Equiparación de Puntajes para el Caso de Grupos No Equivalentes e Itemes Comunes:

Un Análisis de Identificación Parcial

Ernesto San Martín

Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Chile

The Economics School of Louvain, Université catholique de Louvain, Belgium

LIES Laboratorio de Investigación en Estadística Social, Pontificia Universidad Católica de Chile, Chile

Seminario Educación, Facultad de Matemáticas, UC

Trabajo conjunto con Jorge González

NEAT Design

- Dos poblaciones de estudiantes: cada uno escoge rendir o un test X o un test Y.
- Además, ambos grupos rinden una batería de ítemes comunes A.
- Problema: poner en la misma escala los puntajes de ambas poblaciones de estudiantes.

Identificación Puntual

- Primero, etiquetamos a los estudiantes de acuerdo a la prueba que rindieron:

$$Z = \begin{cases} 1, & \text{si estudiante rinde prueba } X; \\ 0, & \text{si estudiante rinde prueba } Y. \end{cases}$$

- Z es una variable aleatoria (por qué?), de hecho es una Bernoulli de parámetro

$$P(Z = 1),$$

  donde esta probabilidad está definida según una proporción, que satisface los axiomas de Kolmogorov (ver Kolmogorov, 1955, capítulo 1).

Identificación Puntual

- Usando el Teorema de Probabilidades Totales se tiene que

$$P(X \leq t \mid A) = P(X \leq t \mid Z = 1, A)P(Z = 1 \mid A) + P(X \leq t \mid Z = 0, A)P(Z = 0 \mid A)$$

$$P(Y \leq t \mid A) = P(Y \leq t \mid Z = 1, A)P(Z = 1 \mid A) + P(Y \leq t \mid Z = 0, A)P(Z = 0 \mid A)$$

- En rojo están las distribuciones no-identificadas, y en azul las que son identificadas.
- Condición de ignorabilidad fuerte:

$$(X, Y) \perp\!\!\!\perp Z \mid A.$$

Identificación Puntual

- La condición de ignorabilidad fuerte es equivalente a

$$P(X \leq t \mid A, Z = 1) = P(X \leq t \mid A, Z = 0) = P(X \leq t \mid A),$$

$$P(Y \leq t \mid A, Z = 1) = P(Y \leq t \mid A, Z = 0) = P(Y \leq t \mid A),$$

- o equivalentemente

$$P(Z = 1 \mid A) = P(Z = 0 \mid A).$$

- La condición de ignorabilidad permite afirmar que, condicionalmente a A, lo que se observa es igual a lo que no se observa (sic).

Identificación Puntual

- Dado que hemos identificado $P(X \leq t \mid A)$ y $P(Y \leq t \mid A)$, entonces podemos identificar

$$P(X \leq t), \qquad P(Y \leq t)$$

pues conocemos $P(A \leq a)$.

De vuelta al problema

- Tenemos que

$$P(X \leq t) = P(X \leq t \mid Z = 1)P(Z = 1) + P(X \leq t \mid Z = 0)P(Z = 0).$$

$$P(Y \leq t) = P(Y \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0).$$

- No tenemos información de por qué un estudiante escoge rendir la prueba X o la prueba Y. Solo sabemos que en algunos contextos, como el chileno, esta decisión es voluntaria.

Modelando la voluntariedad

- Podemos suponer/creer que los estudiantes rinden la prueba que ellos estiman les irá mejor:

$$Z = \mathbb{1}_{\{X>Y\}}.$$

- Así, $Z = 1$ significa que un estudiante escogió la prueba X en lugar de la Y porque asume que le irá mejor en X que en Y.
- Similarmente, $Z = 0$ significa que un estudiante escogió la prueba Y en lugar de la X porque asume que le irá mejor en Y que en X.

Modelando la voluntariedad

- Más específicamente, si $Z = 1$ entonces $X > Y$ y por tanto

$$P(X \leq t \mid Z = 1) \leq P(Y \leq t \mid Z = 1),$$

o equivalentemente

$$P(X > t \mid Z = 1) \geq P(Y > t \mid Z = 1).$$

- Similarmente, si $Z = 0$ entonces $X < Y$ y por tanto

$$P(Y \leq t \mid Z = 0) \leq P(X \leq t \mid Z = 0),$$

o equivalentemente

$$P(Y > t \mid Z = 0) \geq P(X > t \mid Z = 0).$$

Resultado 1 de identificación parcial

- Usando lo anterior se deduce que

$$L_1^y \leq P(Y \leq t) \leq U_1^y,$$

donde

$$L_1^y = P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0),$$

$$U_1^y = P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0).$$

- L_1 corresponde a la proporción de estudiantes que tienen a lo más t puntos ya sea en X o en Y: es correcta esta interpretación?
- El largo de este intervalo es igual a

$$P(X > t \mid Z = 1)P(Z = 1).$$

Resultado 1 de identificación parcial

- De manera similar,

$$L_1^x \leq P(X \leq t) \leq U_1^x,$$

donde

$$L_1^x = P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0),$$

$$U_1^x = P(Z = 0) + P(X \leq t \mid Z = 1)P(Z = 1).$$

Resumen

- Se tiene que

Hyp	$P(X \leq t)$		$P(Y \leq t)$	
	L	U	L	U
$Z = \mathbf{1}_{\{X>Y\}}$	$\alpha(t)\omega + \beta(t)(1-\omega)$	$\alpha(t)\omega + (1-\omega)$	$\alpha(t)\omega + \beta(t)(1-\omega)$	$\beta(t)(1-\omega) + \omega$

- donde

$$\alpha(t) = P(X \leq t \mid Z = 1), \quad \beta(t) = P(Y \leq t \mid Z = 0), \quad \omega = P(Z = 1).$$

Resumen

- Ambos intervalos tienen la misma cota inferior.
- El primer intervalo está contenido en el segundo si y sólo si

$$\frac{P(X > t \mid Z = 1)}{P(Y > t \mid Z = 0)} > \frac{P(Z = 0)}{P(Z = 1)}.$$

Cuantiles

- Para $\alpha \in [0, 1]$, el α-cuantil de Y está dado por

$$q_\alpha(Y) = \inf\{t : P(Y \leq t) \geq \alpha\}.$$

- Consideremos el intervalo

$$L_1^y \leq P(Y \leq t) \leq U_1^y,$$

donde

$$L_1^y = P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0),$$

$$U_1^y = P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0).$$

Cuantiles

- Sea

$$r_1(\alpha) \doteq \inf\{t : P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0) \geq \alpha\},$$

$$s(\alpha) \doteq \{t : P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0) \geq \alpha\}.$$

- Entonces si $t < r_1(\alpha)$, entonces

$$P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0) < \alpha$$

y por tanto

$$P(Y \leq t) \leq \alpha,$$

en consecuencia $q_\alpha(Y) > t$. Se sigue que

$$r_1(\alpha) \leq q_\alpha(Y).$$

Cuantiles

- De manera similar se deduce que

$$q_\alpha(Y) \leq s(\alpha).$$

- Por lo tanto, para $\alpha \in (0,1)$,

$$r_1(\alpha) \leq q_\alpha(Y) \leq s(\alpha),$$

$$r_2(\alpha) \leq q_\alpha(X) \leq s(\alpha),$$

donde

$$r_2(\alpha) \doteq \inf\{t : P(X \leq t \mid Z = 1)P(Z = 1) + P(Z = 0) \geq \alpha\}.$$

Cuantiles

- Más precisamente,

$$q_{\frac{\alpha - P(Z=1)}{P(Z=0)}}(Y \mid Z = 0) \leq q_\alpha(Y) \leq s(\alpha),$$

$$q_{\frac{\alpha - P(Z=0)}{P(Z=1)}}(X \mid Z = 1) \leq q_\alpha(X) \leq s(\alpha),$$

  donde

$$s(\alpha) = \inf\{t : P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0) \geq \alpha\}.$$

Cuantiles

- Supongamos que $P(Z = 0) < P(Z = 1)$. Entonces
- Si $\alpha < P(Z = 0)$, entonces

$$q_{\frac{\alpha - P(Z=1)}{P(Z=0)}}(Y \mid Z = 0) = t_{min}^{Y|Z=0},$$

$$q_{\frac{\alpha - P(Z=0)}{P(Z=1)}}(X \mid Z = 1) = t_{min}^{X|Z=0},$$

y por lo tanto los intervalos

$$\left[t_{min}^{Y|Z=0}, s(\alpha)\right], \qquad \left[t_{min}^{X|Z=0}, s(\alpha)\right]$$

son equivalentes!

Cuantiles

- Si $P(Z = 0) < \alpha < P(Z = 1)$, entonces

$$q_{\frac{\alpha - P(Z=1)}{P(Z=0)}}(Y \mid Z = 0) = t_{min}^{Y|Z=0},$$

y los intervalos

$$\left[t_{min}^{Y|Z=0}, s(\alpha)\right], \qquad \left[q_{\frac{\alpha - P(Z=0)}{P(Z=1)}}(X \mid Z = 1), s(\alpha)\right]$$

son equivalentes.

Cuantiles

- Si $P(Z=0) < P(Z=1) < \alpha$, entonces los intervalos

$$\left[q_{\frac{\alpha - P(Z=1)}{P(Z=0)}}(Y \mid Z=0), s(\alpha) \right], \quad \left[q_{\frac{\alpha - P(Z=0)}{P(Z=1)}}(X \mid Z=1), s(\alpha) \right]$$

son equivalentes.

Problema

- ¿Cómo estudiar

$$s(\alpha) = \inf\{t : P(X \leq t \mid Z = 1)P(Z = 1) + P(Y \leq t \mid Z = 0)P(Z = 0) \geq \alpha\}$$

en función de α y sus relaciones con $P(Z = 0)$ y $P(Z = 1)$?