¿Qué son las variables latentes?

Una interpretación geométrica

Psicometría y Variables Latentes

Ernesto San Martín

[1]Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Chile
[2]The Economics School of Louvain, Université catholique de Louvain, Belgium
[3]LIES Laboratorio Interdisciplinario de Estadística Social, Pontificia Universidad Católica de Chile, Chile

24 de enero de 2018

Laboratorio Interdisciplinario de Estadística Social

Outline

Teoría Clásica de Tests

- La manera estándar de presentar la TCT es la siguiente:

$$Y = \tau + \epsilon,$$

  donde τ, el puntaje verdadero, se asume no correlacionado con el error de medición.

- Además se asume que $E(Y) = E(\tau)$, proveyendo una interpretación frecuentista de este hecho.

- De lo anterior se deduce que

$$Var(Y) = Var(\tau) + Var(\epsilon),$$

  lo cual permite definir la confiabilidad como $\frac{Var(\tau)}{Var(Y)}$.

Teoría Clásica de Tests

- Este marco se utiliza en protocolos de análisis (SIMCE, PISA, TIMMS).
- Se enfatiza con mucha fuerza que dicho marco se ha superado por la moderna Teoría de Respuesta al Item.
- Dicha superación siempre se ilustra/justifica a nivel de estimaciones estadística, y muy poco a nivel conceptual.

Teoría Clásica de Tests

- La TCT asume dos procesos aleatorios secuenciales:
  - La selección de los individuos o estudiantes.
  - La selección de los ítemes que serán respondidos por un individuo.
- Los procesos secuenciales anteriores corresponden a una descomposición marginal-condicional:
  - El proceso marginal corresponde a la generación de una característica individual.
  - El proceso condicional corresponde a la generación de la respuesta al item condicionalmente a la característica individual.
- **El supuesto sustantivo es que la respuesta a los ítemes depende de la característica individual**.

Teoría Clásica de Test y Espacios de Hilbert

- Siguiendo a Novick (1969) y otros, trabajaremos en un espacio que tenga geometría, es decir, que nos permita medir ángulos, y que por tanto el Teorema de Pitágoras sea válido.
- Más específicamente, trabajaremos en el siguiente espacio:

$$L^2(M, \mathcal{M}, P) = \{X \text{ variable aleatoria } \; : \; E(X^2) < \infty\}.$$

- Este es un Espacio de Hilbert, en que los ángulos se miden por medio de un producto interno definido como

$$(X, Y) \doteq E(XY).$$

Teoría Clásica de Test

- HIPÓTESIS 1: La respuesta de un determinado individuo a un determinado item, denotada por Y, mide un rasgo latente θ que caracteriza a dicho individuo.
- El rasgo latente θ puede ser unidimensional o muitidimensional Sin embargo, la TCC asume que θ es **unidimensional**.
- Más formalmente, el proceso conjunto $p(Y, \theta)$ se descompone como

$$p(Y, \theta) \;=\; p(\theta)\, p(Y \mid \theta).$$

Teoría Clásica de Test

Laboratorio Interdisciplinario de Estadística Social

- Variables latentes: una hipótesis que encierra mucha discusión fundamental.
- Por ejemplo, ¿se trata de meros constructos o de entes reales?
- Hay ciertas decisiones de política pública que se basan en estimaciones estadísticas de las variables latentes.

Teoría Clásica de Test

- Sea Y el puntaje obtenido por un individuo que responde un (grupo de) item(es), por lo que sus posibles valores son $0, 1, 2, \ldots, J$.
- Dada la Hipótesis 1, el factor latente θ explica la generación de la respuesta Y, o equivalentemente, Y mide el factor latente θ.
- Esa medición no sólo contiene información de θ, sino eventualmente error de medición:

$$Y = \underbrace{E(Y \mid \theta)}_{\text{Puntaje Verdadero}} + \underbrace{[Y - E(Y \mid \theta)]}_{\text{Error de Medición}}.$$

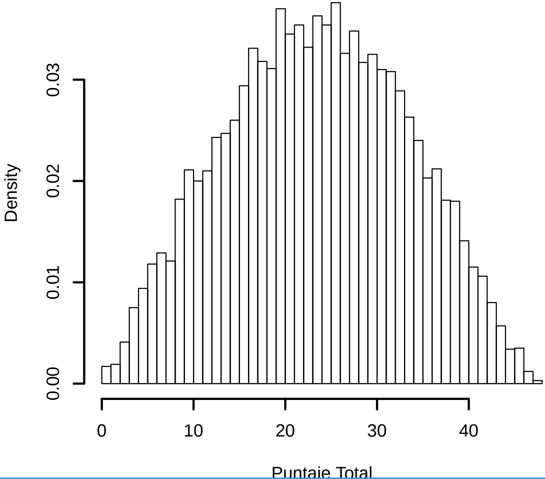
- Esta descomposición es válida aún si θ es multidimensional.
- Además, $E(Y \mid \theta)$ está no correlacionado con $Y - E(Y \mid \theta)$ por construcción.

Puntaje observado – Puntaje verdadero

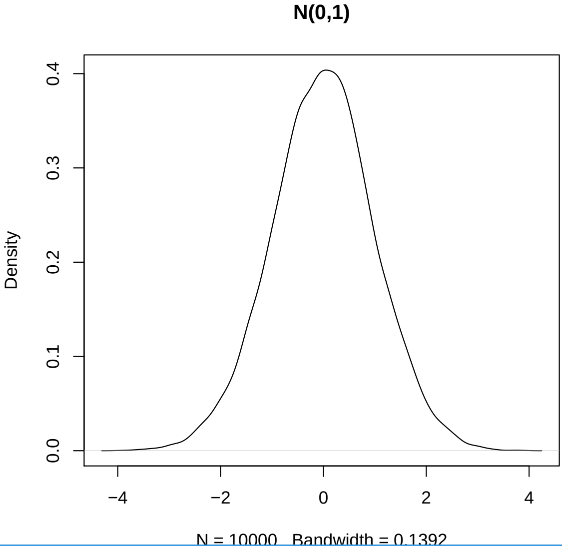
Laboratorio Interdisciplinario de Estadística Social

- El desafío de la TCC es que, bajo ciertas hipótesis, se infiera información del puntaje verdadero (y, en consecuencia, de θ) **a partir del puntaje verdadero**.
- Dicho de otra manera, se busca inducir la distribución de θ a partir del puntaje observado Y.
- Para entrever la dificultad del problema, consideremos los puntajes observados de 10000 estudiantes que responden 48 ítemes dicotómicos.

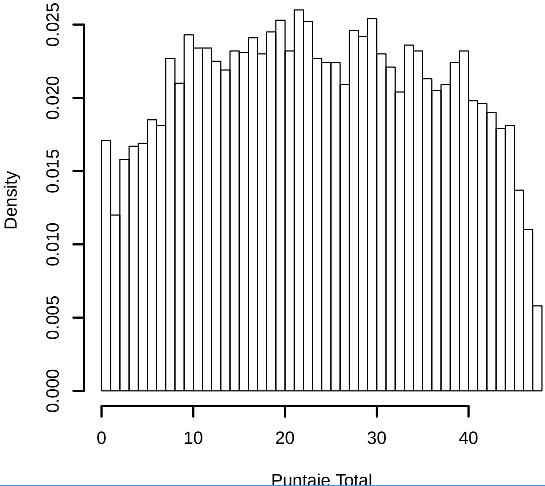
Ejemplo 1: Distribución de Puntajes Observados



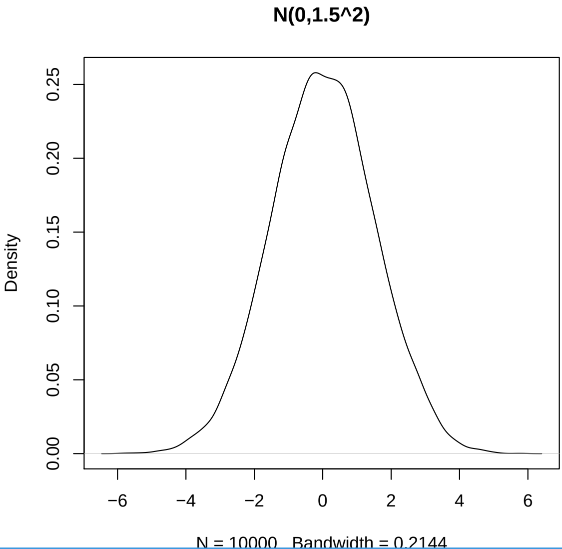
Ejemplo 1: Distribución de Puntajes Verdaderos



Ejemplo 2: Distribución de Puntajes Observados



Ejemplo 2: Distribución de Puntajes Verdaderos



Teoría Clásica de Test

- **Descomposición de la Varianza:** ¿cuánto de la variabilidad observada depende de la variabilidad del puntaje verdadero?
- Usando la sintaxis de varianzas y esperanzas condicionales, se tiene que

$$\sigma_Y^2 \doteq V(Y) \quad = \quad V[E(Y \mid \theta)] + E[V(Y \mid \theta)].$$

- La varianza del puntaje verdadero $V[E(Y \mid \theta)]$ corresponde a la **varianza intra-clase**.
- La varianza del error de medición $E[Var(Y \mid \theta)]$ corresponde a la **varianza entre-clases**.
- Si la variabilidad correspondiente al error de medición es "pequeña", entonces la variabilidad observada proporciona información acerca de la variabilidad del puntaje verdadero.

Teoría Clásica de Test

Laboratorio Interdisciplinario de Estadística Social

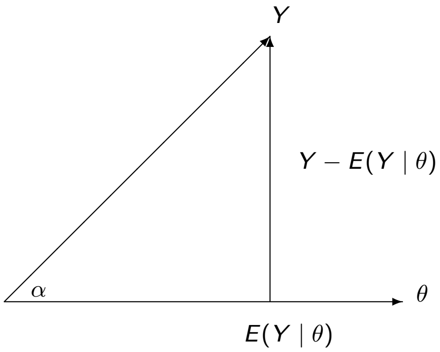
- **Suponiendo que θ es unidimensional**, la afirmación anterior se puede precisar definiendo la **confiabilidad** como la siguiente razón:

$$\eta_{Y|\theta} \doteq \frac{V[E(Y \mid \theta)]}{V(Y)} \doteq \frac{\text{Varianza del Puntaje Verdadero}}{\text{Varianza del Puntaje Observado}} \equiv \frac{\sigma_T^2(\theta)}{\sigma_X^2}.$$

- $0 \leq \eta_{Y|\theta} \leq 1$, es decir, corresponde $\cos^2(\alpha)$, donde α es al ángulo de la proyección ortogonal de Y sobre el espacio generado por θ.

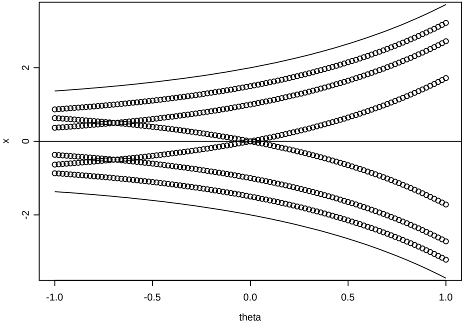
Teoría Clásica de Test

- $\eta^2_{Y|\theta} = \cos^2(\alpha) = \frac{\|E(Y|\theta)\|^2}{\|Y\|^2} = \frac{V[E(Y|\theta))]}{V(Y)}$, donde $\|\cdot\|$ representa el largo de un vector, y



Teoría Clásica de Test

- Si $Y \perp\!\!\!\perp \theta$ entonces $\eta_{Y|\theta} = 0$. Por tanto, $\eta_{Y|\theta} \neq 0$ implica que Y **no** es independiente de θ.
- $\eta_{Y|\theta} = 0$ no implica $Y \perp\!\!\!\perp \theta$ pues, por ejemplo, puede existir la siguiente relación de dependencia:



Teoría Clásica de Test

Laboratorio Interdisciplinario de Estadística Social

- $\eta_{Y|\theta} = 1$ es equivalente a

$$\frac{E[V(Y \mid \theta)]}{V(Y)} = 0,$$

  es decir, la variabilidad del error de medición es nula. Esto a su vez es equivalente a $V(Y \mid \theta) = 0$. Esto último significa que existe una función f tal que $Y = f(\theta)$ con probabilidad 1.

- Por lo tanto, que la confiabilidad $\eta_{Y|\theta}$ sea igual a 1 significa que:
  - No hay error de medición.
  - El puntaje observado Y es una función del rasgo latente θ; es decir, lo observable depende enteramente de lo no-observable.

¿Qué podemos decir de θ?

- Una vez que hemos observado Y, es posible preguntarse (en términos conceptuales) acerca del rasgo latente θ. Siguiendo con el marco conceptual inicial, esto significa explicar θ a partir de Y.

- Por lo tanto, podemos considerar la siguiente descomposición dual:

$$\theta = E(\theta \mid Y) + \{\theta - E(Y \mid \theta)\},$$

  donde $E(\theta \mid Y)$ corresponde a la predicción de θ después de haber observado Y, mientras que $\theta - E(Y \mid \theta)$ corresponde al error de medición.

- Por construcción,

$$E(\theta \mid Y) \perp \{\theta - E(Y \mid \theta)\}.$$

El lado dual de la TCC

- En la jerga psicométrica, $E(\theta \mid Y)$ se conoce como **estimador de Bayes** y se utiliza para "estimar" θ.

- En la TCC no es tradición "estimar" θ, sino sólo la confiabilidad ... que es una función de θ.

El lado dual de la TCC

- Por analogía con el tratamiento estándar, podemos definir la confiabilidad $\eta_{\theta|Y}$, que llamaremos **confiabilidad dual**.
- Así,

$$\eta_{\theta|Y} = \frac{V[E(\theta \mid Y)]}{V(\theta)} \in [0,1].$$

- Si $Y \perp\!\!\!\perp \theta$, entonces $\eta_{\theta|Y} = 0$.
- $\eta_{\theta|Y} = 1$ si y sólo si existe una función g tal que $\theta = g(Y)$ con probabilidad 1, lo que significa que el rasgo latente es una función de lo observable!

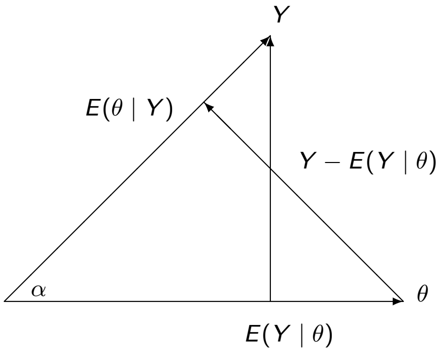
El lado dual de la TCC

Laboratorio Interdisciplinario de Estadística Social

- Pero

$$\cos^2(\alpha) = \eta_{Y|\theta} = \frac{V[E(Y \mid \theta)]}{V(Y)} = \frac{V[E(\theta \mid Y)]}{V(\theta)} = \eta_{\theta|Y}$$

pues



El lado dual de la TCC

- Por lo tanto,

$$\begin{array}{ccc} \eta_{Y|\theta} = 1 & \Longleftrightarrow & \eta_{\theta|Y} = 1 \\ \Updownarrow & & \Updownarrow \\ \text{Existe } f \text{ tal que } Y = f(\theta) & & \text{Existe } g \text{ tal que } \theta = g(Y) \end{array}$$

- Por lo tanto, cuando la confiabilidad es perfecta, el carácter latente θ es una función **desconocida** de lo observable.
- Este resultado ayuda a entender el alcance de la expresión **latente o no-observable**.

El lado dual de la TCC

Laboratorio Interdisciplinario de Estadística Social

- De la igualdad anterior, podemos probar el siguiente resultado:

$$\eta_{Y_1,Y_2|\theta} \geq \eta_{Y_1|\theta}$$

- Es decir, si la predicción de la variable latente está basada sobre más mediciones, entonces la confiabilidad dual mejora.
- Este resultado es una versión menos restrictiva de la Profecía de Spearman-Browne.

Sobre la naturaleza de θ

- Lo anterior permite probar lo siguiente: si la confiabilidad es perfecta entonces
  - Existe una función f tal que $Y = f(\theta)$ con probabilidad 1.
  - Existe una función g tal que $\theta = g(Y)$ con probabilidad 1.
- Es decir, lo latente no es otra cosa que una función desconocida de los observables.

La geometría del AIL

- Consideremos dos observables tales que

$$Y_1 \perp Y_2 \mid \theta,$$

que significa que

$$Y_1 - E(Y_1 \mid \theta) \perp Y_2 - E(Y_2 \mid \theta),$$

que corresponde a la correlación entre Y_1 e Y_2 parcializada por θ.

La geometría del AIL

Laboratorio Interdisciplinario de Estadística Social

- Es importante recordar que siempre es válida la siguiente descomposición:

$$cov(X, Y) = cov[E(X \mid \theta), E(Y \mid \theta)] + cov[X - E(X \mid \theta), Y - E(Y \mid \theta)]$$

para cualquier θ.

- Pero bajo AIL se tiene que

$$cov[X - E(X \mid \theta), Y - E(Y \mid \theta)] = 0$$

y por tanto

$$cov(X, Y) = cov[E(X \mid \theta), E(Y \mid \theta)]$$

La geometría del AIL

- Convención: las variables que usamos generan espacios lineales que serán denotados por la misma variable.
- Así, si Y es una variable tal que $E(Y^2) < \infty$, entonces Y corresponde al espacio lineal generado por Y.

Ortogonalidad minimal

- Decimos que θ_0 es la variable más pequeña que hace ortogonales a Y_1 e Y_2 si
  - θ_0 satisface que $Y_1 \perp Y_2 \mid \theta_0$.
  - Si existe un θ tal que $Y_1 \perp Y_2 \mid \theta$ entonces $\theta \supset \theta_0$.

La geometría del AIL

Laboratorio Interdisciplinario de Estadística Social

- Usando todo lo anterior, tenemos la siguiente caracterización: las siguientes proposiciones son equivalentes,
  - θ_0 es la variable más pequeña que hace ortogonales a Y_1 e Y_2.
  - $E(Y_i \mid \theta_0) = \theta_0$ para $i = 1, 2$.
  - $Y_i^{\perp} \cap \theta_0 = \{0\}$ con $i = 1, 2$.

Discusión

- Uso de espacios de Hilbert.
- Latente $\equiv$ funciones desconocidas de los observables.
- Geometría del AIL.