МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2001
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2001
Proceedings of Thirtieth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 8–11, 2001

НАДЕЖДНОСТ И ТОЧНОСТ НА ОЦЕНКИТЕ ОТ ИЗПИТИ И ТЕСТОВЕ

Пламен Матеев, Евгения Стоименова

Успехът на учениците от всеки изпит зависи както от техните знания и умения, така и от множество други фактори, които са случайни. В статията е представен вероятностен модел на изпит, включващ вероятностни разпределения в три множества – на учениците, на изпитите и на постиженията. Мярка за надеждността на всеки изпит е корелацията между резултатите на изпитваните ученици и техните „действителни знания“. Високата надеждност води до точно оценяване на постиженията на учениците. Изследвано е влиянието на някои от параметрите на задачите върху надеждността.

1. Вероятностен модел на изпити. Нека $\mathcal{T}$ е множество от изпити от определена учебна област, предназначени за някакво множество от ученици $\mathcal{U}$. Всеки изпит се състои от една или няколко задачи, които отразяват знанията на учениците в учебната област. Изборът на изпит от $\mathcal{T}$ се извършва съгласно вероятностно разпределение g в $\mathcal{T}$. Разпределението g е свързано с избора на задачите (напр. по трудност и/или по подобласт) и заедно с $\mathcal{T}$ определя учебната област на изпита.

Всеки ученик $u \in \mathcal{U}$ се характеризира с (латентен) параметър $\theta_u \in \Theta \subset R$, съответен на неговите знания (умения) от областта на измерваните постижения. Целта на изпита е да се получат оценки за неизвестните параметри θ_u на всички изпитвани ученици и евентуално учениците да се наредят според стойностите на параметрите.

Показаният резултат (успехът) на един ученик от един конкретен изпит зависи както от неговите знания, така и от избрания изпит и от някои други фактори, които ще считаме за случайни. Успехът приема стойности в някаква скала – пространството от всички възможни резултати $\mathcal{X} \subset R$. Разглеждаме успеха като случайна величина X в $\mathcal{X}$ с вероятностно разпределение, зависещо от избрания изпит $t \in \mathcal{T}$ и конкретния ученик $u \in \mathcal{U}$:

$$Pr(X = x|u, t) = p_u(x|t).$$

Математическото очакване на $p_u(x|t)$ по всевъзможните изпити от $\mathcal{T}$ е

$$\sum_t p_u(x|t)g(t) = p_u(x).$$

Това очакване не зависи от изпита t и $p_u(x)$ е разпределението на успеха X_u на ученика u в областта на измерваните постижения, определена от $\mathcal{T}$ и g.

Дефиниция 1. *Действителен успех* τ_u *на ученика* $u \in \mathcal{U}$ *ще наричаме математическото очакване на успеха му* X_u *по разпределението* $p_u(x)$*:*

$$\tau_u = E_{p_u(x)} X_u.$$

Действителният успех всъщност е средният успех на ученика u от всички възможни изпити от $\mathcal{T}$ и естествено предполагаме, че е свързан с параметъра θ_u посредством някаква монотонна функция. Действителният успех, както и θ_u, е ненаблюдаем, тъй като не е възможно на ученика да се дадат всички възможни изпити от $\mathcal{T}$. Вместо него за оценка на θ_u се използват резултатите X от един или няколко изпита.

За всеки ученик $u \in \mathcal{U}$ разликата между наблюдавания успех X_u от един изпит и действителния му успех τ_u представлява грешка от изпита:

$$\varepsilon_u = X_u - \tau_u. \tag{1}$$

Това равенство задава класическия модел на изпит (тест) [3].

От тази дефиниция следва, че грешката е случайна величина и нейното условно математическо очакване, когато действителният успех на ученика е τ_u, е нула:

$$E(\varepsilon_u|\tau_u) = E(X_u - \tau_u|\tau_u) = E(X_u|\tau_u) - E(\tau_u|\tau_u) = \tau_u - \tau_u = 0. \tag{2}$$

(Тук математическото очакване е по разпределението $p_u(x)$ в $\mathcal{X}$.) Това означава, че при многократни независими изпитвания на ученика грешката не се натрупва и средният успех от тях е неизместена оценка на действителния му успех.

2. Надеждност на изпит. В първия раздел разглежданията се отнасяха до един ученик. Нека сега да предположим, че разполагаме с *група от ученици* от множеството $\mathcal{U}$, избрани съгласно някакво разпределение f. Тогава съответните им успех от изпит $X = X_u$, действителен успех $\tau = \tau_u$ и грешка от изпита $\varepsilon = \varepsilon_u$ са случайни величини, дефинирани в $\mathcal{U}$.

Надеждността на изпита свързваме с разликите в успеха на различни ученици от един и същи изпит. Има две основни причини за различните резултати. Първата е, че учениците реално имат различен действителен успех τ и съответно различни нива на знанията θ, които се предвижда да се оценят чрез изпита. Ако това е така, изпитът отчита тази истинска разлика и води до правилно нареждане. Втората причина за различието е случайната грешка.

От модела (1) следва, че математическото очакване на действителния успех за групата ученици е равно на математическото очакване на наблюдавания успех, т.е.:

$$E_f(\tau) = E_f(X) - E_f(\varepsilon) = E_f(X). \tag{3}$$

По-нататък ще изпускаме индекса f, но ще имаме предвид, че всички изводи се отнасят за група от ученици, избрани съгласно това разпределение.

От модела (1) следва също, че действителният успех и грешката са независими и корелацията между тях е нула: $corr(\varepsilon, \tau) = 0$.

От независимостта на τ и ε и от равенствата (2) и (3) следва, че дисперсията на успеха на всяка група изпитвани ученици се разлага на две компоненти – дисперсия на действителния успех σ_τ^2 и дисперсия на грешката σ_ε^2:

$$var(X) \equiv \sigma_X^2 = \sigma_\tau^2 + \sigma_\varepsilon^2. \tag{4}$$

Наблюдаваният успех от изпита може да се използва вместо неизвестния дейст-

вителен успех на ученика, ако между тях съществува зависимост.

Дефиниция 2. *Коефициентът на корелационно отношение между действителния и наблюдавания успех наричаме надеждност на изпит:*

$$1 - \frac{var(X|\tau)}{var(X)}. \tag{5}$$

Ще покажем, че надеждността съвпада с квадрата на корелацията между наблюдавания и действителния успех. От независимостта на τ и ε и от разлагането на дисперсията (4) за ковариацията между X и τ намираме: $cov(X,\tau) = cov(\tau+\varepsilon,\tau) = \sigma_\tau^2 + cov(\varepsilon,\tau) = \sigma_\tau^2$. Следователно за корелацията между наблюдавания успех и действителния успех намираме

$$\rho^2(X,\tau) \equiv \frac{cov^2(X,\tau)}{\sigma_X^2\sigma_\tau^2} = \frac{\sigma_\tau^2}{\sigma_X^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2} = 1 - \frac{var(X|\tau)}{\sigma_X^2}, \tag{6}$$

където последното равенство следва от модела (2) и независимостта на τ и ε. От (6) следва, че надеждността е висока, когато дисперсията на грешката е малка. Когато $\rho(X,\tau)$ е близо до 1, наблюдаваният успех от изпита може да се използва вместо неизвестния действителен успех.

Дефиницията на надеждността не позволява директно да се оценява надеждността, тъй като едната променлива (действителният успех) е ненаблюдаема. Въпреки това за нея могат да се правят различни изводи чрез повторни изпити на учениците.

2.1. Корелация между еквивалентни изпити. Да предположим, че за всеки ученик $u \in \mathcal{U}$, освен с успеха от един изпит X_u, разполагаме и с резултатите от още един подобен изпит X'_u. Успехът X'_u и съответните му действителен успех $\tau'_u = E_{p_u(x)}X'_u$ и грешка ε'_u са случайни величини в множеството $\mathcal{U}$.

Дефиниция 3. *Два изпита наричаме еквивалентни, ако за всеки ученик $u \in \mathcal{U}$ е изпълнено: 1) $\tau_u = \tau'_u$ и $var(\varepsilon_u) = var(\varepsilon'_u)$; 2) грешките ε_u и ε'_u са независими и ε_u и τ'_u са също независими.*

Корелацията $\rho(X,X')$ между два резултата от еквивалентни изпити, събрана по всички ученици, може да бъде оценена, тъй като двете променливи X и X' са достъпни за наблюдение. Ще покажем по какъв начин $\rho(X,X')$ е свързана с надеждността $\rho(X,\tau)$.

Нека X и X' са резултатите от два еквивалентни изпити. Нека X удовлетворява модела $X = \tau + \varepsilon$. Очевидно за X' е изпълнено $X' = \tau + \varepsilon'$, тъй като действителният успех τ е един и същ за всички еквивалентни варианти на изпита.

От дефиниция (3) следва, че математическото очакване и дисперсията на успеха X' (по разпределението f в $\mathcal{U}$) са

$$E(X') = E(X) = E(\tau), \qquad var(X') = \sigma_\tau^2 + \sigma_\varepsilon^2 = var(X).$$

Ковариацията и корелацията между X и X' са съответно

$cov(X,X') = E(XX') - E(X)E(X') = E(\tau^2 + \varepsilon'\tau + \varepsilon\tau + \varepsilon'\varepsilon) - E(\tau)^2 = E(\tau^2) - E(\tau)^2 = \sigma_\tau^2$

и

$$\rho(X,X') = \frac{cov(X,X')}{\sqrt{var(X)var(X')}} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}. \tag{7}$$

Последният израз представлява частта на дисперсията на действителния успех

от общата дисперсия. Корелацията $\rho(X, X')$ е голяма, когато дисперсията на грешката σ_ε^2 е малка.

От сравняването на корелацията $\rho(X, X')$ между два еквивалентни изпита, изразена чрез (7), и корелацията $\rho(X, \tau)$ между наблюдавания и действителния успех, изразена чрез (6), следва

$$\rho^2(X, \tau) = \rho(X, X'). \tag{8}$$

По такъв начин надеждността на изпита може да се представи чрез корелацията на два еквивалентни изпита. Двете променливи X и X' във формулата (8) са наблюдаеми, което прави оценяването на надеждността възможно, стига да са изпълнени условията за еквивалентност на двата изпита.

С увеличаване на надеждността на изпита се осигуряват необходимите условия за неговата валидност. Ако е показано, че надеждността е висока, това означава, че успехът на учениците се дължи преди всичко на съществени фактори, различни от грешката на изпита.

3. Вътрешна непротиворечивост на скалата. До тук не предполагахме, че изпитът се състои от отделни задачи. Нека сега предположим, че успехът от изпита е сума на успеха от няколко задачи:

$$X = Y_1 + Y_2 + \cdots + Y_k,$$

където k е броят на задачите от изпита, а Y_i – успехът от i-тата задача. За успеха от всяка задача е изпълнено

$$Y_i = \tau_i + \varepsilon_i, \qquad i = 1, \ldots, k,$$

където τ_i и ε_i са съответно действителният успех и грешката на задачата. Предполагаме, че грешките от задачите са независими.

Дисперсията на сумата на k задачи, натрупана от всички участници в изпита, е равна на

$$\sigma_X^2 = var\left(\sum_i Y_i\right) = \sum_i var\,(Y_i) + 2\sum_{i<j} cov(Y_i, Y_j).$$

Дисперсията на общия успех ще бъде по-малка от сумата на дисперсиите на задачите, ако ковариациите са положителни, т.е. ако задачите мерят един и същ действителен успех. Можем да оценим частта от действителния успех, която се натрупва от задачите, чрез сравняване на сумата от дисперсиите на задачите с дисперсията на общия успех. Чрез нея се дефинира популярния коефициент за оценка на надеждността, наречен α на Кронбах:

$$\alpha \equiv \frac{k}{k-1}\left[1 - \frac{\sum_1^k \sigma_i^2}{\sigma_X^2}\right].$$

В тази формула σ_i^2 е дисперсията на успеха Y_i от i-тата задача, σ_X^2 е дисперсията на общия успех, k е броят на задачите. Коефициентът α приема стойности от 0 до 1.

Да обясним как интерпретира коефициентът α. Ако не получаваме никакъв действителен успех, а само грешка (която е некорелирана между лицата), то дисперсията

на сумата σ_X^2 ще бъде същата, както сумата от дисперсиите $\sum_1^k \sigma_i^2$ на индивидуалните задачи. Следователно, коефициентът α ще е нула. Ако всички задачи са идеално надеждни и мерят едно и също нещо (действителния успех), тогава коефициентът α ще е равен на 1 (тъй като $\sigma_i^2 = \sigma_X^2$).

Всъщност коефициентът α не определя надеждността на изпита, а е само една долна граница за нея. Оценката използва предположението, че корелациите между задачите от изпита са едни и същи: $corr(Y_i, Y_j) = \rho \; \forall i \neq j$. На практика ρ се оценява чрез средната корелация между задачите.

Теорема 1. *Нека успехът X е сума на резултатите от k задачи. Тогава надеждността на скалата, определена с* (5)*, е не по-малка от коефициента α на Кронбах:*

$$\rho^2(X,\tau) = 1 - \frac{\sigma_\varepsilon}{\sigma_X^2} \geq \alpha = \frac{k}{k-1}\left[1 - \frac{\sum_1^k \sigma_i^2}{\sigma_X^2}\right].$$

Доказателство. Успехът от изпита е сума на успеха от отделните задачи: $X = \sum_1^k Y_i$, съответно действителният успех τ е сума на действителния успех от отделните задачи: $\tau = \sum_1^k \tau_i$. От последното следва, че математическото очакване на τ $E(\tau) = \sum_1^k E(\tau_i)$.

Използваме неравенство (следствие от неравенството на Йенсен [2])

$$E(\sum \xi_i)^2 \leq k \sum (E(\xi_i))^2, \tag{9}$$

където $\xi_1, \ldots, \xi_k$ са произволни случайни величини. Равенството се достига, когато всички случайни величини са еднакво разпределени. Тогава ковариацията между всеки две от тях ще е постоянна.

Прилагаме (9) за случайните величини $\xi_i = \tau_i - E(\tau_i)$ и получаваме

$$E(\tau - E(\tau))^2 \leq k \sum E(\tau_i - E(\tau_i))^2,$$

което е еквивалентно на

$$\sigma_\tau^2 = var(\tau) \leq k \sum var(\tau_i).$$

Тъй като грешките от задачите са независими, за дисперсията на успеха получаваме

$$var(X) - \sum var(Y_i) = \sigma_\tau^2 - \sum var(\tau_i) \leq \frac{k-1}{k}\sigma_\tau^2 \tag{10}$$

От друга страна, нека вземем резултатите от още един еквивалентен изпит. Успехът X' от втория изпит удовлетворява модела $X' = \tau + \varepsilon'$. Следователно за ковариацията между X и X' е изпълнено

$$cov(X, X') = var(\tau). \tag{11}$$

От (10) и (11) за надеждността $\rho(X, X')$ получаваме

$$\rho(X, X') = \frac{cov(X, X')}{var(X)} \geq \frac{k}{k-1}\left[1 - \frac{\sum_1^k \sigma_i^2}{\sigma_X^2}\right].$$

Да отбележим, че коефициентът α достига максимална стойност, когато корелацията между всички задачи е една и съща. Тогава надеждността е равна на α.

3.1. Надеждност на композиционна скала. Задачите от една скала измерват един и същи вид постижения. Доста често няколко скали се обединяват в една композиционна скала и съвместно измерват съвкупност от постижения (например алгебра и геометрия в един изпит). Успехът от изпита обикновено е сума на успеха от отделните скали.

Подобно на приноса на задачите в една скала, скалите би трябвало да са в такава връзка, че натрупване на успех от една скала да съответства на натрупване на успех от друга скала, т.е. корелацията между успеха от всеки две скали да е положителна.

Поради различното отношение на задачите от различните скали към композиционната скала вътрешната съгласуваност на всички задачи, оценена с коефициента α, може да е по-ниска от надеждностните на отделните скали. За оценка на надеждността на композиционната скала по-естествено е да се използва вътрешната съгласуваност на задачите в отделните скали, като се отчетат корелациите между скалите. Такава оценка за надеждността на композиционната скала се получава чрез формулата за композиционна надеждност. Формулата определя долна граница за надеждността на изпит, съставен от няколко скали.

Теорема 2. *Нека успехът X е сума на резултатите $X_1, X_2, \ldots, X_n$ от n скали. Тогава надеждността на композиционната скала винаги е по-голяма от*

$$\alpha^* = \frac{\sum_{i=1}^n \sigma_{X_i}^2 \alpha_{X_i} + 2\sum_{i<j} \rho(X_i, X_j)\sigma_{X_i}\sigma_{X_j}}{\sum_{i=1}^n \sigma_{X_i}^2 + 2\sum_{i<j} \rho(X_i, X_j)\sigma_{X_i}\sigma_{X_j}},$$

където k е броят на скалите, $\sigma_{X_i}^2$ е дисперсията на успеха от i-тата скала, α_i е коефициентът на Кронбах на i-тата скала, $\rho(X_i, X_j)$ е корелацията между успеха от i-тата и успеха от j-тата скала.

Доказателство. Ще покажем как се извежда формулата за композиционна надеждност за случая на две скали.

Нека общият успех от изпита е сума от частичните резултати X и Y от две скали. Нека също така X' и Y' са съответно частичните резултати от други две, еквивалентни на първите, скали. Тогава надеждността на изпита, пресметната по два еквивалентни изпита, е

$$\begin{aligned}\rho(X+Y, X'+Y') &= \frac{cov(X+Y, X'+Y')}{var(X+Y)} \\ &= \frac{cov(X, X') + cov(X, Y') + cov(Y, X') + cov(Y, Y')}{var(X+Y)}\end{aligned}$$

$$= \frac{var(X)\frac{cov(X,X')}{var(X)} + 2cov(X,Y) + var(Y)\frac{cov(Y,Y')}{var(Y)}}{var(X+Y)}$$
$$\geq \frac{\sigma_X^2\alpha_X + \sigma_Y^2\alpha_Y + 2\rho(X,Y)\sigma_{X_i}\sigma_{X_j}}{\sigma_X^2 + \sigma_Y^2 + 2\rho(X,Y)\sigma_{X_i}\sigma_{X_j}}.$$

По подобен начин може да се получи долна граница за надеждността и когато скалите не са равностойни, а имат определено тегло в общия успех.

3.2. Стандартна грешка на измерване. Стандартното отклонение σ_ε на грешката в модела (2) се нарича стандартна грешка на измерване. Грешката на измерване се използва за предвиждане на интервала от стойности на успеха на един ученик, които могат да се получат при многократни изпитвания с еквивалентни изпити. Така стандартната грешка на измерване е стандартното отклонение на успеха на учениците около действителния им успех. От това следва, че успехът би трябвало да се смята по-скоро като попадащ в интервал от стойности, отколкото като множество от стойности. Колкото по-тесен е този интервал, толкова по-точно ще е предвиждането.

Стандартната грешка на измерване може да се определи от надеждността и дисперсията на наблюдавания успех по следния начин.

От разлагането на дисперсията на грешката $\sigma_\varepsilon^2 = \sigma_X^2 - \sigma_\tau^2$ от (4) и от представянето на надеждността $\sigma_\tau^2 = \sigma_X^2\rho^2(X,\tau)$ от (6) следва

$$\sigma_\varepsilon^2 = \sigma_X^2 - \sigma_X^2\rho^2(X,\tau) = \sigma_X^2(1-\rho^2(X,\tau)).$$

Следователно за стандартната грешка на измерване получаваме

$$\sigma_\varepsilon = \sigma_X\sqrt{1-\rho^2(X,\tau)}. \tag{12}$$

Във формулата за стандартната грешка на измерване участва надеждността $\rho^2(X,\tau)$, която, както знаем, е неизвестна. Ако вместо нея използваме нейната долна граница – коефициента α, ще получим горна граница за стандартната грешка на измерване:

$$\sigma_\varepsilon \leq \sigma_X\sqrt{1-\alpha}. \tag{13}$$

При нормално разпределение на успеха X стандартната грешка служи за определяне на доверителни интервали за действителния успех.

Когато надеждността на изпита е висока (близо до 1.0), стандартната грешка на измерване е малка и можем да бъдем уверени в точността на изпитните резултати. Обратно, когато надеждността е ниска, стандартната грешка ще е голяма. Поради грешката на изпита трябва много внимателно да се интерпретират малките разлики в резултатите на учениците. Ако надеждността е малка, един ученик трябва да има доста по-висок успех от друг, за да се направи заключение, че той има значително по-високи знания от другия.

Надеждността и стандартната грешка на изпита зависят от основните параметри на всяка задача – валидност и трудност.

Оценките на тези параметри се получават сравнително лесно по извадка от резултатите от „пробни“ изпити. Изпробваните и утвърдени (апробирани) задачи образуват банката от задачи за изпит. Подходящият избор от тази банка осигурява висока надеждност и висока точност на оценяване на изпита. За повече информация за параметрите на задачите и техните оценки от извадки виж [1].

ЛИТЕРАТУРА

[1] Е. Стоименова. Измерителни качества на тестове. НБУ, 2000.
[2] Й. Стоянов, И. Миразчийски, Ц. Игнанов, М. Танушев. Ръководство за упражнения по теория на вероятностите. Наука и изкуство, 1985.
[3] R.S. Traub. Reliability for the Social Sciences. Theory and Applications. SAGE Publications, 1994.

Пламен Матеев
Институт по математика и информатика
ул. Акад. Г. Бончев, бл. 8
1113 София
e-mail: pmat@math.bas.bg

Евгения Стоименова
Институт по математика и информатика
ул. Акад. Г. Бончев, бл. 8
1113 София
e-mail: jeni@math.bas.bg

RELIABILITY AND CORRECT ESTIMATING OF EXAMS AND TESTS

Plamen Mateev, Evgenia Stoimenova

The correct estimating of students' ability depends on reliability of underlying exams. The proposed model includes probability distributions on three basic sets – the set of students, the set of exams, and the set of scores. Reliability of an exam is defined through the correlation ratio between observed scores and "true" scores. High reliability leads to short confidence intervals for the ability parameters. The effect of items parameters on reliability is also discussed.