КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Буров Андрей Васильевич,
магистрант ФГБОУ ВПО «Смоленский государственный университет», г. Смоленск
burov_andrei@inbox.ru

Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена

Аннотация. *Статья посвящена изучению некоторых возможностей применения методов статистической обработки при оценке тестовых материалов единого государственного экзамена. Вопрос об объективности тестирования как способа проверки знаний и умений обучающегося актуален, как для системы образования, так и для общества в целом. Для оценки валидности тестовых материалов единого государственного экзамена применяется один из вариантов измерения валидности тестов с помощью аппарата статистической обработки данных и показывается применение предложенной методики к результатам выполнения учащимися тестовых заданий по русскому языку и математике.*
Ключевые слова: *единый государственный экзамен, оценка, валидность, тест, статистическая обработка данных.*
Раздел: *(01) педагогика; история педагогики и образования; теория и методика обучения и воспитания (по предметным областям).*

Организация и проведение единого государственного экзамена (ЕГЭ) – многоплановая и трудоемкая задача, основная цель которой – оценка уровня знаний выпускников, освоивших образовательные программы среднего (полного) общего образования, при помощи контрольно-измерительных материалов (КИМ), представляющих собой набор заданий примерно одинаковой трудности для каждого варианта теста, выполнение или невыполнение которых позволяет установить (оценить) уровень освоения федерального государственного образовательного стандарта. Но для достижения этой цели необходимо, чтобы полученная на ЕГЭ отметка была адекватной, то есть объективно отражала степень овладения учащимися материалами школьной программы.

Значимость объективного оценивания на едином государственном экзамене определяется следующими условиями.

1. Использование результатов ЕГЭ при конкурсном отборе в вузе, имеющем значение не только для выпускника, но и прежде всего для общества, развитие которого зависит от качества профессиональной подготовки.

2. Необходимостью оценки и совершенствования образования, ведь по результатам единого государственного экзамена все чаще делают вывод об уровне общеобразовательной подготовки школьников, о направлениях развития образования в Российской Федерации.

3. Это трудоемкая задача, требующая привлечения научного аппарата.

Оценка формируется по результатам выполнения заданий контрольно-измерительного материала экзамена. Следовательно, во-первых, оценка должна быть независима от предложенного варианта КИМ (т. е. выставленная отметка при выполнении одного варианта заданий должна приближенно совпадать с оценкой, полученной при решении другого варианта контрольно-измерительных материалов по этому же предмету на одном и том же уровне знания предмета) и, во-вторых, при выполнении одного и того же варианта тестовых заданий учениками разной степени подготовленности полученные баллы должны различаться, отражая уровень подготовки учащихся.

КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

В Российской Федерации широко распространена четырехбалльная (2, 3, 4, 5) шкала оценивания знаний. Данная шкала является порядковой и позволяет сравнивать между собой уровни знаний учащихся. Используемые цифры можно заменить на «неудовлетворительно», «удовлетворительно», «хорошо» и «отлично» или буквенными обозначениями A, B, C, D. Но нельзя говорить о том, что «2» + «2» = «4», т. е. сумма знаний двух двоечников равна сумме знаний ученика, занимающегося на «4», т. к. ранговая шкала не обладает метрическими свойствами.

Для реализации поставленной задачи (объективного оценивания) возникает потребность введения новой шкалы, способной отражать результаты измерений знаний. При этом новая шкала должна обладать метрическими свойствами.

«Сырые» баллы за выполнение заданий без дополнительной интерпретации лишены должной информативности. Сказать, что кто-то верно решил 10 задач математического теста или 30 заданий теста по русскому языку – ничего или почти ничего не сообщить о том, насколько успешно выполнен тест в целом. Баллы, начисленные за правильное выполнение заданий ЕГЭ, являются результатом измерения с определенной точностью уровня подготовленности испытуемых по соответствующему предмету школьной программы. Например, если испытуемый А получил 95 баллов, а испытуемый В получил 85 баллов, то можно говорить, что уровень подготовленности А больше уровня подготовленности В на 10 конкретных единиц. Смысл этой разницы одинаков на разных частях шкалы [1]. Следовательно, введенная балльная шкала обладает метрическими свойствами.

Наряду с введением качественно новой шкалы нужно обсудить проблему валидности теста. Валидность – понятие, характеризующее то, на сколько хорошо тест измеряет изучаемое явление. Содержательная валидность теста обеспечивается благодаря отбору соответствующих заданий на основе систематического просмотра учебников и учебных программ, консультаций специалистов-предметников, учета опыта составления и применения тестовых процедур в нашей стране и за рубежом, применения научно обоснованных процедур отбора заданий.

В представленной статье обосновывается один из вариантов измерения валидности тестов на основе статистической обработки данных и показывается применение полученной методики при анализе некоторого конкретного случая.

Валидность тестов по русскому языку и математике устанавливалась методом внутренней согласованности, для оценки которой использовался метод контрастных групп [2]. Группы формировались из испытуемых с самыми высокими и самыми низкими показателями (баллами) по результатам теста. Результаты выполнения каждого задания теста группой с высокими баллами сравнивались затем с соответствующими результатами группы с низкими баллами.

КИМ по русскому языку состоит из трех частей. Часть А – задания с выбором ответа. К каждому из них даны 4 варианта ответов, из которых только один правильный. Часть В – задания с ответом в краткой форме, ответ нужно сформулировать самостоятельно. Часть С – состоит из одного задания и представляет собой небольшую письменную работу по приведенному тексту (сочинение). КИМ по математике до 2009 года также состоял из трех частей, но с 2010 года разработчики убрали блок заданий части А.

При сопоставлении значения исследуемого признака (валидности заданий теста) двух групп (задания части А и части В оцениваются по принципу «верно» – «неверно», кроме задания В8 экзамена по русскому языку) были построены двухвходовые таблицы с последующим вычислением коэффициента корреляции Gamma (r_γ). Для сравнения результата выполнения заданий части С был вычислен коэффициент корреляции r_γ.

КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Степень корреляционной связи определяется по величине коэффициента корреляции. Будем говорить, что степень корреляционной связи сильная (или тесная), если значение коэффициента корреляции больше 0,70 [3].

Ниже представлен фрагмент таблицы корреляций результатов выполнения заданий по русскому языку 2004 года для двух групп (рис. 1).

Если абсолютное значение коэффициента получалось меньше 0,70 (задания А11, А15, А18), то это значение рассматривалось как незначимое и задание считалось невалидным (требующем переработки или исключения из теста).

Workbook1* - Gamma Correlations (val_2004.sta)

Gamma Correlations (val_2004.sta)
MD pairwise deleted
Marked correlations are significant at p <,05000

Variable	GR_I_A11	GR_I_A12	GR_I_A13	GR_I_A14	**GR_I_A15**	GR_I_A16	GR_I_A17	GR_I_A18
GR_II_A11	-0,73684	1,00000	-1,00000	1,00000	1,00000	1,00000	1,00000	-0,63158
GR_II_A12		1,00000	0,23077	-0,31915	-0,09434	1,00000	0,16129	0,00000
GR_II_A13			1,00000	1,00000	1,00000	-1,00000	1,00000	1,00000
GR_II_A14				1,00000	1,00000	-0,70000	1,00000	1,00000
GR_II_A15					-0,09434	-0,45455	0,16129	1,00000
GR_II_A16						-0,84211	-0,52632	1,00000
GR_II_A17							1,00000	-0,63158
GR_II_A18								-0,16000
GR_II_A19								
GR_II_A20								

Рис. 1. Фрагмент таблицы корреляций результатов выполнения заданий по русскому языку

Итоговые данные представлены в табл. 1, в скобках указано количество заданий теста.

Таблица 1

Количество невалидных заданий тестов по предметам

Русский язык		***Математика***	
Год сдачи	***Количество невалидных заданий ($r_Y < 0,75$)***	***Год сдачи***	***Количество невалидных заданий ($r_Y < 0,75$)***
2004	7 (41)	2004	1 (27)
2005	3 (38)	2005	0 (26)
2006	5 (40)	2006	0 (26)
2007	3 (40)	2007	1 (26)
2008	2 (40)	2008	2 (26)
2009	2 (39)	2009	1 (26)
2010	1 (39)	2010	0 (18)
2011	0 (39)	2011	0 (20)
2012	0 (39)	2012	1 (20)
2013	1 (39)	2013	1 (20)

Из таблицы видно, что количество невалидных (требующих переработки или исключения из теста) заданий с течением времени уменьшается, что позволяет сделать вывод об улучшении качества тестовых материалов.

Регрессионный анализ является одним из наиболее распространенных методов обработки эмпирических данных при изучении зависимостей между величинами в различных научных областях, а задача исследования существующих связей – одна из важнейших задач статистического анализа данных. Процедура регрессинного ана-

КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

лиза состоит в определении аналитического выражения, объясняющего влияние нескольких независимых переменных (факторов) на зависимую компоненту (результативный признак). При этом множество случайных факторов, оказывающих влияние на результативный признак рассматриваются как нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией.

Для значений средних баллов единого государственного экзамена по русскому языку выпускников общеобразовательных учреждений построим линейную регрессионную модель (рис. 2). Величина коэффициента детерминации $R^2 = 0{,}9225$ говорит о том, что изменчивость значений переменной около линии регрессии составляет 1 – 0,9225 от исходной дисперсии. Другими словами, 92% исходной изменчивости в данной модели могут быть объяснены влиянием независимых факторов (в данном случае валидность теста), а 8% остаточной изменчивости остаются необъясненными, т. е. подвержены влиянию случайных факторов (волнение, незнакомая обстановка и участники экзамена, особый контроль за процедурой сдачи и др.).

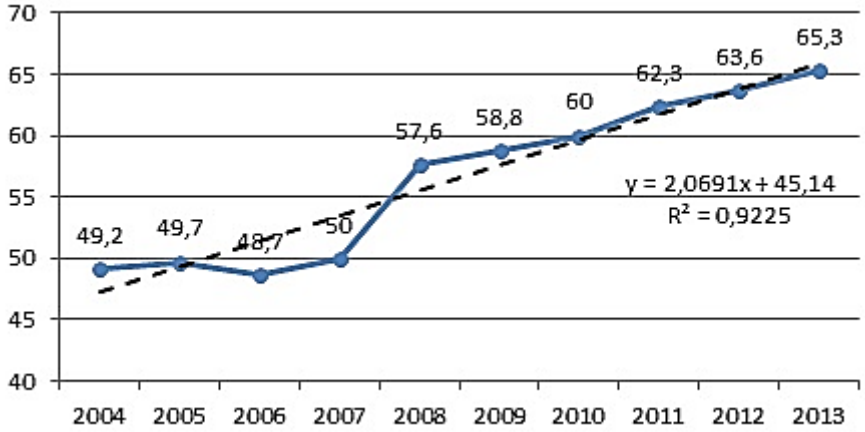


Рис. 2. Линейная модель распределения тестовых баллов по русскому языку

При обосновании статистического вывода следует решить вопрос, являются ли найденные различия существенными или нет. Введем понятие уровня статистической значимости. Уровень статистической значимости – это вероятность того, что мы сочли различия существенными, а они на самом деле случайны. Когда мы указываем, что различия достоверны на 5%-ом уровне значимости, или при $p \leq 0{,}05$, то мы имеем в виду, что вероятность того, что они все-таки недостоверны, составляет 0,05.

Исторически сложилось так, что низшим уровнем значимости является уровень $p \leq 0{,}05$, достаточным – уровень $p \leq 0{,}01$ и высшим – уровень $p \leq 0{,}001$ [3].

В дальнейшем, говоря о значимости коэффициентов, будем иметь в виду статистическую значимость на уровне $p \leq 0{,}05$.

Коэффициенты в уравнении регрессии статистически значимы и это позволяет сделать вывод о том, что построенная модель адекватно описывает распределение средних баллов.

Полученное уравнение задает возрастающую функцию. На графике хорошо видно, что эмпирические значения приближаются к теоретическим и можно сказать, что уровень обучения школьников соответствующему предмету стабилизируется и требования, предъявляемые как к школьному образованию, так и к содержанию тестовых материалов приближаются друг к другу.

КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Распределение средних тестовых баллов по математике лучше всего описывает не линейная модель (R^2 = 0,2227), а полиномиальная ($y = 0{,}2413x^2 - 2{,}6075x + 50{,}642$) степени 2 ($R^2$ = 0,4667). На графике видно (рис. 3), что до 2009 года наблюдается снижение уровня сдачи экзамена по математике, а с 2010 года увеличение результата. Несомненно, на уровне знаний выпускников отразилась не только грамотность составления экзаменационных материалов. На это оказали влияние и другие, внешние, причины, которыми могут быть сложность предмета, социально-политические факторы и др.

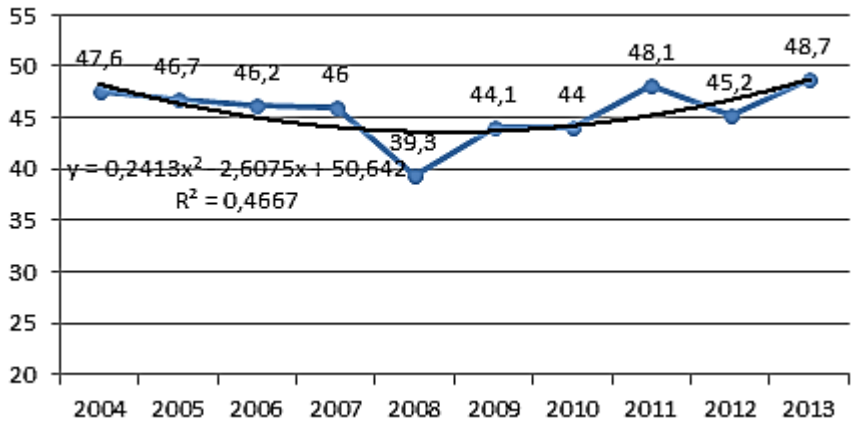


Рис. 3. Модель распределения тестовых баллов по математике

Нужно отметить, что коэффициенты в полиномиальном уравнении регрессии статистически значимы, величина коэффициента R^2 = 0,4667 говорит о том, что построенная модель на достаточно хорошем уровне описывает распределение средних баллов по математике.

Понятно, что в формирование окончательного балла единого государственного экзамена вносят свой вклад выполнение всех заданий всех частей теста. Попытаемся выяснить, какой вклад какая из частей вносит наибольший вклад в формирование итогового балла. В таблице 2 представлены доли выполнения участниками экзамена каждой из частей теста по русскому языку и математике и средние баллы по предметам. Построим линейную регрессионную модель зависимости итогового балла от результатов выполнения заданий частей теста.

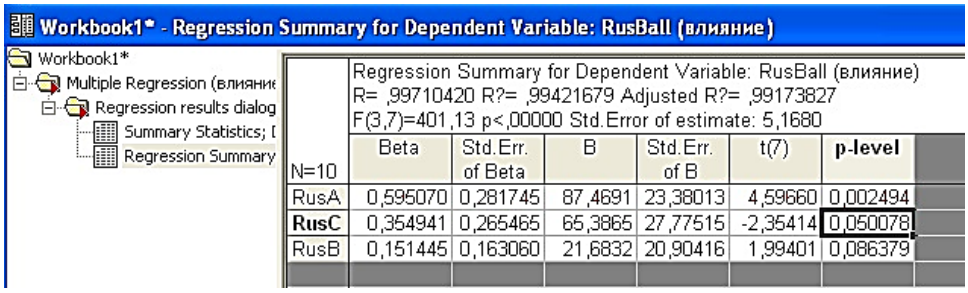
Таблица 2

Средние результаты выполнения частей КИМ и средний балл по предмету

Русский язык					*Математика*				
Год сдачи	***A***	***B***	***C***	***Балл***	***Год сдачи***	***A***	***B***	***C***	***Балл***
2004	0,608	0,284	0,542	49,2	2004	0,620	0,183	0,041	47,6
2005	0,663	0,422	0,527	49,7	2005	0,631	0,278	0,046	46,7
2006	0,682	0,362	0,495	48,7	2006	0,682	0,263	0,057	46,2
2007	0,694	0,387	0,565	50	2007	0,725	0,275	0,036	46
2008	0,689	0,373	0,536	57,6	2008	0,723	0,314	0,041	39,3
2009	0,698	0,423	0,550	58,8	2009	0,814	0,287	0,038	44,1
2010	0,730	0,466	0,565	60	2010		0,547	0,221	44
2011	0,566	0,375	0,292	62,3	2011		0,727	0,075	48,1
2012	0,758	0,574	0,624	63,6	2012		0,689	0,048	45,2
2013	0,741	0,650	0,659	65,3	2013		0,682	0,066	48,7

КОНЦЕПТ
научно-методический электронный журнал
ART 14017 УДК 371.27:31
Буров А. В. Применение методов статистической обработки данных к оценке валидности тестовых материалов единого государственного экзамена // Концепт. – 2014. – № 01 (январь). – ART 14017. – 0,4 п. л. – URL: http://e-koncept.ru/2014/14017.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

В уравнении регрессии построенной модели свободный член равен 0 и уравнение регрессии имеет вид $Ball = k_1 \cdot A + k_2 \cdot B + k_3 \cdot C$, где k_1, k_2, k_3 – коэффициенты при независимых переменных. В данной модели влияние коэффициентов в уравнении регрессии на формирование итогового балла можно увидеть на рисунке 4 (столбец Beta). Если расположить коэффициенты в порядке убывания влияния, то наибольшее влияние оказывает коэффициент при независимой переменной решения задач части А, затем части С и В. Сами же коэффициенты указаны в столбце В.

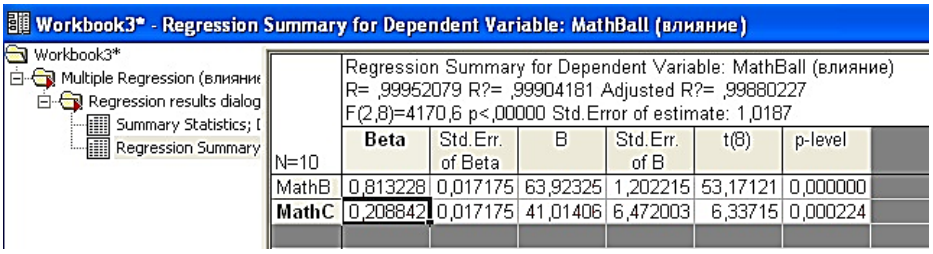


Regression Summary for Dependent Variable: RusBall (влияние)
R= ,99710420 R?= ,99421679 Adjusted R?= ,99173827
F(3,7)=401,13 p<,00000 Std.Error of estimate: 5,1680

N=10	Beta	Std.Err. of Beta	B	Std.Err. of B	t(7)	p-level
RusA	0,595070	0,281745	87,4691	23,38013	4,59660	0,002494
RusC	0,354941	0,265465	65,3865	27,77515	-2,35414	0,050078
RusB	0,151445	0,163060	21,6832	20,90416	1,99401	0,086379

Рис. 4. Коэффициенты уравнения регрессии зависимости среднего балла от выполнения заданий частей теста по русскому языку

В построенной регрессионной модели зависимости балла по математике от выполнения заданий отдельных частей теста на формирование итогового результата оказало влияние выполнение заданий части В (рис. 5). Напомним, что, начиная с 2010 года разработчики КИМ убрали часть А из тестов по математике. Задания же части С оказывают влияние на формирование итогового балла намного меньше, что связано, в первую очередь, с их сложностью – не всякий выпускник приступает к выполнению заданий части С по математике.



Regression Summary for Dependent Variable: MathBall (влияние)
R= ,99952079 R?= ,99904181 Adjusted R?= ,99880227
F(2,8)=4170,6 p<,00000 Std.Error of estimate: 1,0187

N=10	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
MathB	0,813228	0,017175	63,92325	1,202215	53,17121	0,000000
MathC	0,208842	0,017175	41,01406	6,472003	6,33715	0,000224

Рис. 5. Коэффициенты уравнения регрессии зависимости среднего балла от выполнения заданий частей теста по математике

Итак, в работе мы применили метод внутренней согласованности и коэффициентов корреляции Gamma (r_γ) для оценки валидности задании тестовых материалов по математике и русскому языку. Установили, что количество невалидных заданий, выявленных с помощью представленной методики, уменьшается с каждым годом, что позволяет говорить об улучшении качества тестовых материалов.

Пытаясь оценить влияние выполнения заданий отдельных частей экзаменационной работы на итоговый результат сделали соответствующие выводы. Но не стоит забывать, что эти выводы сделаны для достаточно большой выборки, что характеризует саму выборку (в среднем), но не отдельный элемент данной группы.

Построили регрессионные модели для значений средних баллов единого государственного экзамена, рассчитанных по результатам выполнения выпускниками тестовых

материалов по математике и русскому языку в 2004–2013 годах и на основании соответствующих расчетов можно сделать вывод о том, что уровень обучения школьников по этим предметам стабилизируется и требования, предъявляемые как к школьному образованию, так и к содержанию тестовых материалов приближаются друг к другу.

Ссылки на источники

1. Нейман Ю. А., Хлебников В. А. Как оценивается уровень подготовленности учащихся по результатам единого государственного экзамена. – М.: Центр тестирования Минобразования РФ, 2003 – 48 с.
2. Анастази А., Урбина С. Психологическое тестирование. – СПб.: Питер, 2007. – 688 с.
3. Сидоренко Е. В. Методы математической обработки в психологии. – СПб.: ООО «Речь». 2003 – 350 с.

Andrey Burov,
master student, Smolensk State University, Smolensk
burov_andrei@inbox.ru

ISSN 2304-120X

Application of data statistical processing methods for estimation items validity at universal state exam

Abstract. The paper is dedicated to the study of data statistical processing methods for estimation items validity at universal state exam. The question of objectivity testing as a means of assessment of knowledge is topical question for society. For estimation of validity testing items at universal state exam we apply some methods of statistical processing of data. And in this article we propose one of mechanism for processing and control data of deliverables testing items by pupils at universal state exam.

Key words: universal state exam, estimation, validity, test, statistical processing of data.
References: 1–3 – Russian Sources.

Рекомендовано к публикации:
Усачевым В. И. доцентом кафедры прикладной информатики ФГБОУ ВПО «Смоленский государственный университет»;
Горевым П. М., кандидатом педагогических наук, главным редактором журнала «Концепт»