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 2. 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, тестування, шкалювання, спостережуваний бал, кореляція, надійність, роздільна здатність, однорідність, латентна змінна, модель Раша, характеристичні криві, логіт.*

При побудові нормативно-орієнтованих тестів велике значення приділяється вивченню не тільки субтестових зв'язків, а й кореляцій між окремими тестовими завданнями, а також між ними та індивідуальними тестовими оцінками). Оскільки первинні змінні тут можуть представляти різні шкали, коефіцієнти кореляції між ними мають обчислюватися за відповідними формулами [9; 13][1].

Аналіз кореляційних матриць результатів виконання тестових завдань дозволяє визначити невдалі завдання, що негативно корелюють з більшістю інших завдань, а відтак вимірюють дещо інше, ніж латентну змінну, для вимірювання якої призначено тест. Дуже часто причиною такої ситуації є відсутність предметної чистоти вмісту завдань. Такі завдання мають видалятись із тесту [13].

Часто як альтернативу (38), (39) (див. Ч. 1[2]) надійність тесту визначають через середній коефіцієнт кореляції ($\bar{\rho}$) між усіма (M) завданнями тесту:

$$\mathrm{Re}\, l = \frac{M \cdot \bar{\rho}}{1 + (M-1) \cdot \bar{\rho}}, \qquad (48)^3$$

де $$\bar{\rho} = \frac{1}{M \cdot N} \cdot \sum_{j=1}^{M} \sum_{n=1}^{N} \rho_{jn},$$

ρ_{jn} – j n-й елемент кореляційної матриці.

Найпростішим способом оцінювання дисперсії випадкової складової $s^2(\varepsilon)$ у моделі паралельних тестів є розрахунок дисперсії різниці результатів за тотожними завданнями (метод Рюлона). Однак неминуча розбіжність завдань спотворює результат.

Кронбах вирішив цю задачу та довів, що надійність дихотомічних тестів можна обчислити за формулою Кьюдера–Річардсона:

$$\mathrm{Re}\, l = \frac{M}{M-m} \cdot \left(1 - \frac{\sum_{j=1}^{M/m} p_j \cdot q_j}{s_Y^2} \right). \qquad (49)$$

[1] Нумерацію джерел наведено у продовження частини 1 статті, надрукованої у № 4 за 2014 рік.
[2] Частину 1 статті надруковано у № 4 за 2014 рік.
[3] Нумерацію формул наведено як продовження їх нумерації у частині 1 статті, надрукованій у № 4 за 2014 рік.

Як правило, для однорідних тестів оцінки надійності, отримані з використанням методів Рюлона та Кьюдера–Ричардсона, відрізняються несуттєво. Оцінки надійності, що лежать у діапазоні $0{,}70 \le \mathrm{Re}l \le 0{,}79$, вважаються задовільними; $0{,}80 \le \mathrm{Re}l \le 0{,}89$ – добрими; $0{,}90 \le \mathrm{Re}l$ – відмінними.

Однорідність тесту означає, що за нульовою гіпотезою математичні очікування різниць балів, отриманих за кожну пару паралельних завдань, мають бути рівними [13]:

$$H_0 : E\{\Delta_j^{(1)}\} = E\{\Delta_j^{(2)}\}, \quad j = \overline{1,(M/m)-1}, \tag{50}$$

де $\Delta_j^{(v)}$ – різниця балів двох суміжних завдань: j-го та $(j+1)$-го, у v-му варіанті субтесту;

M/m – кількість завдань у кожному субтесті.

Конкретно:

$$\frac{\Delta_j^{(2)} - \Delta_j^{(1)}}{s(\Delta_j^{(2)} - \Delta_j^{(1)})} = \frac{\Delta_j^{(2)} - \Delta_j^{(1)}}{\sqrt{s^2(\Delta_j^{(1)}) + s^2(\Delta_j^{(2)})}} \approx u_{1,2}^{(j)} \in \mathcal{N}(0,\,1), \tag{51}$$

де $\mathcal{N}(0,1)$ – нормальний розподіл стандартизованої та нормованої випадкової змінної, що має нульове математичне очікування та одиничну дисперсію.

Для тесту в цілому:

$$\sum_{j=1}^{M-1} \frac{(\Delta_j^{(2)} - \Delta_j^{(1)})^2}{s^2(\Delta_j^{(1)}) + s^2(\Delta_j^{(2)})} = \chi^2_{M-1}, \tag{52}$$

де χ^2_{M-1} – квантіль розподілу χ – квадрат з $(M-1)$ степенями свободи.

За умови $\chi^2_{M-1} \le \chi^2_{крит}(\alpha;\nu = M-1)$, де α – рівень значущості; ν – число степенів свободи для критичного χ^2 – гіпотеза щодо однорідності завдань тесту приймається. Якщо ця умова не виконується, можна знайти неоднорідні завдання шляхом перевірки виконання (51) для кожної пари завдань.

Формули (51), (52) справедливі для даних, віднесених до метричної шкали. Для дихотомічних оцінок розрахунки проводяться за формулами:

$$\sqrt{\frac{w \cdot w_1 \cdot w_2}{w_1 \cdot w_2}} \cdot \left(\frac{w_1^{(1)}}{w_1} - \frac{w_1^{(2)}}{w_2} \right) = u_{1,2}^{(1)} \in \mathcal{N}(0,\,1), \tag{53}$$

де $w_1^{(1)}$ і $w_1^{(2)}$, $w_2^{(1)}$ і $w_2^{(2)}$ – кількість правильних відповідей, відповідно, на перше та друге завдання, відповідно, у першій та другій половинах тесту;

$$w_1^{(1)} + w_1^{(2)} = w_1; \quad w_2^{(1)} + w_2^{(2)} = w_2; \quad w_1 + w_2 = w.$$

Для тесту в цілому сума квадратів величин типу (53) всіх варіантів пар завдань розподілена за законом χ^2:

$$\sum_{l=1}^{m-1} \sum_{j=1}^{M/m} [u_{l,l+1}^{(j)}]^2 = \chi^2_{\nu}, \tag{54}$$

де m – кількість паралельних субтестів;

$\nu = (M-1)\cdot(m-1)$ – число степенів свободи.

Завдання тесту вважаються однорідними на рівні значущості (α) за виконання умови $\chi^2_{\nu} \le \chi^2_{крит}(\alpha,\nu)$. Якщо ця умова не виконується, для знаходження причини цього слід перевірити всі пари завдань на виконання умови: $u_{l,l+1}^{(j)} \le u_{крит}$, де

$u_{крит}$ – аргумент функції Лапласа, яка дорівнює $(1-\alpha)/2$. Його можна обчислити, наприклад, за допомогою функції Excel НОРМСТОБР.

Роздільну здатність простого тесту (індекс дискримінаційності) в найпростішому варіанті оцінюють за т. зв. методом "контрастних груп". Порівнюється частка правильних відповідей (p_{1j}) у частини випробовуваних, які посіли перші місця в рейтингу (сильні), та (p_{0j}) – у частини випробовуваних аналогічної чисельності, які посіли останні місця у рейтингу (слабкі):

$$rd_j = p_{1j} - p_{0j}, \tag{55}$$

де rd_j – індекс дискримінаційності j-го завдання.

Для завдання, з яким впоралися всі сильні випробовувані та не впорався ніхто зі слабких, індекс дискримінаційності дорівнює 1. Таке завдання має максимальний роздільний ефект. Завдання, для яких $0 \le rd \le 0{,}2$, рекомендують уточнити, а для яких $rd < 0$, рекомендують вилучити з тесту.

Іншим способом оцінювання дискримінаційної здатності тестового завдання є розрахунок точкового коефіцієнта бісеріальної кореляції (c_{pb}) [4; 13]. Він використовується у тих випадках, коли один набір значень заданий у дихотомічній шкалі, а інший – в інтервальній (метричній). Саме таку ситуацію маємо при розрахунку зв'язку між результатами виконання кожного завдання (дихотомічна шкала) та первинним балом випробовуваного (інтервальна шкала):

$$\rho_{pb} = \frac{\overline{Y_1} - \overline{Y_0}}{s_Y} \cdot \sqrt{\frac{N_1 \cdot N_0}{N \cdot (N-1)}}, \tag{56}$$

де $\overline{Y_1}$ і $\overline{Y_0}$ та $\boldsymbol{N}_1$ і $\boldsymbol{N}_0$ – середні первинні бали та чисельність випробовуваних, що впоралися та, відповідно, не впоралися із завданням;

s_Y – стандартне відхилення для індивідуальних балів усіх випробовуваних;

$\boldsymbol{N}_1 + \boldsymbol{N}_0 = \boldsymbol{N}$.

Чим вище c_{pb}, тим більш ефективно тестове завдання розділяє випробовуваних за їхньою підготовкою.

Той факт, що наведені формули в більшості задовольняють нормальному розподілу, висуває певні вимоги щодо складності[4] завдань.

Вміст тесту не може бути тільки легким, середнім або важким, він має бути однорідним. Переважна кількість легких завдань утворює ілюзію наявності знань у випробовуваних, тому що ними перевіряються мінімальні знання. Крива розподілу балів у цьому випадку матиме правобічну (від'ємну) асиметрію. З іншого боку, за наявності переважної кількості завдань підвищеної важкості переважатиме лівостороння (позитивна) асиметрія. Переважання середніх за складністю завдань дає найближчий до теорії результат, але зменшує валідність тесту.

Характеристику симетрії функцій розподілу первинних балів звичайно використовують разом із показником ексцесу. Додатне значення ексцесу свідчить на користь малої роздільної здатності тесту, а від'ємне – великої. В комплексі асиметрія та ексцес визначають критерії валідності результатів тестування [14]. Якщо завдання має середній рівень важкості та високу роздільну здатність, то воно придатне для включення до нормативно-орієнтованого тесту.

Існують рекомендації щодо вибору складності завдань за коефіцієнтом серіальної кореляції [4]. Для інтервалу $0{,}3 \le \rho_{pb} \le 0{,}4$ рівень важкості завдань має бути в інтервалі від 0,4 до 0,6. При значеннях $\rho_{pb} \ge 0{,}6$ можна застосовувати більш широкий діапазон складності.

---

[4] У спеціальній літературі, присвяченій тестуванню, розрізнять поняття "важкість" та "складність". Ступінь складності учбового матеріалу характеризується реальною (об'єктивною) насиченістю учбового завдання і формою його викладення, а ступінь важкості означає співвідношення викладеного учбового матеріалу до раніше засвоєного учбового матеріалу та інтелектуальних можливостей учнів.

Зрозуміло, що в середньому вміст традиційного тесту має істотним чином варіювати залежно від рівня підготовленості групи, на вимірювання знань якої спрямований тест. Дійсно, за відсутності точки абсолютного відліку один й той самий випробуваний може виглядати по-різному на тлі сильної або слабої групи. У цьому полягає суттєвий недолік використання в якості результатів тестування первинних балів.

Були спроби вирішити це питання шляхом використання шкали первинних процентилів, у якій первинні бали відповідають квантилям емпіричного розподілу первинних балів. Але оскільки розподіл процентилів є прямокутним, цей підхід не набув поширення [10].

Певною точкою відліку для порівняння може служити математичне очікування середнього значення первинних балів, отриманих групою випробовуваних, а виставленим балам – абсолютна відстань персональних балів від середнього арифметичного. На цьому основана конвертація первинних балів у т. зв. стандартну Z-шкалу, або процес лінійної стандартизації.

Межі інтервалів первинних балів для тестової Z-шкали визначають таким чином:

$$Z_i = \frac{Y_i - \overline{Y}}{s_Y}; \quad \frac{Z_i - \overline{Z}}{s_Z} \Rightarrow Y_i = \overline{Y} + \frac{s_Y}{s_Z} \cdot (Z_i - \overline{Z}), \tag{57}$$

де Y_i – шукана межа інтервалу первинних балів;

$\overline{Y}$ – середній арифметичний[5] первинний бал групи з N випробовуваних;

s_Y – стандартне відхилення первинного балу випробовуваного;

Z_i – межа інтервалу у стандартній тестовій шкалі;

s_Z – стандартне відхилення у стандартній тестовій шкалі;

$\overline{Z}$ – середнє тестової шкали.

Відомо, що лінійні перетворення, у тому числі (57), не змінюють тип первинного розподілу, тому якщо первинні бали мали нормальний розподіл, то у Z-шкалі вони матимуть стандартний нормальний розподіл $\mathcal{N}(0, 1)$, тобто такий, що:

$$\overline{Z} = 0, \qquad s_Z^2 = 1. \tag{58}$$

Таким чином, для первинного балу, що точно відповідає середньому значенню групи, Z-оцінка дорівнюватиме нулю. Від'ємні її значення вказують на результати нижче середнього, а позитивні – на такі, що знаходяться вище середнього значення первинних балів у групі.

Z-оцінки легко перерахувати у відсотки та інтерпретувати у термінах процентилів (т. зв. шкала нормалізованих процентилів). У таких шкалах пентилі утворюють 5-бальну шкалу, децилі – 10-бальну, а сентилі – 100-бальну.

Межі Z-оцінок складають практично від – 3 до +3, що замало для забезпечення високої роздільної здатності тесту і призводить до потреби ведення розрахунків до другого знаку після коми. Крім того, від'ємні числа не дуже прийнятні для використання. Тому у практиці більш поширеними є шкали, отримані шляхом додаткового лінійного перетворення первинної Z-шкали:

$$Z_1 = \overline{Z}_1 + s_{Z_1} \cdot Z, \tag{59}$$

де $\overline{Z}_1$ – нове (бажане) значення середнього;

s_{Z_1} – нове (бажане) значення стандартного відхилення.

Як приклад можна навести шкали:

– 100-бальну $\mathcal{N}(50, 10^2)$ T- шкалу: $Z_T = 50 + 10 \cdot Z$; (60)

– 200-бальну $\mathcal{N}(100, 15^2)$ шкалу IQ: $ZIQ = 100 + 15 \cdot Z$; (61)

– 1000-бальну $\mathcal{N}(500, 100^2)$ шкалу $CEEB$: $Z_{CEEB} = 500 + 100 \cdot Z$. (62)

[5] Тут маємо протиріччя з репрезентативною теорією вимірювань.

Коли реальні розподіли первинних балів не принципово відрізняються від нормального, існує можливість підігнати їхню криву розподілу під нормальну (емпірична стандартизація). Для цього трансформацію первинних балів у стандартні здійснюють шляхом знаходження процентильних меж груп у первинному розподілі, які відповідають межам груп у нормальному розподілі стандартної шкали. Алгоритм: для кожного первинного показника (індивідуального балу) визначають суму його кумульованої частоти та половини кількості випробовуваних, які отримали відповідний бал. Далі цю суму ділять на кількість усіх випробовуваних у групі та, використовуючи її як аргумент функції Лапласа, знаходять (наприклад, за допомогою функції Excel НОРМСТОБР) відповідну частку площі під нормальною стандартною функцією розподілу, що відповідає цьому балу.

Залежно від середнього значення та масштабу (стандартного відхилення), додатково до (60)–(62) використовують такі тестові шкали з відкритими крайніми інтервалами [16]:

- шкалу стенайнів – 9 інтервалів по $0{,}5 \cdot s_Z$ ($\overline{Z} = 5$);
- шкалу стенів (Кетела) – 10 інтервалів по $0{,}5 \cdot s_Z$ ($\overline{Z} = 5{,}5$);
- 11-бальну шкалу – 11 інтервалів по $0{,}5 \cdot s_Z$ ($\overline{Z} = 6$).

Ці шкали вважаються більш придатними для використання у традиційних екзаменах, а шкали (60)–(62) – для застосування у комп'ютерних тестах. Повторимося, що за сутністю всі стандартні шкали є порядковими, і зіставляти виміряні за ними бали можна з натяжкою, опираючись на загальний принцип їх побудови.

Якщо ж розподіл первинних балів принципово не є нормальним, то він залишиться таким і у отриманих лінійним перетворенням Z-оцінок. Тобто при використанні Z-шкали для критеріально-орієнтованих тестів усі властивості оцінок мають будуватись на непараметричних статистиках, що менш точно, ніж при використанні параметричних статистик (наприклад, нормального розподілу).

Надійність тестів знижується зі зростанням помилки вимірювання. Однією з характеристик останньої є ймовірність вгадування правильної відповіді. Як вже зазначалося, для тестів закритого типу вона зменшується зі збільшенням числа дистракторів. Ймовірність вгадування правильної відповіді розраховують за формулою [17]:

$$P = \sum_{i=\left[\frac{Q \cdot M}{100}\right]+1}^{M} C_M^i \left(\frac{1}{n}\right)^i \left(1-\frac{1}{n}\right)^{M-i}, \qquad (63)$$

де M – число питань у тесті;

Q – відсоткова межа кількості правильних відповідей, досягнення якої дозволяє вважати тестування успішно пройденим;

n – число відповідей, пропонованих на кожне завдання.

У разі, коли кількість пропонованих відповідей на питання в різних завданнях не однакова, формула (63) має більш складний вигляд:

$$P = \sum_{j=\left[\frac{Q \cdot M}{100}\right]+1}^{M} \overline{P_j}, \qquad (64)$$

де $\overline{P_j}$ – ймовірність вгадування відповідей на j питань, яка обчислюється таким чином:

$$\overline{P_j} = \sum_{t_1=1}^{n_1} \cdots \sum_{t_{r-1}=1}^{n_{r-1}} C_{n_1}^{t_1} P_1^{t_1} (1-P_1)^{n_1-t_1} \cdot \ldots \cdot C_{n_{r-1}}^{t_{r-1}} P_{r-1}^{t_{r-1}} (1-P_{r-1})^{n_{r-1}-t_{r-1}} \cdot C_{n_r}^{t_r} P_r^{t_r} (1-P_r)^{n_r-t_r}, \qquad (65)$$

де r – число груп, що об'єднують питання з однаковою ймовірністю вгадування P_i, $0 < P_i < 1$;

n_i – кількість питань в i-й групі $(i = \overline{1,r})$, причому $\sum_{i=1}^{r} n_i = M$;

$t_r = j - (t_1 + t_2 + \ldots + t_{r-1})$; якщо $t_r > k_r$, то вважається, що $C^{t_r}_{k_r} = 0$.

Формули (63)–(65) дозволяють визначити, скільки відповідей слід пропонувати на кожне питання при складанні тесту, щоб імовірність успішно пройти тестування, вгадавши правильні відповіді, була P. Наприклад, для $M = 10$ і $Q = 0{,}75$: $k = 2$, $P < 0{,}2$; $k = 3$, $P < 0{,}02$; $k = 4$, $P < 0{,}004$. За норму беруть $P < 0{,}05$.

Щоб мотивувати випробовуваних до відмови від вгадування, можна нараховувати бали згідно з тією або іншою спеціальною процедурою. Наприклад, за вірну відповідь ставити +1 бал, за невірну відповідь –1 бал, а за відмову від відповіді 0 балів.

Одним зі шляхів вирішення проблеми вгадування вважають введення поправки на вгадування. Теоретичне обґрунтування розрахунку поправок на вгадування [3; 18] у СТТ будують на припущеннях, що середня кількість відповідей у тестовому завданні складає $k = 3\ldots5$, а значення $p = 0{,}3\ldots0{,}2$ повністю обумовлені вгадуванням, тобто частка вгаданих відповідей приблизно дорівнює $p_0 = 1/k$. Однак вважати, що випробовувані, які отримали високі індивідуальні бали, так само вгадували, як слабо підготовлені, є некоректним. Тому застосовують нелінійну залежність поправки від частки q неправильних відповідей:

$$Y = \left[p - \frac{1}{k} \cdot \left(\frac{k \cdot q}{k-1} \right)^d \right] \cdot M, \qquad (66)$$

де d – параметр, що визначає тип моделі поправки на вгадування.

Звичайно використовують $d = 1$ чи 2. Від'ємні значення Y замінюють нулем.

Іншим засобом боротьби із вгадуванням є використання серії завдань з альтернативами для одного елементу знань. У цьому випадку завдання складається із серії декількох завдань з вибором. Завдання вважається виконаним, якщо на всі завдання в серії отримана вірна відповідь.

Найкраще захищені від вгадування завдання з множинними правильними відповідями (політомічні тести). Але оцінювання виконання завдання з множинними правильними відповідями є більш складним, ніж оцінювання завдання з вибором однієї вірної відповіді. У роботі [10] за повністю виконане завдання з вибором декількох вірних відповідей пропонують давати 1 бал, і 0 балів, якщо має місце принаймні одна невірна відповідь. У [19] пропонують використовувати штрафи, тобто за повністю правильне рішення давати 3 бали, за кожну помилку знімати один бал. Якщо помилок більше трьох, то давати 0 балів. У результаті процедура об'єктивного вимірювання результатів тестування замінюється процедурою суб'єктивної ідентифікації цих результатів за правилами, встановленим викладачем або розробником тесту. Це веде до зниження точності вимірювань такого тесту.

В роботах [20–22] наведено процедури призначення первинних балів на основі теорії розпізнавання образів. При цьому форми закритих тестових завдань ("множинний вибір", "відповідність" і "впорядкований список") і похідні від них форми зведено до двох базових форм відповідей: "безліч" (неврегульована множина елементів) або "список" (впорядкована множина).

Для відповідей типу "безліч" отримано міру, яка визначає різницю між множиною, що характеризує відповідь випробовуваного, та множиною, що характеризує еталонну відповідь:

$$d = 1 - \frac{k}{w + v - k}, \qquad (67)$$

де k – число збігів відповідей випробовуваного з еталонними відповідями;
w – число відповідей типу "Так", наданих випробовуваним;
v – число відповідей типу "Так" у еталонному пулі відповідей.

Квадрат величини (67) автори [20] називають абсолютною невпорядкованістю відповіді випробовуваного:

$$Q = d^2 . \tag{68}$$

Перехід до B-бальної шкали здійснюється за формулою:

$$B = \log_2 \frac{S}{Q} = \log_2 \frac{2^B \cdot Q_{вм}}{Q} , \tag{69}$$

де S – коефіцієнт, що залежить від числа градацій шкали та максимальної оцінки (верхньої межі) $Q_{вм}$, що не спричиняє зниження балу оцінки [21]. Наприклад, для 5-ти бальної традиційної шкали $Q_{вм}$ – величина, що відповідає відмітці "5" – "відмінно".

Перехід між шкалами балів ($B_1 \rightarrow B_2$) можливий за формулою:

$$B_2 = \log_2 \left(\frac{S_2}{S_1} \cdot 2^{B_1} \right) . \tag{70}$$

Для відповідей типу "список" невпорядкованість відповіді визначається як:

$$Q = 1 - (1 - Q_1) \cdot (1 - Q_2), \tag{71}$$

де Q_1 – оцінка невпорядкованості відповіді випробовуваного відносно еталону за номенклатурою елементів; розраховується з використанням формул (67), (68);

Q_2 – оцінка невпорядкованості відповіді випробовуваного відносно еталону за впорядкованістю елементів, яка будується таким чином:

- якщо випробовуваний включив у відповідь дистрактори, їхнім елементам у списку відповідей присвоюється однакове значення λ;
- якщо число відповідей, обраних випробовуваним, більше числа елементів еталонного списку, то еталонний список доповнюється з кінця числами λ у кількості, потрібній для того, щоб урівняти його за числом елементів з числом відповідей випробовуваного;
- якщо число відповідей, обраних випробовуваним, більше числа елементів еталонного списку, то еталонний список доповнюється з кінця числами λ у кількості, потрібній для того, щоб урівняти його за числом елементів x відповіддю;
- якщо число відповідей, обраних випробовуваним, більше числа елементів еталонного списку, то еталонний список доповнюється з кінця числами λ у кількості, потрібній для того, щоб урівняти його за числом елементів x відповіддю;
- якщо число відповідей, обраних випробовуваним, менше числа елементів еталонного списку, то список, сформований випробовуваним, доповнюється з кінця числами λ у кількості, потрібній для того, щоб урівняти його за числом елементів з еталоном;
- вирівняні за кількістю списки поелементно порівнюються:

$$\Delta_{lk}^{(i)} = \begin{cases} +1 , & якщо \quad a_{il} > a_{ik} ; \\ -1 , & якщо \quad a_{il} < a_{ik} , \; l < k ; \\ 0 , & якщо \quad a_{il} = a_{ik} , \end{cases} \tag{72}$$

де i – номер завдання;
l – номер елемента у списку відповідей випробовуваного;
k – номер елемента у списку еталонних відповідей;
- розраховується "відстань" між списками за Кендаллом:

$$d = \frac{1}{2} - \frac{1}{M \cdot (M-1)} \cdot \sum_{l<k} \Delta_{lk}^{(i)} \cdot \Delta_{lk}^{(j)} , \qquad (73)$$

де M – кількість завдань

– обчислюють $Q_2 = d^2$.

Комп'ютерні технології відкривають можливості формування тестових завдань поточного тесту фіксованої довжини шляхом вибору з бази завдань великого обсягу. Якщо вибір є випадковим, а завдання близькі за вмістом, то йдеться про випадкові паралельні тести. Відповідно оцінка дисперсії j-ї первинної оцінки для всієї сукупності паралельних дихотомічних тестів дорівнює:

$$s_Y^2 = \frac{Y_j \cdot (M - Y_j)}{M-1} . \qquad 74)$$

На відміну від нормативно-орієнтованих тестів критеріально-орієнтовані тести не можуть бути суто паралельними, тобто такими, що мають однакові дисперсії або кореляції з генеральними оцінками [4]. Це ускладнює визначення надійності даного типу тестів. А триразове проведення одного і того ж самого тесту, як цього вимагає теорія, важко реалізувати. У роботі [4] наведено підхід, що дозволяє оцінювати надійність критеріально-орієнтованих тестів різної форми (дизайну) з меншими витратами. Підхід ґрунтується на т. зв. теорії "генералізації", розробленій для прийняття рішень на основі експертних оцінок[6]. У термінах цієї теорії коефіцієнт надійності тесту визначають як т. зв. коефіцієнт генералізації. Наприклад, для тесту, що містить однакові завдання множинного вибору для всіх випробовуваних ("однофасетний дизайн № 2"), він складає:

$$\rho_{\bar{Y}}^2 = \frac{s_T^2}{s_T^2 + (s_j^2 + s_e^2)/M} , \qquad (75)$$

де s_T^2 – дисперсія дійсних (генеральних) оцінок випробовуваних, що представлені математичним очікуванням $E(T_j)$;

s_j^2 – дисперсія складностей завдань;

s_e^2 – дисперсія помилки вимірювання за усіма завданнями.

Символ $\bar{Y}$ у позначенні $\rho_{\bar{Y}}^2$ означає, що дисперсію визначають для середнього арифметичного балів, отриманих за всі завдання. Внаслідок цього вона зменшується у M разів.

Для обчислення складових формули (75), слідуючи Хойту, можна використати двофакторний дисперсійний аналіз (*2 Way ANOVA*), де в якості факторів (джерел дисперсії, або незалежних змінних – *Independent Variable*) виступають випробовувані та завдання. Кожний випробовуваний та кожне завдання представляють рівні відповідного фактора, а комірки блоку суміжності, або дисперсійного комплексу (виділено темним у табл. 1), містять первинні бали (реалізації залежної змінної – *Depended Variable*).

Процедуру *ANOVA* включено до багатьох популярних статистичних програмних комплексів, серед яких: *Пакет аналізу MS Excel*, *STATISTICA* і *SPSS*. Але треба обирати той програмний модуль, що відповідає базовим припущенням методу [16], тобто *двофакторний ANOVA з випадковими факторами без повторів*. Для цього прикладу найбільше підходить *MS Excel*, оскільки його модуль *"Двухфакторный дисперсионный анализ без повторений"* працює безпосередньо з таблицею суміжності за обумовлених вимог. Результати, отримані за його допомогою, наведено у табл. 2.

[6] Цілком придатна для використання при проведенні державних іспитів і конкурсного відбору (коли тестування не використовується або неможливе).

Таблиця 1

Приклад результатів дихотомічного тесту [4]

№ випробо-вуваного	№ завдання						Первинний бал
	1	2	3	4	5	6	
1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	1
3	1	0	1	1	1	0	4
4	1	1	1	1	1	1	6
5	1	1	1	1	1	1	6
6	0	0	1	0	0	0	1
7	0	0	1	1	1	0	3
8	0	0	0	1	0	0	1
9	1	0	1	1	1	0	4
10	0	1	0	1	0	1	3
p	0,4	0,3	0,6	0,7	0,6	0,3	

Таблиця 2

Результати двофакторного дисперсійного аналізу даних табл. 1 засобами *MS Excel*

Складова варіації	Сума квадратів (*SS*)	Степені свободи (*df*)	Середній квадрат (*MS*)	Критерій Фішера (*F*)	*P–значення*	*F–крит.*
Випробовувані (S_n^2)	6,81667	9	0,75741	5,099751	9,266E-05	2,095755
Завдання (S_j^2)	1,48333	5	0,29667	1,997506	0,0972535	2,422085
Помилка (S_e^2)	6,68333	45	0,14852			
Разом	14,9833	59				

Величина S_n^2 наближає величину S_T^2 і може бути розрахована за формулою [4]:

$$s_n^2 = (MS_n - MS_r)/M. \qquad (76)$$

Відповідно, для складових дисперсії помилки маємо:

$$s_j^2 = (MS_j - MS_r)/N \text{ і } s_\varepsilon^2 = MS_r. \qquad (77), (78)$$

Використовуючи (67)–(69), отримуємо:

$$\rho_{\bar{Y}}^2 = \frac{(0{,}75741 - 0{,}14852)/6}{(0{,}75741 - 0{,}14852)/6 + [(0{,}29667 - 0{,}14852)/10 + 0{,}14852]/6} = 0{,}79.$$

Висновки

1. ССТ є добре розвиненою зі статистичного погляду.
2. У ССТ результати вимірювання знань випробовуваних залежать від характеристик тестових завдань, що фактично визначаються всім контингентом випробовуваних.
3. Хоча наведені результати статистичних основ ССТ далеко не повні, вони можуть стати у нагоді викладачам-науковцям при оцінюванні придатності до використання тестових завдань власної конструкції чи при виборі відповідних комп'ютерних програм.

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 психологии : [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. Чередниченко О. Ю. 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.
18. Ким В. С. Тестирование учебных достижений : [монография] / В. С. Ким. – Уссурийск : Издательство УГПИ, 2007. – 214 с.
19. Аванесов В. С. Форма тестовых заданий / В. С. Аванесов. – М. : Центр тестирования, 2005. – 156 с.
20. Печников А. Н. Модели и процедуры оценки результатов компьютерного тестирования знаний [Электронный ресурс] / А. Н. Печников, А. О. Туровская, Р. Р. Туктаров. – Режим доступа : http://ifets.ieee.org/russian/depository/v16_i4/html/6.htm
21. Печников А. Н. Теоретические основы психолого-педагогического проектирования автоматизированных обучающих систем / Печников А. Н. – Петродворец : ВВМУРЭ им. А. С. Попова, 1995. – 322 с.
22. Кумаритов А. М., Дубенко Ю. В. Методы и алгоритмы контроля знаний и оценки эффективности автоматизированных обучающих систем на производственном предприятии / А. М. Кумаритов, Ю. В. Дубенко. – Аудит и финансовый анализ. – 2009. – № 3. – 12 с.

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. Часть 2. Классический подход

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, разрешающую способность, стандартную ошибку и надежность тестовых измерений.

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

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

Statistical Tools for Measuring the Quality of Education. Part 2. Classical Approach

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.*