Polo del Conocimiento

Pol. Con. (Edición núm. 92) Vol. 9, No 5
Mayo 2024, pp. 729-742
ISSN: 2550 - 682X
DOI: https://doi.org/10.23857/pc.v9i5.7153

Equiparación de puntuaciones mediante máxima verosimilitud y valores plausibles

Equation of scores using maximum likelihood and plausible values

Equação de pontuações usando máxima verossimilhança e valores plausíveis

Héctor Salomón Mullo-Guaminga [I]
hmullo@espoch.edu.ec
https://orcid.org/0000-0001-8448-4652

Nathaly Geovanna Paredes-Ayala [II]
nathaly.paredes@espoch.edu.ec
https://orcid.org/0009-0002-4187-7557

David Vitelio Ayala-Cárdenas [III]
vitelio.ayala@espoch.edu.ec
https://orcid.org/0009-0006-6964-7779

Valeryn Anahí Belalcázar-Loor [IV]
valeryn.belalcazar@espoch.edu.ec
https://orcid.org/0009-0004-7730-1538

Correspondencia: hmullo@espoch.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

*** Recibido:** 10 de marzo de 2024 ***Aceptado:** 20 de abril de 2024 *** Publicado:** 09 de mayo de 2024

I. Ingeniero en Estadística Informática, Máster Universitario en Estadística Aplicada, Doctor en Estadística Matemática y Aplicada, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
II. Estudiante de Ingeniería en Estadística en la Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
III. Estudiante de Ingeniería en Estadística en la Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
IV. Estudiante de Ingeniería en Estadística en la Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.

Resumen

La investigación se centra en desarrollar un proceso de simulación y estimación de la habilidad en datos para dos formas de prueba que se ajusten al modelo de teoría de respuesta al ítem de Rasch. Se busca minimizar el error cuadrado medio en la estimación de la habilidad de sustentantes de la prueba. Se considera aspectos cruciales como el porcentaje de ítems ancla, la distribución de las dificultades de los ítems y el sesgo de la distribución. Utilizando estos datos, se estima la habilidad mediante la máxima verosimilitud y valores plausibles. Un total de 768 escenarios, evaluando el error absoluto medio y el error cuadrático. Los resultados obtenidos muestran 15 escenarios que minimizaron las medidas de error, en general, esto ocurrió cuando la distribución de la dificultad de los ítems es normal estándar, además, cuando la estimación de las habilidades se desarrolla en un ambiente de desconocimiento de las dificultades de los ítems. La conclusión más importante es que el mejor método de estimación de la habilidad es el de máxima verosimilitud, donde las dificultades de los ítems siguen una distribución normal.

Palabras Clave: Equiparación; Psicometría; Valores plausibles; Teoría de Respuesta al Ítem.

Abstract

The research focuses on developing a process of simulation and estimation of ability in data for two forms of testing that fit the Rasch item response theory model. The aim is to minimize the mean square error in estimating the ability of test takers. Crucial aspects such as the percentage of anchor items, the distribution of item difficulties and the skewness of the distribution are considered. Using this data, skill is estimated using maximum likelihood and plausible values. A total of 768 scenarios, evaluating the mean absolute error and the squared error. The results obtained show 15 scenarios that minimized the error measurements, in general, this occurred when the distribution of the difficulty of the items is standard normal, in addition, when the estimation of the skills is developed in an environment of ignorance of the difficulties of the items. The most important conclusion is that the best method for estimating ability is maximum likelihood, where item difficulties follow a normal distribution.

Keywords: Equation; Psychometry; Plausible values; Item Response Theory.

Resumo

A pesquisa se concentra no desenvolvimento de um processo de simulação e estimativa de habilidade em dados para duas formas de testes que se enquadram no modelo da teoria de resposta ao item de Rasch. O objetivo é minimizar o erro quadrático médio na estimativa da habilidade dos candidatos. São considerados aspectos cruciais como a percentagem de itens âncora, a distribuição das dificuldades dos itens e a assimetria da distribuição. Usando esses dados, a habilidade é estimada usando máxima verossimilhança e valores plausíveis. Um total de 768 cenários, avaliando o erro médio absoluto e o erro quadrático. Os resultados obtidos mostram 15 cenários que minimizaram as medidas de erros, em geral, isso ocorreu quando a distribuição da dificuldade dos itens é padrão normal, além disso, quando a estimativa das habilidades é desenvolvida em um ambiente de desconhecimento das dificuldades de os itens. A conclusão mais importante é que o melhor método para estimar a habilidade é a máxima verossimilhança, onde as dificuldades dos itens seguem uma distribuição normal.

Palavras-chave: Equação; Psicometria; Valores plausíveis; Teoria de Resposta ao Item.

Introducción

Las evaluaciones de aprendizaje a gran escala forman parte del campo teórico de la psicometría, específicamente se enmarcan en las teorías de los test, los cuales construyen modelos teóricos y metodológicos (Muñiz, 2010). El primer modelo que se formuló se conoce como la Teoría Clásica de los Test (TCT), luego surgió nuevas teorías y técnicas de medición que han superado la perspectiva clásica, si bien existen varios modelos, la Teoría de Respuesta al Ítem (TRI) es la más reconocida (Attorresi et al., 2009). La TRI agrupa varias líneas de investigación psicométricas, donde el factor común de estos desarrollos es que "establecen una relación entre el comportamiento de un sujeto frente a un ítem y el rasgo responsable de esta conducta (rasgo latente)" (Attorresi et al., 2009, p. 180). El análisis de las respuestas en una prueba que propone la TRI es radicalmente diferente a la TCT, pues se enfoca "en los componentes constituyentes de la misma (es decir, los ítems) en vez del resultado global de la medición" (Leenen, 2014, p. 41).

Dentro de los modelos de la TRI, se encuentra el modelo (Rasch 1960) que utiliza la siguiente función para modelar la probabilidad de que un sustentante con habilidad θ, responda correctamente un ítem con dificultad δ

$$Prob(Y = 1) = \frac{exp(\theta - \delta)}{1 + exp(\theta - \delta)},$$

donde Y es la puntuación del sustentante que toma valores de 0 (respuesta incorrecta) o 1 (respuesta correcta).

La TRI es apropiada para analizar instrumentos como las pruebas estandarizadas. Para ello, es importante definir el rasgo latente que subyace en la prueba y que se intenta estimar. El rasgo latente es un constructo teórico de carácter cognitivo, procedimental o actitudinal que no puede ser medido directamente debido a que no es observable explícitamente, el cual se estima a través de los indicadores que conforman un instrumento de evaluación.

En las evaluaciones estandarizadas a gran escala se aplican instrumentos de evaluación a los sustentantes de poblaciones de interés. En estas a menudo es necesario comparar los resultados de los sustentantes en dos formas de prueba que evalúan el mismo constructo, esto se puede realizar mediante una equiparación entre las dos formas de prueba utilizando TRI cuando los grupos evaluados de cada forma no son equivalentes (es decir, los grupos son de diferentes poblaciones) y se cuenta con covariables (variables de estratificación utilizadas en el proceso de muestreo o diseño muestral) e ítems anclas (es decir, ítems que son comunes en las dos formas de prueba). En el proceso de equiparación de puntuaciones en evaluaciones, las diferencias de las puntuaciones se deben a: i) diferencias en la dificultad de las pruebas de cada ciclo y ii) diferencias en las habilidades de los sustentantes de cada ciclo. Una definición formal de equiparación dado por Braun y Holland (1982) es la siguiente:

> Sean R y S dos pruebas que generan ambos datos de puntuación X e Y, respectivamente. Se dice que X e Y se equiparan en la población T por $\varphi : X \mapsto Y$ (transformación de equiparación) si $F_Y(y) = F_{\varphi(x)}(y)$, donde F_Y y $F_{\varphi(x)}$ representan la función de distribución de Y y $\varphi(x)$ respectivamente.

Para la estimación de las puntuaciones (o habilidad) se puede realizar mediante el método de máxima verosimilitud ponderada o utilizando valores plausibles. Esto cuando se tiene dos formas de prueba aplicados a dos grupos no equivalentes, donde se cuenta con ítems ancla entre formas (Grupo Ítems I_2). La siguiente tabla ejemplifica lo anterior

Tabla 1: *Diseño de prueba con matriz incompleta*

	Grupo Ítems I_1	Grupo Ítems I_2	Grupo Ítems I_3
Grupo sustentantes 1	I_1	I_2	
Grupo sustentantes 2		I_2	I_3

Elaborado por: *Dirección de Análisis Psicométrico*
Fuente: *Fuentealba, 2020*

La matriz incompleta se utiliza para cubrir la longitud de la estructura de evaluación, este es un tema importante para tener en cuenta en el proceso de equiparación de dos formas de prueba. En el mismo sentido, la utilización de máxima verosimilitud o valores plausibles en el paquete TAM es un tema de interés que se relaciona directamente con la utilización de una matriz incompleta para cubrir la longitud de la estructura de evaluación para un mismo constructo.

Según Wu (2022) los valores plausibles son sorteos aleatorios de la distribución posterior, donde la distribución posterior es aquella que se obtiene después de que los sustentantes rindieron la evaluación, mediante la información de la distribución anterior (distribución de los sustentantes en el rango de la habilidad) y los resultados de la evaluación combinados.

En este sentido el objetivo del presente trabajo de investigación es determinar el mejor método de estimación de la habilidad entre la máxima verosimilitud y valores plausibles en el contexto de una equiparación de dos formas de prueba aplicados a grupos de prueba independientes. Se buscará el mejor método mediante el cálculo del error absoluto medio y el error cuadrático medio.

Metodología

Para alcanzar el objetivo de la investigación se simula datos de respuesta al ítem que se ajustan al modelo de Rasch. Estos datos son matrices una por forma de prueba con 1000 sustentantes y 100 ítems, se considera una tasa de no respuesta de ítems del 4.8%. La simulación de los datos se desarrolla en el software estadístico R mediante el paquete “MIRT” desarrollado por Chalmers et al. (2023). La equiparación mediante máxima verosimilitud [(Myung, 2003), (Eliason, 1993), (Hambleton et. al 1985) y (Banerjee, 2008)] y valores plausibles (Thompson, 2009), y la comparación de los métodos se realiza mediante el paquete “TAM” de autoría de Kiefer, Robitzsch y Wu (2016).

Simulación de Datos Rasch: En este punto se considerar tres aspectos importantes en el diseño de pruebas como son: i) el porcentaje de ítems ancla en relación con número total de ítems del

instrumento de evaluación, ii) la distribución de probabilidad de las dificultades de los ítems, y iii) el sesgo de la distribución.

Estimación de las habilidades de los sustentantes: Para este punto se toma en cuenta tres aspectos que son: i) el número de valores plausibles extraídos de la distribución posterior de las habilidades, ii) se considera modelos con dificultades fijas (es decir conocidas en la simulación de los datos) y modelos con dificultades no conocidas, y iii) el rango de la distribución de la habilidad, para la extracción de los valores plausibles.

En la simulación y estimación de la información se hizo fluctuar los parámetros descritos en los dos párrafos anteriores con los siguientes valores:

Simulación de Datos Rasch:

- Porcentaje de ítems ancla: $\{10\%, 20\%, 30\%\}$.
- Distribución de probabilidad de las dificultades de los ítems: $N(0,1)$ y $U(-3,3)$.
- Sesgo de la distribución: Para la distribución normal estándar se considerar un parámetro de sesgo de $\{0, 0.3, 0.6, 0.9\}$.

Estimación de las habilidades de los sustentantes:

- Número de valores plausibles extraídos: $\{1, 3, 5, 7\}$.
- Modelo TRI con dificultades: $\{conocidas, desconocidas\}$.
- Rango de la distribución de la habilidad, para la extracción de los valores plausibles: $\{0.5, 1.2, 4, 5\}$.

A partir de los parámetros anteriores se establecieron 768 escenarios, donde se simularon datos Rasch (dos formas de prueba) y se estimó la habilidad mediante máxima verosimilitud y valores plausibles, considerando una calibración concurrente (en una sola matriz), es decir, se calibró a partir del diseño de prueba con matriz incompleta (ver Tabla 1.). Se comparó el ajuste de los modelos mediante el error cuadrático medio y la desviación media absoluta, posteriormente se toma los primeros 10 mejores ajustes para la discusión de los resultados.

Resultados

De la simulación y estimación de las habilidades en los 768 escenarios se obtuvo el error cuadrático medio y la desviación media absoluta para la estimación de la habilidad mediante

máxima verosimilitud y valores plausibles versus las habilidades iniciales. A continuación (ver Tabla 2.) se presenta los 15 mejores escenarios que minimizaron las medidas de error, en general, esto ocurrió cuando la distribución de la dificultad de los ítems es normal estándar. Además, cuando la estimación de las habilidades se desarrolla en un ambiente de desconocimiento de las dificultades de los ítems. En relación con los demás parámetros se tiene que, el porcentaje de ítems anclas puede variar del 10% al 30%, el parámetro del sesgo de la distribución de la dificultad puede variar entre 0 a 0.9, el número de valores plausibles debe ser de por lo menos 5 y el rango de valores plausibles de 2 0 4.

Lo anterior indica que, i) podemos estimar de manera precisa las habilidades de los sustentantes sin conocer a priori las dificultades, sin embargo, es necesario que la distribución de las dificultades sea una norma estándar sin importar su sesgo; ii) el porcentaje de ítems ancla debe ser inferior o igual al 30% de los ítems de la forma de prueba; iii) el número de valores plausibles para la estimación es de por lo menos 5; y iv) el rango de los valores plausibles puede ser de 2 o 4.

Ahora, se interpreta el primer modelo de la Tabla 2. Dado un conjunto de datos que siguen un modelo de Rasch con dificultades de los ítems distribuidos normalmente (normal estándar) sin sesgo con 10 ítems ancla de 100 presentes, con un 4.8% de no respuestas de los ítems. La estimación por valores plausibles de las habilidades de los sustentantes con 5 valores plausibles y con un rango de 2, tiene un valor de 0.1382 de la raíz cuadrada del error cuadrático medio y una desviación absoluta media de 0.1045. Esto quiere decir que el escenario en estudio es el más preciso en la estimación de la habilidad, además se espera que en promedio se cometa un error máximo de 0.1045 al momento de estimar la habilidad, tomando en cuenta que regularmente la habilidad fluctúa en el intervalo $(-6, 6)$ este error es pequeño. Del mismo modo, en la estimación por máxima verosimilitud se tiene un RMSE de 0.0512 y una MAD de 0.0406, evidentemente números más pequeños en comparación de los valores plausibles (esto se repitió en todos los escenarios). Por lo tanto, la mejor opción de equiparación de dos formas de prueba en el contexto descrito en este trabajo es mediante máxima verosimilitud, sin embargo, los resultados de valores plausibles son alentadores y con alta precisión.

Tabla 2: *Primeros 15 mejores escenarios en la estimación de la habilidad mediante máxima verosimilitud y valores plausibles*

Escenario	Porcentaje Anclas	Parámetro de sesgo	Número de valores plausibles	Rango valores plausibles	RMSE.VP	MAD.VP	RMSE.MV	MAD.MV
1	0,1	0	5	2	0,1382	0,1045	0,0512	0,0406
2	0,3	0,9	5	2	0,1389	0,1043	0,0545	0,0428
3	0,3	0	5	2	0,1397	0,1035	0,0524	0,0408
4	0,2	0,3	5	2	0,1403	0,1043	0,0510	0,0398
5	0,2	0,9	5	2	0,1405	0,1065	0,0536	0,0421
6	0,1	0,9	7	4	0,1409	0,1121	0,0510	0,0395
7	0,1	0,6	5	2	0,1411	0,1040	0,0519	0,0409
8	0,2	0,3	7	4	0,1416	0,1119	0,0510	0,0398
9	0,3	0,9	7	2	0,1417	0,1063	0,0545	0,0428
10	0,3	0	7	4	0,1419	0,1119	0,0524	0,0408
11	0,1	0,6	7	4	0,1421	0,1125	0,0519	0,0409
12	0,1	0,9	5	2	0,1424	0,1069	0,0510	0,0395
13	0,1	0,3	7	4	0,1424	0,1132	0,0504	0,0392
14	0,1	0	7	2	0,1425	0,1065	0,0512	0,0406
15	0,1	0,3	5	2	0,1426	0,1067	0,0504	0,0392

Fuente: *(Autores, 2022)*
Elaborado: *Autores*

En la Tabla 3. se muestra los 15 peores escenarios que maximizaron las medidas de error, en general, ocurre esto cuando la distribución de la dificultad de los ítems es uniforme en el intervalo $(-3, 3)$. Además, cuando la estimación de las habilidades se desarrolla conociendo las dificultades de los ítems. Para los demás parámetros se tiene que, el porcentaje de ítems anclas puede variar libremente al igual que el parámetro del sesgo de la distribución de la dificultad, el número de valores plausibles debe ser de uno, y el rango de valores plausibles de 5.

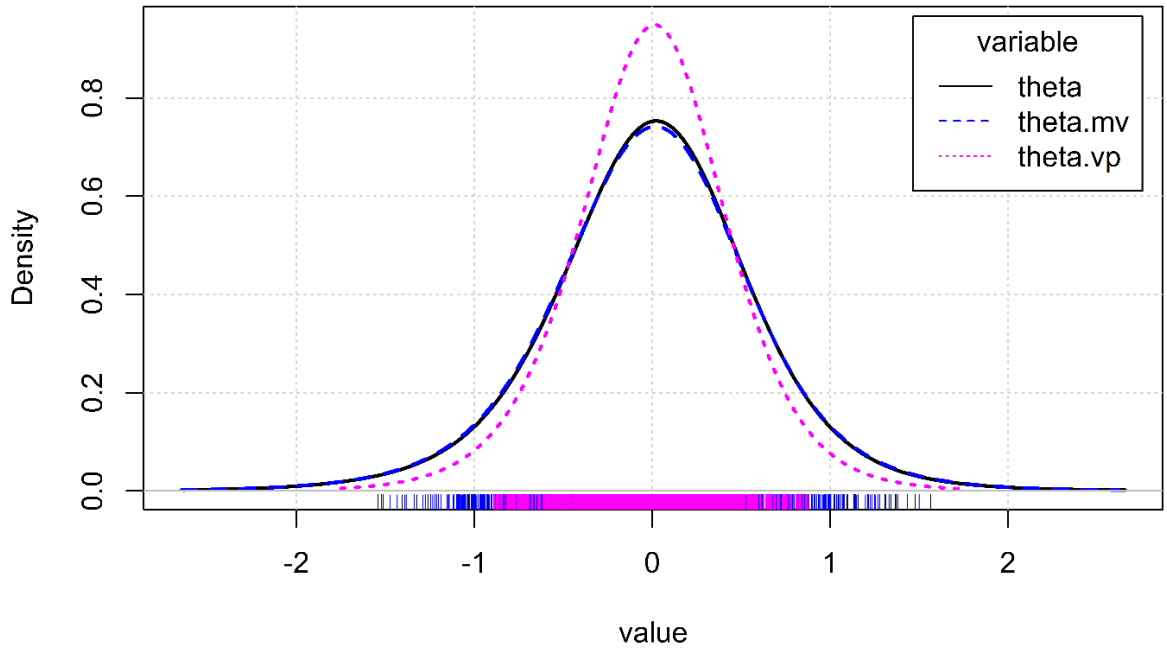
Tabla 3: *Últimos 15 peores escenarios en la estimación de la habilidad mediante máxima verosimilitud y valores plausibles*

Modelo	Porcentaje Anclas	Parámetro de sesgo	Número de valores plausibles	Rango valores plausibles	RMSE.VP	MAD.VP	RMSE.MV	MAD.MV
1	0,3	0,6	1	5	3,0037	2,9524	0,5772	0,4482
2	0,3	0,9	1	5	2,9747	2,9228	0,5501	0,4274
3	0,2	0,6	1	5	2,9640	2,9220	0,3536	0,2738
4	0,2	0	1	5	2,9577	2,9078	0,4310	0,3422
5	0,3	0,3	1	5	2,9525	2,9044	0,5772	0,4447
6	0,3	0	1	5	2,9363	2,8875	0,5465	0,4233
7	0,2	0,6	1	5	2,8937	2,8419	0,4304	0,3391
8	0,2	0,9	1	5	2,8908	2,8399	0,4336	0,3367
9	0,2	0,3	1	5	2,8904	2,8391	0,4549	0,3546
10	0,1	0,6	1	5	2,8198	2,7674	0,2407	0,1915
11	0,1	0	1	5	2,8094	2,7558	0,2392	0,1907
12	0,1	0,9	1	5	2,6878	2,6334	0,2247	0,1785
13	0,1	0,3	1	5	2,6664	2,6124	0,2248	0,1812
14	0,3	0,6	1	5	2,6482	2,5996	0,2092	0,1709
15	0,3	0,9	1	5	2,5343	2,4731	0,0601	0,0469

Fuente: *(Autores, 2022)*
Elaborado: *Autores*

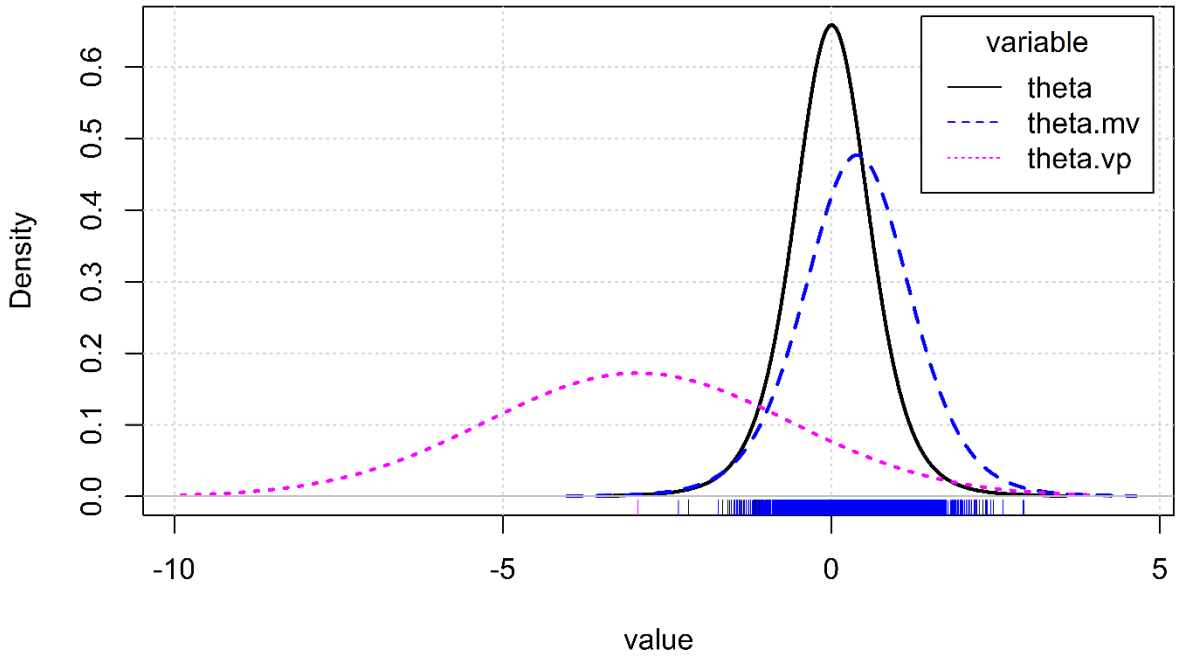
En las Figuras 1. y 2. se muestran la función densidad de las habilidades de los sustentantes (theta) y las estimaciones mediante máxima verosimilitud (theta.mv) y valores plausibles (theta.vp). Lo anterior para el mejor y peor escenario. En el caso del mejor escenario se mira un buen ajuste de la curva de theta.mv hacia theta, mientras que la curva de theta.vp no tiene a ubicarse sobre la curva de theta. Obviamente en el peor escenario las curvas de theta.mv y theta.vp no presentan un buen ajuste a la curva de theta, sin embargo, el ajuste de theta.mv es mejor (esto ocurrió en los 768 escenarios). Por lo tanto, en sintonía con lo indicado en los anteriores párrafos la estimación por máxima verosimilitud es la mejor opción en cualquier escenario de equiparación.

Figura 1: *Mejor escenario: Distribución de densidad de las variables habilidad (theta), habilidad estimada mediante máxima verosimilitud (theta.mv) y habilidad estimada mediante valores plausibles (theta.vp).*



Elaborado: *Autores*

Figura 2: *Peor escenario: Distribución de densidad de las variables habilidad (theta), habilidad estimada mediante máxima verosimilitud (theta.mv) y habilidad estimada mediante valores plausibles (theta.vp).*



Elaborado: *Autores*

Pol. Con. (Edición núm. 92) Vol. 9, No 5, Mayo 2024, pp. 729-742, ISSN: 2550 - 682X

Conclusiones

Se desarrolló un proceso de simulación de datos de dos formas de prueba que se ajustan al modelo de teoría de respuesta al ítem de Rasch. Donde se consideró aspectos importantes como i) el porcentaje de ítems ancla, ii) la distribución de probabilidad de las dificultades de los ítems, y iii) el sesgo de la distribución. A partir de estos datos, se estimó la habilidad de los sustentantes mediante máxima verosimilitud tomando en cuenta un modelo con dificultades conocidas o desconocidas. Además, se estimó la habilidad a través de valores plausibles, considerando el i) número de valores plausibles extraídos de la distribución posterior de las habilidades, y ii) el rango para la extracción. En total se trabajó con 768 escenarios donde se calculó el error absoluto medio y el error cuadrático medio. Se llegó a las siguientes conclusiones, teniendo en mente que se busca el escenario que minimice el error absoluto medio y el error cuadrático medio (es decir, la mejor estimación de la habilidad de los sustentantes):

- Para la equiparación de dos formas de prueba no es importante el porcentaje de ítems ancla (lógicamente llegando máximo al 30%). Tampoco es importante sesgos leves de la distribución de las dificultades de los ítems.
- En general, cuando se construye un instrumento de evaluación se desconoce las dificultades y su distribución. De los resultados de esta investigación se tiene que no es importante conocer las dificultades, sin embargo, estos deben seguir una distribución normal estándar. Esto es importante en la estimación de las habilidades mediante máxima verosimilitud.
- En la estimación de la habilidad por valores plausibles, es recomendable extraer 5 o más valores plausibles y considerar un rango de extracción de 2 o 4.
- Los peores escenarios muestran que esto ocurre cuando la distribución de la dificultad de los ítems es uniforme en el intervalo $(-3, 3)$. Al respecto, al construir una prueba de evaluación no es buena idea elegir ítems que generen una distribución uniforme de las dificultades en la prueba, más bien, deben generar una distribución normal estándar.

Algunos autores han estudiado el problema de la estimación de la habilidad en un entorno de equiparación, al respecto Seong (1990) muestra que el aumento del número de puntos de cuadratura mejoró la precisión de la estimación de los parámetros del ítem. Kim y Nicewander (1993) exploraron los estimadores de máxima verosimilitud [MLE (θ)], probabilidad ponderada [WLE (θ)], modal bayesiano [BME (θ)], esperado a posteriori [EAP (θ)], ellos muestran que las estimaciones de la habilidad de estos estimadores son razonablemente imparciales para el rango de

habilidades correspondiente a la dificultad de una prueba, y que sus errores estándar eran relativamente pequeños. Estos autores presentan evidencia a favor de los resultados de este trabajo. Por otro lado, algunos autores presentan resultados que contrastan con los de este trabajo. Guamán y Sepa (2023) indican que las estimaciones más precisas, son con la estimación mediante valores plausibles, sin embargo, en esta investigación no se demuestra que los datos se ajustan a un modelo psicométrico de Rasch, por lo tanto, los resultados no son confiables. Lord (1953) menciona que la estimación optima de la habilidad, mediante ítems de opción múltiple con un nivel de dificultad del ítem algo más fácil que el punto medio entre 0.5 y 1.

Dentro de las limitaciones de esta investigación están, el estudio de la precisión de la estimación de la habilidad cuando el tamaño muestral aumenta. Además, solo se consideraron dos estimadores de la habilidad. Al respecto se podría considerar en la investigación futura otros estimadores como el algoritmo EM (Bock y Aitkin, 1981). También, se podría estudiar métodos para reducir el sesgo cuando se desconoce las dificultades de los ítems (Zhan, 2005).

Referencias

1. Attorresi, H. F., Lozzia, G. S., Abal, F. J. P., Galibert, M. S., & Aguerri, M. E. (2009). Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. Revista Argentina de clínica psicológica, 18(2), 179-188.
2. Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. The Journal of Machine Learning Research, 9, 485-516.
3. Bock, RD, Aitkin, M. Estimación de máxima verosimilitud marginal de los parámetros del ítem: aplicación de un algoritmo EM. Psicometrika 46, 443–459 (1981). https://doi.org/10.1007/BF02293801
4. Braun, H., & Holland, P. (1982). Observed-score test equating: a mathematical analysis of some ETS equating procedures. In P. Holland & D. Rubin (Eds.), Test equating (Vol. 1, pp. 9–49). New York: Academic Press.
5. Chalmers, P., Pritikin, J., & Oguzha, O. (2023). MIRT: Multidimensional Item Response Theory. R Package Version 3.6.0.
6. Eliason, SR (1993). Estimación de máxima verosimilitud: Lógica y práctica (Nº 96). Sabio.

7. Guamán, E., & Sepa M. (2023). Comparación En La Estimación De La Habilidad De Sustentantes Entre La Teoría De Respuesta Al Ítem Vs La Metodología De Valores Plausibles.
8. Hambleton, R. K., Swaminathan, H., Hambleton, R. K., & Swaminathan, H. (1985). Estimation of Ability. Item Response Theory: Principles and Applications, 75-99.
9. Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test analysis modules. R Package Version 1.995-0.
10. Kim, JK, Nicewander, WA Estimación de capacidad para pruebas convencionales. Psicometrika 58, 587–599 (1993). https://doi.org/10.1007/BF02294829
11. Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Investigación en educación médica, 3(9), 40-55.
12. Lord, FM Una aplicación de intervalos de confianza y de máxima verosimilitud a la estimación de la capacidad de un examinado. Psicometrika 18, 57–76 (1953). https://doi.org/10.1007/BF02289028
13. Muñiz Fernández, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo: Revista del Colegio Oficial de Psicólogos.
14. Myung, I. J. (2003). Tutorial on maximum likelihood estimation. Journal of mathematical Psychology, 47(1), 90-100.
15. Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.
16. Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied psychological measurement, 14(3), 299-311.
17. Thompson, N. A. (2009). Ability estimation with item response theory. Assessment Systems Corporation, 20.
18. Wu, M. (2022). A Course on Test and Item Analyses. https://www.edmeasurementsurveys.com/IRT/index.html
19. Zhang, J. (2005). Bias correction for the maximum likelihood estimate of ability. ETS Research Report Series, 2005(2), i-39.