УДК 378
ББК 74.58

В.В. Братищенко

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЭКЗАМЕНАЦИОННЫХ ОЦЕНОК

Предлагается традиционные оценки результатов тестирования применять для изучения качества обучения по оценкам преподавателей. Описываются методики вычисления статистических оценок показателей валидности, дискриминативности и надежности. Приведены значения показателей для оценок одного потока студентов.

Ключевые слова: качество учебного процесса, валидность, дискриминативность, надежность, критерий хи-квадрат, дисперсионный анализ, альфа Кронбаха, коэффициент корреляции.

V.V. Bratishenko

STATISTIC ANALYSIS OF EXAMINATION GRADES

The author suggests applying traditional assessment of test results for studying quality of educational process based on grades given by tutors. The article describes methods of calculating statistic estimates of validity, discrimination and reliability indicators, and demonstrates the indicators for assessing one stream of students.

Keywords: quality of educational process, validity, discrimination, reliability, chi square test, analysis of variance, Cronbach's alpha, correlation coefficient.

Традиционно по экзаменационным оценкам вычисляются показатели успеваемости (доля успевающих) и качества обучения (доля оценок «хорошо» и «отлично»). Эти показатели не позволяют выявить расхождения в методиках экзаменационного оценивания различных предметов преподавателей, оценить значимость расхождений и выявить «проблемные» методики с целью их улучшения. В данной работе для такого анализа предлагается использовать различные известные статистические процедуры.

В классической теории тестирования [1] для оценки качества тестов широко применяются следующие характеристики:

– валидность — свойство оценки правильно характеризовать знания и навыки студента по предмету;

– дискриминативность — способность теста разделять тестируемых по уровню знаний;

– надежность — характеристика точности измерения.

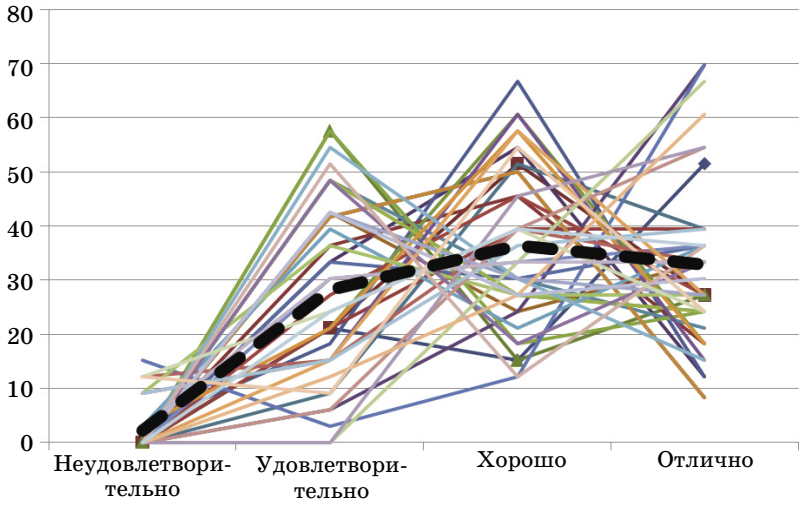
Статистические оценки перечисленных характеристик с некоторыми оговорками могли бы быть применены к оценкам преподавателей.

Одним из способов оценки валидности мог бы быть некоторый, независимый от преподавателя, механизм сравнения знаний студента с эталоном. Такой механизм интенсивно создается в высшей школе в виде системы профессионального тестирования (ФЭПО). Валидность можно оценивать по корреляции между оценками профессионального тестирования и оценками преподавателей.

На рисунке приведено эмпирическое распределение оценок потока студентов по разным предметам. Распределение оценок по всем предметам («неудовлетворительно» — 2,3%, «удовлетворительно» — 28,3%, «хорошо» — 36,4%, «отлично» — 32,8%) достаточно сильно отличает-

http://eizvestia.isea.ru
2011. № 3
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ

ся от распределения по отдельным предметам. Следует сразу объяснить незначительную долю оценок «неудовлетворительно». Приведенные статистические исследования относятся к оценкам, полученным за все время обучения. При этом оценки студентов, отчисленных за неуспеваемость, не попадают в выборку.



Распределение оценок по предметам и по всем предметам (пунктирная линия), %

Анализ частот оценок является очевидной простой методикой изучения одной из сторон дискриминативности оценивания — соответствия оценок общепринятой шкале оценок. Такой анализ можно свести к сравнению эмпирического распределения с некоторым стандартным распределением оценок. Например, в европейской системе ECTS принято следующее распределение оценок: 10% — Exellent (превосходно), 25% — Very good (очень хорошо), 30% — Good (хорошо), 25% — Satisfactory (удовлетворительно), 10% — Sufficient (достаточно). Это распределение имеет моду в точке Good (хорошо). Таким же свойством обладает и распределение оценок, усредненное по всем предметам (пунктирная линия на рисунке). Определение стандартного распределения для оценок российской высшей школы — это самостоятельная задача, выходящая за рамки данной статьи.

Для выработки единого подхода к оценке знаний студентов в рамках одного вуза предлагается сравнивать распределение оценок для каждого предмета с эмпирическим распределением по всем предметам и выполнять проверку гипотезы о совпадении распределений, например по критерию хи-квадрат. Для этого можно вычислять следующую статистику, имеющую распределение $\chi^2(3)$ хи-квадрат с тремя степенями свободы:

$$t_i = \sum_{j=2}^{5} \frac{(e_{ij} - m_i \bar{p}_j)^2}{m_i \bar{p}_j},$$

где e_{ij} — количество оценок $j \in \{2, 3, 4, 5\}$ для i-го предмета, $i = 1, \ldots, n$; n — количество предметов;

$m_i = \sum_{j=2}^{5} e_{ij}$ — количество оценок по i-му предмету;

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2011. № 3
http://eizvestia.isea.ru

ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
2011. № 3
http://eizvestia.isea.ru

$\overline{p}_j = \sum_{i=1}^{n} e_{ij} \Big/ \sum_{j=2}^{5} \sum_{i=1}^{n} e_{ij}$ — частота оценки j эмпирического распределения по всем предметам.

В табл. 1 приведены вычисления статистики для разных предметов, упорядоченные по возрастанию значений критерия (показаны первые и последние предметы в последовательности). Последние предметы имеют значительный перекос в распределении оценок.

Таблица 1

Значения статистики хи-квадрат для различных предметов

Предмет	Оценка, %				t_i	$P\{\chi^2(3) > t_i\}$
	Неудовлетворительно	Удовлетворительно	Хорошо	Отлично		
Лингвистическое обеспечение информационных систем	3,0	30,3	33,3	33,3	0,16	0,983 581 2
Предметно-ориентированные экономические информационные системы	0,0	30,3	33,3	36,4	1,11	0,775 436 1
Основы организации цифровых систем обработки информации	0,0	24,2	39,4	36,4	1,25	0,741 391 2
...						
Проектирование информационных систем	9,1	15,2	60,6	15,2	15,9	0,001 165 9
Философия	0,0	6,1	24,2	69,7	21,6	0,000 077 9
Экономическая теория (микроэкономика 1)	0,0	0,0	33,3	66,7	21,7	0,000 072 0

Для изучения влияния дисциплины на оценку можно применить дисперсионный анализ [4]. Пусть x_{ik} — оценка k-го студента ($k = 1, \ldots, m$) по i-му предмету ($i = 1, \ldots, n$). Сравниваются усредненные выборочные дисперсии по предметам

$$M_2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \overline{x}_{i*})^2, \qquad \overline{x}_{i*} = \frac{1}{m} \sum_{j=1}^{m} x_{ij}$$

с межгрупповой дисперсией

$$M_1 = \frac{m}{n-1} \sum_{i=1}^{n} (\overline{x}_{i*} - \overline{x}_{**})^2, \qquad \overline{x}_{**} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij},$$

которые в случае отсутствия влияния предмета на оценку являются оценками дисперсии оценок студентов. Статистика $F = M_1 / M_2$, при условии одинакового нормального распределения и независимости вариаций среди оценок одного предмета, будет иметь распределение Фишера со степенями свободы $n - 1$ и $mn - n$.

В исследованиях по статистике отмечается, что дисперсионный анализ устойчив по отклонению от нормальности, однородности дисперсии, асимметрии распределения [2]. Вычисления по оценкам одного потока студентов ($n = 37$ предметов и $m = 33$ студента) дают следующие значения:

$M_1 = 3{,}14$ и $M_2 = 0{,}63$, которые при выполнении гипотезы об отсутствии влияния фактора являются оценками дисперсии оценок. Уровень значимости соответствующего критического значения статистики $F = 4{,}95$ составляет $2{,}7 \cdot 10^{-17}$. Даже с учетом отклонений от классических условий применения дисперсионного анализа следует признать гипотезу об отсутствии влияния предмета на оценки не прошедшей статистическую проверку. Это свидетельствует о различиях в методиках преподавания и оценивания знаний.

Приведенные статистические процедуры позволяют изучить «перекосы» в оценках по отдельным предметам без учета согласованности оценок, полученных одним студентом по разным предметам. Очевидно, что использование согласованности позволило бы строить более точные статистические оценки. Основанием для этого является тот факт, что для подавляющего большинства студентов не играет особой роли специфика предмета: способный студент примерно одинаково справляется и с профильными, и с непрофильными дисциплинами. Кроме этого, возможно выполнять исследования по однородной группе дисциплин, например по дисциплинам гуманитарного цикла.

Для оценки дискриминативности на основе согласованности оценок, полученных по разным предметам, можно применять коэффициенты корреляции. Парные корреляции не совсем подходят для оценки отдельного предмета, поэтому предлагается использовать коэффициенты корреляции оценок по предмету со средними оценками студентов:

$$y_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}, \quad k = 1, \ldots, m. \tag{1}$$

Согласованность оценок x_{ik}, $k = 1, \ldots, m$ по i-му предмету со средними оценками (1) можно оценивать с помощью оценки коэффициента корреляции:

$$\bar{\hat{r}}_i = \frac{\frac{1}{m}\sum_{k=1}^{m} x_{ik}y_k - \left(\frac{1}{m}\sum_{k=1}^{m} x_{ik}\right)\left(\frac{1}{m}\sum_{k=1}^{m} y_k\right)}{\bar{\hat{\sigma}}_i \bar{\hat{\sigma}}_y},$$

$$\bar{\hat{\sigma}}_i^2 = \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - \left(\frac{1}{m}\sum_{k=1}^{m} x_{ik}\right)^2,$$

$$\bar{\hat{\sigma}}_y^2 = \frac{1}{m}\sum_{k=1}^{m} y_k^2 - \left(\frac{1}{m}\sum_{k=1}^{m} y_k\right)^2.$$

Чем ближе коэффициент корреляции к 1, тем выше согласованность. В табл. 2 приведены коэффициенты корреляции предметов одного потока студентов. Некоторые предметы, показавшие плохую согласованность по распределению оценок (например, «Проектирование информационных систем»), оказались ближе к ряду средних оценок. Дополнительно в табл. 2 приведена оценка $\bar{\hat{\sigma}}_{r_i}$ среднеквадратического отклонения, характеризующая точность оценки коэффициентов корреляции. Эта оценка вычислялась на основе дисперсии коэффициента корреляции [3]:

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ
(Байкальский государственный университет экономики и права)
2011. № 3
http://eizvestia.isea.ru

$$\bar{\hat{\sigma}}_r^{\,2} \to D[r] = \frac{r^2}{4m}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right),$$

где $\mu_{ij} = M\left[(X - M[X])^i (Y - M[Y])^j\right]$ — смешанные центральные моменты.

Таблица 2

Значения оценок коэффициента корреляции рядов оценок различных предметов с рядом средних оценок

Название строки	$\bar{\hat{r}}_i$	$\bar{\hat{\sigma}}_{r_i}$
Математика	0,864	0,033
Математическая экономика	0,850	0,035
Численные методы	0,826	0,050
...		
Проектирование информационных систем	0,757	0,062
...		
Лингвистическое обеспечение информационных систем	0,694	0,106
...		
Предметно-ориентированные экономические информационные системы	0,624	0,116
...		
Философия	0,405	0,150
Экономическая теория (микроэкономика 1)	0,399	0,129

Для оценки дискриминативности тестов применяется метод контрастных групп. Из общей совокупности испытуемых выделяют две подгруппы — лучшую и худшую. Тогда индекс дискриминативности может быть определен как разность между долями испытуемых, правильно выполнивших задание в этих двух подгруппах. Индекс дискриминативности изменяется от −1 (в лучшей группе никто не справился, в худшей — все справились) до +1 (в лучшей все справились, в худшей не справился никто).

В применении к традиционным оценкам высшей школы можно для определения индекса I_i дискриминативности i-го предмета воспользоваться разностью $I_i = (\bar{x}_i' - \bar{x}_i'')/2$ средней оценки

$$\bar{x}_i' = \frac{1}{m'} \sum_{j \in J'} x_{ij}$$

множества J' оценок лучших студентов в количестве m' и средней оценки

$$\bar{x}_i'' = \frac{1}{m''} \sum_{j \in J''} x_{ij}$$

множества J'' оценок худших студентов в количестве m''. Деление на два не принципиально. В данном случае оно выполнено для приведения коэффициента дискриминативности к интервалу от −1 до 1, так как средние оценки лежат в интервале от 3 до 5. Результаты вычислений приведены в табл. 3. Группы лучших и худших студентов определялись по средним оценкам студентов. Таблицы 2 и 3 достаточно близки по содержанию, по крайней мере в отношении наиболее «проблемных» предметов.

http://eizvestia.isea.ru
2011. № 3
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2011. № 3
http://eizvestia.isea.ru

Таблица 3

Значения оценок коэффициента дискриминативности различных предметов

Предмет	I_i
Математическая экономика	0,954 545
Математика	0,909 091
...	
Численные методы	0,772 727
...	
Лингвистическое обеспечение информационных систем	0,681 818
Проектирование информационных систем	0,681 818
...	
Предметно-ориентированные экономические информационные системы	0,590 909
...	
Философия	0,227 273
Экономическая теория (микроэкономика 1)	0,227 273

Для измерения надежности тестов часто используют показатель альфа Кронбаха. Этот показатель можно вычислять и по оценкам преподавателей:

$$\alpha = \frac{n}{n-1}\left(\frac{\bar{\sigma}_Z^2 - \sum_{i=1}^{n}\bar{\sigma}_i^2}{\bar{\sigma}_Z^2}\right).$$

В альфа Кронбаха сравниваются дисперсия $\bar{\sigma}_Z^2 = n^2\bar{\sigma}_Y^2$ суммы баллов студентов

$$z_k = \sum_{i=1}^{n} x_{ik} = ny_k, \quad k = 1, \ldots, m$$

и сумма дисперсий оценок $\bar{\sigma}_i^2$. В случае совпадения оценок студента по каждому предмету со средней по студенту (максимальная согласованность оценок по разным предметам) альфа Кронбаха будет равна 1. Когда никакой согласованности нет, оценка студента по предмету является случайной величиной, независимой от других его оценок, дисперсия суммы будет равна сумме дисперсий и альфа Кронбаха будет равна 0. В классической статистической теории тестирования тест считается надежным, если $\alpha > 0{,}8$. Для оценок потока студентов, рассмотренного в этой статье, $\alpha = 0{,}969$.

Результаты вычислений демонстрируют применимость предложенных методик для изучения валидности, дискриминативности и надежности экзаменационных оценок. Однако любые статистические выводы не допускают чисто механического применения в силу своей стохастической природы. Тем более это справедливо в отношении оценок преподавателей. Статистические показатели оценок преподавателей следует использовать для выявления экзаменов с «плохими» характеристиками, а исследование недостатков преподавания и оценивания знаний нужно выполнять другими методами.

Список использованной литературы

1. Аванесов В.С. Композиция тестовых заданий / В.С. Аванесов. — М.: АДЕПТ, 1998. — 216 с.
2. Дисперсионный анализ: Предположения и последствия их нарушения. — URL: http://www.statsoft.ru/home/textbook/modules/ stanman#assumptions.
3. Крамер Г. Математические методы статистики / Г. Крамер. — М.: Мир, 1975. — 648 с.
4. Хьютсон А. Дисперсионный анализ / А. Хьютсон; пер с англ. А.Г. Кругликова. — М.: Статистика, 1971. — 88 с.

Referenses

1. Avanesov V.S. Kompozitsiya testovykh zadanii / V.S. Avanesov. — M.: ADEPT, 1998. — 216 s.
2. Dispersionnyi analiz: Predpolozheniya i posledstviya ikh narusheniya. — URL: http://www.statsoft.ru/home/textbook/modules/ stanman#assumptions.
3. Kramer G. Matematicheskie metody statistiki / G. Kramer. — M.: Mir, 1975. — 648 s.
4. Kh'yutson A. Dispersionnyi analiz / A. Kh'yutson; per s angl. A.G. Kruglikova. — M.: Statistika, 1971. — 88 c.

Информация об авторе

Братищенко Владимир Владимирович — кандидат физико-математических наук, доцент, начальник информационного управления, Байкальский государственный университет экономики и права, г. Иркутск, e-mail: bvv@isea.ru.

Author

Bratishenko Vladimir Vladimirovich — PhD in Physical and Mathematical Sciences, Associate Professor, Head of Information Office, Baikal State University of Economics and Law, Irkutsk, e-mail: bvv@isea.ru.

http://eizvestia.isea.ru
2011. № 3
ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ
(Байкальский государственный университет экономики и права)