones que podemos detectar entre variables observadas.

Los MVL son conocidos de muchas formas distintas: *modelos de ecuaciones estructurales, modelos mixtos o modelos de efectos aleatorios, modelos jerárquicos o multinivel, modelos de respuesta al item, modelos de factor común, modelos de clases latentes y modelos de perfil latente.*

Modelos de variables latentes como método de asignación

Recordemos que en general en ciencias sociales se utiliza—para bien o para mal—la definición de Stevens (1946) según la cual medición es asignación numérica.

En ciencias sociales, y en psicometría en particular, los modelos de variables latentes han sido utilizados de forma prominente como un método para hacer esta asignación. En este sentido, los modelos de variables latentes son comunmente entendidos como "modelos de medición".

Like most statistical techniques, latent variable modeling is not an isolated statistical number crunching endeavor but part of a research procedure embedded in a set of more or less closely associated ideas, norms, and practices regarding the proper treatment of data in scientific research.

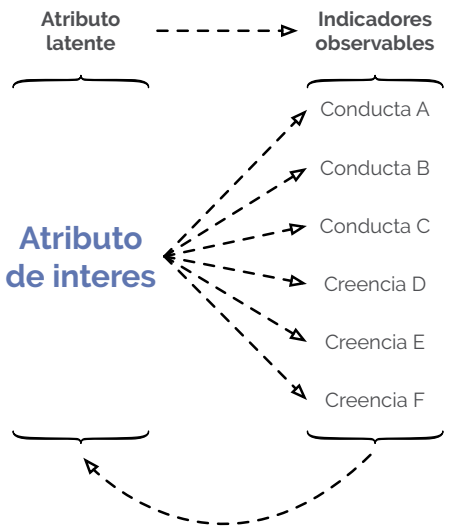
— Borsboom, 2011

Medición de variables latentes mediante variables observadas

Modelando el atributo y los indicadores

- En general, las ciencias sociales están interesadas en estudiar y medir atributos que son comunmente considerados como “no observables”, en oposición al sentido en que, por ejemplo, distancia es considerada “observable”.
- Atributos como “inteligencia”, “personalidad”, “ansiedad” y “conocimiento” son considerados como “no directamente observables” y usualmente no podemos manipularlos.
- Sin embargo, creemos que si bien estos atributo no son observables directamente, asumimos que se manifiestan a través de indicadores visibles.

Hipotetizamos que el atributo latente explica cambios en indicadores observables.



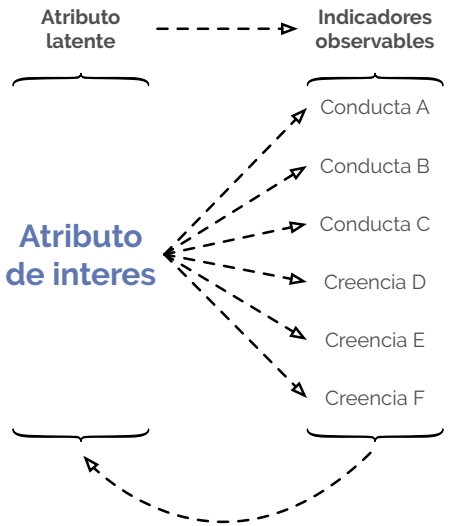
Dado esa hipótesis, hacemos inferencias sobre el atributo latente en base a esos indicadores.

Medición con variables latentes: un modelo probabilístico

Modelando el atributo y los indicadores

- El uso de modelos de variables latentes como modelos de medición implica necesariamente que existe una relación probabilística (en oposición a determinística) entre el atributo latente y los indicadores observables.
- Esto implica que no existe una relación única entre los valores de la variable latente y los valores que observamos en nuestros indicadores.

Hipotetizamos que el atributo latente explica cambios en indicadores observables.



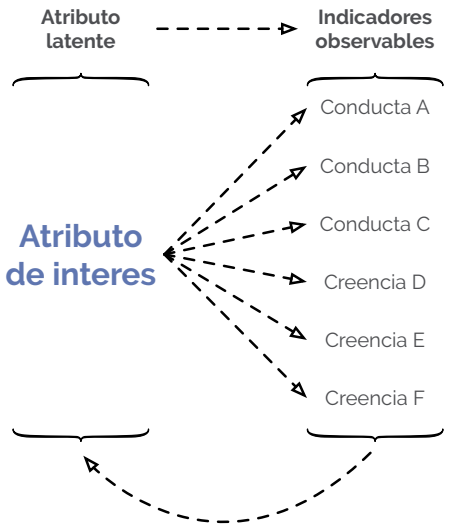
Dado esa hipótesis, hacemos inferencias sobre el atributo latente en base a esos indicadores.

Medición con variables latentes: un modelo probabilístico

Modelando el atributo y los indicadores

- En concreto, esto significa que creemos que cada estado de la variable latente puede provocar múltiples estados en los indicadores. Algunos pueden ser más probables que otros, pero no es posible afirmar con certeza qué estado de la variable latente dio origen a un patrón de indicadores observados.
- Esto nos lleva a un ...

Hipotetizamos que el atributo latente explica cambios en indicadores observables.



Dado esa hipótesis, hacemos inferencias sobre el atributo latente en base a esos indicadores.

PRIMER SUPUESTO CONCEPTUAL

¿Por qué creemos que nuestra variable latente tiene una relación probabilística con los indicadores observables?

1. Modelos probabilísticos

Los modelos de variables latentes son probabilísticos, ya que modelamos la probabilidad de observar estados particulares de las variables observadas (e.g., el patrón de respuestas de un estudiante a una prueba o un cuestionario).

Observamos respuestas a preguntas como "2 + 3 =", "¿Te gusta ir a fiestas?", "¿Usted tiene problemas para dormir?", etc. Pero no creemos inmediatamente que todo el que las responde correcta o afirmativamente sabe sumar, es extrovertido o tiene depresión.

Consideramos que la evidencia que recolectamos tiende a ser un indicador de que algo es más o menos probable, no una fuente de predicción certera, dado un estado de la variable latente.

¿Pero de dónde viene esta aleatoreidad?

Existen dos posibles interpretaciones respecto a la potencial fuente de aleatoreidad en los modelos psicométricos: la interpretación de *sujetos estocásticos* y la interpretación de *muestreo aleatorio*.

1. Modelos probabilísticos

Los modelos de variables latentes son probabilísticos, ya que modelamos la probabilidad de observar estados particulares de las variables observadas (e.g., el patrón de respuestas de un estudiante a una prueba o un cuestionario).

Observamos respuestas a preguntas como “2 + 3 =”, “¿Te gusta ir a fiestas?”, “¿Usted tiene problemas para dormir?”, etc. Pero no creemos inmediatamente que todo el que las responde correcta o afirmativamente sabe sumar, es extrovertido o tiene depresión.

Consideramos que la evidencia que recolectamos tiende a ser un indicador de que algo es más o menos probable, no una fuente de predicción certera, dado un estado de la variable latente.

¿Pero de dónde viene esta aleatoreidad?

Existen dos posibles interpretaciones respecto a la potencial fuente de aleatoreidad en los modelos psicométricos: la interpretación de *sujetos estocásticos* y la interpretación de *muestreo aleatorio*.

1. Modelos probabilísticos

Los modelos de variables latentes son probabilísticos, ya que modelamos la probabilidad de observar estados particulares de las variables observadas (e.g., el patrón de respuestas de un estudiante a una prueba o un cuestionario).

Observamos respuestas a preguntas como "2 + 3 =", "¿Te gusta ir a fiestas?", "¿Usted tiene problemas para dormir?", etc. Pero no creemos inmediatamente que todo el que las responde correcta o afirmativamente sabe sumar, es extrovertido o tiene depresión.

Consideramos que la evidencia que recolectamos tiende a ser un indicador de que algo es más o menos probable, no una fuente de predicción certera, dado un estado de la variable latente.

¿Pero de dónde viene esta aleatoreidad?

Existen dos posibles interpretaciones respecto a la potencial fuente de aleatoreidad en los modelos psicométricos: la interpretación de *sujetos estocásticos* y la interpretación de *muestreo aleatorio*.

1. Modelos probabilísticos

Los modelos de variables latentes son probabilísticos, ya que modelamos la probabilidad de observar estados particulares de las variables observadas (e.g., el patrón de respuestas de un estudiante a una prueba o un cuestionario).

Observamos respuestas a preguntas como "2 + 3 =", "¿Te gusta ir a fiestas?", "¿Usted tiene problemas para dormir?", etc. Pero no creemos inmediatamente que todo el que las responde correcta o afirmativamente sabe sumar, es extrovertido o tiene depresión.

Consideramos que la evidencia que recolectamos tiende a ser un indicador de que algo es más o menos probable, no una fuente de predicción certera, dado un estado de la variable latente.

¿Pero de dónde viene esta aleatoreidad?

Existen dos posibles interpretaciones respecto a la potencial fuente de aleatoreidad en los modelos psicométricos: la interpretación de *sujetos estocásticos* y la interpretación de *muestreo aleatorio*.

1. Modelos probabilísticos

Los modelos de variables latentes son probabilísticos, ya que modelamos la probabilidad de observar estados particulares de las variables observadas (e.g., el patrón de respuestas de un estudiante a una prueba o un cuestionario).

Observamos respuestas a preguntas como "2 + 3 =", "¿Te gusta ir a fiestas?", "¿Usted tiene problemas para dormir?", etc. Pero no creemos inmediatamente que todo el que las responde correcta o afirmativamente sabe sumar, es extrovertido o tiene depresión.

Consideramos que la evidencia que recolectamos tiende a ser un indicador de que algo es más o menos probable, no una fuente de predicción certera, dado un estado de la variable latente.

¿Pero de dónde viene esta aleatoreidad?

Existen dos posibles interpretaciones respecto a la potencial fuente de aleatoreidad en los modelos psicométricos: la interpretación de *sujetos estocásticos* y la interpretación de *muestreo aleatorio*.

Probabilidad por sujetos estocásticos

Esta interpretación coloca la fuente de aleatoreidad a nivel individual.

En este escenario suponemos que el comportamiento de una persona tiene variabilidad inherente.

Vale decir, enfrentados a exactamente la misma situación, nuestras respuestas pueden cambiar debido a un componente aleatorio.

Esta variación del individuo justificaría inferencias del tipo:

> *Una persona cuyo estado en la variable latente es X tiene un 30 % de probabilidad de responder esta pregunta correcta o afirmativamente.*

Probabilidad por sujetos estocásticos

Esta interpretación coloca la fuente de aleatoreidad a nivel individual.

En este escenario suponemos que el comportamiento de una persona tiene variabilidad inherente.

Vale decir, enfrentados a exactamente la misma situación, nuestras respuestas pueden cambiar debido a un componente aleatorio.

Esta variación del individuo justificaría inferencias del tipo:

> *Una persona cuyo estado en la variable latente es X tiene un 30 % de probabilidad de responder esta pregunta correcta o afirmativamente.*

Probabilidad por sujetos estocásticos

Esta interpretación coloca la fuente de aleatoreidad a nivel individual.

En este escenario suponemos que el comportamiento de una persona tiene variabilidad inherente.

Vale decir, enfrentados a exactamente la misma situación, nuestras respuestas pueden cambiar debido a un componente aleatorio.

Esta variación del individuo justificaría inferencias del tipo:

> *Una persona cuyo estado en la variable latente es X tiene un 30 % de probabilidad de responder esta pregunta correcta o afirmativamente.*

Probabilidad por sujetos estocásticos

Esta interpretación coloca la fuente de aleatoreidad a nivel individual.

En este escenario suponemos que el comportamiento de una persona tiene variabilidad inherente.

Vale decir, enfrentados a exactamente la misma situación, nuestras respuestas pueden cambiar debido a un componente aleatorio.

Esta variación del individuo justificaría inferencias del tipo:

Una persona cuyo estado en la variable latente es X tiene un 30 % de probabilidad de responder esta pregunta correcta o afirmativamente.

Probabilidad por muestreo aleatorio

Esta interpretación coloca la fuente de aleatoridad a nivel poblacional.

En este escenario, creemos que en una población, sujetos que comparten el mismo nivel del atributo (la variable latente), pueden responder de forma distinta.

Por ejemplo, de dos estudiantes con el mismo nivel de conocimiento de matemáticas, uno puede que conozca cómo definir números primos y el otro no.

Este tipo de diferencias entre miembros de la población justificaría inferencias del tipo:

Un 30 % de las personas cuyo estado en la variable latente es X responderán esta pregunta correcta o afirmativamente.

Probabilidad por muestreo aleatorio

Esta interpretación coloca la fuente de aleatoridad a nivel poblacional.

En este escenario, creemos que en una población, sujetos que comparten el mismo nivel del atributo (la variable latente), pueden responder de forma distinta.

Por ejemplo, de dos estudiantes con el mismo nivel de conocimiento de matemáticas, uno puede que conozca cómo definir números primos y el otro no.

Este tipo de diferencias entre miembros de la población justificaría inferencias del tipo:

Un 30 % de las personas cuyo estado en la variable latente es X responderán esta pregunta correcta o afirmativamente.

Probabilidad por muestreo aleatorio

Esta interpretación coloca la fuente de aleatoridad a nivel poblacional.

En este escenario, creemos que en una población, sujetos que comparten el mismo nivel del atributo (la variable latente), pueden responder de forma distinta.

Por ejemplo, de dos estudiantes con el mismo nivel de conocimiento de matemáticas, uno puede que conozca cómo definir números primos y el otro no.

Este tipo de diferencias entre miembros de la población justificaría inferencias del tipo:

Un 30 % de las personas cuyo estado en la variable latente es X responderán esta pregunta correcta o afirmativamente.

Probabilidad por muestreo aleatorio

Esta interpretación coloca la fuente de aleatoridad a nivel poblacional.

En este escenario, creemos que en una población, sujetos que comparten el mismo nivel del atributo (la variable latente), pueden responder de forma distinta.

Por ejemplo, de dos estudiantes con el mismo nivel de conocimiento de matemáticas, uno puede que conozca cómo definir números primos y el otro no.

Este tipo de diferencias entre miembros de la población justificaría inferencias del tipo:

Un 30 % de las personas cuyo estado en la variable latente es X responderán esta pregunta correcta o afirmativamente.

Fuentes de variabilidad

Si bien la interpretación individual parece ser común y es en general atractiva en ciencias sociales, es preferible basarse en la interpretación poblacional.

Estrictamente hablando, modelos tradicionales de variables latentes en general no implican la interpretación individual, y se requieren de supuestos adicionales para sostenerla.

La recomendación de preferir la interpretación muestral puede ser encontrada en Lord & Novick (1968), Holland (1990), Borsboom (2005) y artículos que discuten el supuesto de *homogeneidad local* que sería requerido por la interpretación de sujetos estocásticos.

Fuentes de variabilidad

Si bien la interpretación individual parece ser común y es en general atractiva en ciencias sociales, es preferible basarse en la interpretación poblacional.

Estrictamente hablando, modelos tradicionales de variables latentes en general no implican la interpretación individual, y se requieren de supuestos adicionales para sostenerla.

La recomendación de preferir la interpretación muestral puede ser encontrada en Lord & Novick (1968), Holland (1990), Borsboom (2005) y artículos que discuten el supuesto de *homogeneidad local* que sería requerido por la interpretación de sujetos estocásticos.

Fuentes de variabilidad

Si bien la interpretación individual parece ser común y es en general atractiva en ciencias sociales, es preferible basarse en la interpretación poblacional.

Estrictamente hablando, modelos tradicionales de variables latentes en general no implican la interpretación individual, y se requieren de supuestos adicionales para sostenerla.

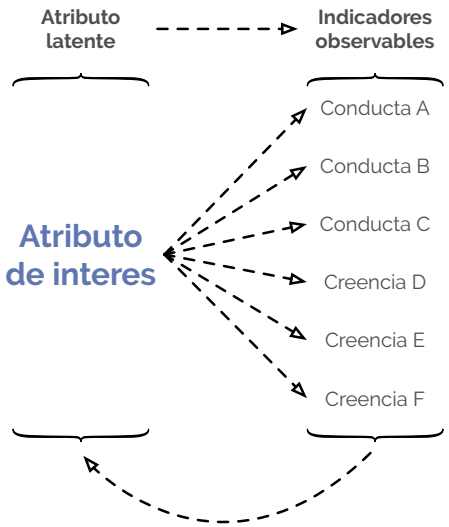
La recomendación de preferir la interpretación muestral puede ser encontrada en Lord & Novick (1968), Holland (1990), Borsboom (2005) y artículos que discuten el supuesto de *homogeneidad local* que sería requerido por la interpretación de sujetos estocásticos.

Medición con variables latentes: un modelo probabilístico

Modelando el atributo y los indicadores

- En resumen, estos modelos conllevan un grado de incertidumbre respecto al valor que puede ser inferido de la variable latente en base un patrón determinado de respuestas.
- Idealmente el uso de un modelo probabilístico debiera ser justificada conceptualmente, pudiendo explicar en términos substantivos el origen de esta incertidumbre.
- Los modelos de variables latentes nos permiten cuantificar el grado de incertidumbre asociado a nuestras inferencias sobre el estado de la variable latente. Esto es central en términos de su interpretación como modelos de medición ya que esta incertidumbre cuantificada es interpretada como nuestra estimación del error de medición.

Hipotetizamos que el atributo latente explica cambios en indicadores observables.



Dado esa hipótesis, hacemos inferencias sobre el atributo latente en base a esos indicadores.

¿TODO CLARO HASTA AQUÍ?

¿TODO CLARO HASTA AQUÍ?

¿Seguros?

Una pausa filosófica: ¿Qué variables son observables y cuáles no observables?

Hasta ahora hemos usado la distinción entre variables observables y no observables basándonos en la intuición que “inteligencia”, “ansiedad” y “conocimiento” son distintas a varibles observables como “largo” y “ancho” o como “sexo” y “nacionalidad”.

¿Creemos que existe alguna diferencia esencial entre variables observables y variables no observables (i.e., una diferencia ontológica)?

¿O es simplemente una diferencia respecto a qué tanta información tenemos o que tanta confianza adscribimos a nuestros datos (i.e., una diferencia epistemológica)?

Una pausa filosófica: ¿Qué variables son observables y cuáles no observables?

Borsboom (2011) argumenta que:

- *"there is no reason to make an ontological distinction between latent and observed variables; hence, all variables are ontologically on par."*
- *"What differs between situations where one treats variables as latent or observed is the degree to which the researcher assumes variable structures to be epistemically accessible."*
- *"To treat variables as observed is to assume full accessibility; that is, inferences from data to variable structure are assumed to be without error."*

Yo estoy de acuerdo con él (en esto).

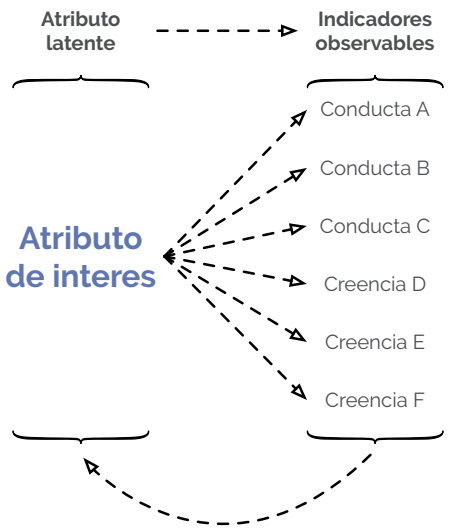
FIN DE LA PAUSA FILOSÓFICA

Estructura de la variable latente y tipos de variables observadas

Modelando el atributo y los indicadores

- El punto clave es que consideramos a estos atributos como no observables directamente pero creemos que se manifiestan a través de indicadores observables.
- Nosotros asumimos la existencia de atributos latentes, y hacemos hipótesis respecto a su estructura.
- **Nuestros modelos estadísticos consideran tanto el tipo de estructura que adscribimos al atributo como también la forma en que modelamos los indicadores**.

Hipotetizamos que el atributo latente explica cambios en indicadores observables.



Dado esa hipótesis, hacemos inferencias sobre el atributo latente en base a esos indicadores.

SEGUNDO SUPUESTO CONCEPTUAL

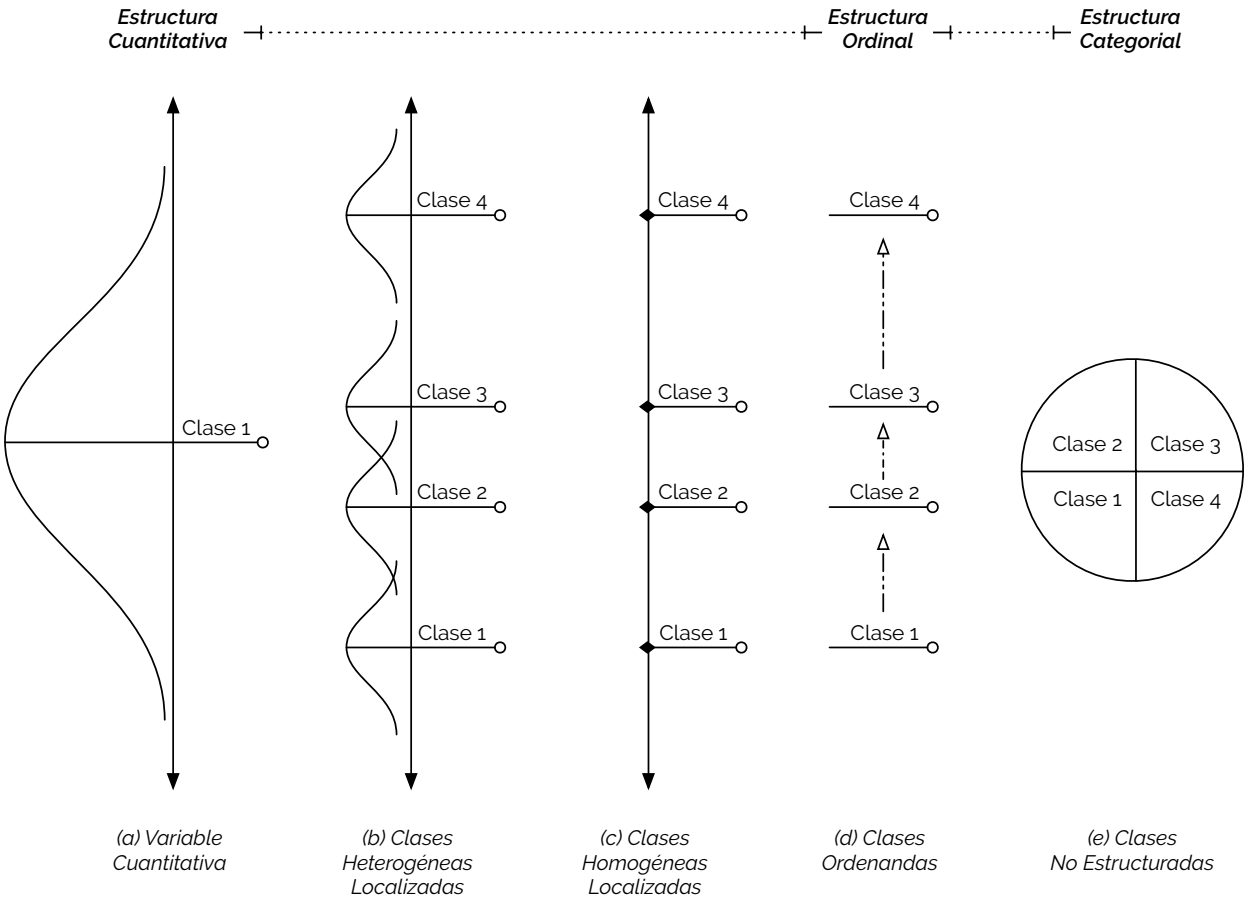
¿Cuál es la estructura de la variable latente?

¿Qué estructura tiene el atributo que queremos medir?

- ¿Entendemos la personalidad como una taxonomía que distingue entre tipos cualitativamente distintos o como el producto de algunas variables cuantitativas?
- ¿Entendemos diferencias en el conocimiento de matemática de las personas como distancias en una cantidad o como estados discretos de entender o no enteder ciertas ideas?

Estructura de la variable latente

Modelo del atributo que queremos estudiar



¿Podemos hacer distinciones de igualdad o desigualdad? ¿Podemos hacer distinciones de orden? ¿Y de distancias?

Tipos de indicadores de la variable latente

Modelo de las variables observables

- Supongamos que ya tenemos una teoría respecto a la estructura de nuestro atributo latente... **tenemos ahora que considerar también que podemos hacer modelar ese atributo mediante distintos tipos de indicadores**.
- Puedo querer medir una variable cuantitativa con indicadores cuantitativos, ordinales o categóricos.
- Asimismo, puedo medir una variable categórica con indicadores cuantitativos, ordinales o categóricos.
- Pero... ¿por qué es importante considerar la estructura de la variable latente y los tipos de indicadores?

Una taxonomía de modelos de variables latentes (típicos)

Modelo del atributo y modelo del indicador

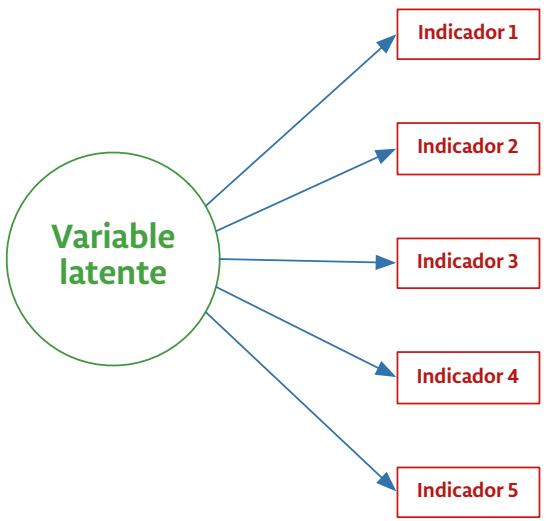
El considerar la estructura (supuesta) del atributo latente y el tipo de indicador obsevable es una manera común de organizar algunos de los tipos de modelos de variables latentes más típicos.

Variables Manifiestas	Variables latentes	
	Cuantitativas	Categóricas
Cuantitativas	Análisis factorial	Análisis de perfil latente
Categóricas	Análisis de rasgo latente Teoría de respuesta al ítem	Análisis de clases latentes

Esta taxonomía no es en ningún caso exhaustiva, sino que recoge los nombres más usados para ciertos modelos de variables latentes que son populares dentro de algunas disciplinas.

¿Representando modelos distintos?

Una manera común de representar modelos de variables latentes es utilizando diagramas acíclicos dirigidos.



Los círculos representan variables latentes.
Así se representan los atributos que estamos midiendo con los indicadores observados.

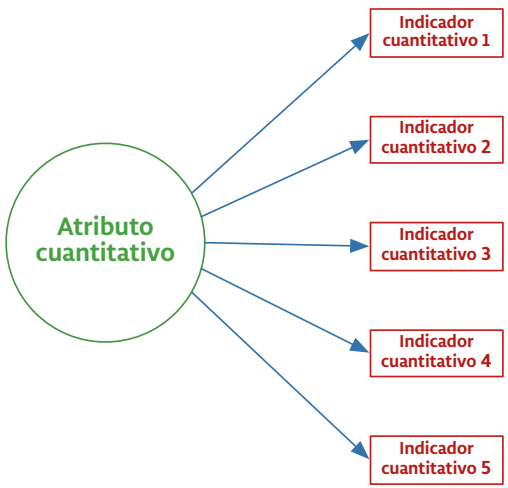
Las flechas rectas representan regresiones.
Así se representan las relaciones entre el atributo latente y los indicadores observados.

Los rectángulos representan variables observadas.
Así se representan los indicadores que usaremos para basar nuestras inferencias sobre el atributo.

Estos diagramas representan gráficamente (de forma no exhaustiva) modelos matemáticos de variables latentes.

¿Representando modelos distintos?

Este diagrama representa un modelo de análisis factorial confirmatorio.



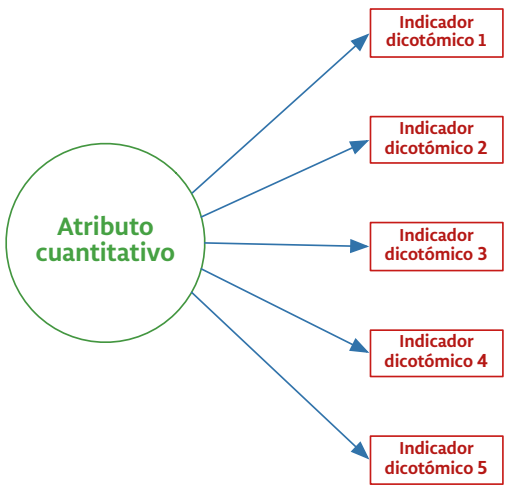
El atributo cuantitativo es llamado factor.

Las relaciones son llamadas cargas (*loadings*).

Los rectángulos representan variables observadas.
Estas pueden tener interceptes asociados, pero normalmente son centradas en cero en analisis factoriales.

¿Representando modelos distintos?

Este diagrama representa un modelo de teoría de respuesta al ítem.



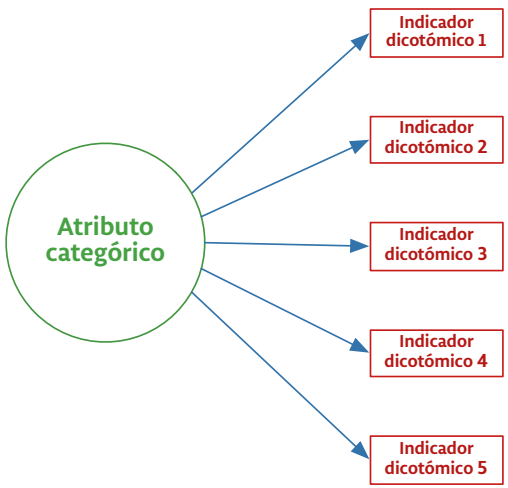
El atributo cuantitativo es llamado rasgo o habilidad.

Las relaciones son llamadas discriminación.

Los interceptos asociados a las variables observadas son llamadas dificultades.

¿Representando modelos distintos?

Este diagrama representa un modelo de análisis de clases latentes.



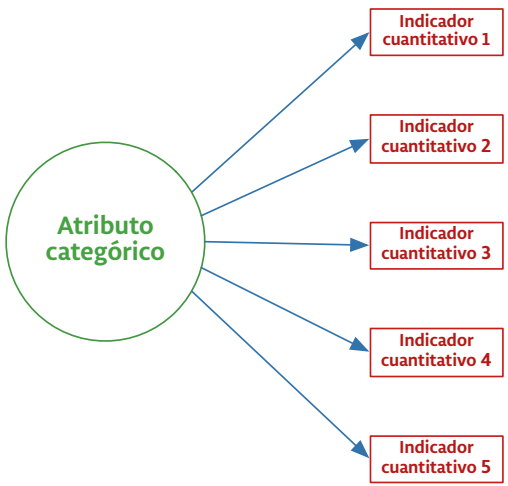
El atributo cuantitativo es llamado clase latente.
El diagrama no expresa que es necesario definir cuantas clases están siendo consideradas.

Las relaciones en este caso indican que los interceptos varían para cada clase.

Los interceptos corresponden a la probabilidad de observar una respuesta afirmativa o correcta.

¿Representando modelos distintos?

Este diagrama representa un modelo de análisis de perfiles latentes.



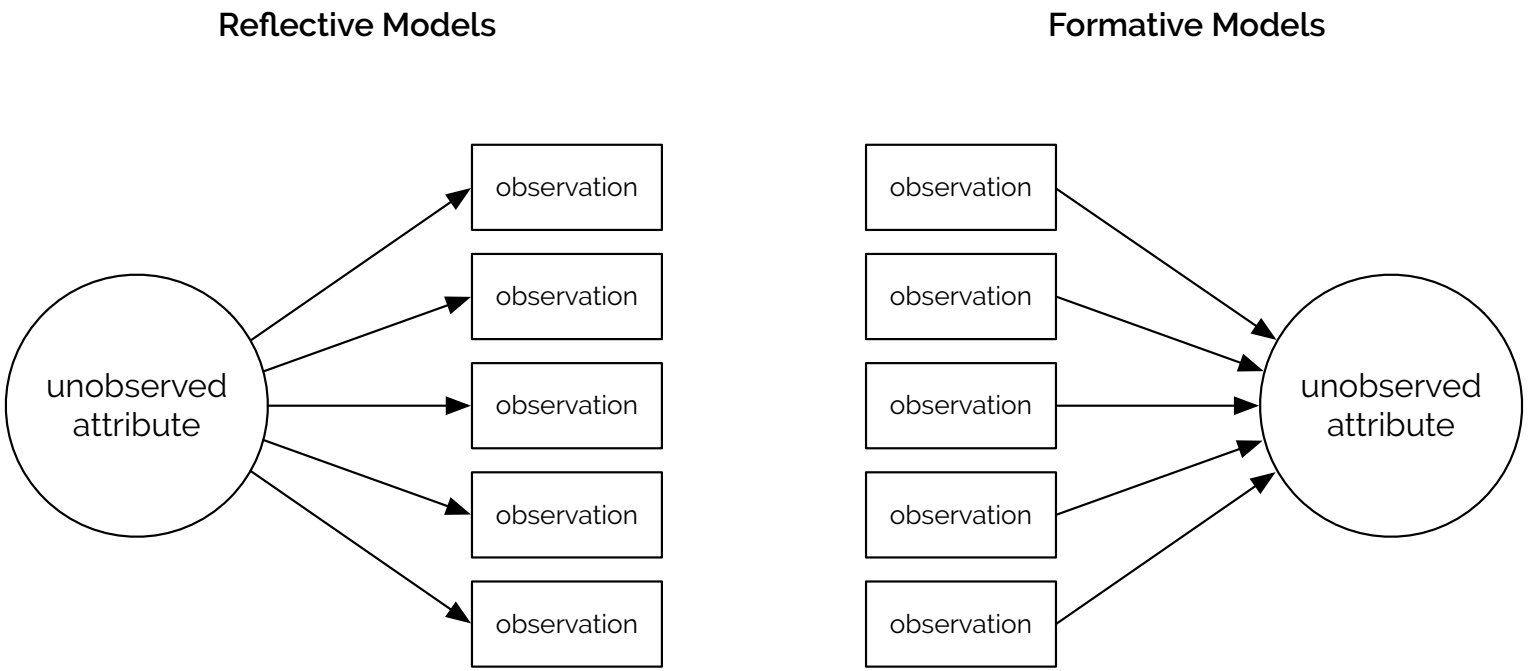
El atributo cuantitativo es llamado clase latente.
El diagrama no expresa que es necesario definir cuantas clases están siendo consideradas.

Las relaciones en este caso indican que los interceptos varían para cada clase.

Los interceptos corresponden a la promedios del indicador en cada clase.

¿TODO CLARO HASTA AQUÍ?

Otra pausa: intepretando las flechas

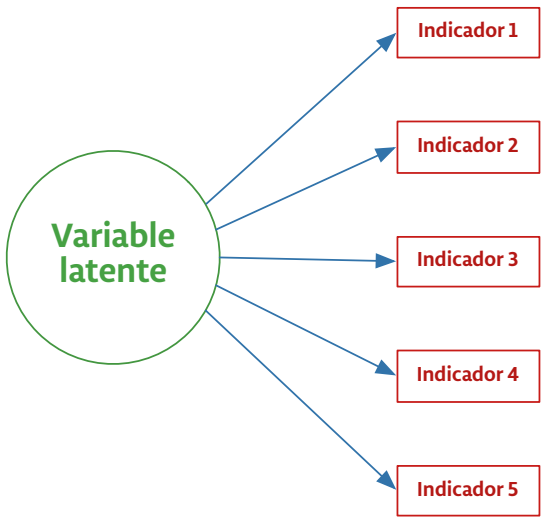


¿Qué pasa cuando la dirección de las flechas se invierte?

FIN DE LA PAUSA

¿Representando modelos distintos?

Todos los modelos de variables latentes representados en las láminas anteriores pueden ser interpretados como modelos de medición.



Los círculos representan variables latentes.
Así se representan los atributos que estamos midiendo con los indicadores observados.

Las flechas rectas representan regresiones.
Así se representan las relaciones entre el atributo latente y los indicadores observados.

Los rectángulos representan variables observadas.
Así se representan los indicadores que usaremos para basar nuestras inferencias sobre el atributo.

Las comunalidades entre estos modelos no se agotan en que puedan ser representados gráficamente de forma similar.

Si bien distintos modelos están basados en distintos supuestos, una comunalidad central entre estos modelos (y en general de cualquier modelo que involucre variables latentes) es que **están basados en el supuesto de independencia condicional**.

SUPUESTO DE INDEPENDENCIA CONDICIONAL

El supuesto estadístico central
detrás de los modelos de variables latentes.

Independencia condicional

Independencia condicional es un supuesto que requiere que las variables observadas sean independientes entre sí cuando se condiciona en la variable latente.

Dicho de otro modo, toda correlación entre los ítemes (variable observadas) es producto de la variable latente.

Este es un supuesto compartido por la mayoría de los modelos de variables latentes.

Un ejemplo de Lazarsfeld y Henry (1968) puede ilustrar este concepto.

Imaginen que 1,000 personas respondieron estas dos preguntas en un cuestionario:

¿Usted lee la revista "A"?

¿Usted lee la revista "B"?

Independencia condicional

Independencia condicional es un supuesto que requiere que las variables observadas sean independientes entre sí cuando se condiciona en la variable latente.

Dicho de otro modo, toda correlación entre los ítemes (variable observadas) es producto de la variable latente.

Este es un supuesto compartido por la mayoría de los modelos de variables latentes.

Un ejemplo de Lazarsfeld y Henry (1968) puede ilustrar este concepto.

Imaginen que 1,000 personas respondieron estas dos preguntas en un cuestionario:

¿Usted lee la revista "A"?

¿Usted lee la revista "B"?

Independencia condicional

Independencia condicional es un supuesto que requiere que las variables observadas sean independientes entre sí cuando se condiciona en la variable latente.

Dicho de otro modo, toda correlación entre los ítemes (variable observadas) es producto de la variable latente.

Este es un supuesto compartido por la mayoría de los modelos de variables latentes.

Un ejemplo de Lazarsfeld y Henry (1968) puede ilustrar este concepto.

Imaginen que 1,000 personas respondieron estas dos preguntas en un cuestionario:

¿Usted lee la revista "A"?

¿Usted lee la revista "B"?

Independencia condicional

Independencia condicional es un supuesto que requiere que las variables observadas sean independientes entre sí cuando se condiciona en la variable latente.

Dicho de otro modo, toda correlación entre los ítemes (variable observadas) es producto de la variable latente.

Este es un supuesto compartido por la mayoría de los modelos de variables latentes.

Un ejemplo de Lazarsfeld y Henry (1968) puede ilustrar este concepto.

Imaginen que 1,000 personas respondieron estas dos preguntas en un cuestionario:

> ¿Usted lee la revista "A"?
>
> ¿Usted lee la revista "B"?

Independencia condicional

	Lee revista A	No lee revista A	Total	Pr
Lee revista B	*260*	*240*	*500*	$Pr(\mathrm{B}) = 0{,}5$
No lee revista B	*140*	*360*	*500*	$Pr(\neg \mathrm{B}) = 0{,}5$
Total	*400*	*600*	*1000*	
Pr	$Pr(\mathrm{A}) = 0{,}4$	$Pr(\neg \mathrm{A}) = 0{,}6$		

Al mirar esta tabla podemos ver que estas variables parecen estar correlacionadas.

Si ambas preguntas fueran independientes, entonces:

$$\Pr(\mathrm{AB}) = \Pr(\mathrm{A}) \times \Pr(\mathrm{B})$$

Pero este no es el caso:

$$0{,}26 \neq 0{,}4 \times 0{,}5$$

Pero podemos hipotetizar que esta relación puede ser explicada por una tercera variable... *nivel educacional*.

Independencia condicional

	Lee revista A	No lee revista A	Total	Pr
Lee revista B	*260*	*240*	*500*	$Pr(\mathrm{B}) = 0{,}5$
No lee revista B	*140*	*360*	*500*	$Pr(\neg\mathrm{B}) = 0{,}5$
Total	*400*	*600*	*1000*	
Pr	$Pr(\mathrm{A}) = 0{,}4$	$Pr(\neg\mathrm{A}) = 0{,}6$		

Al mirar esta tabla podemos ver que estas variables parecen estar correlacionadas.

Si ambas preguntas fueran independientes, entonces:

$$\Pr(\mathrm{AB}) = \Pr(\mathrm{A}) \times \Pr(\mathrm{B})$$

Pero este no es el caso:

$$0{,}26 \neq 0{,}4 \times 0{,}5$$

Pero podemos hipotetizar que esta relación puede ser explicada por una tercera variable... *nivel educacional*.

Independencia condicional

	Lee revista A	No lee revista A	Total	Pr
Lee revista B	*260*	*240*	*500*	$Pr(\mathrm{B}) = 0{,}5$
No lee revista B	*140*	*360*	*500*	$Pr(\neg\mathrm{B}) = 0{,}5$
Total	*400*	*600*	*1000*	
Pr	$Pr(\mathrm{A}) = 0{,}4$	$Pr(\neg\mathrm{A}) = 0{,}6$		

Al mirar esta tabla podemos ver que estas variables parecen estar correlacionadas.

Si ambas preguntas fueran independientes, entonces:

$$\Pr(\mathrm{AB}) = \Pr(\mathrm{A}) \times \Pr(\mathrm{B})$$

Pero este no es el caso:

$$0{,}26 \neq 0{,}4 \times 0{,}5$$

Pero podemos hipotetizar que esta relación puede ser explicada por una tercera variable... *nivel educacional*.

Independencia condicional

	Lee revista A	No lee revista A	Total	Pr
Lee revista B	*260*	*240*	*500*	$Pr(\mathrm{B}) = 0{,}5$
No lee revista B	*140*	*360*	*500*	$Pr(\neg\mathrm{B}) = 0{,}5$
Total	*400*	*600*	*1000*	
Pr	$Pr(\mathrm{A}) = 0{,}4$	$Pr(\neg\mathrm{A}) = 0{,}6$		

Al mirar esta tabla podemos ver que estas variables parecen estar correlacionadas.

Si ambas preguntas fueran independientes, entonces:

$$\Pr(\mathrm{AB}) = \Pr(\mathrm{A}) \times \Pr(\mathrm{B})$$

Pero este no es el caso:

$$0{,}26 \neq 0{,}4 \times 0{,}5$$

Pero podemos hipotetizar que esta relación puede ser explicada por una tercera variable... *nivel educacional*.

Independencia condicional

	Lee revista A	No lee revista A	Total	Pr
Lee revista B	*260*	*240*	*500*	$Pr(\mathrm{B}) = 0{,}5$
No lee revista B	*140*	*360*	*500*	$Pr(\neg\mathrm{B}) = 0{,}5$
Total	*400*	*600*	*1000*	
Pr	$Pr(\mathrm{A}) = 0{,}4$	$Pr(\neg\mathrm{A}) = 0{,}6$		

Al mirar esta tabla podemos ver que estas variables parecen estar correlacionadas.

Si ambas preguntas fueran independientes, entonces:

$$\Pr(\mathrm{AB}) = \Pr(\mathrm{A}) \times \Pr(\mathrm{B})$$

Pero este no es el caso:

$$0{,}26 \neq 0{,}4 \times 0{,}5$$

Pero podemos hipotetizar que esta relación puede ser explicada por una tercera variable... *nivel educacional*.

Independencia condicional

Nivel educacional = 1

	A	¬A	Total
B	*240*	*60*	*300*
¬B	*160*	*40*	*200*
Total	*400*	*100*	*500*

$$\Pr(AB|Ed = 1) = \Pr(A|Ed = 1) \times \Pr(B|Ed = 1)$$

$$\frac{240}{500} = \frac{400}{500} \times \frac{300}{500}$$

$$0{,}48 = 0{,}8 \times 0{,}6$$

Nivel educacional = 2

	A	¬A	Total
B	*20*	*80*	*100*
¬B	*80*	*320*	*400*
Total	*100*	*400*	*500*

$$\Pr(AB|Ed = 2) = \Pr(A|Ed = 2) \times \Pr(B|Ed = 2)$$

$$\frac{20}{500} = \frac{100}{500} \times \frac{100}{500}$$

$$0{,}04 = 0{,}2 \times 0{,}2$$

Independencia condicional

Nivel educacional = 1

	A	¬A	Total
B	*240*	*60*	*300*
¬B	*160*	*40*	*200*
Total	*400*	*100*	*500*

Nivel educacional = 2

	A	¬A	Total
B	*20*	*80*	*100*
¬B	*80*	*320*	*400*
Total	*100*	*400*	*500*

$$\Pr(AB|Ed = 1) = \Pr(A|Ed = 1) \times \Pr(B|Ed = 1)$$

$$\frac{240}{500} = \frac{400}{500} \times \frac{300}{500}$$

$$0{,}48 = 0{,}8 \times 0{,}6$$

$$\Pr(AB|Ed = 2) = \Pr(A|Ed = 2) \times \Pr(B|Ed = 2)$$

$$\frac{20}{500} = \frac{100}{500} \times \frac{100}{500}$$

$$0{,}04 = 0{,}2 \times 0{,}2$$

Independencia condicional

Nivel educacional = 1

	A	¬A	Total
B	*240*	*60*	*300*
¬B	*160*	*40*	*200*
Total	*400*	*100*	*500*

Nivel educacional = 2

	A	¬A	Total
B	*20*	*80*	*100*
¬B	*80*	*320*	*400*
Total	*100*	*400*	*500*

$$
\begin{aligned}
N(AB) = &\big(Pr(A|Ed=1) \times Pr(B|Ed=1) \times N(Ed=1)\big) + \\
&\big(Pr(A|Ed=2) \times Pr(B|Ed=2) \times N(Ed=2)\big) \\
260 = &\big(400/500 \times 300/500 \times 500\big) + \big(100/500 \times 100/500 \times 500\big) \\
260 = &\big(0{,}8 \times 0{,}6 \times 500\big) + \big(0{,}2 \times 0{,}2 \times 500\big) \\
260 = &\big(0{,}48 \times 500\big) + \big(0{,}04 \times 500\big) \\
260 = &\ 240 + 20
\end{aligned}
$$

Independencia condicional

En este ejemplo usamos una tercera variable observada para condicionar la relación entre las dos variables indicadoras, pero en el contexto de modelos de variables latentes las variables indicadoras son condicionadas en las variables latentes incluidas en cada modelo.

En análisis de clases latentes se asume que la independencia de las variables observadas es condicional a la membresía en una clase.

Esto significa asumir que, dentro de cada clase, las variables observadas (e.g. síntomas o respuestas) son estadísticamente independientes.

En análisis de teoría de respuesta al item se asume que la independencia de las variables observadas es condicional al nivel de la variable latente.

Esto significa asumir que, para cada nivel de la variable latente, las variables observadas (e.g. síntomas o respuestas) son estadísticamente independientes.

Independencia condicional

En este ejemplo usamos una tercera variable observada para condicionar la relación entre las dos variables indicadoras, pero en el contexto de modelos de variables latentes las variables indicadoras son condicionadas en las variables latentes incluidas en cada modelo.

En análisis de clases latentes se asume que la independencia de las variables observadas es condicional a la membresía en una clase.

Esto significa asumir que, dentro de cada clase, las variables observadas (e.g. síntomas o respuestas) son estadísticamente independientes.

En análisis de teoría de respuesta al item se asume que la independencia de las variables observadas es condicional al nivel de la variable latente.

Esto significa asumir que, para cada nivel de la variable latente, las variables observadas (e.g. síntomas o respuestas) son estadísticamente independientes.

¿Qué hemos visto de modelos de variables latentes?

- Los MVL son un tipo de modelos estadísticos que nos permiten hacer inferencias respecto a *una variable aleatoria no observada* a partir de las relaciones que podemos detectar entre variables observadas.
- Existen muchos tipos de modelos de variables latentes.
- Todos tienen a la base ciertos supuestos conceptuales: expresan una relación probabilística y asumen cierta estructura de los atributos latentes.
- Todos comparten el supuesto estadistico de la independencia condicional.
- Todos pueden ser expresados utilizando gráficos acíclicos.
- Sin embargo, a pesar de estas similitudes...

With regard to the choice of form for the latent variable structure, there appears to be a strong influence of one's statistical upbringing; for instance, those who are accustomed to working with factor models seem to conceptualize theoretical constructs as continuous dimensions more or less automatically.

Indeed, one sometimes wonders whether psychologists are sufficiently well-informed on the fact that psychological attributes may not necessarily behave as linearly ordered dimensions (e.g., like the factors in a factor model) and that making this assumption while it is false may seriously distort the interpretation of research findings.

— Borsboom, 2011

Unfortunately, this statement applies equally well today and also to researchers from the biometric, econometric, psychometric and other statistical communities.

For instance, bio-statisticians typically attribute the invention of linear mixed models to Laird & Ware (1982), although Laird and Ware themselves referred to the work of the mathematical statistician Harville (1977).

Interestingly, neither Harville nor Laird and Ware appeared to be aware of equivalent models introduced by the econometrician Swamy (1970, 1971) in an Econometrica paper and a Springer book with the title 'Statistical Inference in Random Coefficient Regression Models'.

Even more remarkably, such a lack of communication is also evident within specific statistical communities. For instance, factor analysts and item–response theorists rarely cite each other, although their work is closely related and often published in the same journal, Psychometrika.

— Skrondal & Rabe-Hesketh, 2007

Traduciendo entre múltiples lenguajes

La segmentación entre modelos oculta equivalencias conceptuales y formales entre modelos debido a notaciones alternativas y jergas propias de cada disciplina.

- Regresión logistica multinivel (HLM).

$$\text{logit}(\Pr(y_{ij} = 1|u_j)) = \beta_1 item_{1j} + \cdots + \beta_I item_{Ij} + u_{0j}$$

$$u_{0j} \sim N(0, \tau^2)$$

- Análisis factorial confirmatorio unidimensional generalizado.

$$y^*_{ij} = \beta_i + \lambda\eta_j + e_{ij}$$

$$\eta_j \sim N(0, \psi), cov(\eta_j, e_{ij}) = 0$$

donde

$$y_{ij} = \begin{cases} 1, & \text{if } y^*_{ij} > 0 \\ 0, & \text{if } y^*_{ij} <= 0 \end{cases}$$

- Modelo de teoría de respuesta al ítem de un parametro logístico.

$$\text{logit}(\Pr(Y_{ip} = 1|\theta_p)) = \theta_p - \delta_i$$

Problemas de la fragmentación

- La segmentación entre las distintas tradiciones conlleva múltiples problemas, incluyendo el desconocimiento de estructuras alternativas para conceptualizar atributos.
- Asimismo, la segmentación lleva a la "reinvención" de técnicas ya disponibles en otra disciplinas.
- La falta de flexibilidad al momento de modelar problemas, en ocasiones debido a las restricciones de softwares especializados producto de tradiciones académicas específicas.
- ¿Cómo solucionar esta desconexión?

MARCOS GENERALES DE MODELOS DE VARIABLES LATENTES

Todo se convierte en un caso especial.

Grandes marcos de variables latentes

- Si bien contínua habiendo altos grados de segmentación en el uso de variables latentes, desde hace cerca de 50 años que se han hecho esfuerzos por consolidar esta variedad de modelos bajo un solo marco.
- Lazarsfeld & Henry presentaron uno de los primeros marcos de este tipo, el que denominaron “análisis de estructura latente” en los años sesenta.
- Un siguiente paso imporante hacia la consolidación fue la introduccion de modelos de ecuaciones estructurales propuesto por Jöreskog en los años setenta, y la subsecuente expansión de este marco a variables categóricas por Muthén en los años ochenta.

Marcos de variables latentes "actuales"

- Estos marcos generales han sido a su vez expandido y reformulado en marcos expandidos como el de *modelos generalizados de ecuaciones estructurales* propuesto por Muthén (2002) y el marco de *modelos lineales latentes generalizados mixtos* de Skrondal y Rabe-Hesketh (GLLAMM por su sigla en inglés; 2004).
- Desde estos marcos, los modelos de variables latentes típicos son considerados como casos especiales, que pueden ser expandidos, combinados y adaptados dependiendo de las demandas específicas de cada problema.

Marcos generales y software estadístico

- Estos marcos han contribuido a la creación de software estadístico sumamente flexible, como por ejemplo Mplus (desarrollado por Muthén), el paquete GLLAMM en Stata (desarollado por Rabe-Hesketh y Skrondal) o LatentGold (desarrollado por Vermunt and Magidson).
- Asimismo, la comprención de estos modelos como tipos de modelos lineales generalizados permite el uso de lenguages estadísticos generales tales como R, Stata y SAS para la estimación de modelos a medida sin requerir de software especializado.

El marco GLLAMM (Skrondal and Rabe-Hesketh, 2007)

Generalized **L**inear **L**atent and **M**ixed **M**odels

Los modelos Lineales Latentes Generalizados Mixtos tienen dos componentes principales:

1. Modelo de respuesta
2. Modelo estructural

El modelo de respuesta

El modelo de respuesta corresponde a un modelo lineal generalizado donde:

$$\begin{aligned} y &= \text{es una respuesta} \\ \mathbf{x}, \mathbf{z} &= \text{son vectores de covariables} \\ \boldsymbol{\eta} &= \text{el vector de todas las variables latentes en el modelo} \end{aligned}$$

De forma tal que el valor esperado condicional de la respuesta y que sigue una distribucion condicional $f(\cdot)$, dado $\mathbf{x}$, $\mathbf{z}$ y $\boldsymbol{\eta}$, está "vinculado" con el predictor lineal ν a través de la función de vínculo $g(\cdot)$.

$$\text{Distribución condicional: } y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\eta} \sim f(\cdot)$$

$$\text{Función de vínculo (\textit{link function}): } g(\mathrm{E}[y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\eta}]) = \nu$$

El modelo de respuesta

Algunas combinaciones comunes de funciones de vínculo y distribuciones de probabilidad incluyen:

- El vínculo de identidad y la distribución normal para respuesta cuantitativas continuas (e.g., regresión múltiple común, HLM)
- El vínculo logit y la distribución binomial (e.g., IRT dicotómico)
- El vínculo adjacent category logit y la distribución multinomial (e.g., IRT crédito parcial)
- El vínculo logit multinomial y la distribución multinomial (e.g., Clases latentes)
- El vínculo log y la distribución de Poisson para frecuencias (e.g., Análisis de sobrevivencia)

El predictor lineal

La tercera y última parte del modelo de respuesta es el predictor lineal:

$$\text{Distribución condicional: } y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\eta} \sim f(\cdot)$$

$$\text{Función de vínculo (\textit{link function}): } g(E[y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\eta}]) = \nu$$

$$\textbf{Predictor lineal: } \nu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sum_{m=1}^{M} \eta_{mj}\mathbf{z}'_{mij}\boldsymbol{\lambda}_m$$

Donde:

- $\mathbf{x}'_{ij}$ son covariables asociadas a efectos fijos $\boldsymbol{\beta}$.
- η_{mj} es la variable latente número m en el modelo.
- $\mathbf{z}'_{mij}$ son covariables asociadas a la variable latente m.
- $\boldsymbol{\lambda}_m$ son parámetros asociados a la variable m, usualmente cargas (*loadings*).

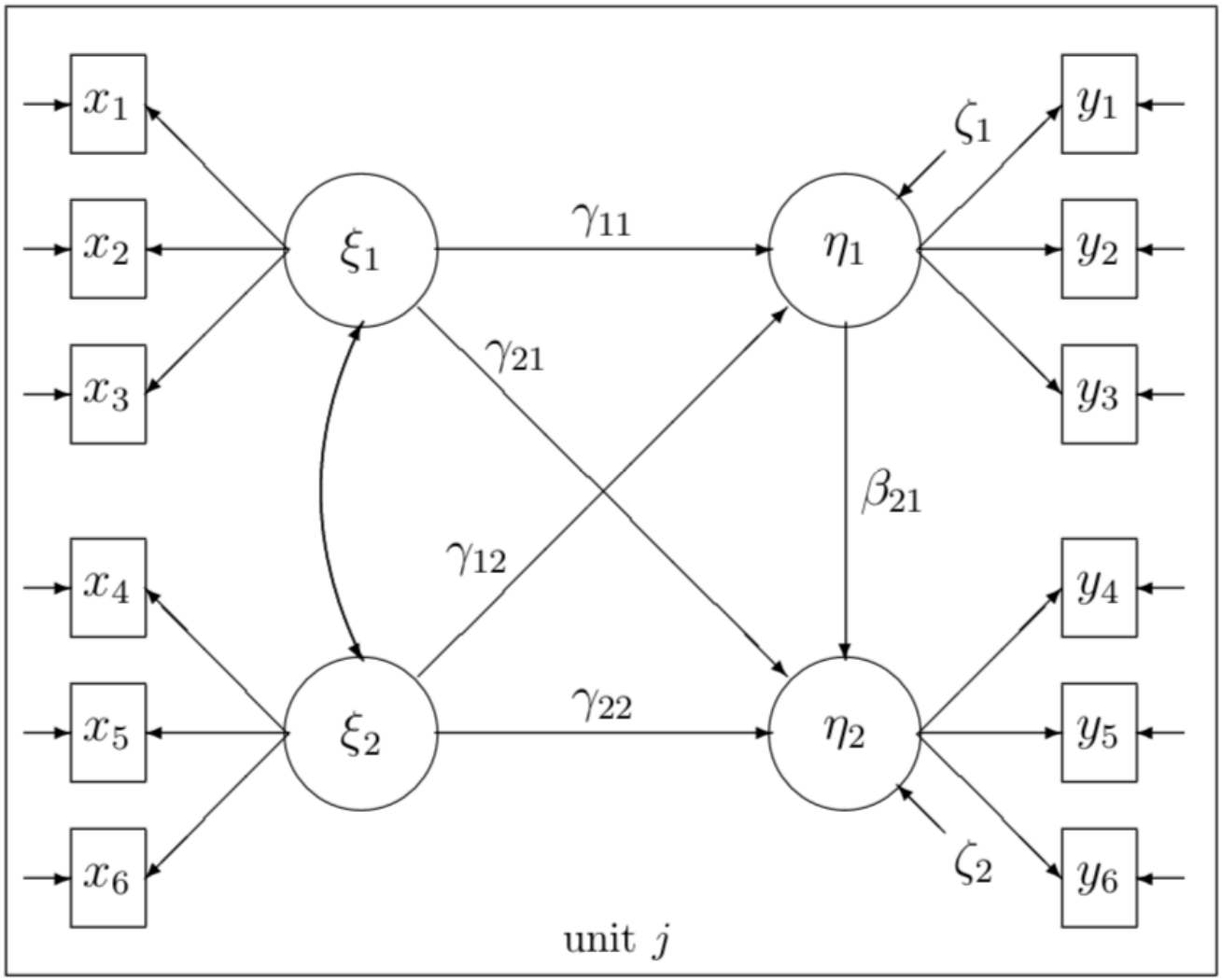
Modelo estructural

Variables latentes cuantitativas:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\mathbf{w} + \boldsymbol{\zeta}$$

Variables latentes categóricas:

$$\pi_{jc} = \frac{\exp(\mathbf{w}_j' \boldsymbol{\rho}^c)}{\sum_d \exp(\mathbf{w}_j' \boldsymbol{\rho}^d)}$$


x_1
x_2
x_3
x_4
x_5
x_6
ξ_1
ξ_2
γ_{11}
γ_{21}
γ_{12}
γ_{22}
η_1
η_2
ζ_1
ζ_2
β_{21}
y_1
y_2
y_3
y_4
y_5
y_6
unit j

¿Qué hemos visto de modelos de variables latentes?

- Los MVL son un tipo de modelos estadísticos que nos permiten hacer inferencias respecto a *una variable aleatoria no observada* a partir de las relaciones que podemos detectar entre variables observadas.
- Existen muchos tipos de modelos de variables latentes.
- Todos tienen a la base ciertos supuestos conceptuales: expresan una relacion probabilística y asumen cierta estructura de los atributos latentes.
- Todos comparten el supuesto estadístico de la independencia condicional.
- Todos pueden ser expresados utilizando gráficos acíclicos.
- Pueden ser entendidos como partes de grandes marcos analíticos tales como SEM generalizado o GLLAMM.

Referencias

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge; New York: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203-219. Borsboom, D. (2008). Latent variable theory.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological methods*, *5*(2), 155.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin Company.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Pub. Co.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A., & Rabe-Hesketh, S. O. P. H. I. A. (2007). Latent variable modelling: a survey. *Scandinavian Journal of Statistics*, *34*(4), 712-745.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Variables Latentes: Introducción y conceptos básicos

David Torres Irribarra

Escuela de Psicología
Pontificia Universidad Católica de Chile

30 de enero de 2018