УДК 001.891:371.263

М. Є. СІНИЦЬКИЙ,
кандидат фіз.-мат. наук, доцент,
доцент кафедри інформаційних систем і технологій,
Національна академія статистики, обліку та аудиту

Статистичні інструменти вимірювання якості освіти. Частина 1. Класичний підхід

Представлено огляд статистичних основ тестології. Описано базові задачі, математичні моделі та основні розрахункові формули класичної (CTT) та сучасної (IRT) теорій тестування, що дозволяють зі статистичних позицій оцінити правильність побудови, роздільну здатність, стандартну помилку та надійність тестових вимірювань.

Ключові слова: *якість, освіта, тест вання, шкалювання, спостережуваний бал, кореляція, надійність, роздільна здатність, однорідність, латентна змінна, модель Раша, характеристичні криві, логіт.*

Постановка проблеми. Ключовою проблемою реформування освіти вважають підвищення її якості. Певною перешкодою на цьому шляху стає необізнаність широкого кола викладачів щодо сучасних інструментів вимірювання якості освітньої діяльності та якості результату освіти. Частково це пояснює "рубіжний" підхід до поняття "якість освіти", яке трактують як "рівень здобутих особою знань, умінь, навичок, інших, що відображає її компетентність відповідно до стандартів вищої освіти" [1].

Аналіз останніх досліджень і публікацій. Стаття представляє аналіз досліджень і публікацій зі статистичних основ тестування. Світова спільнота вже багато років використовує та удосконалює тестування як універсальний метод контролю якості викладання і навчальних досягнень. Багато хто вбачає у тестуванні фактор, що стимулює до навчання [2]. Адже за кордоном роботодавці широко використовують дані державних і приватних служб тестування [3; 4], які є "довідниками" рівнів навченості та компетентності випускників шкіл, коледжів і вищих навчальних закладів (ВНЗ).

В Україні існує усвідомлення перспективності тестування як "основи сучасного моніторингу якості освіти і встановлення соціальної справедливості в сфері освіти" [3], але тестування як обов'язковий інструмент поточного та рубіжного контролю навчального процесу поки що не використовується. Більше того, при прийомі до ВНЗ певна вага надається середньому арифметичному балу шкільного атестату, що з точки зору теорії вимірювань є нонсенсом.

Стаття ставить за **мету** інтенсифікувати впровадження сучасних статистичних методик та програмних продуктів у практику планування та реалізації навчального процесу задля удосконалення його якості.

Виклад основного матеріалу. Слід зазначити, що традиційні шкільні оцінки, до яких всі звикли, належать до порядкової шкали, а вона дозволяє лише іменувати та ранжувати елементи множини [5]. Аналогічна ситуація має місце у практиці експертного оцінювання, зокрема при агрегації думок експерттів, побудові узагальнених показників і рейтингів тощо, тобто при роботі з об'єктами статистики нечислової природи.

Існує декілька підходів до вирішення цієї проблеми. Один з них – використання досягнень репрезентативної теорії вимірювань (РТВ), що базується на алгоритмах аналізу даних, результат роботи яких не змінюється при будь-якому допустимому перетворенні шкали вимірювання (тобто є інваріантним щодо цього перетворення) [6]. Нагадаємо, що вимірювальна шкала – це математична характеристика змінної, що визначає відповідний математико-статистичний ме-

тод роботи з нею. Перед тим, як виконувати ту чи іншу математичну операцію, навіть елементарну, слід бути певним, що відповідна величина була виміряна в адекватній шкалі [4–9].

У тестуванні первинні дані вимірюються у номінальній (класифікаційній) шкалі. Для неї допустимими є всі взаємно однозначні перетворення, а величини можна категоризувати; обчислювати їх відносну (відсоткову) частоту та оцінювати ймовірність присутності вимірюваної ознаки у спостережуваній вибірці); моду; коефіцієнти зв'язку (асоціації, контингенції, взаємної зв'язаності). Прикладом номінальної шкали може служити дихотомічна шкала [9]. До неї часто звертаються, оскільки статистичні розрахунки в ній є достатньо простими.

Використовувана у вітчизняній освіті шкала відміток є порядковою, порівняння в ній дозволяє лише ранжувати учнів за ступенем прояву певних знань, але не пояснює, наскільки відрізняються ці знання та чи досягнуто мету навчання.

Допустимими статистичними показниками у порядковій шкалі є медіана, квантилі, коефіцієнт рангової кореляції. Правомірне проведення дисперсійного аналізу.

Додаткову можливість, а саме, порівняння різниць і визначення відповідності однакових відстаней у градаціях вимірюваної ознаки однаковим інтервалам шкали надає інтервальна шкала. Вона має точку початку відліку та одиницю вимірювання (метрику), але для кожної змінної ці параметри обираються довільно за домовленістю. Це не дозволяє в інтервальній шкалі вимірювати абсолютні величини. Тобто можна говорити, наскільки більше або наскільки менше виражена вимірювана властивість, але не у скільки разів вона більша або менша.

В інтервальній шкалі можна обчислювати моду, медіану, квантилі, рангові критерії, вибіркову середню, дисперсію, стандартне відхилення, коефіцієнт кореляції, застосовувати метод найменших квадратів (МНК) та інші способи знаходження екстремумів. Однак, як відмічено у [7], хоча значення, отримані за інтервальною шкалою, в значній мірі схожі зі звичайними дійсними числами, вони все ж визначаються не однозначно, а лише з точністю до перетворень, що зберігають структуру інтервалів між початковими числами.

Найуніверсальнішою є шкала відносин. Вона дозволяє оцінювати не тільки різниці двох величин, а й кожну величину окремо, і вимагає наявності на шкалі не довільної, а природної нульової точки, визначеної на множині дійсних чисел (наприклад, як в абсолютній шкалі температур). У шкалі відносин можна створювати еталони, вимірювати майже всі фізичні величини, але вона непридатна для використання у психологічних і соціальних вимірюваннях.

У шкалі відносин допустимими є всі арифметичні операції та правомірні всі поняття і методи математичної статистики.

Перші дві шкали іменують якісними (неметричними, слабкими), а інші – кількісними (метричними, сильними). Домінує думка, що одним з найважливіших аргументів на користь віднесення шкали до метричних шкал є розподіл результатів за нормальним законом.

Очевидно, що статистичні висновки можуть бути адекватні реальності тільки тоді, коли вони не залежать від того, яку одиницю вимірювання використовує дослідник, тобто коли вони інваріантні щодо допустимого перетворення шкали. В рамках РТВ було показано, що:

- у порядковій шкалі як середні можна використовувати тільки члени варіаційного ряду (порядкові статистики), зокрема медіану, але не середнє арифметичне, середнє геометричне тощо;
- у шкалі інтервалів з усіх середніх можна застосовувати тільки середнє арифметичне;
- у шкалі відносин з усіх середніх стійкими щодо порівняння є тільки степеневі середні та середнє геометричне.

Таким чином, згідно з РТВ правильним є порівняння шкільних атестатів за значеннями медіан відміток. Для використання середніх арифметичних потрібен перехід до інтервальної шкали.

Більш важким і цілком доступним сьогодні є використання для оцінювання "привабливості" абітурієнта полідисциплінарних тестів із залученням для оброблення результатів факторного аналізу та батовимірного шкалювання. Цей підхід може бути застосований і для рейтингового оцінювання поточних досягнень студентів.

Іншим напрямком є відмова від людини-екзаменатора як від засобу вимірювання рівня знань. Адже еталони оцінюваних об'єктів людина створює у своїй уяві, і вони залежать від безлічі факторів, а тому є нестабільними. При формуванні шкали оцінок людиною існує велика частка суб'єктивізму, оскільки тут багато що залежить від досвіду, інтуїції, компетентності і професіоналізму викладача. Крім того, вимоги, що висуваються різними викладачами до рівня знань студентів, коливаються в дуже широких межах. Тому реальні знання студента часто не отримують об'єктивного відображення, що знижує позитивний вплив екзаменаційної оцінки на пізнавальну діяльність і якість учбового процесу в цілому.

Тестування, яке при правильному застосуванні дає результати на інтервальній або порядковій шкалах, дозволяє формалізувати уявлення та не допускає суб'єктивності в оцінюванні знань. Аби результати тестування були валідними (адекватними вимірюваним знанням і вмінням) та надійними (порівняними і такими, які можна повторно перевірити), мають виконуватися певні вимоги до якості завдань та тесту в цілому. Аналіз їх виконання є важливою частиною підготовки та відбору завдань до застосування у педагогічному вимірюванні. Інакше використання тестів може бути не менш суб'єктивним, ніж звичайний екзамен [10].

Розрізняють класичну теорію тестування (Classical Test Theory, CTT) [4; 9; 11] та теорію, звану на Заході Item Response Theory (IRT) [4; 12]. Остання є переважною за кількістю наукових досліджень та опублікованих тестів. На її основі створюються адаптивні навчальні та контрольні системи багатьох університетів світу.

Звичайно розглядають два типи тестів відповідно до задач, які вони вирішують. Якщо йдеться про ранжування учасників тестування та упорядкування їх за рівнем підготовки, таке тестування називають відносним або нормативно орієнтованим. У ньому задача оцінювання рівня підготовки кожного учасника не ставиться, а потрібно тільки порівняти учасників один з одним або зі спеціально відібраною групою осіб, прийнятою за нормативну вибірку. Остання має бути репрезентативною, тобто адекватно представляти всю множину осіб, на яких розрахований певний тест. Досягають цього стратифікацією учасників за рівнем та умовами навчання (місто, село, звичайна школа, ліцей, гімназія тощо). Обсяг нормативної вибірки оцінюють добутком загального числа можливих пар страт на 100–300 осіб [13]. Результати тестування (математичне очікування, дисперсія тощо) на утвореній у такий спосіб вибірці називають тестовими нормами. Процес визначення норм називається стандартизацією тесту. Зіставлення результатів тестування з усередненими стабільними нормами дозволяє ослабити залежність тестового балу як від характеристик страти, так і від змісту навчальної дисципліни, і тим самим стандартизувати результат. Процедура стандартизації увійшла як обов'язкова у деякі міжнародні рекомендації та стандарти, наприклад ОСТ Т 1.1 Російської федерації 2001 р.

Процес встановлення норм займає досить багато часу, іноді до декількох років. Завдання, що включаються до стандартних тестів, ретельно відбираються за параметрами складності (від 0,3 до 0,7) та роздільної здатності (більше 0,3). Надійність тесту, на відміну від валідності, можна оцінити за допомогою статистичних методів, зокрема за кореляцією результатів або двох послідовних тестувань, або одноразового тестування, в якому завдання тесту поділені на дві еквівалентні частини. Нормативно-орієнтовані тести зазвичай використовуються для конкурсного відбору абітурієнтів під час вступу до ВНЗ (коефіцієнт надійності має бути не менше 0,9) [14].

Другий тип тестів – абсолютні, або критеріально-орієнтовані. Вони мають на меті оцінювання персонального рівня підготовленості кожного учасника тестування у чітко визначеній та достатньо вузькій області знань (змістовній

області)[1], безвідносно до оцінок інших учасників. Наприклад, крітеріально-орієнтований тест використовують для прийняття рішення щодо того, чи можуть студенти продовжувати навчання. Тобто вміст таких тестів повинен відповідати сукупності знань та вмінь, передбачених визначеними вимогами (навчальною програмою, державним стандартом тощо). Первинні результати трансформують у шкалу, градації якої відображають ступінь набуття потрібної сукупності знань та вмінь. Встановлюється певний критерій (оцінка мінімальної компетентності), що дозволяє приймати рішення (за принципом "зараховано – не зараховано") стосовно переходу до наступного навчального модулю. Критерій має бути однозначно ув'язаний з основоположними поняттями та найважливішими вміннями тієї області знань, для якої він розроблявся. Критерій має будуватися на умові мінімізації ймовірності пропустити непідготовлену особу або не пропустити достатньо підготовлену. Тобто йдеться про перевірку певної статистичної гіпотези, результативні бали тестування для якої виступають критерієм її прийняття чи відкидання з визначеними ймовірностями допустити помилки першого чи другого роду [4; 13].

На відміну від відносного тестування, де бажаними є значна дисперсія результатів (що є доказом доброї роздільної здатності цього типу тесту) та їх нормальний розподіл, що досягається вирівнюванням завдань за складністю, у абсолютному тестуванні, навпаки, використовуються завдання різної складності (частіше – такої, що рівномірно зростає), які дозволяють виявити глибину засвоєння матеріалу в певній області знань. Дисперсія результатів абсолютного тестування є невеликою, і тому вони не придатні для рейтингового оцінювання. Мають місце також значні відхилення від нормального розподілу.

Критеріально-орієнтовані тести призначені для використання при ліцензуванні та сертифікації, у поточному та рубіжному контролі.

У масовому тестуванні, особливо з використанням комп'ютерної техніки, зазвичай переважають завдання закритої форми, тобто з готовими (однією чи декількома) правильними відповідями, які потрібно вибрати, а іноді – упорядкувати згідно із завданням (зустрічаються і комбінації)[2]. Завдання з вибором містять один чи більше число (частіше – до п'яти) дистракторів (правдоподібних, але невірних відповідей). Правильний підбір дистракторів є частиною процедури розроблення якісного тесту. Якість дистракторів перевіряється за рівномірністю розподілу частот вибору дистракторів випробовуваними (дистракторний аналіз) [10; 13]. Зрозуміло, що збільшення числа дистракторів зменшує ймовірність вгадування правильної відповіді.

Як відомо [9], теорія СТТ, як і рання теорія факторного аналізу, будувалась К. Пірсоном на дихотомічних оцінках (завдання виконано – 1, не виконано – 0). Множину відповідей n-го випробовуваного на j-те завдання $\{a_{nj}\}$ записують у вигляді $A(N \times M)$ матриці, рядки, якої містять пул балів конкретних випробовуваних за всі завдання тесту, а стовпці відповідають тестовим завданням (індикаторам навченості). Елементи дихотомічної матриці розглядають як величини випадкові, що мають біноміальний розподіл[3], тобто набувають значення 1 з імовірністю P_{nj}, а значення 0 – з імовірністю $Q_{nj} = 1 - P_{nj}$. Їх математичне очікування та дисперсія дорівнюють:

$$E\{a_{nj}\} = 1 \cdot P_{nj} + 0 \cdot Q_{nj} = P_{nj}, \quad (1)$$

$$D\{a_{nj}\} = E\{a_{nj}^2\} - E^2\{a_{nj}\} = 1^2 \cdot P_{nj} + 0^2 \cdot Q_{nj} - P_{nj}^2 = P_{nj} \cdot Q_{nj}. \quad (2)$$

[1] Що ще більше звужується (до т. зв. домену [4]) внаслідок неможливості покриття всієї змістовної області тестовими завданнями.

[2] Завдання типу "відповідність", коли випробовуваному потрібно упорядкувати два списки таким чином, що б вони відповідали один одному, і "впорядкований список", коли випробовуваному потрібно упорядкувати список у певному порядку. Існують також завдання з градуйованими відповідями, тобто такими, що є більш-менш правильними.

[3] Розподіл величини a_{nj} наближається до нормального, якщо за 1 приймається факт досягнення певного рівня навченості.

Статистикою n-го випробовуваного у i-му тесті (індекс i для спрощення опущено), що асоціюється з мірою його успіху, вважають набраний ним первинний бал (Y_n), який зазвичай розраховують як суму балів, отриманих за всі завдання:

$$Y_n = \sum_{j=1}^{M} a_{nj} \ , \tag{3}$$

де $n = \overline{1, N}$; N – кількість випробовуваних;

$j = \overline{1, M}$; M – кількість завдань.

Первинні бали мають узагальнений біноміальний розподіл. Математичне очікування $E\{Y_n\}$ та дисперсія $s^2\{Y_n\}$ первинного балу кожного n-го випробовуваного складають [13]:

$$E\{Y_n\} = E\{\sum_{j=1}^{M} a_{nj}\} = \sum_{j=1}^{M} E\{a_{nj}\} = \sum_{j=1}^{M} P_{nj} = P_n \ , \tag{4}$$

$$s^2\{Y_n\} = \sum_{j=1}^{M} s_{jn}^2 + 2 \cdot \sum_{j<k} \rho_{jkn} \cdot s_{jn} \cdot s_{kn} = \sum_{j=1}^{M} s_{jn}^2 + 2 \cdot \sum_{j<k} E\{a_{nj} \cdot a_{nk}) - E\{a_{jn}\} \cdot E\{a_{kn}\} =$$

$$= \sum_{j=1}^{M} P_{nj} \cdot Q_{nj} + 2 \cdot \sum_{j<k} (P_{jkn} - P_{jn} \cdot P_{kn}) = M \cdot \sum_{j=1}^{M} P_{nj} - \sum_{j=1}^{M} P_{nj}^2 + 2 \cdot \sum_{j<k} (P_{jkn} - P_{jn} \cdot P_{kn}) \ , \tag{5}$$

де остання сума охоплює всі можливі пари завдань j і k, для яких $j < k$, і представляє коваріації всіх пар завдань;

ρ_{jkn} – коефіцієнт кореляції j-го і k-го завдань, виконаних n-м випробовуваним;

s_{jn}, s_{kn} – стандартні відхилення відповідей n-го випробовуваного відповідно на j-те та k-те завдання;

P_{jkn} – імовірність того, що n-й випробовуваний вирішив завдання j і k.

Величина P_n характеризує рівень підготовленості n-го випробовуваного. Її оцінкою служить відношення первинного балу (Y_n) до максимальної кількості балів ($Y_n^{(\max)}$), яку мав змогу набрати випробовуваний. За дихотомічного оцінювання $Y_n^{(\max)} = M$, тому маємо:

$$P_n \approx \frac{Y_n}{Y_n^{(\max)}} = \frac{Y_n}{M} \ . \tag{6}$$

З (5) видно, що дисперсія результатів тесту зростає зі збільшенням коваріацій завдань, а це означає, що випробовувані, які правильно (неправильно) відповідають на j-те завдання, так само правильно (неправильно) відповіли на k-те завдання, що можливо, коли завдання мають приблизно однакову важкість. Саме на цьому будують стратегію нормативно орієнтованих тестів. І навпаки, за наявністю завдань від легких до важких, що дають некорелюючі результати, дисперсія тесту буде найменшою, що й потрібно для критеріально орієнтованих тестів.

На практиці дуже часто різні ймовірності P_{nj} замінюють однією усередненою величиною:

$$P_n = \frac{1}{M} \cdot \sum_{j=1}^{M} P_{nj} \ . \tag{7}$$

Це спрощує розрахунки математичного очікування та дисперсії первинного балу:

$$E\{Y_n\} = M \cdot P_n \,, \tag{8}$$

$$s^2\{Y_n\} = M \cdot P_n \cdot Q_n \,, \tag{9}$$

але дисперсія первинного балу при цьому є максимальною з можливих [13]. При нехтуванні коваріаціями відповідей на завдання краще використовувати формулу (5).

За показник важкості j-го завдання у СТТ обирають частку правильних відповідей (p_j)[4], що наближує імовірність правильної відповіді:

$$p_j \approx w_j / N \,, \tag{10}$$

де N – чисельність групи випробовуваних;

w_j – число правильних відповідей на j-те завдання тесту, або первинний бал j-го завдання:

$$w_j = \sum_{n=1}^{N} a_{nj} \,. \tag{11}$$

Відповідно, показник легкості j-го завдання розраховують як частку неправильних відповідей (q_j) або як:

$$q_j = 1 - p_j \,. \tag{12}$$

Аналогічно можна оцінювати важкість та легкість комплексу завдань (p_{j+l}), наприклад, j-го та l-го:

$$p_{j+l} = (w_j + w_l) / N \; i \; q_{+,l} = 1 - p_{j+l} \,, \tag{13}$$

де w_l – число правильних відповідей на l-те завдання тесту.

Первинний бал завдання (w_j), як і первинний бал випробовуваного (Y_n), розподілений за узагальненим біноміальним законом. Його математичне очікування ($E\{w_j\}$) та дисперсія $(s^2\{w_j\})$ дорівнюють:

$$E\{w_j\} = p_j^{5} \text{ і } s^2\{w_j\} = p_j \cdot q_j \,. \tag{14}$$

Кореляцію результатів виконання пари завдань, оцінюваних дихотомічно, можна знайти (обмеження див. [4]) за формулою т. зв. *фі* – коефіцієнта Пірсона[6], що слідує з коефіцієнта кореляції добутку моментів:

$$\varphi_{j,l} = \frac{p_{j,l} - p_j \cdot p_l}{\sqrt{q_j \cdot p_j \cdot q_l \cdot p_l}} \,, \tag{15}$$

де $p_{j,l}$ – сумісна частка випробовуваних, що правильно відповіли на j-те та l-те завдання.

Зі зростанням кількості завдань та випробовуваних біноміальний розподіл первинних результатів швидко наближується до нормального, тому довірчі інтервали для Y_n і w_j можна знайти з використанням інтегральної теореми Лапласа [13]. Наприклад, імовірність попадання первинного балу випробовуваного у діапазон $Y_n \mp \Delta Y_n$ дорівнюватиме:

---

4 За логікою це мала б бути частка неправильних відповідей.

5 Те, що ця оцінка слідує з природи розподілу, а не є результатом обчислення середнього арифметичного, є великою перевагою дихотомічної шкали.

6 Має сенс використовувати за ручних розрахунків.

$$P_n\{(Y_n - \Delta Y_n) < E\{Y_n\} < (Y_n + \Delta Y_n)\} = 1 - \alpha \ , \tag{16}$$

де

$$Y_n - \Delta Y_n = \frac{M \cdot Y_n}{M + l^2} + \frac{l^2}{2} - l \cdot \sqrt{\frac{Y_n \cdot (M - Y_n)}{M} + \frac{l^2}{2}} \ ;$$

$$Y_n + \Delta Y_n = \frac{M \cdot Y_n}{M + l^2} + \frac{l^2}{2} + l \cdot \sqrt{\frac{Y_n \cdot (M - Y_n)}{M} + \frac{l^2}{2}} \ ; \tag{17}$$

l – значення аргументу функції Лапласа, за якого $2 \cdot \Phi(l) = 1 - \alpha$;

α – рівень значущості (за звичай 0,05).

Для первинного бала завдання формула є аналогічною, потрібно тільки замінити Y_n на w_j, а M на N.

Нормативно-орієнтовані тести були першими, що отримали фундаментальне статистичне обґрунтування, покладене в основу СТТ.

СТТ побудовано на базових постулатах [4; 11; 15]:

1. Результат вимірювання (Y_i) в i-му тесті у n-го випробовуваного (індекс n далі для спрощення опущено) деякої дійсної величини освіченості (T_i) обтяжений адитивною випадковою помилкою вимірювання (ε_i):

$$Y_i = E(Y_i) + \varepsilon_i = T_i + \varepsilon_i \ , \tag{18}$$

де $E(Y_i)$ – математичне очікування випадкової величини Y_i.

Величини T_i та ε_i зазвичай невідомі. Величину Y_i називають спостережуваним (індивідуальним) балом випробовуваного за i-ю тестовою формою, а T_i – його дійсним балом.

2. Дійсна та випадкова складові результату вимірювання не корелюють:

$$Cov(T_i, \varepsilon_i) = 0 \ , \tag{19}$$

тобто:

$$s^2(Y_i) = s^2(T_i) + s^2(\varepsilon_i) \ . \tag{20}$$

3. Математичне очікування помилки вимірювання дорівнює нулю:

$$E(\varepsilon_i) = 0 \ . \tag{21}$$

4. Помилки вимірювання двох будь-яких тестів не корелюють:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \ . \tag{22}$$

Окрім цього, в основу ССТ покладено два визначення – моделі паралельних і еквівалентних (не суворо паралельних) тестів (субтестів). Паралельні тести містять завдання, що мають подібний вміст з однієї й тієї самої загальної дидактичної одиниці та приблизно однакові важкість і варіацію результатів випробовуваних. Суворо кажучи (теоретично), вони мають відповідати вимогам (18) – (22), при цьому дійсні компоненти для кожної вибірки випробовуваних, що відповідають на обидва тести, та їх дисперсії повинні бути однаковими (т. зв. T-еквівалентність):

$$T_i = T_j \tag{23}$$

$$s^2(\varepsilon_i) = s^2(\varepsilon_j) \tag{24}$$

Згідно з припущенням (23), існує деяка однозначно визначена латентна (скритна, неспостережувана) змінна (характеристика) освіченості θ, яка відповідає умовам:

$$Y_i = \theta + \varepsilon_i , \tag{25}$$

$$E(\theta) = E(Y_i) , \tag{26}$$

$$s^2(\theta) = Cov(Y_i, Y_j), \ \ i \neq j , \tag{27}$$

$$s^2(\varepsilon_i) = s^2(Y_i) - Cov(Y_i, Y_j), \ \ i \neq j . \tag{28}$$

Таким чином, виходячи з припущень моделі паралельних тестів, шукана латентна змінна η та її статистичні характеристики можуть бути розраховані за параметрами розподілу як мінімум двох тестових оцінок.

Для моделі еквівалентних тестів постулюється, що вона має відповідати всім вимогам паралельних тестів за винятком одного – дійсні компоненти одного тесту не обов'язково повинні дорівнювати дійсним компонентам іншого паралельного тесту, але відрізнятися вони можуть на одну і ту ж саму константу (т. зв. істотна *T*-еквівалентність):

$$T_i = T_j + \lambda_{ij.} \tag{29}$$

де λ_{ij} – деяке число у лінійному метричному просторі.

При цьому: $$T_i = \lambda_{ij0} + \lambda_{ij1} \cdot T_j, \lambda_{ij0} > 0, \lambda_{ij1} > 0 , \tag{30}$$

де $\lambda_{ij0} > 0$, $\lambda_{ij1} > 0$,

тобто два тести, вимірюючи знання в одній і тій самій області, даватимуть лінійно пов'язані результати (т. зв. *T*-однорідність).

Іншими словами, модель *T*-однорідних тестів передбачає, що існує латентна змінна η, яка однозначно визначається перетворенням:

$$\eta = T_i + \lambda_i , \ \ \lambda_i \in R \tag{31}$$

і при цьому мають місце властивості (27) і (28).

Усі параметри моделі еквівалентних тестів можна визначити за результатами не менше трьох тестів, для яких виконуються припущення (29) і (30).

Якщо побудувати шкалу η так, щоб:

$$E(\eta) = 0 \ \ \text{і} \ \ s^2(\eta) = 1 , \tag{32}$$

то матимемо:

$$\lambda_{i1} = \sqrt{\frac{Cov(Y_i, Y_j) \cdot Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)}}, \ i \neq j; \ \ i \neq k; \ \ j \neq k , \tag{33}$$

$$s^2(\varepsilon_i) = s^2(Y_i) - \lambda^2_{i1} , \tag{34}$$

$$s^2(T_i)/s^2(Y_i) = 1 - s^2(\varepsilon_i)/s^2(Y_i) . \tag{35}$$

Праву частину виразу (35) ототожнюють з надійністю (Re*l*) тесту [4]:

$$\text{Re}l = 1 - s^2(\varepsilon_i)/ s^2(Y_i) = [s^2(Y_i) - s^2(\varepsilon_i)]/ s^2(Y_i) = s^2(T_i)/s^2(Y_i) = \rho_{TY} . \tag{36}$$

Тобто надійність тесту вимірюють коефіцієнтом кореляції дійсних і спостережуваних балів (ρ_{TY}) – чим він вище, тим точніше спостережуваний бал оцінює (генералізує) латентну змінну.

Знаючи надійність, можна оцінити стандартну помилку тестових вимірювань, тобто стійкість тестових результатів до дії випадкових чинників:

$$s(\varepsilon_i) = s(Y_i) \cdot \sqrt{1 - \text{Re}l_i} . \tag{37}$$

Якщо вважати, що випадкові помилки вимірювання (ε_i) розподілені за нормальним законом, то можна очікувати, що приблизно 68% спостережуваних оцінок випробовуваних знаходиться в інтервалі $T \pm s(\varepsilon_i)$, а 95% – в інтервалі $T \pm 2 \cdot s(\varepsilon_i)$.

Для інтервальної шкали та шкали відносин звичайним способом визначення Re*l* є розрахунок коефіцієнта парної кореляції Пірсона (добутку моментів) двох субтестових оцінок (Y_i та Y_k), що відповідає моделі паралельних тестів [11]:

$$\mathrm{Re}\, l = \hat{\rho}(Y_i, Y_k)$$

Статистичну значушість (відмінність від нуля) $\hat{\rho}(Y_i, Y_k)$ при кількості випробовуваних *(N)* менше 50 перевіряють за допомогою прямого та зворотного *Z*–перетворень Фішера. Пряме перетворення:

$$z = Z(\hat{\rho}) = \frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}}\ , \qquad (39)$$

забезпечує величині $\hat{\rho}$ стандартний нормальний розподіл, тому перевірка гіпотези щодо незначущості парного коефіцієнта кореляції ототожнюється з перевіркою гіпотези щодо незначущості величини z – H_0: $z = 0$ шляхом порівняння величини:

$$t(1-\alpha/2) = z/\sqrt{M-3}\ , \qquad (40)$$

з критичним значенням $t_{крит}$, представленим 1– α/2 квантилем стандартного нормального розподілу. Якщо $|t| > t_{кр}$, то гіпотеза щодо незначущості парного коефіцієнту кореляції відхиляється з імовірністю помилитися α.

Довірчий інтервал для величини $\hat{\rho}$ отримують за допомогою зворотного Z-перетворення Фішера. В термінах функцій MS Excel це виглядає так:

$$\gamma_{\hat{\rho}} = ФИШЕРОБР\{ФИШЕР(\hat{\rho}) \pm$$
$$\pm НОРМОБР(1-\frac{\alpha}{2})/\sqrt{M-3}] - \hat{\rho}/[2\cdot(M-1)]\}\ . \qquad (41)$$

Якщо M > 50, розрахунок довірчого інтервалу можна здійснювати за формулою:

$$\gamma_{\hat{\rho}} = \hat{\rho} \pm 2/\sqrt{M-1}\ . \qquad (42)$$

Надійність сумарної оцінки $S = Y_1 + ... + Y_m$ паралельних субтестів знаходять за формулою Спірмена–Брауна:

$$\mathrm{Re}\, l(S) = \frac{m\ .\ \mathrm{Re}\, l(Y_i)}{1+(m-1)\cdot \mathrm{Re}\, l(Y_i)}\ , \qquad (43)$$

де m – число паралельних субтестів у складі тесту.

Важливо, що, відстежуючи залежність Re*l* від *m*, можна визначити оптимальну кількість паралельних субтестів, а відтак, і загальну кількість завдань (*M*) для досягнення потрібного значення Re*l*(*S*).

Для еквівалентних субтестів формула (24) не прийнятна. Їхню надійність оцінюють як:

$$\mathrm{Re}l = Cov(Y_i, Y_j)/s^2(Y_i),\ i \neq j\ . \qquad (44)$$

Для істотно *T*-еквівалентних субтестів у якості нижньої межі надійності їхньої сумарної оцінки $S = Y_1 + ... + Y_m$ приймають коефіцієнт α – Кронбаха:

$$\mathrm{Re}l \geq \alpha = \frac{M}{M-m} \cdot \left(1 - \frac{\sum_{i=1}^{M/m} s^2(Y_i)}{s^2(S)} \right), \qquad 45)$$

де $s^2(Y_i)$ – дисперсія результатів виконання i-го тесту;

$s^2(S)$ – дисперсія сумарної тестової оцінки.

Чим вище коефіцієнт α, тим більше частка дисперсії тестових оцінок, що пояснюється латентною змінною.

За результатами паралельних тестів можна побудувати лінійну регресію первинних балів, отриманих випробовуваними у першому субтесті (Y_i), на результати другого субтесту (Y_j), рівняння якої наближатиме дійсні оцінки випробовуваних [4; 13; 16; 17]:

$$\hat{Y}_i - \overline{Y}_i = \hat{\rho}_{Y_iY_j} \cdot \frac{s_{Y_i}}{s_{Y_j}} \cdot (Y_j - \overline{Y}_j), \qquad (46)$$

де $\hat{Y}_i$ – прогнозоване (теоретичне) значення, розраховане за рівнянням регресії, для відповідного значення Y_i;

$\hat{\rho}_{Y_iY_j}$ – вибірковий коефіцієнт лінійної кореляції Пірсона;

s_{Y_i} і s_{Y_j} – стандартні відхилення відповідних субтестів.

У термінах методу найменших квадратів (МНК) рівняння (46) має вигляд:

$$\hat{Y}_i = \hat{a}_1 \cdot Y_j + \hat{a}_0, \qquad (47)$$

в якому вільний член $\hat{a}_0$ легко зводиться до нуля, якщо первинні бали перевести у Z-шкалу (див. нижче).

У роботах [4; 17] доводиться, що величина $\hat{Y}_i$ є точковою оцінкою латентної змінної (T_i), а довірчі інтервали регресії з визначеною імовірністю включають цю величину. Крім того, описаний прийом є ефективним для перерахунку в єдину шкалу балів, отриманих за тестами, тотожними за тематикою, але різної важкості (T-еквівалентними) [13].

Продовження статті читайте в наступному номері.

Список використаних джерел

1. Закон України "Про вищу освіту" від 01.07.2014 № 1556-VII. – ст.1. – [Електронний ресурс]. – Режим доступу : http://zakon4.rada.gov.ua/laws/show/1556-18
2. Морев И. А. Образовательные информационные технологии. Часть 5: Методическая система стимулирования обучаемости средствами дидактического тестирования : [монография] / И. А. Морев. – Владивосток : Изд-во Дальневост. ун-та, 2004. – 120 с.
3. Ефремова Н. Ф. Тестовый контроль в образовании : [учебное пособие] / Н. Ф. Ефремова. – М. : Университетская книга; Логос, 2005. – 368 с.
4. Крокер Л. Введение в классическую и современную теорию тестов : [учебник] / Л. Крокер, Дж. Алгина ; пер. с англ. Н. Н. Найденовой, В. Н. Смилкина, М. Б. Челышковой ; под общ. ред. В. И. Звонникова, М. Б. Челышковой. – М. : Логос, 2010. – 668 с.
5. Литвак Б. Г. Экспертная информация: Методы получения и анализа / Б. Г. Литвак. – М. : Радио и связь, 1982. – 184 с.
6. Орлов А. И. Нечисловая статистика / А. И. Орлов. – М. : МЗ-Пресс, 2004. – 513 с.
7. Толстова Ю. Н. Измерение в психологи : [учебное пособие] / Ю. Н. Толстова. – М. : КДУ, 2007. – 288 с.

8. Гусев А. Н. Измерение в психологии : [общий психологический практикум] / А. Н. Гусев, Ч. А. Измайлов, М. Б. Михалевская. – [2-е изд.]. – М. : Смысл, 1998. – 286 с. – (Серия «Практикум». Вып. 2).
9. Глас Дж. Статистические методы в педагогике и психологии / Дж. Глас, Дж. Стэнли. – М. : Прогресс, 1976. – 495 с.
10. Звонников В. И. Современные средства оценивания результатов обучения : [учеб. пособие для студ. высш. учеб. заведений] / В. И. Звонников, М. Б. Челышкова. – М. : Издательский центр «Академия», 2007. – 224 с.
11. Steyer R. Classical (Psychometric) Test Theory [Electronic resource] / R. Steyer. – Access Mode : http://metheval.uni-jena.de/materialien/ publikationen/ctt.pdf/
12. Rash G. On Objectivity and Specificity of the Probabilistic Basis for Testing [Electronic resource] / G. Rash. – Access Mode : http://www.rasch.org/memo196x.pdf
13. Нейман Ю. М. Введение в теорию моделирования и параметризации педагогических тестов / Ю. М. Нейман, В. А. Хлебников. – М. : Прометей, 2000. – 168 с.
14. Булах І. Є. Створюємо якісний тест : [навч. посіб.] / І. Є. Булах, М. Р. Мруга. – К. : Майстер-клас, 2006. – 169 с.
15. Модели тестирования знаний и методы оценки надежности полученных результатов / О. Ю. Чередниченко, С. И. Ершова, О. В. Янголенко, Т. Н. Запорожец // Восточно-европейский журнал передовых технологий. – 2011. – № 6/4 (58). – С. 35–40.
16. Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных / А. Д. Наследов. – М. : Речь, 2003. – 400 с.
17. Как составить тест // Слойер К. Математические фантазии. – М. : Мир, 1993. – С. 116–118.

М. Е. СИНИЦКИЙ,
кандидат физ.-мат. наук, доцент,
доцент кафедры информационных систем и технологий,
Национальная академия статистики, учета и аудита

Статистические инструменты измерения качества образования

Представлен обзор статистических основ тестологии. Описаны базовые задачи, математические модели и основные расчетные формулы классической (CTT) и современной (IRT) теорий тестирования, которые позволяют со статистических позиций оценить правильность построения, разрешающая способность, стандартную ошибку и надежность тестовых измерений.

Ключовые слова: *качество, образование, тестирование, шкалирование, наблюдаемый бал, корреляция, надежность, разрешающая способность, однородность, латентная переменная, модель Раша, характеристические кривые, логит.*

M. E. SINYTSKYI,
PhD (Phys.-Math.), Associate Professor,
Associate Professor of Department for Information Systems and Technologies,
National Academy of Statistics, Accounting and Audit

Statistical Tools for Measuring the Quality of Education

The article presents an overview of the statistical grounds of testology. The purpose of this paper is to explain the inexperienced readers, such as teachers of economic disciplines, opportunities of improvement of quality of education with the utilization of objective and impartial tools of students' achievement measurement, known as tests.

The first part of the article overviews the shortcomings of the traditional system of evaluation of educational achievements that is built around the use of ordinal scales. The

limitations imposed to the possibilities of statistical processing of the raw data by the type of scale are shown. Basic tasks, the corresponding mathematical models, statistical characteristics and sub-test score evaluation reliability formulas are described.

The second part of the article describes approaches to the determination of reliability, uniformity and resolution of the test, built on the analysis of the correlation between students' answers to the identical questions asked. Options of conversion of primary points to a quantitative scale are provided. Ways of lowering the probability of correctly guessed answers are shown. The approach to processing of results of complex test structures is given and the possibility of utilization of two-factor analysis of variance (2 Way ANOVA) for dichotomous tests reliability estimation is demonstrated.

The third and the fourth parts of the article are devoted to the modern theory of tests (IRT).

The third part provides an analysis of shortcomings of the CTT, which were the main focus of efforts to overcome of IRT supporters during the last 60 years. The theoretical basis for building a Rach model and its subsequent developments is described. The methodology of estimation of properties of the test by its characteristic curves and parameters of its information function is illustrated. The basic equation, the correspondent solution of which gives an estimate of the probability of obtaining a certain personal score of a test is formulated. The fourth part of the article provides various methods of finding a solution of the basic equation for the 1PL and 2PL – models and data preparation for a correct use. Several software packages, both considered to be classical tools as well as brand new ones, are overviewed. An example of ranking of NASOA students' achievements obtained by traditional evaluation and IRT approach is given.

Keywords: *quality, education, testing, scaling, observed scores, correlation, reliability, resolution, uniformity, latent variable model of Rush, characteristic curves, logit.*