Sławomir Sapanowski
Okręgowa Komisja Egzaminacyjna w Łodzi

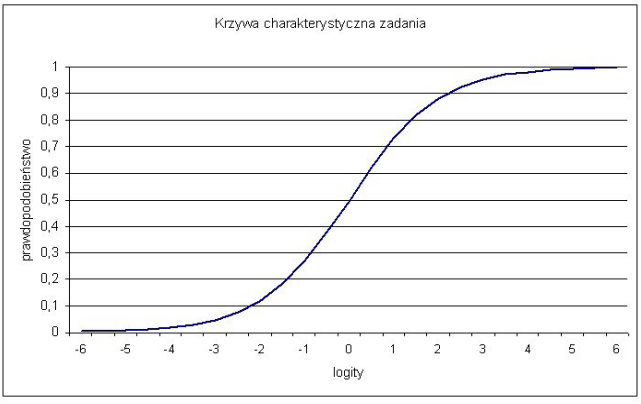
Oszacowanie umiejętności „teta" oraz wyskalowanie osi w metodzie IRT dla potrzeb obliczania parametrów zadań

W ostatnim czasie wśród ekspertów zajmujących się analizą wyników egzaminów zewnętrznych ogromną karierę robi teoria odpowiadania na pozycje testowe (IRT, *Item response theory*). Założenia tej metody są powszechnie dostępne i znane. Mimo to przypomnijmy zasadnicze tezy tworzące podstawy tego sposobu analizy.

1. Wynik otrzymany na egzaminie zależy od szeregu czynników nazywanych ukrytą cechą (*latent trait*), którą w zasadzie utożsamiamy z poziomem umiejętności i wiedzy ucznia przystępującego do egzaminu.
2. Występuje lokalna niezależność zadań. Założenie to jest spełnione, gdy odpowiedź na konkretne zadanie w teście nie wpływa na odpowiedzi na inne zadania.
3. Trzecia teza dotyczy związku między możliwością udzielenia poprawnej odpowiedzi na zadanie, a poziomem umiejętności ucznia. Zależność ta opisywana jest za pomocą funkcji charakterystycznej zadania (*item characteristic curve* - ICC).

$$f(\Theta)=c+\frac{1-c}{1+\exp(-a(\theta-b))}$$

Jest to funkcja, która wskazuje, że wraz ze wzrostem umiejętności rośnie prawdopodobieństwo udzielenia poprawnej odpowiedzi.



Oś pionowa reprezentuje prawdopodobieństwo uzyskania sukcesu, a oś pozioma - różnicę między umiejętnością ucznia a trudnością zadania.
Pełną informację na temat kształtu krzywej charakterystycznej zawierają trzy parametry:

a. moc różnicująca zadania,
b. trudność zadania,
c. poziom zgadywania (jest to tzw. model trójparametryczny albo model Birnbauma).

Jeśli znamy te wielkości, to położenie krzywej zależne jest jedynie od nieznanego parametru Θ (teta). Jego oszacowanie, a co za tym idzie wyskalowanie poziomej osi wykresu ICC, ma kluczowe znaczenie dla porównywania zadań i testów stosowanych w kolejnych latach. Kilkukrotnie pojawiały się już takie próby[1], ale niestety nie ma do tej pory jednolitego standardu, który pozwalałby na swobodną wymianę zdań między osobami zajmującymi się pomiarem dydaktycznym. Artykuł ten jest propozycją szacowania wielkości Θ, a także próbą wprowadzenia takiego standardu.

Ciekawy wykład na temat szacowania wartości Θ przedstawił N. Verhelst w artykule *Probabilistyczna teoria wyniku zadania testowego* zamieszczonym w *Biuletynie Badawczym CKE* nr 9/2007. Autor przedstawia w nim metodę maksymalnego prawdopodobieństwa (ML, *maksimum likelihood*) oraz estymator Warma[2], które służą do określenia wartości Θ dla pojedynczego ucznia. Szczególnie estymator Warma dobrze spełnia swoje zadanie, ponieważ jego obciążenie[3] jest niewielkie (dla zakresu, w którym wartości funkcji informacyjnej testu są większe od 2). Proponowane przez Verhelsta estymatory posiadają niestety wady. Przede wszystkim są one skomplikowane pod względem zastosowanego aparatu matematycznego oraz dają jednoznaczne rezultaty jedynie dla modelu jednoparametrycznego (Rascha) i dwuparametrycznego.

Przy obu estymatorach okazuje się, że oszacowana wielkość theta zależy jedynie od wyniku punktowego testu a nie od konkretnego układu odpowiedzi. Dotyczy to modelu Rascha oraz dwuparametrycznego modelu logistycznego (2PLM). Nie dotyczy to jednak modelu trójparametrycznego. [podkr. Własne] (Verhelst, 2007, s. 63)

W arkuszach egzaminacyjnych (sprawdzian i egzamin gimnazjalny) znajduje się gros zadań zamkniętych, które są podatne na zgadywanie. Parametr c (poziom zgadywania) dla krzywej charakterystycznej różni się od zera. Powoduje to

[1] *Omawiane zagadnienie nie jest zbyt proste. W rzeczywistości istnieje kilka sposobów szacowania wartości theta na podstawie uzyskanych odpowiedzi, z których każdy ma swoje zalety i wady* (N. Verhelst, 2007, s. 53).

[2] Szacowaną wielkość Warma definiuje się jako tę wartość theta, przy której iloczyn dwóch funkcji osiąga maksymalną wartość. Jedną funkcją jest funkcja prawdopodobieństwa, drugą pierwiastek kwadratowy funkcji informacyjnej (zob. N. Verhelst, 2007).

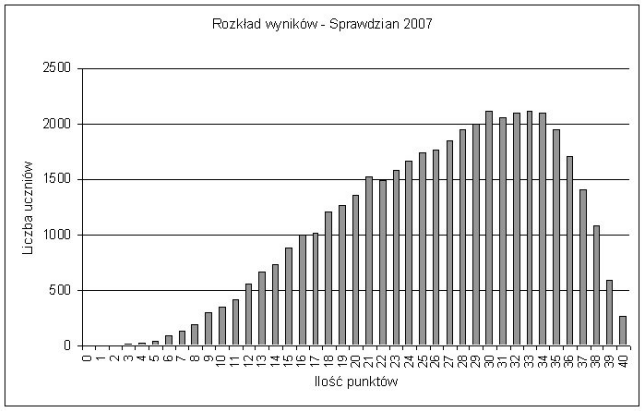
[3] obciążenie - różnica pomiędzy oczekiwaną oszacowaną wartością a prawdziwą wartością theta (zob. N. Verhelst, 2007).

konieczność stosowania do analizy tychże zadań modelu Birnbauma (3PLM). Jakie zatem istnieje i czy istnieje rozwiązanie? Spróbujmy odpowiedzieć na te zasadnicze pytania. Z prostego przekształcenia wzoru funkcji logistycznej otrzymujemy:

$$\Theta\text{-}b=\ln\left(\frac{P(s)}{1-P(s)}\right) \quad (*)$$

gdzie P(s) to prawdopodobieństwo sukcesu (udzielenia przez ucznia poprawnej odpowiedzi).

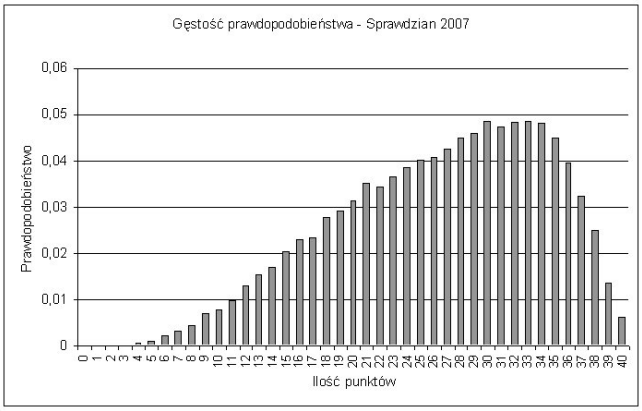
A zatem problem szacowania wartości Θ sprowadza się do obliczenia wartości ilorazu występującego pod znakiem logarytmu. Aby to uczynić, przyjrzyjmy się rozkładowi punktowemu wyników Sprawdzianu 2007.



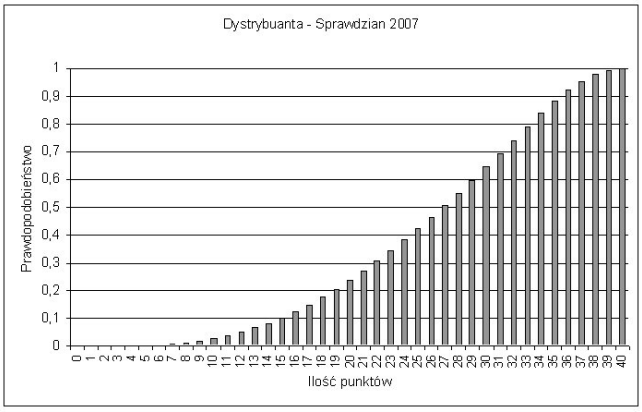
Jest to empiryczny rozkład dla całej populacji w OKE Łódź.

Częstość występowania określonej wartości cechy statystycznej w całej zbiorowości określa prawdopodobieństwo występowania tej wartości w ogóle. Tym samym rozkład cechy statystycznej skonstruowany na podstawie badania pełnego, a więc dotyczący całej zbiorowości, jest tożsamy z rozkładem prawdopodobieństwa wartości, jakie ta cecha przyjmuje (Rószkiewicz, 2002, s. 77).

Przekształćmy wykres tak, aby przedstawiał rozkład prawdopodobieństwa. W tym celu wystarczy podzielić liczbę uczniów w kolejnych kolumnach przez ilość wszystkich zdających. Spowoduje to, że suma długości wszystkich „słupków" będzie wynosiła 1.



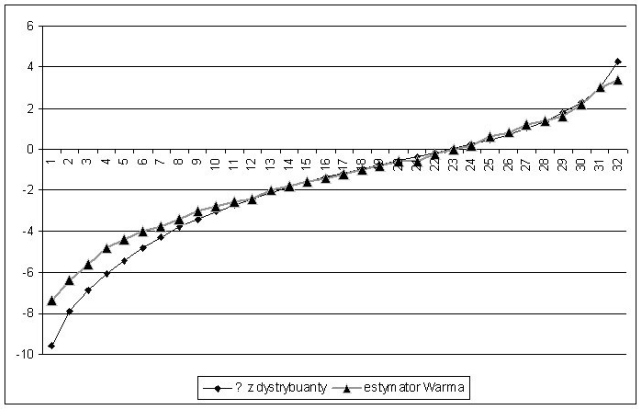
Zauważmy, że kształt rozkładu pozostał taki sam, a zmianie uległa skala osi pionowej. Cóż nam to daje? Otóż dysponując gęstością prawdopodobieństwa, możemy (w przypadku zmiennej skokowej) łatwo skonstruować dystrybuantę prawdopodobieństwa.



Oczywiste jest, że rezultatowi 40 pkt. musimy przypisać wartość prawdopodobieństwa równą 1, ponieważ przystępujący do sprawdzianu jakiś konkretny wynik osiągnąć musiał - jest to zdarzenie pewne. Na przykład, wysokość kolumny dla wartości 28 punktów wynosi 0,55. Oznacza to, że prawdopodobieństwo uzyskania przez ucznia ze sprawdzianiu wyniku 28 lub mniej punktów wynosi 0,55. Tak więc wynikowi 28 punktów przyporządkowujemy wartość P(s)=0,55.

Obliczając wartość wyrażenia (*) dla podanego przykładu (28 pkt.), otrzymujemy: $\Theta \approx 0{,}2$. Postępując podobnie dla pozostałych wyników, otrzymamy szacunkowe wielkości Θ dla całej populacji.

Rodzi się pytanie, czy taka estymacja Θ nie jest sprzeczna z wynikami otrzymanymi za pomocą estymatora Warma. Przyjrzyjmy się poniższemu wykresowi. Ze względu na metodologię szacowania Θ metodą Warma, usunięto z analizy zadania wielopunktowe, w związku z czym skala uległa skróceniu do 32 pkt.



Zauważamy, że w zakresie od 11 do 31 pkt. wykresy praktycznie pokrywają się (dokładniejsze obliczenia wskazują na różnice rzędu kilku setnych). Czym zatem można wytłumaczyć rozbieżności na końcach skali?

Jak zauważa sam Verhelst:

Estymator Warma zdradza jedynie małe (nieistotne) obciążenie w obrębie przedziału wokół punktu maksymalnej informacji. Poza tym przedziałem (...) przy niskich wartościach theta obciążenie jest dodatnie, przy wyższych wartościach jest ono ujemne. Rezultat tego obciążenia jest taki, że zmienność szacowanych wielkości Warma będzie na ogół mniejsza aniżeli zmienność rzeczywistych wartości theta. [podkr. Własne] Ten efekt znany jest jako skurczenie (Verhelst, 2007, s. 63).

Ponadto estymator Warma nie daje jednoznacznych oszacowań dla modelu Birnbauma, a zauważmy, że udział zadań zamkniętych (z założenia punktowanych 0-1) w naszej 32-punktowej skali jest znaczny i wynosi ponad 60 %, i są to zadania opisywane trójparametrycznie.

Zdający, który uzyskał wynik z przedziału 0-6 pkt., w znacznej części osiągnął to dzięki zgadywaniu, a nie faktycznej wiedzy. Dlatego też oszacowanie Θ Warma w tym zakresie jest wyższe niż Θ z dystrybuanty.

Mam nadzieję, że podana metoda szacowania wartości Θ stanie się standardem przy analizie zadań metodą IRT. Ma to ogromne znaczenie podczas tworzenia banku zadań, który jest z konieczności związany z konstruowaniem arkuszy egzaminacyjnych. Taki bank powinien zawierać propozycje zadań z parametrami (dyskryminacja, trudność, poziom zgadywania) obliczanymi według tej samej, zunifikowanej skali, której do tej pory nikt nie proponował.

Pozostaje oczywiście dopracowanie szczegółów, jak np. obciążenie proponowanego estymatora w zależności od wartości funkcji informacyjnej. Ale jest to temat na kolejny artykuł i kolejną Konferencję Diagnostyki Edukacyjnej.

Bibliografia:

1. Rószkiewicz M., *Statystyka*, Efekt, Warszawa 2002.
2. Verhelst N., *Probabilistyczna teoria wyniku zadania testowego,* [w:] *Biuletyn Badawczy CKE Egzamin. Klasyczna i probabilistyczna teoria testu. Nr 9/2007.* CKE, Warszawa 2007.