М.П. Карпенко, *доктор технических наук, профессор*
А.В. Абрамова
В.А. Басов, *кандидат физико-математических наук*
А.В. Слива , *кандидат технических наук, старший научный сотрудник*

Методы математической статистики для выявления нарушений конфиденциальности тестовых баз недобросовестными студентами

Статья посвящена разработки метода выявления нарушения конфиденциальности тестовых баз вуза недобросовестными студентами. Предложен подход, позволивший сформулировать и решить задачи математической статистики, дающий возможность определить, использовал ли студент при тестировании знаний заранее известные ему верные ответы или нет. Представлены примеры, иллюстрирующие эффективность предлагаемого метода.

Ключевые слова: *высшее образование, математическое ожидание, дисперсия, выборка, объем выборки, статистика, статистическая гипотеза, критерий Крамера-Уэлча, объективная оценка знаний, тестовая база, конфиденциальность.*

Обеспечение объективности – одна из ведущих проблем оценивания знаний студентов. Компьютерное тестирование полностью исключает отношение к личности студента со стороны преподавателя как искажающий оценку субъективный фактор. Однако практика высшего образования показывает, что недобросовестные студенты стремятся различными способами получить в свое распоряжение ответы на вопросы тестовых баз. При этом зачастую усилия вуза по защите информации не приносят успеха. Это приводит к нарушениям объективности оценивания, ставит недобросовестного студента, использующего краденые ответы на тесты, в преимущественное положение по сравнению с добросовестными учащимися. Если такие ситуации не пресечь, то это влечет за собой не только снижение качества обучения, но и способствуют формированию у студентов пренебрежительного отношения к этическим нормам и правилам поведения, снижает уровень воспитательной работы как неотъемлемой части образовательного процесса.

В этой связи в настоящей работе ставится задача выявления нарушения конфиденциальности тестовых баз путем периодической замены части тестовых

вопросов на новые и сравнения между собой результатов ответов на совокупности «старых» и «новых» вопросов.

Постановка задачи

По некоторой учебной дисциплине имеется две тестовых базы достаточно большого объема:

A – «старая база»;

B – «новая база».

Предполагается, что базы A и B эквивалентны по сложности, т.е. при тестировании, в среднем, доля правильных ответов студентов на вопросы из базы A и вопросы из базы B будут приблизительно равны.

Есть предположение, что ответы на старую базу A стали известны студентам (т. е. конфиденциальность базы нарушена и студенты получают «подсказку»). В этом случае отвечать на вопросы из базы A студенты будут «заведомо лучше», чем на вопросы из базы B. Как подтвердить или опровергнуть факт наличия «подсказки» путем анализа результатов тестирования на основе баз A и B, если про базу B точно известно, что по ней «подсказок» нет.

Математическая постановка задачи

Предполагаем, что с точки зрения контроля знаний базы A и B эквивалентны, т. е. при тестировании по обеим базам в отсутствие подсказок будут равны усредненные доли правильных ответов студентов, при достаточно большом количестве протестированных по обеим базам (более 100 чел.).

Каждому студенту выдается тестовое задание, состоящее из 40 вопросов, в которое случайным образом отбирается по 20 вопросов из баз A и B. По таким тестовым заданиям тестируется n студентов (n – объем выборки, который следует принять ≈ 100).

Введем две случайные величины:

X – доля правильных ответов студента на вопросы базы A из 20 вопросов, вошедших в предъявленное ему тестовое задание;

Y – доля правильных ответов студента на вопросы базы B из 20 вопросов, вошедших в предъявленное ему тестовое задание.

Обозначим:

MX – математическое ожидание случайной величины X (доли правильных ответов, на вошедшие в тестовое задание 20 вопросов из базы A),

MY – математическое ожидание случайной величины Y (доли правильных ответов, на вошедшие в тестовое задание 20 вопросов из базы B).

Тогда поставленная задача сводится к последовательному решению двух задач проверки статистических гипотез.

Задача 1

Проверить статистическую гипотезу равенства математических ожиданий случайных величин X и Y – гипотезу MX = MY при конкурирующей гипотезе MX ≠ MY.

Если подтвердится статистическая гипотеза MX = MY, то это будет означать, что конфиденциальность базы A не нарушена, на чем анализ возможности нарушения конфиденциальности заканчивается.

Если окажется, что различие MX и MY – статистически значимое, то требуется рассмотрение двух вариантов.

Первый вариант: в среднем ответы по базе A могут оказаться статистически значимо хуже ответов по базе B.

Но поскольку мы предположили в постановке задачи анализа возможности нарушения конфиденциальности базы A, что базы A и B эквивалентны, то этот случай исключается из рассмотрения.

Поэтому переходим к рассмотрению второго варианта: в среднем ответы по базе A могут оказаться статистически значимо лучше ответов по базе B, т. е. возникает подозрение, что конфиденциальность базы A нарушена. Это подозрение можно проверить, решив следующую задачу.

Задача 2

Проверить статистическую гипотезу MX = MY при конкурирующей гипотезе MX > MY). Тем самым подтвердить или опровергнуть предположение о нарушении конфиденциальности базы A.

Теоретические предпосылки решения задач 1 и 2

Многочисленные исследования показывают [1], что результаты педагогических измерений знаний учащихся распределены по нормальному закону. Таким образом, поставленная задача – это определение факта наличия статистической значимости различия математических ожиданий двух нормальных распределений X и Y, т. е. MX и MY.

В этом случае производится проверка гипотезы MX = MY. Чаще всего для решения такой задачи применяется t-критерий Стьюдента. Однако в нашем случае этот подход не годится, поскольку он работает только в случае равных или известных дисперсий, а в нашем случае равенство дисперсий случайных величин X и Y заранее неизвестно, как и сами значения дисперсий случайных величин X и Y. Поэтому следует предполагать возможность того, что они не равны.

В таком случае для проверки указанной гипотезы следует использовать критерий Крамера-Уэлча [2], основанный на статистике

$$T = \frac{\sqrt{mn}(\mu X - \mu Y)}{\sqrt{nS(X) + mS(Y)}},$$

где S(X) и S(Y) – выборочные значения дисперсии X и Y; μX и μY – выборочные средние значения, полученные соответственно по выборкам объема *m* и *n*.

В нашем случае каждая выборка, как уже отмечалось выше, представляет блок из 40 тестовых заданий, из которых 20 – из базы A и 20 из базы B.

Таким образом, размеры выборок для получения указанных выборочных статистических характеристик случайных величин X и Y равны ($m = n$).

Тогда статистика, используемая для критерия Крамера-Уэлча, примет вид

$$T = \frac{\sqrt{n}(\mu X - \mu Y)}{\sqrt{S(X) + S(Y)}}. \quad (1)$$

В [3] показано, что закон распределения T при больших n (порядка 100–200 вполне достаточно) близок к нормальному. Отсюда вытекает правила принятия решения по критерию Крамера-Уэлча.

Отметим важное для дальнейшего изложения свойство статистики Крамера-Уэлча: в [4] показано, что точность аппроксимации статистики Крамера-Уэлча предельным стандартным нормальным распределением вполне удовлетворительна даже при небольших объемах выборок порядка 10 и различных (необязательно нормальных) функциях распределений случайных величин X и Y. Это существенно расширяет возможности применения рассматриваемого критерия.

Решение задачи 1

Проверяется статистическая гипотеза MX = MY при конкурирующей гипотезе MX ≠ MY.

В этом случае, согласно [5], Ткр (критическое значение критерия) определяется из уравнения

$$\Phi(\text{Ткр}) = (1-\alpha)/2, \quad (2)$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz$ – функция Лапласа.

Суть Ткр заключается в том, что если при выбранном уровне значимости α будет

$$|T| \le \text{Ткр},$$

то гипотеза равенства математических ожиданий MX и MY принимается.

Если

$$|T| > \text{Ткр},$$

то принимается конкурирующая гипотеза MX ≠ MY.

Пример 1

Пусть по тестовым заданиям, включающим 20 вопросов из базы A и 20 вопросов из базы B (всего 40 вопросов каждому студенту), протестированы $n = 150$ студентов.

При этом расчеты выборочных значений математических ожиданий X и Y дали результат μX = 0,75 и μY = 0,6, а расчет выборочных значений дисперсий дал S(X) = 0,12 и S(Y) = 0,09.

Согласно (1), получим

$$T = \frac{\sqrt{n}(\mu X - \mu Y)}{\sqrt{S(X) + S(Y)}} = \frac{\sqrt{150}(0,75 - 0,6)}{\sqrt{0,12 + 0,09}} = 4,008919 \approx 4,01.$$

Проведем расчет для разных значений уровня значимости α.

Для уровня значимости α = 0,05, согласно (2), для значения Ткр получим уравнение: Ф(Ткр) = (1–α)/2 = (1–0,05)/2 = 0,475.

По таблицам функции Лапласа для значения Ф(Ткр) = 0,475 получим Ткр = 1,96.

Поскольку 4,01>1,96, то гипотеза равенства MX и MY отвергается и принимается альтернативная гипотеза – MX ≠ MY.

Ужесточим уровень значимости, уменьшив его до α = 0,01. Тогда, согласно (2), для значения Ткр получим уравнение

$$Ф(Ткр) = (1–α)/2 = (1\text{-}0,01)/2 = 0,495.$$

Согласно таблицам функции Лапласа, для этого значения Ф(Ткр) получим Ткр = 2,58. Поскольку 4,01>2,58, то гипотеза равенства MX и MY также отвергается и принимается альтернативная гипотеза – MX ≠ MY.

При дальнейшем ужесточенном уровня значимости, например, до α = 0,001, также будет принята альтернативная гипотеза MX ≠ MY, поскольку для этого уровня значимости легко подсчитать Ф(Ткр) = 0,4995, и согласно таблицам функции Лапласа, определить соответствующее значение Ткр = 3,3<4,009.

Таким образом, в рассмотренном примере гипотеза равенства ответов студентов на вопросы из баз A и B отвергнута. Поэтому необходима проверка подозрения на нарушение конфиденциальности базы A (задача 2).

Решение задачи 2

Проверяется статистическая гипотеза MX = MY при конкурирующей гипотезе MX > MY.

В этом случае, согласно [5], Ткр определяется из уравнения

$$Ф(Ткр) = (1–2α)/2, \qquad (3)$$

где Ф(x) – функция Лапласа.

Если при выбранном уровне значимости α будет

$$T \le Ткр,$$

то гипотеза равенства математических ожиданий MX и MY принимается.

Если T>Ткр, то принимается конкурирующая гипотеза MX > MY.

ПРИМЕЧАНИЕ. Если все-таки окажется, что T < 0 (т. е. μX < μY, означающее, что в среднем по выборке студенты лучше отвечали на вопросы новой базы B, чем старой базы A), то это будет означать, что наше исходное предположение об эквивалентности баз A и B по сложности может оказаться неверным и требует проверки.

Пример 2

Продолжим рассмотрение ситуации тестирования, изложенной в примере 1. В этих условиях решение задачи 2 сводится к проверке гипотезы MX = MY при конкурирующей гипотезе MX > MY.

Начнем с уровня значимости α = 0,05. Согласно (3), получим Ф(Ткр) = (1–2α)/2 = 0,45, откуда по таблицам функции Лапласа получим Ткр = 1,64.

Поскольку T = 4,01 > Tкр = 1,64, то при уровне значимости α = 0,05 принимается конкурирующая гипотеза MX > MY.

Поэтому в данном случае с доверительной вероятностью 1–0,05 = 0,95 можно утверждать, что имеет место нарушение конфиденциальности базы A.

Рассмотрение более жестких уровней значимости, например, α = 0,01 и α = 0,001 дает соответственно значения Tкр = 2,33 и Tкр = 3,1, т. е. на уровнях доверительной вероятности 0,99 и 0,999 также принимается статистическая гипотеза, означающая нарушение конфиденциальности базы A.

Таким образом, предложенная методика позволяет выявить возможные нарушения конфиденциальности тестовой базы A при безусловной конфиденциальности базы B и эквивалентности баз A и B по сложности. Методика может быть достаточно легко реализована в виде компьютерной программы, которая будет в состоянии осуществлять оперативный контроль нарушения конфиденциальности тестовых баз. Это, в свою очередь, помимо обеспечения наказания недобросовестных студентов и соответственно повышения объективности контроля знаний, позволит снижать расходы на обновление тестовых баз, не заменяя остающиеся актуальными по содержанию тестовые базы до момента их хищения.

Литература

1. Бодряков В.Ю., Фомина Н.Г. Простая вероятностно-статистическая модель количественной оценки уровня знаний учащихся // Вестник Ивановского государственного энергетического университета. 2008. № 7.

2. Крамер Г. Математические методы статистики: Пер. с англ. 2-е изд. М.: Мир, 1975.

3. Орлов А.И. Устойчивость в социально-экономических моделях. М.: Наука, 1979.

4. Орлов А.И. Эконометрика: Учебник. М.: Экзамен, 2002.

5. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998.

VIRTUAL TECHNOLOGIES

Karpenko M.P., *Doctor in Technical Sciences, professor*
Abramova A.V.
Basov V.A., *Candidate of Physical and Mathematical Sciences*
Sliva A.V., *Candidate of Technical Sciences, Senior Staff Scientist*

Methods of Mathematical Statistics for Detecting Test Databases Confidentiality Violations by Unscrupulous Students

The article is devoted to develop a method to identify a breach of confidentiality of University's test databases by unscrupulous students. It was proposed the approach

which allows to formulate and solve tasks of mathematical statistics, giving the opportunity to determine whether the student while testing knew in advance the right answers or not. The examples for illustration the efficiency of the proposed method are presented.

Key words: *higher education, mathematical expectation, dispersion, sampling, sample size, statistics, statistical hypothesis, the criterion Kramer-Welch, objective assessment of knowledge, test database, confidentiality.*