ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ т. 9, № 4(39), с. 219–238

С. В. Зыков, А. А. Незнанов, О. В. Максименкова

Критерии отклонения распределения случайных величин от нормального в математическом обеспечении программных систем поддержки измерений в образовании

Аннотация. В статье обсуждается задача проверки гипотез о типе распределения данных, получаемых при измерениях в образовании, в программных системах. Приведён обзор критериев проверки нормальности, имеющих дискретные аппроксимации, что делает их пригодными для реализации в программных системах. Обсуждается место и необходимость применения указанных критериев при автоматизации измерений в образовании. Результаты обзора положены в основу алгоритма подбора критерия или группы критериев в программной системе, ориентированной на измерения в образовании

Ключевые слова и фразы: измерения в образовании, математическое обеспечение, распределение результатов измерений, нормальность распределения, программная система.

Введение

Образование сегодня немыслимо без программных систем, поддерживающих административные процессы, процессы обучения и контроля знаний. Данные, продуцируемые и накапливаемые в этих системах, – ценный источник знаний как об участниках образовательного процесса, так и о самом процессе.

Анализ учебных данных (educational data mining, EDM), возникший и активно прогрессирующий последнее десятилетие, в части анализа результатов измерений в образовании (в особенности – тестов) плодотворно развивался уже более полувека. Широко известны, например, разнообразные методы тестологии [1–6].

Статья подготовлена в результате проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) и с использованием средств субсидии в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

10.25209/2079-3316-2018-9-4-219-238

Данная работа посвящена обсуждению базового математического обеспечения модулей анализа результатов систем поддержки измерений в образовании (ИвО). Предметом особого внимания являются критерии проверки отклонения выборок от нормального распределения. Такой интерес связан с тем, что многие методы анализа результатов ИвО опираются на предположение о нормальности распределения исследуемых данных. Например, построение Z-и T-оценок методами классической теории тестирования [7, 8] или требование асимптотической нормальности для применения теории тестлетов [9, 10]. Кроме того, исследователи в области тестологии указывают на необходимость установления нормальности распределения первичных баллов [11, 12] или предполагают, что процедура измерений организована так, чтобы результаты максимально «соответствовали» нормальному распределению [13].

1. О современном состоянии области в условиях развития непараметрических критериев принятия статистических гипотез

Нормальное распределение и соответствие ему некоторой случайной величины играет огромную роль при выборе инструментов параметрической статистики, имеющих ограниченную область применения. Параметрические методы постепенно «вымываются» из практики многих областей именно из-за своих ограничений, которые становятся незначимыми с ростом вычислительной мощности компьютерной техники, что позволяет исходно строить вероятностную модель без соответствующих ограничений, хотя и с повышенной вычислительной сложностью решения модели.

В области ИвО как минимум вся классическая тестология, а также базовые модели IRT (начиная с модели Раша) в современной теории тестирования используют параметрические модели. Поэтому понятно желание как минимум уметь сравнивать результаты расчётов с применением различных моделей, например, с непараметрическими моделями IRT, развиваемыми с конца 1990-х годов [14]. Современные непараметрические критерии принятия статистических гипотез с позиции практика хорошо изложены в [15].

При этом в настоящее время проработка критериев проверки нормальности остаётся в целом востребованной темой в изысканиях по математической статистике. Например, бурно развивается область проверки нормальности распределения на основе эмпирических функций

вычисления моментов [16]. Достаточно подробную классификацию современных методов и имитационные исследования можно найти в работе [17]. К сожалению, как и на протяжении всей истории данной области, разработка численных методов, которые можно применять непосредственно при разработке автоматических и автоматизированных программных систем, существенно отстаёт от теории.

2. Задача о типе распределение результатов ИвО

Вопросы о типе распределений результатов ИвО возникли ещё в середине XX века. Широко известны работы Лорда [18] и Кука [19], связанные с вопросами распределения первичных баллов и ошибок измерений [20]. Исследование Кука – это попытка воспроизведения более раннего исследования Лорда, искавшего подтверждение гипотезам:

(*1*) распределение первичных баллов склонно иметь отрицательную асимметрию для лёгких тестов и положительную для сложных;

(*2*) для симметрично распределённых первичных баллов характерен меньший эксцесс, чем у нормального распределения.

Куку удалось воспроизвести только результаты, касающиеся первой гипотезы, а эксперименты Лорда по второй гипотезе у Кука не воспроизвелись, более того, расхождению результатов не нашлось вразумительного объяснения. Можно утверждать, что критерий асимметрии и эксцесса может применяться только для исследований адекватности сложности теста уровню обученности группы испытуемых, но для уверенного применения стандартизированных шкал требуются более мощные критерии.

Знаковая работа Миццери[21] о нормальности данных в образовании (в том числе и результаты ИвО) описывает результаты применения методов робастной статистики к результатам педагогических измерений. Миццери пришёл к выводам о наличии сложных взаимосвязей в данных и формировании групп, обладающих различными тенденциями. Данное исследование положило начало ряду направлений, например, проверке нормальности в сравнительных исследованиях в образовании и медицине [22], шкалированию в масштабных тестированиях в условиях отсутствия нормальности данных [23].

Наши исследования показали [24], что применение развитого математического аппарата других предметных областей, например, медицинской статистики, к анализу данных в образовании также требуют исследования нормальности распределения данных. Отдельной проблемой является случай установления нормальности малой

выборки, он выходит за рамки указанных робастных и описательных исследований и требует отдельного рассмотрения.

Поскольку обзор критериев проверки нормальности распределений проводится для подбора математического обеспечения системы измерений в образовании, существенным преимуществом критерия полагается простота вычисления его статистики и (или) оценок статистики или наличие соответствующих аппроксимаций. Например, из рассмотрения исключён энтропийный критерий Ла Брека [25], позволяющий проверять сложные гипотезы нормальности без необходимости вычисления значений μ и σ. Поскольку при работе с результатами тестирований мы не сталкиваемся с ситуациями, когда вычисление указанных параметров затруднено или невозможно, автоматизация данного критерия представляется избыточной. Кроме того, его вычисления опираются на таблицу процентных точек.

Отметим также, что отдельные простые в вычислении критерии также оставлены за рамками данной работы. Указанные критерии требуют дополнительных исследований мощности против различных альтернатив. Например, корреляционный критерий Филлибена [26] является достаточно мощным и не уступает на симметричных альтернативах критерию Шапиро-Уилка. Критерий прост в вычислениях и не связан с таблицей процентных точек, но его поведение на асимметричных альтернативах требует дополнительного исследования. Аналогичная ситуация – с энтропийным критерием Васичека [27], не нуждающемся в таблице коэффициентов, но показавшим свою эффективность только против равномерных и экспоненциальных альтернатив.

3. Обзор критериев проверки нормальности распределений применительно к измерениям в образовании

Ранее в приложении к результатам тестирования Максименковой совместно с Подбельским [28] подробно рассматривались совместный критерий проверки асимметрии и эксцесса, Шапиро-Уилка и Эппса-Палли [29], предложенные в ГОСТ 5479-2002 [30] для малых выборок. Показано, что при уровне значимости $\alpha = 0,05$, согласно [31], можно полагать оценки, полученные критериями Шапиро-Уилка и Эппса-Палли, приемлемыми для результатов ИвО малого объёма при $n > 20$.

3.1. Критерий Хегази-Грина

Серия критериев, объединённая под названием критерия Хегази-Грина, предложена в [32] и основана на статистиках

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\left|y_{(i)} - \eta_i\right| \quad \text{и} \quad T_2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_{(i)} - \eta_i\right)^2,$$

где

$$y_{(i)} = \frac{x_i - \bar{x}}{s}, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \sum_{i=1}^{n} x_i, \quad \eta_i = \Phi^{-1}(m_i),$$

Φ – функция распределения $N(0;1)$, m_i – количество элементов выборки и $P_i = \frac{i}{n+1}$ – вероятность попадания в i-й интервал разбиения.

Вычисление критерия построено на анализе значений процентных точек для статистик T_1, T_2 и не имеет пригодных для автоматизации аппроксимаций для различных уровней значимости. Для уровня значимости $\alpha = 0,01$ предложена аппроксимация

$$T_1(0,01) = 0,7195 - 0,1751\ln(n) + 0,0108(\ln(n))^2,$$
$$T_2(0,01) = 0,0178 + \frac{2,8736}{n} - \frac{8,2894}{n^2},$$

а для уровня значимости $\alpha = 0,05$—

$$T_1(0,05) = 0,6027 - 0,1481\ln(n) + 0,009(\ln(n))^2,$$
$$T_2(0,05) = 0,0126 + \frac{1,9227}{n} - \frac{5,00677}{n^2}.$$

Заметим однако, что в работе [33] показано, что критерий мощнее критериев Шапиро-Уилка и Эппса-Палли при различении с другими куполообразными распределениями и на малой выборки даёт существенное смещение.

3.2. Критерий Гири

Удобный с точки зрения вычислений и автоматизации критерий предложен в работе [34] Гири. Статистика критерия:

$$d = \frac{1/n\sum_{i=1}^{n}|x_i - \bar{x}|}{\sqrt{1/n\,(x_i - \bar{x})}}.$$

Согласно [35] статистики критерия Гири асимптотически нормальна при $n \geqslant 40$ со средним $E(d)$ и дисперсией $D(d)$, на меньших выборках

критерий проявляет нестабильность. Аппроксимации для $E(d)$ и $D(d)$ основаны на результатах работы [36] и приведены в [37]:

$$E(d) = 0,797885 + \frac{0,199471}{n} + \frac{0,024934}{n^2} - \frac{0,031168}{n^3} - \frac{0,008182}{n^4},$$
$$D(d) = \frac{0,045070}{n} - \frac{0,124648}{n^2} + \frac{0,084859}{n^3} + \frac{0,006323}{n^4}.$$

Для практических расчётов в [38] рекомендована формула

$$\sum_{i=1}^{n} |x_i - \bar{x}| = 2\left(\sum x' - \bar{x}n'\right),$$

где x' – значения x, превышающих $\bar{x}$, а n' – их количество.

Поскольку критерий является двусторонним, вычисляются квантили

$$d\left(\alpha/2\right) = E(d) + \sqrt{D(d)}\, u_{\alpha/2} \quad \text{и} \quad d\left(1 - \alpha/2\right) = E(d) + \sqrt{D(d)}\, u_{1-\alpha/2},$$

где u_α – α-квантиль стандартного распределения; α – уровень значимости. Гипотеза о нормальности распределения принимается, если $d\left(\frac{\alpha}{2}\right) \le d \le d\left(1 - \frac{\alpha}{2}\right)$.

Указанные аппроксимации серьёзно критикуются в работе [39], где отмечено, что распределения статистики являются асимметричными и плохо аппроксимируются нормальным законом с указанными параметрами. При этом в [39] для табличных значений критерия Гири подтверждена его высокая мощность в отношении гипотезы нормальности по отношению к конкурирующим гипотезам. Данные аппроксимации в первоисточнике [40] представлены бесконечными рядами, значимость отброшенных в [38] остаточных членов не исследуется. В более поздней работе Гири [37] существенно уточнил коэффициенты в формулах аппроксимации:

$$E(d) = 0,7978845608 + \frac{0,19947114}{n} + \frac{0,02493389}{n^2} - \frac{0,03116737}{n^3},$$
$$D(d) = \frac{0,04507034}{n} - \frac{0,07957747}{n^2} + \frac{0,03978874}{n^3}.$$

В работе [41] Деагостино и Росман привели результаты имитационного моделирования мощности критерия Гири на малых выборках ($n = 20, 50, 100$). Авторами использована аппроксимация из работы [36], не учитывающая более поздних поправок. Симуляцией подтверждено превосходство альтернативных критериев, в частности критерия Шапиро-Уилка над критерием Гири как для симметричных, так и

для смещённых альтернатив. Деагостино и Росман отмечают, что высокая мощность данного критерия для симметричных альтернатив с показателем эксцесса менее трёх и его вычислительная простота позволяют считать критерий Гири достаточно удобным в использовании.

3.3. Критерий Дэвида-Хартли-Пирсона

Дэвидом, Хартли и Пирсоном в работе [42] предложен двусторонний критерий со статистикой: $U = {}^{R}\!/_{s}$, где $R = x_{\max} - x_{\min}$ – размах выборки, s – стандартное отклонение. Гипотеза о нормальности принимается, если $U_1(\alpha) < U < U(\alpha)$, где α – уровень значимости.

Согласно работе Томсона для больших n ($n \to \infty$) имеет место удобная для вычислений аппроксимация:

$$2\sqrt{1 - {}^{1}\!/_{n}} \le {}^{R}\!/_{s} \le \sqrt{2(n-1)}, \quad n = 2, 4, \ldots, 2k,$$
$$2\sqrt{1 - {}^{1}\!/_{n}} \le {}^{R}\!/_{s} \le \sqrt{2(n-1)}, \quad n = 1, 3, \ldots, 2k-1.$$

Исследованиями показана более низкая мощность критерия Дэвида-Хартли-Пирсона по отношению к критерию Гири [39]. Поскольку критерий Дэвида-Хартли-Пирсона уступает по мощности критериям Гири, Эппса-Палли и Шапиро-Уилка, а удобная оценка определена только для больших объёмов выборок, его использование нецелесообразно на малых выборках, а для больших существуют более точные, например, критерий Колмогорова-Смирнова.

3.4. Критерий Шпигельхальтера

Статистика комбинированного критерия Шпигельхальтера основана на комбинации статистик критериев Гири и Дэвида- Хартли-Пирсона:

$$T' = \left\{ \frac{1}{(C_n U)^{n-1}} + \frac{1}{d^{n-1}} \right\}^{\frac{1}{n-1}}, \quad C_n = \frac{(n!)^{\frac{1}{n-1}}}{2n} \quad \text{и} \quad U = \frac{R}{s},$$

где U — статистика критерия Дэвида-Хартли-Пирсона, d — статистика критерия Гири. Гипотеза о нормальности распределения принимается, если $T' < T'(\alpha)$.

Шпигельхальтером в [43] показана удовлетворительная мощность критерия против симметричных альтернатив. Там же приведены критические значения статистики критерия на уровнях значимости 0,05 и 0,1 для выборок объемом n = 5, 10, 15, 20, 50, 100. Исследования для несимметричных альтернатив описаны в [44].

Существенным недостатком для автоматизации является плохая аппроксимируемость известными распределениями критических значений статистики критерия Шпигельхартера. Значения для указанных выше объёмов выборок рассчитаны в [43] на основе множества нормально распределённых выборок, для получения значений статистики в промежуточных точках предлагается связываться с автором критерия.

3.5. Критерий Локка-Спурье

Большинство рассмотренных выше критериев имеет высокую эффективность против симметричных альтернатив, в работе [45] Локком и Спурье был предложен эффективный критерий проверки нормальности против асимметричных альтернатив. Статистики критерия имеют вид:

$$T_{1n} = \frac{1}{C_n^3} \sum_{i=1}^{n} \omega_i x_i \quad \text{и} \quad T_{2n} = \frac{1}{C_n^3} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} V_{ij} (x_j - x_i)^2,$$

где $\omega_i = C_{i-1}^2 - 2\,(n-1)\,(i-1) + C_{n-1}^2$, $V_{ij} = i + j - n - 1$ и x_i — i-я порядковая статистика.

Для статистик критерия имеют место аппроксимации. При $n \geq 5$ статистика $T' = \frac{T_{1n} - E(T_{1n})}{\sqrt{D(T_{1n})}}$ распределена как нормальная случайная величина с критическими значениями $T_{1n} = E\,(T_{1n}) + \sqrt{D\,(T_{1n})}\,u_\alpha$, где u_α — α-квантиль стандартного нормального распределения. При $n \geq 10$ нормальное приближение для T_{2n} имеет вид

$$T_{2n} = E\,(T_{2n}) + \sqrt{D\,(T_{2n})}\,u_\alpha.$$

Гипотеза нормальности распределения вероятностей случайной величины отклоняется, если $T_{1n}\,(T_{2n}) > T_{1n}\,(\alpha)\,(T_{2n}\,(\alpha))$, где $(1-\alpha)$ — уровень значимости.

Для критерия Локка-Спурье в [45] показана эффективность для несимметричных альтернатив с «мягкими» хвостами. Исследования для несимметричных альтернатив с обоими видами хвостов представлено в работе [46].

3.6. Критерий Оя

В работе [47] Оя, основываясь на работах [48, 49], предложил критерий проверки нормальности, статистики которого построены

на комбинациях порядковых статистик

$$T_1 = \frac{1}{C_n^3} \sum_{1 \le i < j < k \le n} \frac{x_k - x_j}{x_k - x_i} \quad \text{и} \quad T_2 = \frac{1}{C_n^4} \sum_{1 \le i < j < l \le n} \frac{x_k - x_j}{x_l - x_i}.$$

При справедливости гипотезы нормальности $E(T_1) = 0,5$ и $E(T_2) = 0,298746$, дисперсии приведены в работе [49]. В таблице дисперсии для выборок $n \le 30$ приведены с пропусками. Приближённо нормально распределены статистики

$$\tilde{T}_1(\alpha) = 0,5 + \sqrt{D(T_1)}\, u_\alpha \quad \text{и} \quad \tilde{T}_2(\alpha) = 0,298746 + \sqrt{D(T_2)}\, u_\alpha,$$

где u_α – α-квантиль стандартного нормального распределения. Рекомендован комбинированный критерий $\chi^2 = \tilde{T}_1^2 + \tilde{T}_2^2$, который имеет распределение χ^2 с $f = 2$ степенями свободы.

Более просты в вычислительном плане предложенные в работе [47] модифицированные критерии

$$T_1' = \sum_{1 \le i < j \le n} a_{ij} \log(x_j - x_i) \quad \text{и} \quad T_2' = \sum_{1 \le i < j \le n} b_{ij} \log(x_j - x_i),$$

где $a_{ij} = \frac{1}{C_n^3}(i + j - n - 1)$, $b_{ij} = \frac{1}{C_n^4}\left(2(n-j)(i-1) - C_{n-j}^2 - C_{i-1}^2\right)$. Если гипотеза нормальности справедлива, то

$$\begin{aligned}
&E(T_1') = 0, \qquad E(T_2') = 0,4523,\\
&D(T_1') = \frac{1}{C_n^3}(0,11217(n-4)(n-3) + 0,118899(n-3) + 2,8979),\\
&D(T_2') = \frac{1}{C_n^4} 0,0874(n-6)(n-5)(n-4)\\
&\qquad + 0,0435(n-5)(n-4) + 5,342(n-4) + 8,8552),\\
&T_1'(\alpha) = \sqrt{D((T_1'))}\, u_\alpha,\\
&T_2'(\alpha) = 0,4523 + \sqrt{D((T_2'))}\, u_\alpha.
\end{aligned}$$

Может быть использован критерий

$$\chi^2 = \left(\tilde{T}_1\right)^2 + \left(\tilde{T}_2\right)^2 = \frac{(T_1' - E(T_1'))^2}{D(T_1')} + \frac{(T_2' - E(T_2'))^2}{D(T_2')}.$$

4. Обсуждение критериев

Аппроксимации статистик критериев Локка-Спурье и Оя опираются на критерий χ^2. Существуют и другие специальные критерии

проверки нормальности, сводящиеся к использованию критерия χ^2, но оставленные за границами данного исследования, например, критерий де Агостино [**50**, **51**]. В указанных критериях речь идёт об использовании модификации общего критерия χ^2 для проверки нормальности распределения, описание и таблицы которого могут быть найдены, например, в [**38**].

В исследованиях мощности [**39**, **41**] критерия Гири имеются ограничения, а также отсутствует единое мнение об областях его применимости. Поэтому несмотря на его вычислительную простоту мы полагаем, что:

1) нецелесообразно использовать критерий Гири в алгоритмах анализа результатов ИвО, так как на малых выборках его мощность уступает критерию Шапиро-Уилка;
2) желательно провести имитационное исследование критерия Гири с уточнёнными по работе [37] коэффициентами для симметричных и несимметричных альтернатив на выборках объёма $n \leqslant 100$.

Учитывая серьёзные ограничения критерия Шпигельхальтера, связанные с невозможностью аппроксимирования его статистики, и наличие расчётных значений только для шести объёмов выборок было принято следующее решение.

(*1*) Не включать исследования, основанные на критерии Шпигельхальтера в подсистемы анализа результатов ИвО.

(*2*) Провести дополнительное исследование и с использованием синтетических выборок вычислить промежуточные процентные точки значений статистики критерия.

5. Алгоритм выбора критерия проверки нормальности распределения

Как показывает материал раздела 3, процесс выбора критерия зависит как минимум от двух серьёзных решений, которые должна уметь принимать информационная система поддержки измерений в образовании. Первое связано с объёмом данных, получаемых от системы. Критерии, рассмотренные в обзоре, закрывают вопросы, связанные с малочисленными измерениями, когда объём выборки не превышает 200. Второе решение принимается по результатам проверки отклонения от нормальности, то есть по результату вычисления асимметрии. То есть мы имеем дело, с точки зрения информационной системы, с адаптивным алгоритмом выбора критерия (для наглядности алгоритм

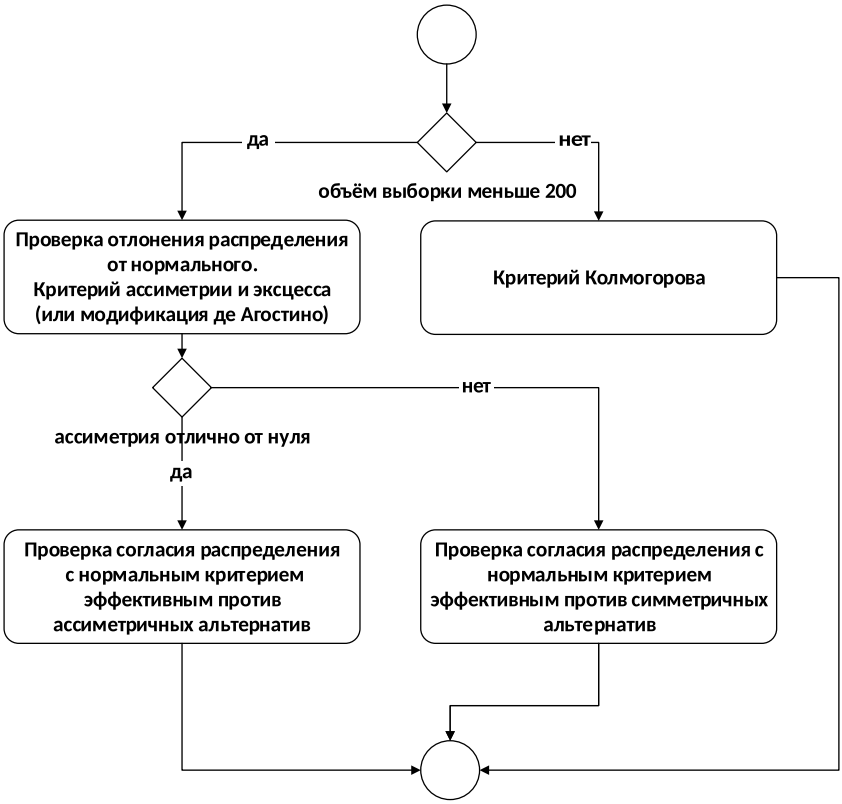

Рис. 1. Алгоритм выбора критерия проверки нормальности распределения

представлен простой UML- диаграммой на рис. 1) для использования в программных системах ИвО.

Дополнительно отметим, что критерии проверки нормальности против симметричных и асимметричных альтернатив могут быть подразделены по альтернативам с разными видами хвостов (например, тяжёлыми хвостами) [52]. Однако применительно к измерениям в образовании уточнение алгоритма с точки зрения хвостов требует проведения отдельного исследования по установлению наиболее распространённых видов распределений, описывающих эмпирические распределения величин.

Заключение

В заключении отметим, что данный обзор серьёзно детализирован численными методами, позволяющими вычислять статистики критериев проверки нормальности непосредственно при реализации в программной системе, так и за счёт подключения существующих внешних решателей. Так, практически все перечисленные в работе критерии имеют реализации в виде R- пакетов, доступных в репозитории CRAN. Кроме того, в обзор включены первоисточники критериев, поскольку авторами замечено, что многократное перецитирование работ в данной области привело к досадным ошибкам в вычислительных формулах и коэффициентах. Все перечисленное сделано с целью предложить читателю готовые, обоснованные проектные решения, связанные с подбором математического обеспечения систем поддержки измерений в образования.

Список литературы

[1] R. L. Brennan. “Generalizability theory and classical test theory”, *Applied Measurement in Education*, **24** (2011), pp. 1–21. ↑219

[2] D. N. M. de Gruijter, L. J. T. Van der Kamp. *Statistical test theory for education and psychology*, Chapman and Hall/CRC, 2007, 280 p. ↑219

[3] H. Gulliksen. *Theory of Mental Tests*, John Wiley & Sons, 1950, 504 p. ↑219

[4] В. С. Аванесов. «Item Response Theory: Основные понятия и положения. Статья первая», *Педагогические измерения*, 2007, №3, с. 3–36. ↑219

[5] R. K. Hambleton, H. Swaminathan, H. J. Rogers. *Fundamentals of item response theory*, SAGE, Newbury Park, 1991, 184 p. ↑219

[6] H. Wainer, E. T. Bradlow, X. Wang. *Testlet response theory and its application*, Cambridge University press, New York, 2007. ↑219

[7] S. E. R. Kurpius, M. E. Stafford. *Testing and measurement. A user-friendly guide*, SAGE, New York, 2011, 184 p. ↑220

[8] Ю. М. Нейман, В. А. Хлебников. *Введение в теорию моделирования и параметризации педагогических тестов*, Прометей, М., 2000, 168 с. ↑220

[9] X. Wang, E. T. Bradlow, H. Wainer. *A general Bayesian model for testlets: theory and applications*, Educational testing service, Princeton, NJ, 2002. ↑220

[10] M. C. S. Paap, C. A. W. Glas, B. P. Veldkamp. *An overview of research on the testlet effect: associated features, implications for test assembly, and the impact of model choice on ability estimates*, LSAC Research Report Research Report 13-03, Law School Admission Council, Enschede, 2013. ↑220

[11] М. Б. Челышкова. *Разработка педагогических тестов на основе современных математических моделей*, Учебное пособие, Исследовательский центр, М., 1995, 32 с. ↑220

[12] М. Б. Челышкова. *Теория и практика конструирования педагогических тестов*, Логос, М., 2002, 432 с. ↑220

[13] А. А. Бодалев, В. В. Столин, В. С. Аванесов. *Общая психодиагностика*, Речь, Санкт-Петербург, 2000, 440 с. ↑220

[14] B. W. Junker, K. Sijtsma. "Nonparametric item response theory in action: an overview of the special", *Applied Psychological Measurement*, **25**:3 (2001), pp. 211–220. ↑220

[15] M. Neuhauser. *Nonparametric statistical tests: a computational approach*, Chapman and Hall/CRC, 2011, 248 p. ↑220

[16] N. Henze, S. Koch. "On a test of normality based on the empirical moment generating function", *Stat. Papers*, 2017, pp. 1–13. ↑221

[17] D. Szynal. "On two families of tests for normality with empirical description of their performnaces", *Discussiones Mathematicae Probability and Statistics*, **34**:1–2 (2014), pp. 169–185. ↑221

[18] F. M. Lord. "A survey of observed test-score distributions with respect to skewness and kurtosis", *Educational and Psychological Measurement*, **15**:4 (1955), pp. 383–389. ↑221

[19] D. L. Cook. "A replication of Lord's study on skewness and kurtosis of observed test-score distributions", *Educational and Psychological Measurement*, **19**:1 (1959), pp. 81–87. ↑221

[20] F. M. Lord. "An empirical study of the normality and independence of errors of measurement in test scores", *Psychometrica*, **25**:1 (1960), pp. 91–104. ↑221

[21] T. Micceri. "The unicorn, the normal curve, and other inprobable creatures", *Psychological Bulletin*, **105**:1 (1989), pp. 156–166. ↑221

[22] J. Rochon, M. Gondan, M. Kieser. "To test or not to test: preliminary assessment of normality when comparing two independend samples", *BMC Medical Research Methodology*, **12**:81 (2012), pp. 11. ↑221

[23] A. D. Ho, C. C. Yu. "Descriptive statistics for modern test score distributions: skewness, kurtosis, discreteness, and ceiling effects", *Educational and Psychological Measurement*, **75**:3 (2015), pp. 365–388. ↑221

[24] O. Maksimenkova, A. Neznanov, M. Skryabin. «On MOOCs quality estimation: a case of modern nonparametric superiority and noninferiority statistical tests», *eLearning Stakeholders and Researchers Summit 2017*, Материалы международной конференции (Moscow, 2017), ред. Е. Ю. Кулик, Национальный исследовательский университет "Высшая школа экономики", М., 2017, с. 165–174. ↑221

[25] J. LaBrecque. "Goodness-of-fit tests based on nonlinearity in probability plots", *Technometrics*, **19**:3 (1977), pp. 293–306. ↑222

[26] J. J. Filliben. "The probability plot correlation coefficient test for normality", *Technometrics*, **17**:1 (1975), pp. 111–117. ↑222

[27] O. Vasicek. “A test for normality based on sample entropy”, *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**:1 (1976), pp. 54–59. ↑222

[28] О. В. Максименкова, В. В. Подбельский. «О применении некоторых специальных критериев проверки нормальности распределения результатов педагогических измерений», *Законодательная и прикладная метрология*, 2014, №6, с. 28–32. ↑222

[29] N. Henze. “An approximation to the limit distribution of the Epps-Pulley test statistic for normality”, *Metrika*, **37**:1 (1990), pp. 7–18. ↑222

[30] *ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения*, ИПК Издательство стандартов, М., 2002, 27 с. ↑222

[31] Б. Ю. Лемешко, С. Б. Лемешко. «Сравнительный анализ критериев проверки отклонения распределения от нормального закона», *Метрология*, 2005, №2, с. 3–23. ↑222

[32] Y. A. S. Hegazy, J. R. Green. “Some new goodness-of-fit tests using order statistics”, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **24**:3 (1975), pp. 299–308. ↑223

[33] Б. Ю. Лемешко, А. П. Рогожников. «Исследование особенностей и мощности некоторых критериев нормальности», *Метрология*, 2009, №4, с. 3–24. ↑223

[34] R. C. Geary. “The ratio of the mean deviation to the standard deviation as a test of normality”, *Biometrika*, **27**:3/4 (1935), pp. 310–332. ↑223

[35] E. S. Pearson. “A comparison of β_1 and Mr Geary’s ωn criteria”, *Biometrika*, **27**:3/4 (1935), pp. 333–352. ↑223

[36] R. C. Geary. “Note on the correlation between β_1 and ω'”, *Biometrika*, **27**:3/4 (1935), pp. 353–363. ↑224

[37] R. C. Geary. “Testing for normality”, *Biometrika*, **34**:3/4 (1947), pp. 209–242. ↑224

[38] А. И. Кобзарь. *Прикладная математическая статистика*, Для инженеров и научных работников, Физматлит, М., 2006, 816 с. ↑224,228

[39] Б. Ю. Лемешко. *Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход*, НГТУ, Новосибирск, 2011, 888 с. ↑224,225,228

[40] R. C. Geary. “Moments of the ratio of the mean deviation to the standard deviation for normal samples”, *Biometrika*, **28**:3/4 (1936), pp. 295–307. ↑224

[41] R. B. D’Agostino, B. Rosman. “The power of Geary’s test of normality”, *Biometrika*, **61**:1 (1974), pp. 181–184. ↑224,228

[42] H. A. David, H. O. Hartley, E. S. Pearson. “The distribution of the ratio, in a single normal sample, of range to standard deviation”, *Biometrika*, **41**:3/4 (1954), pp. 482–493. ↑

[43] D. J. Spiegelhalter. “A test for normality against symmetric alternatives”, *Biometrika*, **64**:2 (1977), pp. 415–418. ↑225,226

[44] D. J. Spiegelhalter. “An omnibus test for normality for small samples”, *Biometrika*, **67**:2 (1980), pp. 493–496. ↑225

[45] C. Locke, J. D. Spurrier. “The use of U-statistics for testing normality against nonsymmetric alternatives”, *Biometrika*, **63**:1 (1976), pp. 143–147. ↑226

[46] C. Locke, J. D. Spurrier. “The use of U-statistics for testing normality against alternatives with both tails heavy or both tails light”, *Biometrika*, **64**:3 (1977), pp. 638–640. ↑226

[47] H. Oja. “New tests for normality”, *Biometrika*, **70**:1 (1983), pp. 297–299. ↑227

[48] H. Oja. “Two location and scale goodness-of-fit tests”, *Biometrika*, **68**:3 (1981), pp. 637–640. ↑

[49] C. E. Davids, D. Quade. “U-statistics for skewness or symmetry”, *Communication is Statistics-Theory and Methods*, **7**:5 (1978), pp. 413–418. ↑227

[50] R. B. D’Agostino. “An omnibus test of normality for moderate and large size samples”, *Biometrika*, **58**:2 (1971), pp. 341–348. ↑228

[51] R. B. D’Agostesto. “Small sample probability points for the D test of normality”, *Biometrika*, **59**:1 (1972), pp. 219–221. ↑228

[52] Б. Ю. Лемешко, С. Б. Лемешко, С. Н. Постовалов. «Сравнительный анализ мощности критериев согласия при близких альтернативах. II. Проверка сложных гипотез», *Сибирский журнал индустриальной математики*, **11**:4(36) (2008), с. 78–93. ↑229

Пример ссылки на эту публикацию:

С. В. Зыков, А. А. Незнанов, О. В. Максименкова. «Критерии отклонения распределения случайных величин от нормального в математическом обеспечении программных систем поддержки измерений в образовании». *Программные системы: теория и приложения*, 2018, **9**:4(39), с. 219–238.
10.25209/2079-3316-2018-9-4-219-238
http://psta.psiras.ru//read/psta2018_4_219-238.pdf

Об авторах:

Сергей Викторович Зыков

д.т.н., доцент, доцент департамента программной инженерии ФКН НИУ ВШЭ

e-mail: szykov@hse.ru

Алексей Андреевич Незнанов

к.т.н., доцент, старший научный сотрудник международной научно-учебной лаборатории интеллектуальных систем и структурного анализа ФКН НИУ ВШЭ

e-mail: aneznanov@hse.ru

Ольга Вениаминовна Максименкова

м. н. с. международной научно-учебной лаборатории интеллектуальных систем и структурного анализа ФКН НИУ ВШЭ

e-mail: omaksimenkova@hse.ru

Sergey Zykov, Aleksey Neznanov, Olga Maksimenkova. *Tests for normality as mathematical support for educational measurement software.*

Abstract. The paper discusses estimation of normality of data collected from educational software. It reviews tests for normality which have discrete approximations and may be directly implemented in program systems. The paper discusses the necessity of these tests to educational data and their place in automation of education. The results of the review are used a base of a test selection algorithm, which can be used in an educational measurement software system. (*In Russian*).

Key words and phrases: educational measurement, assessment, mathematical support, test for normality, measurement results distribution, program system, programming, software.

References

[1] R. L. Brennan. "Generalizability theory and classical test theory", *Applied Measurement in Education*, **24** (2011), pp. 1–21. ↑219

[2] D. N. M. de Gruijter, L. J. T. Van der Kamp. *Statistical test theory for education and psychology*, Chapman and Hall/CRC, 2007, 280 p. ↑219

[3] H. Gulliksen. *Theory of Mental Tests*, John Wiley & Sons, 1950, 504 p. ↑219

[4] V. S. Avanesov. "Item Response Theory: Basic terms and concepts. The first paper", *Pedagogicheskiye izmereniya*, 2007, no.3, pp. 3–36. ↑219

[5] R. K. Hambleton, H. Swaminathan, H. J. Rogers. *Fundamentals of item response theory*, SAGE, Newbury Park, 1991, 184 p. ↑219

[6] H. Wainer, E. T. Bradlow, X. Wang. *Testlet response theory and its application*, Cambridge University press, New York, 2007. ↑219

[7] S. E. R. Kurpius, M. E. Stafford. *Testing and measurement. A user-friendly guide*, SAGE, New York, 2011, 184 p. ↑220

[8] Yu. M. Neyman, V. A. Khlebnikov. *Introduction to item response theory*, Prometey, M., 2000, 168 p. ↑220

[9] X. Wang, E. T. Bradlow, H. Wainer. *A general Bayesian model for testlets: theory and applications*, Educational testing service, Princeton, NJ, 2002. ↑220

[10] M. C. S. Paap, C. A. W. Glas, B. P. Veldkamp. *An overview of research on the testlet effect: associated features, implications for test assembly, and the impact of model choice on ability estimates*, LSAC Research Report Research Report 13-03, Law School Admission Council, Enschede, 2013. ↑220

[11] M. B. Chelyshkova. *Educational tests development based on modern mathematical models*, Uchebnoye posobiye, Issledovatel'skiy tsentr, M., 1995, 32 p. ↑220

[12] M. B. Chelyshkova. *Theory and practice of educational tests development*, Logos, M., 2002, 432 p. ↑220

10.25209/2079-3316-2018-9-4-219-238

[13] A. A. Bodalev, V. V. Stolin, V. S. Avanesov. *General psychodiagnostics*, Rech', Sankt-Peterburg, 2000, 440 p.↑220

[14] B. W. Junker, K. Sijtsma. "Nonparametric item response theory in action: an overview of the special", *Applied Psychological Measurement*, **25**:3 (2001), pp. 211–220. ↑220

[15] M. Neuhauser. *Nonparametric statistical tests: a computational approach*, Chapman and Hall/CRC, 2011, 248 p.↑220

[16] N. Henze, S. Koch. "On a test of normality based on the empirical moment generating function", *Stat. Papers*, 2017, pp. 1–13. ↑221

[17] D. Szynal. "On two families of tests for normality with empirical description of their performnaces", *Discussiones Mathematicae Probability and Statistics*, **34**:1–2 (2014), pp. 169–185. ↑221

[18] F. M. Lord. "A survey of observed test-score distributions with respect to skewness and kurtosis", *Educational and Psychological Measurement*, **15**:4 (1955), pp. 383–389. ↑221

[19] D. L. Cook. "A replication of Lord's study on skewness and kurtosis of observed test-score distributions", *Educational and Psychological Measurement*, **19**:1 (1959), pp. 81–87. ↑221

[20] F. M. Lord. "An empirical study of the normality and independence of errors of measurement in test scores", *Psychometrica*, **25**:1 (1960), pp. 91–104. ↑221

[21] T. Micceri. "The unicorn, the normal curve, and other inprobable creatures", *Psychological Bulletin*, **105**:1 (1989), pp. 156–166. ↑221

[22] J. Rochon, M. Gondan, M. Kieser. "To test or not to test: preliminary assessment of normality when comparing two independend samples", *BMC Medical Research Methodology*, **12**:81 (2012), pp. 11. ↑221

[23] A. D. Ho, C. C. Yu. "Descriptive statistics for modern test score distributions: skewness, kurtosis, discreteness, and ceiling effects", *Educational and Psychological Measurement*, **75**:3 (2015), pp. 365–388. ↑221

[24] O. Maksimenkova, A. Neznanov, M. Skryabin. "On MOOCs quality estimation: a case of modern nonparametric superiority and noninferiority statistical tests", *eLearning Stakeholders and Researchers Summit 2017*, Materialy mezhdunarodnoy konferentsii (Moscow, 2017), ed. Ye. Yu. Kulik, Natsional'nyy issledovatel'skiy universitet "Vysshaya shkola ekonomiki", M., 2017, pp. 165–174.↑221

[25] J. LaBrecque. "Goodness-of-fit tests based on nonlinearity in probability plots", *Technometrics*, **19**:3 (1977), pp. 293–306. ↑222

[26] J. J. Filliben. "The probability plot correlation coefficient test for normality", *Technometrics*, **17**:1 (1975), pp. 111–117. ↑222

[27] O. Vasicek. "A test for normality based on sample entropy", *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**:1 (1976), pp. 54–59. ↑222

[28] O. V. Maksimenkova, V. V. Podbel'skiy. "O primenenii nekotorykh spetsial'nykh kriteriyev proverki normal'nosti raspredeleniya rezul'tatov pedagogicheskikh izmereniy", *Zakonodatel'naya i prikladnaya metrologiya*, 2014, no.6, pp. 28–32.↑222

[29] N. Henze. "An approximation to the limit distribution of the Epps-Pulley test statistic for normality", *Metrika*, **37**:1 (1990), pp. 7–18. ↑222

[30] *GOST R ISO 5479-2002. Statistical methods. Tests for departure of the probability distribution from the normal distribution*, IPK Izdatel'stvo standartov, M., 2002, 27 p.↑222

[31] B. Yu. Lemeshko, S. B. Lemeshko. "Comparative study of test of distribution deviation from the normal", *Metrologiya*, 2005, no.2, pp. 3–23. ↑222

[32] Y. A. S. Hegazy, J. R. Green. "Some new goodness-of-fit tests using order statistics", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **24**:3 (1975), pp. 299–308. ↑223

[33] B. Yu. Lemeshko, A. P. Rogozhnikov. "The study of properties and the power of some tests for normality", *Metrologiya*, 2009, no.4, pp. 3–24. ↑223

[34] R. C. Geary. "The ratio of the mean deviation to the standard deviation as a test of normality", *Biometrika*, **27**:3/4 (1935), pp. 310–332. ↑223

[35] E. S. Pearson. "A comparison of β_1 and Mr Geary's ωn criteria", *Biometrika*, **27**:3/4 (1935), pp. 333–352. ↑223

[36] R. C. Geary. "Note on the correlation between β_1 and ω'", *Biometrika*, **27**:3/4 (1935), pp. 353–363. ↑224

[37] R. C. Geary. "Testing for normality", *Biometrika*, **34**:3/4 (1947), pp. 209–242. ↑224

[38] A. I. Kobzar'. *Applied mathematical statistics*, For engineers and scientists, Fizmatlit, M., 2006, 816 p.↑224,228

[39] B. Yu. Lemeshko. *Data analysis, simulation and study of probability regularities. Computer approach*, NGTU, Novosibirsk, 2011, 888 p.↑224,225,228

[40] R. C. Geary. "Moments of the ratio of the mean deviation to the standard deviation for normal samples", *Biometrika*, **28**:3/4 (1936), pp. 295–307. ↑224

[41] R. B. D'Agostino, B. Rosman. "The power of Geary's test of normality", *Biometrika*, **61**:1 (1974), pp. 181–184. ↑224,228

[42] H. A. David, H. O. Hartley, E. S. Pearson. "The distribution of the ratio, in a single normal sample, of range to standard deviation", *Biometrika*, **41**:3/4 (1954), pp. 482–493. ↑

[43] D. J. Spiegelhalter. "A test for normality against symmetric alternatives", *Biometrika*, **64**:2 (1977), pp. 415–418. ↑225,226

[44] D. J. Spiegelhalter. "An omnibus test for normality for small samples", *Biometrika*, **67**:2 (1980), pp. 493–496. ↑225

[45] C. Locke, J. D. Spurrier. "The use of U-statistics for testing normality against nonsymmetric alternatives", *Biometrika*, **63**:1 (1976), pp. 143–147. ↑226

[46] C. Locke, J. D. Spurrier. "The use of U-statistics for testing normality against alternatives with both tails heavy or both tails light", *Biometrika*, **64**:3 (1977), pp. 638–640. ↑226

[47] H. Oja. "New tests for normality", *Biometrika*, **70**:1 (1983), pp. 297–299. ↑227

[48] H. Oja. "Two location and scale goodness-of-fit tests", *Biometrika*, **68**:3 (1981), pp. 637–640. ↑

[49] C. E. Davids, D. Quade. "U-statistics for skewness or symmetry", *Communication is Statistics-Theory and Methods*, **7**:5 (1978), pp. 413–418. ↑227

[50] R. B. D'Agostino. "An omnibus test of normality for moderate and large size samples", *Biometrika*, **58**:2 (1971), pp. 341–348. ↑228

[51] R. B. D'Agostesto. "Small sample probability points for the D test of normality", *Biometrika*, **59**:1 (1972), pp. 219–221. ↑228

[52] B. Yu. Lemeshko, S. B. Lemeshko, S. N. Postovalov. “Sravnitel’nyy analiz moshchnosti kriteriyev soglasiya pri blizkikh al’ternativakh. II. Proverka slozhnykh gipotez”, *Sibirskiy zhurnal industrial’noy matematiki*, **11**:4(36) (2008), pp. 78–93. ↑229

Sample citation of this publication:

Sergey Zykov, Aleksey Neznanov, Olga Maksimenkova. “Tests for normality as mathematical support for educational measurement software”. *Program Systems: Theory and Applications*, 2018, **9**:4(39), pp. 219–238. (*In Russian*).

10.25209/2079-3316-2018-9-4-219-238

http://psta.psiras.ru//read/psta2018_4_219-238.pdf