УДК 519.248

СТАТИСТИЧЕСКИЙ АНАЛИЗ ИНДИВИДУАЛЬНЫХ ЗАДАНИЙ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ

Кацман Юлий Янович,
канд. техн. наук, доцент кафедры вычислительной техники
Института кибернетики ФГАОУ ВПО «Национальный исследовательский
Томский политехнический университет»,
Россия, 634050, г. Томск, пр. Ленина, д. 30. E-mail: katsman@tpu.ru

Актуальность *работы обусловлена необходимостью непрерывного повышения качества образовательных программ, реализуемых в Томском политехническом университете (ТПУ).*
Цель работы: *показать связь оценок образовательных успехов с характеристиками тестов, контролирующих знания студентов; проанализировать качество тестов, с помощью которых осуществляется мониторинг знаний студентов по теории вероятностей; убедиться в высоком качестве предлагаемых вариантов индивидуальных заданий либо, в противном случае, получить достоверную информацию о конкретных вариантах, требующих улучшения.*
Методы исследования. *Особенности применяемых тестов (малое число задач, ограниченный объем выборки) делают некорректным использование современной теории тестов. В работе использовались статистические методы анализа результатов тестирования. Среди используемых методов можно отметить точечное и интервальное оценивание, кластерный анализ, однофакторный анализ: ранговые критерии и дисперсионный анализ. В работе использовались ранговые методы: тест Крускела–Уоллиса и медианный тест. Обычно после получения статистически значимой оценки F теста, желательно было бы знать парные различия между всеми группами. Тест Шеффе был использован для определения значимой разности между средними значениями групп в дисперсионном анализе. Все исследования проведены с использованием различных модулей программы Statistica 6.1.*
Результаты. *Статистический анализ индивидуальных заданий показал отсутствие параллельности (равносильности) тестов в ряде вариантов индивидуальных заданий. Результаты, представленные в табличном и графическом виде, показали, что все варианты индивидуальных заданий (тестов) можно разделить по сложности на 3 кластера. Применяемые в работе статистические методы показали высоко значимое различие (непараллельность) тестов разных кластеров. В работе предложен способ обеспечения параллельности тестов.*

Ключевые слова:
Статистический анализ, мониторинг знаний, тестирование, диаграмма рассеяния, выборочная (точечная) характеристика, ранг, медиана, кластер.

На протяжении ряда лет повышению качества основных образовательных программ (ООП) в рамках реализации концепции CDIO в Томском политехническом университете уделяется повышенное внимание. В работе [1] одной из основных задач совершенствования образовательного процесса и ООП указывается «оптимизация процедур мониторинга качества ООП для их непрерывного совершенствования…». Вопрос оценки качества методических контролирующих материалов является достаточно трудоемкой и сложной задачей, которая является актуальной как для новых дисциплин, так и для тех, обучение по которым производится в течение ряда лет [2, 3]. Оценка качества обучения актуальна не только в российском образовательном пространстве, но и в зарубежных странах [4, 5].

Большинство известных работ [6–10] посвящены мониторингу и оценке тестов школьников, что, как правило, подразумевает большие объемы выборки (тысячи наблюдений) и так называемые бинарные тесты, в которых правильный ответ кодируется 1, а неправильный 0. В работе [11], посвященной тестированию студентов каждое из 12 заданий состоял из 2 подзаданий. Таким образом, правильный ответ на два подзадания оценивался в 2 балла, одного – в 1 балл и неправильный ответ – в 0 баллов. За исключением данного факта (трехбалльная оценка теста, вместо двухбалльной) в представленной работе, как и в работе [12], было обеспечено большое число наблюдений, что позволило оценить результаты экзаменов по химии с помощью математического аппарата классической теории тестов.

Можно отметить еще один подход, когда тестирование либо оценка достижений испытуемого осуществляется с помощью автоматизированных систем контроля [13, 14]. Наряду с неоспоримыми преимуществами данный подход имеет и ряд недостатков: *во-первых*, каждому испытуемому (тестируемому) необходим персональный компьютер, *во-вторых*, не для всех типов заданий применим автоматический (автоматизированный) контроль.

Промежуточный контроль качества знаний студентов по теории вероятностей проводился с помощью индивидуальных контрольных заданий (контрольных работ). В процессе изучения дисциплины студентам предлагалось решить четыре контрольных задания, позволяющих оценить качество усвоения соответствующих теоретических модулей. Каждое задание включало, как правило, три задачи, одна из которых соответствовала предыдущей теме. Основное отличие данных заданий от традиционного тестирования заключается в следующем:

- количество задач в одном задании мало, т. к. контрольная работа (тестирование) проводится в течение одного/двух академических часов;
- задачи необходимо решать, что соответствует заданиям типа С (ЕГЭ), причем правильность решения оценивалась в баллах (не 0/1);

- количество различных вариантов достаточно велико (35–40), чтобы максимально исключить возможность несамостоятельного решения;
- количество студентов в группе 15–25, число групп в потоке 4–6, так что общий объем выборки (наблюдений) не превышал 100;
- так как все задания проверяются одним преподавателем, полностью исключить субъективный подход не представляется возможным.

В данной работе проведен статистический анализ параллельности вариантов (билетов) индивидуальных контрольных заданий по теории вероятностей, с помощью которых оценивалось качество знаний, усвоенных студентами. Обычно при анализе качества контролирующих материалов большое внимание уделяется обеспечению параллельности вариантов задания [15, 16]. При этом, если применение современной теории тестов – Item Response Theory (IRT) [17] для оценки латентных факторов требует обеспечить для одного теста минимальную выборку от 200 до 1000 наблюдений, классическая статистическая теория позволяет получить оценки параметров, ограничиваясь значительно меньшим количеством опытов.

По результатам первой контрольной работы минимальная оценка (3 балла) давалась за попытку решить хотя бы одну задачу, максимальная оценка (10 баллов) – за правильное решение трех задач. Все результаты первой контрольной работы по теории вероятностей для 229 наблюдений были обработаны в лицензионной программе Statistica 6.1. На первом этапе обработки из анализа были исключены нерепрезентативные варианты, по которым было 3 и менее наблюдений.

Далее предполагалось, что предлагаемые варианты параллельны (равносильны), и тогда оценки студентов должны быть адекватны их знаниям, а не сложности билетов. С этой целью для каждого варианта были рассчитаны точечные и интервальные оценки, что с учетом случайных факторов предполагало приблизительное равенство средних баллов и дисперсий для каждого варианта. Реальные оценки для каждого варианта представлены на рис. 1 в виде диаграмм рассеяния.

Представленные результаты наглядно свидетельствуют о неодинаковой сложности (непараллельности) различных вариантов контрольной работы. На рис. 1 наблюдаются как очень сложные варианты, например 16 – средний балл равен 3,5 (ни одна задача из трех не решена правильно), так и очень простые 14, 29 – средний балл выше 9 (решены все три задачи с небольшими ошибками).

Разбиение вариантов на группы сложности можно провести с помощью методов кластерного анализа, например метода *k-средних* (*k-means*) [18]. Однако в данном случае разбиение всех наблюдений на кластеры осуществляется только по одной переменной – баллам, поэтому для обеспечения примерного равенства количества наблюдений в каждой группе и однородности наблюдений внутри группы мы отсортировали все варианты по среднему баллу:

- Cluster_1 – сложные задания (средний балл менее 6,5);
- Cluster_2 – задания средней сложности (средний балл больше 6,5 и меньше/равен 7,5);
- Cluster_3 – легкие задания (средний балл больше 7,5).

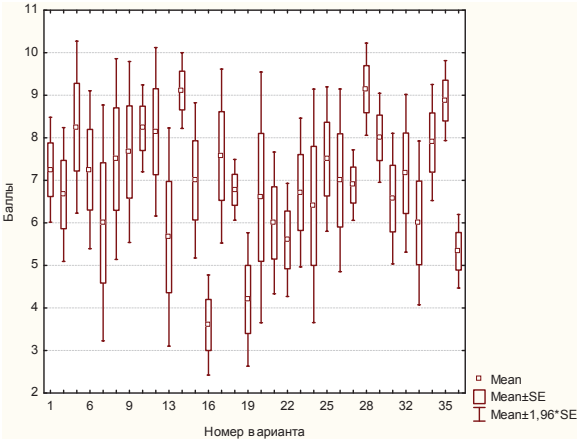

Рис. 1. *Диаграммы рассеяния для различных вариантов контрольной работы*

Fig. 1. *Scatter plot for different variants of test*

Запустив модуль *Описательные статистики*, получили точечные оценки для всех наблюдений и каждого кластера отдельно. На основании результатов, представленных в табл. 1, можно сделать следующие выводы:

- максимальное различие в оценках для 1 и 3 кластеров составляет менее 3 баллов;
- практически все точечные характеристики для второго кластера и всей совокупности наблюдений совпадают;
- для всех наблюдений и второго кластера 50 % полученных результатов превышают 7 баллов; в то же время для первого кластера 50 % результатов не превышает 6 баллов, а для третьего кластера 50 % оценок превышают 9 баллов;
- дисперсии для всех наблюдений и всех кластеров можно считать практически равными (отношение дисперсий менее 2);
- анализ коэффициентов скоса и эксцесса (Skewness & Kurtosis) свидетельствует, что распределения баллов в каждой группе несимметрично и существенно отличается от гауссова распределения.

Таблица 1. *Точечные характеристики переменной (баллы) для различных групп*

Table 1. *Sample characteristics of a variable (balls) for different groups*

Переменная Variable	Descriptive Statistics (Контрольные по ТВ_2014)						
	Valid N	Mean	Median	Variance	Std.Dev.	Skewness	Kurtosis
All Groups	214	7,0327	7,0000	5,6468	2,3763	−0,4079	−0,9088
Cluster_1	58	5,5000	6,0000	5,9035	2,4297	0,4066	−1,0748
Cluster_2	83	6,9759	7,0000	4,1945	2,0480	−0,3420	−0,3416
Cluster_3	73	8,3151	9,0000	3,6910	1,9212	−1,2133	1,0298

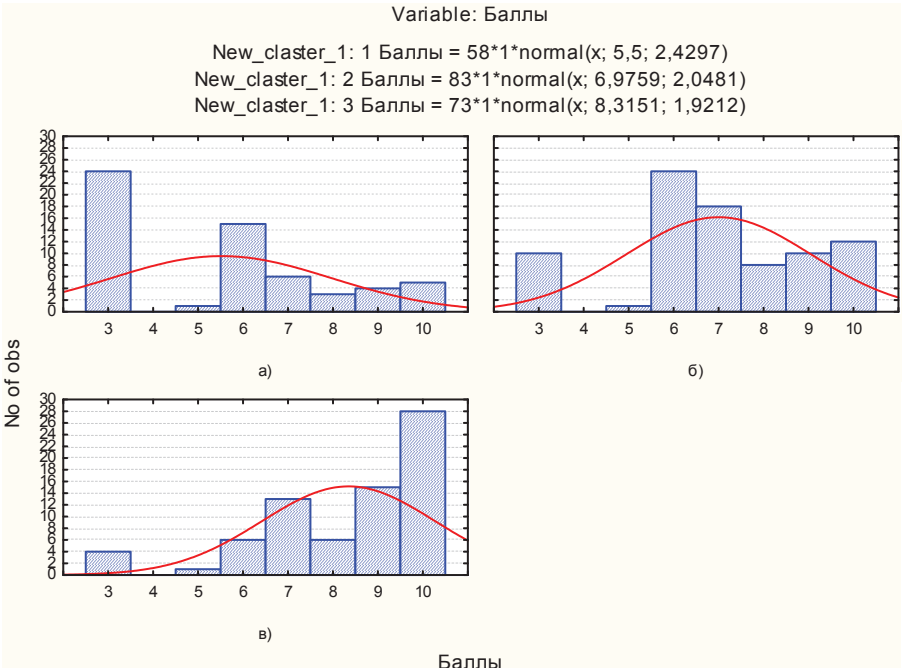

Рис. 2. *Категоризованные диаграммы оценок для различных групп сложности*

Fig. 2. *Categorized assessments diagrams for different complexity groups*

Таким образом, возникает вопрос: существенно ли различие средних баллов для каждого кластера или его можно объяснить случайными вариациями? Тогда разбиение вариантов заданий по группам сложности безосновательно, и варианты заданий можно считать параллельными (равносильными). Построим гистограммы распределения баллов для различных кластеров (рис. 2).

Категоризованные гистограммы (рис. 2) подтверждают, что распределения баллов по контрольным заданиям для 1, 2 и 3 кластеров существенно различны. Так, если для 1 кластера большинство студентов набрали 6 и менее баллов, то для кластера 3 работы ~70 % студентов получили более 6 баллов, для 2 кластера распределение по баллам более равномерное. Для каждого кластера на рис. 2, *а–в* приведены (сплошной линией) гауссовы распределения. Сравнение гистограммы и гауссова распределения показало, что ни для одного кластера распределение баллов нельзя считать нормальным.

Эффективное решение выше сформулированной задачи будем искать в рамках теории ранговых критериев [19, 20]. Данный подход обусловлен тем, что, *во-первых*, распределение баллов для всех наблюдений и каждой группы в отдельности не является нормальным, а, *во-вторых*, ранговые процедуры, свободные от распределения, являются более мощными по сравнению с параметрическими методами.

Таким образом, дальнейшие исследования проведем в рамках однофакторного анализа. В нашем случае фактором является сложность вариантов индивидуальных заданий, имеющая три уровня (1–3 кластеры). Откликом же является оценка, полученная студентом за выполнение конкретного задания. Прежде чем оценивать количественное влиянии фактора на полученную оценку (баллы), логично вначале ответить на вопрос: а есть ли такое влияние вообще? Нельзя ли объяснить расхождение (разброс) в баллах для различных уровней факторов действием чистой случайности? Ведь полученные оценки оказываются различными даже при неизменном значении фактора. Возможно, по той же причине (изменчивость, случайность) можно объяснить и различие между результатами в различных группах?

Выскажем статистическое предположение – все данные принадлежат одному и тому же распределению, а наблюдаемые различия случайны и не значимы – нулевая гипотеза.

К достоинствам теории ранговых критериев можно отнести следующее, если в результате проверки рангового критерия (критериев) можно однозначно (статистически) утверждать, что верна гипотеза H_0, т. е. влияние фактора незначимо, тогда на этом однофакторный анализ закончен. Если же верна альтернативная гипотеза H_1 – влияние фактора значительное, то однофакторный анализ следует продолжить и перейти от качественной оценки к количественной.

Для проверки нулевой гипотезы H_0 запустим процедуры однофакторного рангового анализа в пакете Statistica. Во всех дальнейших исследованиях зададим уровень значимости α=0,05. Результаты теста Крускела–Уоллиса для оценок индивидуальных заданий по теории вероятностей приведены в табл. 2.

Таблица 2. *Тест Крускела–Уоллиса*

Table 2. *Kruskal–Wallis test*

Dependent: Баллы Balls	Kruskal–Wallis ANOVA by Ranks; Баллы (Контрольные по ТВ_2014) Independent (grouping) variable: New_claster_1 Kruskal–Wallis test: H (2, N=214)=45,40989 p=0,0000		
	Code	Valid N	Sum of Ranks
1	1	58	4074,00
2	2	83	8571,00
3	3	73	10360,00

Обозначения: ***Code*** *– уникальный код группы (число);* ***Sum of Ranks*** *– сумма рангов;* ***p*** *– вероятность принятия гипотезы* H_0*;* ***Valid N*** *– число наблюдений в группе;* ***H*** *– статистика Крускела–Уоллиса.*

Notations: ***Code*** *is the unique code of a group (number);* ***Sum of Ranks*** *is the sum of ranks;* ***p*** *is the probability of accepting hypothesis* H_0*;* ***Valid N*** *is a number of observations in a group;* ***H*** *is the statistics of Kruskal–Wallis.*

Анализируя суммы рангов по группам (кластерам), представленным в табл. 2, можно говорить о влиянии уровня фактора на оценки студентов. Результаты подтверждают, что максимальная оценка наблюдается в третьем кластере, а минимальная – в первом.

В статистике Крускела–Уоллиса вычисляется сумма квадратов разностей средних рангов в группе и среднего ранга по всей выборке. Тогда, если верна гипотеза H_0 и влияние фактора незначимо, значение статистики мало, а соответствующая вероятность велика. В нашем случае H=45,40989, так что нулевую гипотезу можно принять с вероятностью p=0,0000. Поскольку заданный нами уровень значимости много больше (α=0,05), то нулевую гипотезу (варианты заданий параллельны, и кластеризация не влияет на оценки) следует отвергнуть в пользу альтернативной гипотезы H_1 – влияние фактора существенное.

Проведем ранговое тестирование тех же данных, используя независимый от предыдущего метода *медианный тест* (*критерий*) [19]. Известно, что статистика медианного теста при нулевой гипотезе асимптотически подчиняется распределению χ^2 с k–1 степенями свободы. Полученные результаты приведены в табл. 3.

Таблица 3. *Медианный тест*

Table 3. *Median test*

Dependent: Баллы Balls	Median Test, Overall Median=7,00000; Баллы (Контрольные по ТВ_2014) Independent (grouping) variable: New_claster_1 Chi-Square=30,76912 df=2 p=0,0000			
	1	2	3	Total
<= Median: observed	46,0000	53,00000	24,0000	123,0000
expected	33,3364	47,70561	41,9579	
obs.-exp.	12,6636	5,29439	–17,9579	
> Median: observed	12,0000	30,00000	49,0000	91,0000
expected	24,6636	35,29439	31,0421	
obs.-exp.	–12,6636	–5,29439	17,9579	
Total: observed	58,0000	83,00000	73,0000	214,0000

В верхней части таблицы приведены количества рангов в группах, которые были меньше или равны медиане. В нижней части таблицы – аналогичные значения, превышающие значение медианы.

Отчет по статистике медианного теста (табл. 3) позволяет проанализировать полученные результаты на качественном уровне. По значению разности предсказанных и полученных значений можно сделать следующие выводы:

- верхняя половина таблицы – максимальное значение указывает на кластер, имеющий минимальные оценки (максимум сложности);
- нижняя половина таблицы – максимальное значение соответствует кластеру, имеющему максимальные оценки – (минимум сложности).

Что же касается количественной оценки медианного теста, то нулевую гипотезу можно принять с вероятностью p=0,0000, что много меньше уровня значимости α. Следует принять альтернативную гипотезу H_1 – влияние фактора существенное.

Так как проведенный ранговый однофакторный анализ подтвердил гипотезу о значимом влиянии фактора, количественную оценку этого влияния получим в рамках дисперсионного анализа. Подробный отчет проведенных исследований представлен в табл. 4.

Статистика Фишера F=28,60658 незначимо отличается от единицы с вероятностью p=0,000000, что значительно меньше уровня значимости. Следовательно, отвергается нулевая гипотеза в пользу альтернативной гипотезы H_1 – влияние фактора существенно.

Так как параллельно с дисперсионным анализом в системе Statistica можно получить оценки эффектов обработки, представим их графически на рис. 3.

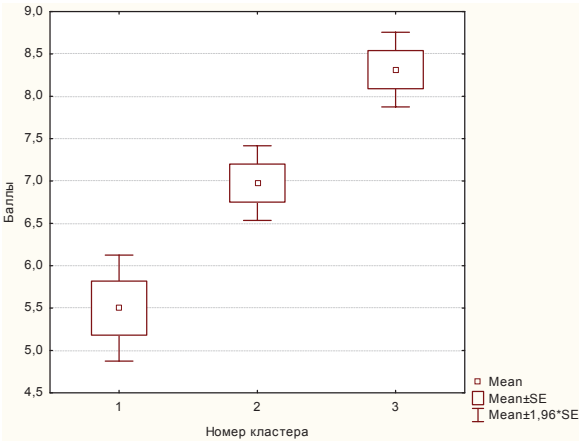

Рис. 3. *Диаграммы рассеяния для всех кластеров*

Fig. 3. *Scatter plot for all clusters*

Приведенные результаты свидетельствуют о существенном различии точечных и интервальных характеристик для различных групп. Отметим, что наряду со средними значениями можно проа-

Таблица 4. *Результаты дисперсионного анализа*

Table 4. *Results of ANOVA*

Переменная Variable	Analysis of Variance (Контрольные по ТВ_2014) Marked effects are significant at p<0,05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Баллы / Balls	256,5658	2	128,2829	946,2052	211	4,484385	28,60658	0,000000

Обозначения: ***SS (Sum of Squares) Effect*** *– сумма квадратов факторов (вторая оценка дисперсии) умноженная на $k-1$;* ***df Effect*** *– число степеней свободы фактора;* ***MS (Mean Square) Effect*** *– средний квадрат фактора;* ***SS Error*** *– сумма квадратов (оценка дисперсии) умноженная на $N-k$;* ***df Error*** *– число степеней свободы наблюдений равная $N-k$;* ***MS Error*** *– оценка дисперсии;* ***F*** *– значение статистики Фишера;* ***p*** *–вероятность принятия гипотезы H_0.*

Notations: ***SS*** *is the Sum of Squares Effect (the second estimation of dispersion) multiplied by $k-1$;* ***MS*** *is the Mean Square Effect;* ***df Error*** *is the number of degrees of observation freedom equals $N-k$;* ***SS Error*** *is the sum of squares (dispersion estimation multiplied by $N-k$);* ***F*** *is the value of Fisher statistic;* ***df Effect*** *is the number of factor degrees of freedom;* ***MS Error*** *is the dispersion estimation;* ***p*** *is the probability of accepting H_0 hypothesis.*

нализировать такие групповые параметры, как дисперсия, медиана, нижний и верхний квартили, минимальное и максимальное значения и т. п.

На этом этапе исследования в рамках однофакторного анализа можно считать законченными, но, анализируя последние данные (рис. 3), можно попытаться ответить на вопрос: какие пары групп вариантов заданий можно считать значимо различными? Для ответа на этот вопрос проведем сравнения средних по методу Шеффе для различных пар уровней факторов. Результаты теста представлены в табл. 5.

Таблица 5. *Шеффе S-метод множественных сравнений [21]*

Table 5. *Scheffe S-method of multiple comparisons [21]*

New_claster_1	Scheffe Test; Variable: Баллы (Контрольные по ТВ_2014) Marked differences are significant at p<0,05000		
	{1} M=5,5000	{2} M=6,9759	{3} M=8,3151
1 {1}		0,000341	0,000000
2 {2}	0,000341		0,000557
3 {3}	0,000000	0,000557	

В результате проверки гипотезы о незначимом различии средних баллов различных пар кластеров справедливости нулевой гипотезы можно принять с вероятностью более чем в 100 раз меньшей, чем заданный уровень значимости (α=0,05).

Статистический анализ проведенного мониторинга индивидуальных заданий по теории вероятностей (дисциплина преподавалась на протяжении ряда лет) доказал, что варианты тестовых заданий непараллельны. На основании проведенных исследований были сделаны окончательные выводы о качестве предлагаемых тестовых заданий:

- Из 40 тестовых заданий 8 вариантов были исключены из анализа, т. к. число опытов каждого задания составляет не более 3 (выборка непрезентативна); по этой причине из анализа были исключены варианты: 3, 5, 11, 31, 37, 38, 39 и 40;
- Cluster_1 включает задания высокой сложности; в этот кластер входили 9 вариантов, средняя оценка по которым составляет менее 6,5 баллов: 7, 13, 16, 19, 21; 22, 24; 33, 36;
- Cluster_2 включает задания средней сложности; в этот кластер входили 13 вариантов, средняя оценка по которым составляет более 6,5 и равна (меньше) 7,5 баллов: 1, 2, 6, 8, 15, 18, 20, 23, 25, 26, 27, 30, 32;
- Cluster_3 включает легкие задания; в этот кластер входили 10 вариантов, средняя оценка по которым составляет более 7,5 баллов: 4, 9, 10, 12, 14, 17, 28, 29, 34, 35.

Проведенные исследования показали, что Cluster_2 можно расширить за счет новых вариантов, а именно: из вариантов Cluster_3 исключить 1 или 2 легких задания, учитывая средний набранный балл, заменив их более сложными задачами из Cluster_1, опять же учитывая набранный средний балл. С новыми вариантами тестов необходимо провести эксперименты (мониторинг знаний студентов) и полученные результаты сравнить на параллельность с данными Cluster_2.

В заключение можно еще раз отметить, что мониторинг качества ООП в большой степени определяется качеством методических индивидуальных контролирующих материалов (тестов). Одной из важнейших характеристик вариантов тестов является их параллельность. В работе показано, что даже для заданий, используемых на протяжении ряда лет, задача обеспечения параллельности (одинаковой сложности) тестов является актуальной. Предложенные в работе статистические методы позволяют успешно решить эту задачу, что продемонстрировано на примере контрольных заданий по теории вероятностей.

СПИСОК ЛИТЕРАТУРЫ

1. Решение научно-методической конференции «Уровневая подготовка специалистов: международная концепция CDIO и Стандарты ООП ТПУ» 26–30 марта 2013. URL: http://www.portal.tpu.ru:7777/science/konf/methodconf/results/2013/resolution.pdf (дата обращения: 12.05.2014).
2. Болотов В.А., Вальдман И.А., Ковалева Г.С. Российская система оценки качества образования: чему мы научились за 10 лет? // Тенденции развития образования: проблемы управления и оценки качества образования: Материалы VIII Междунар. научно-практ. конф. – М.: Университетская книга, 2012. – С. 22–31.
3. Агранович М.Л. Можно ли сопоставить результаты ЕГЭ и ГИА. Сравнение показателей, рассчитанных на основе разных тестовых испытаний // Вопросы образования. – 2014. – № 1. – С. 80–91.
4. Clarke M. What Matters Most for Student Assessment Systems: Framework Paper. SABER–Student Assessment Working Paper № 1, 2012. Washington, DC, World Bank. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1/ (дата обращения: 23.07.2014).
5. Ramirez M.J. Disseminating and Using Student Assessment Information in Chile. SABER – Student Assessment Working Paper № 3, 2012. Washington, DC, World Bank. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/17474/682360WP00PUBL0WP30READ0web04006012.pdf?sequence=1/ (дата обращения: 23.07.2014).
6. Майоров А.Н. Мониторинг в образовании. Изд. 3-е, испр. и доп. – М.: Интеллект-Центр, 2005. – 424 с.
7. Челышкова М.Б. Теория и практика конструирования педагогических тестов. – М.: Логос, 2002. – 432 с.
8. Дружинин В.Н. Экспериментальная психология. Изд. 2-е. – СПб.: Питер, 2011. – 320 с.
9. Илюхин Б.В. Оценка качества образования и принцип разумной достаточности // Народное образование. – 2012. – № 6. – С. 118–126.
10. Горлов П.И., Илюхин Б.В. Как построить систему оценки качества образования? // Журнал руководителя управления образованием. – 2012. – № 6. – С. 41–46.
11. Тестовая технология контроля знаний студентов по химии / М.Г. Минин, Н.Ф. Стась, Е.В. Жидкова, О.Б. Родкевич // Известия Томского политехнического университета. – 2005. – Т. 308. – № 4. – С. 231–235. URL: http://www.lib.tpu.ru/fulltext/v/Bulletin_TPU/2005/v308/i4/53.pdf (дата обращения: 12.05.2014).
12. Статистический анализ качества тестов, применяемых для контроля знаний по химии / М.Г. Минин, Н.Ф. Стась, Е.В. Жидкова, О.Б. Родкевич // Известия Томского политехнического университета. – 2007. – Т. 310. – № 1. – С. 282–286. URL: http://www.lib.tpu.ru/fulltext/v/Bulletin_TPU/2007/v310/i1/60.pdf (дата обращения: 12.05.2014).
13. Крец И.В., Хаустов П.А., Кацман Ю.Я. Автоматическая система проверки задач по олимпиадному программированию // Молодежь и современные информационные технологии: Сборник трудов VI Всеросс. научно-практ. конф. студентов, аспирантов и молодых ученых. – Томск, 26–28 февраля 2008. – Томск: СПб Графикс, 2008. – С. 299–300.
14. Лепустин А.В., Кацман Ю.Я. Автоматизированная система тестирования качества обучения «Эмпирик» // Современные техника и технологии: Труды XIV Междунар. научно-практ. конф. студентов, аспирантов и молодых ученых. – Томск, 24–28 марта 2008. – Томск: ТПУ, 2008. – С. 245–247.
15. Suen H.K., Lei P.W. Classical versus Generalizability theory of measurement. URL: http://suen.educ.psu.edu/~hsuen/pubs/Gtheory.pdf (дата обращения: 12.05.2014).
16. Илюхин Б.В., Пермяков О.Е. Проблемы обеспечения качества приема и направления совершенствования системы конкурсного отбора поступающих в вузы Российской Федерации // Известия Томского политехнического университета. – 2007. – Т. 310. – № 1. – С. 269–275.
17. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960. – 216 p.
18. Халафян А.А. STATISTICA 6. Статистический анализ данных. 3-е изд. – М.: ООО «Бином-Пресс», 2007. – 512 с.
19. Гаек Я., Шидак Э. Теория ранговых критериев. – М.: Наука, 1971. – 374 с.
20. Кендалл М., Стюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 900 с.
21. Справочник по прикладной статистике в 2-х т. Т. 1 / пер с англ. под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, 1989. – 510 с.

Поступила 31.07.2014 г.

UDC 519.248

STATISTICAL ANALYSIS OF INDIVIDUAL TASKS ON PROBABILITY THEORY

Yuliy Ya. Katsman,
Cand. Sc., National Research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, 634050, Russia. E-mail: katsman@tpu.ru

The relevance *of the work is caused by the need to improve continuously the quality of educational programs at Tomsk Polytechnic University.*

The main aim of the study *is to show the relation of assessments of educational progress with the characteristics of tests checking students' knowledge; to analyze the quality of the tests for monitoring students' probability theory knowledge; to ensure the quality of the proposed variants of individual tasks or to obtain reliable information on specific variants requiring improvement.*

The methods used in the study. *Features of tests used (a few number of tasks, the limited size of the sample) make the use of Item Response Theory (IRT) invalid. The author has used statistical methods for analyzing test results. Among the methods used the following ones can be noted: sample and interval estimation, cluster analysis, one-way factor analysis: ranking criteria and dispersion analysis. The author used ranking methods: Kruskal–Wallis ANOVA and Median test. Usually, after obtaining a statistically significant F test from the ANOVA, one wants to know which means contributed to the effect; that is, which groups are particularly different from each other. Scheffe's test was used to determine the significant differences between group means in an analysis of variance setting. All investigations were carried out using various modules of the program Statistica 6.1.*

The results. *Statistical analysis showed that tests in some variants of individual tasks are not parallel (equal). The results given in tabular and graphical forms, showed that all the variants of individual tasks (tests) could be divided in three clusters in compliance with complexity of the variants. The statistical methods applied in-process showed highly significant difference (nonparallelism) of tests in different clusters. The paper proposes a method for providing parallel tests.*

Key words:
Statistical analysis, knowledge monitoring, testing, scatter plot, sample characteristic, rank, median, cluster.

REFERENCES

1. *Reshenie nauchno-metodicheskoy konferentsii «Urovnevaya podgotovka spetsialistov: mezhdunarodnaya kontseptsiya CDIO i Standarty OOP TPU»* [The solution of the scientific-methodical conference. Level training: international CDIO concept and standards OOP TPU]. 26–30 March 2013. Available at: http://www.portal.tpu.ru:7777/science/konf/methodconf/results/2013/resolution.pdf (accessed 12 May 2014).
2. Bolotov V.A., Valdman I.A., Kovaleva G.S. Rossiyskaya sistema otsenki kachestva obrazovaniya: chemu my nauchilis za 10 let? [The Russian system of education quality assessment: what have we learned for 10 years?]. *Tendentsii razvitiya obrazovaniya: problemy upravleniya i otsenki kachestva obrazovaniya. Materialy* VIII *Mezhdunarodnoy nauchno-prakticheskoy konferentsii* [Mat. 8th Intern. Scien.-Pract. Conf. Trends in Education: Problems of management and evaluation of education quality]. Moscow, 2012, pp. 22–31.
3. Agranovich M.L. Mozhno li sopostavit rezultaty EGE i GIA. Sravnenie pokazateley, rasschitannykh na osnove raznykh testovykh ispytaniy [Is it possible to compare the results of the CSE and the GIA. Comparing the indicators calculated on the basis of different test runs]. *Educational Studies*, 2014, no. 1, pp. 80–91.
4. Clarke M. What Matters Most for Student Assessment Systems: a Framework Paper. SABER–Student Assessment Working Paper, no 1, 2012. Washington, DC, World Bank. Available at: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1/ (accessed 23 July 2014).
5. Ramirez M.J. Disseminating and Using Student Assessment Information in Chil SABER – Student Assessment Working Paper, no. 3, 2012. Washington, DC, World Bank. Available at: https://openknowledge.worldbank.org/bitstream/handle/10986/17474/682360WP00PUBL0WP30READ0web04006012.pdf?sequence=1/ (accessed 23 July 2014).
6. Mayorov A.N. *Monitoring v obrazovanii* [Monitoring in education]. Moscow, Intellekt-Tsentr Publ., 2005. 424 p.
7. Chelyshkova M.B. *Teoriya i praktika konstruirovaniya pedagogicheskikh testov* [Theory and practice of designing pedagogical tests]. Moscow, Logos Publ., 2002. 432 p.
8. Druzhinin V.N. *Eksperimentalnaya psikhologiya* [Experimental psychology]. St. Petersburg, Piter Publ., 2011. 320 p.
9. Ilyukhin B.V. Otsenka kachestva obrazovaniya i printsip razumnoy dostatochnosti [Assessment of the quality of education and reasonable sufficiency principle]. *Narodnoe obrazovanie*, 2012, no. 6, pp.118–126.
10. Gorlov P.I., Ilyukhin B.V. Kak postroit sistemu otsenki kachestva obrazovaniya? [How to build a system for evaluating the quality of education?] *Zhurnal rukovoditelya upravleniya obrazovaniem*, 2012, no.6, pp. 41–46.
11. Minin M.G., Stas N.F., Zhidkova E.V., Rodkevich O.B. Testovaya tekhnologiya kontrolya znaniy stydentov po khimii [Test technology for controlling students' knowledge in chemistry]. *Bulletin of the Tomsk Polytechnic University*, 2005, vol. 308, no. 4, pp. 231–235. Available at: http://www.lib.tpu.ru/fulltext/v/Bulletin_TPU/2005 /v308/i4/53.pdf (accessed 12 May 2014).
12. Minin M.G., Stas N.F., Zhidkova E.V., Rodkevich O.B. Statisticheskiy analiz kachestva testov, primenyaemykh dlya kontrolya znaniy po khimii [Statistical analysis of the quality of tests used for controlling knowledge in chemistry]. *Bulletin of the Tomsk Polytechnic University*, 2007, vol. 310, no. 1, pp. 282–286. Available at: http://www.lib.tpu.ru/fulltext/v/Bulletin_TPU/2007/v310/i1/60.pdf (accessed 12 May 2014).
13. Krets I.V., Khaustov P.A., Katsman Yu.Ya. Avtomaticheskaya sistema proverki zadach po olimpiadnomu programmirovaniyu [Automatic verification of tasks of the olympiad programming]. *Sbornik trudov VI Vserossiyskoy nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchenykh «Molodezh i sovremennye informatsionnye tekhnologii»* [Proc. 6th All-Russian scientific practical conference of students and young scientists]. Tomsk, 26–28 February 2008. pp. 299–300.
14. Lepustin A.V., Katsman Yu.Ya. Avtomatizirovannaya sistema testirovaniya kachestva obucheniya «Empirik» [Automated system for testing the quality of teaching «Empiricist»]. *Trudy XIV Mezhdunarodnoy nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchenykh «Sovremennye tekhnika i tekhnologii»* [Proc. 14th Int. Scientific-practical Conf. for students, graduate students and young scientists. Modern equipment and technology]. Tomsk, 24–28 March 2008. Tomsk, TPU Publ. House, 2008. pp. 245–247.
15. Suen H.K., Lei P.W. *Classical versus Generalizability theory of measurement*. Available at: http://suen.educ.psu.edu/~hsuen/pubs/Gtheory.pdf (accessed 12 May 2014).
16. Ilyukhin B.V., Permyakov O.E. Problemy obespecheniya kachestva priema i napravleniya sovershenstvovaniya sistemy konkursnogo otbora postupayushchikh v vuzy Rossiyskoy Federatsii [Quality assurance issues and directions in perfecting the system of competitive selection of applicants in the Russian Federation]. *Bulletin of the Tomsk Polytechnic University*, 2007, vol. 310, no. 1, pp. 269–275.
17. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. 216 p.
18. Khalafyan A.A. *STATISTICA 6. Statisticheskiy analiz dannykh* [STATISTICA 6. Statistical analysis of data]. Moscow, Binom-Press, 2007, 512 p.
19. Gaek Ya., Shidak E. *Teoriya rangovykh kriteriev* [Theory of rank tests]. Moscow, Nauka Publ., 1971, 374 p.
20. Kendall M., Styuart A. *Statisticheskie vyvody i svyazi* [The advanced theory of statistics]. Moscow, Nauka Publ., 1973. 900 p.
21. *Spravochnik po prikladnoy statistike* [Handbook of applicable mathematics]. Ed. by Lloyd E., Lederman Yu., Tyurin Yu.N. Moscow, Finansy i statistika Publ., 1989. 510 p.

Received: 31 July 2014.