



Е. А. Ровба, В. К. Бойко, Ю. А. Войтукевич, В. Е. Лявшук, С. В. Петров

КАЧЕСТВО ОБРАЗОВАНИЯ: DATA-MINING БАЗ ДАННЫХ РЕЗУЛЬТАТОВ ЦЕНТРАЛИЗОВАННОГО ТЕСТИРОВАНИЯ

Представлены результаты применения методов data-mining для измерения качества абитуриентов и студентов, а также снижения затрат на привлечение целевой группы на непопулярные специальности. Введен новый количественный показатель качества абитуриента, рассчитанный методом принципиальных компонент на основе данных централизованного тестирования и названный индикатором системности знаний абитуриента. Показано использование индикатора системности знаний для определения целевой аудитории для прямой почтовой рассылки писем. Для оценки достоверности и определения величины эффекта рассылки построена статистическая бутстреп-процедура. Оценен количественный эффект от примененной методики профориентации и рассчитана стоимость привлечения одного абитуриента при помощи прямой рассылки писем.

Ключевые слова: методы data-mining, централизованное тестирование, индикатор системности знаний, статистическая бутстреп-процедура, эффективность университетского маркетинга.



E. A. Rouba, V. K. Boiko, Yu. A. Vaitukevich, V. E. Liaushuk, S. V. Petrov

The quality of education: using data-mining of the centralized testing results

The paper presents results of research of entrants and students quality by means of data-mining. The authors introduce a new quantitative indicator of a university entrant quality. The indicator is derived from results of the centralized testing by using the principal components and classification analysis. The quality indicator is considered as a systematic knowledge indicator. The results were used in marketing of the university. It is shown how an effect of the direct mailing was exposed and how the mail efficiency was estimated. The target group was defined a priori by using a systematic knowledge indicator. Bootstrapping was used to reveal the direct mailing effect from a superposition of other marketing efforts towards the university entrants and to validate the effect. The efficiency of the applied method to enroll the target group of entrants was equal. Specific cost to attract the target entrant by the direct mail was calculated.

Keywords: data-mining, national centralized testing, a systematic knowledge indicator, bootstrapping, efficiency marketing of university.

Массовый характер высшего образования обусловил возникновение новых проблем. Одна из них заключается в том, что вузы регулярно сталкиваются с ситуацией недобора абитуриентов по определенным направлениям подготовки. Как правило, эти направления социально значимы (например, учителя математики или физики, инженеры для промышленности и т. д.). Чтобы сохранить рабочие места преподавателей

и выполнить госзаказ по подготовке специалистов, руководство вуза часто решает зачислять на непопулярные специальности всех подавших заявления. В результате на такие специальности, как теоретическая физика и математика, идут и те, кто не мыслит жизнь вне науки, и те, кому количество баллов в сертификате не позволило претендовать на что-либо более модное. Впоследствии если первых будет больше, чем вто-

рых, то вторые будут отчислены за неуспеваемость. Если этого не произойдет, то вуз снизит стандарты качества обучения, и сильные студенты уйдут в другие вузы или опустятся на уровень слабых. Оба варианта ведут к снижению способности университета удовлетворять потребности непосредственных потребителей (студентов) и других заинтересованных сторон (работодателей и общества в целом).

Управлять потоками абитуриентов при помощи ценовых факторов вузу не позволяет государственное ценообразование в сфере образовательных услуг. Поэтому руководство вузов стремится управлять спросом при помощи мер стимулирования. В целом это достаточно разработанное и обоснованное в теоретическом плане направление деятельности. Однако практическая реализация эффективных механизмов управления такой деятельностью является, как правило, оригинальным ноу-хау каждого вуза и примером для остальных.

Первая задача, с которой сталкивается вуз при попытке управления спросом на свои специальности, связана с определением целевой аудитории на базе объективных показателей. Парадоксально, но оценка по централизованному тестированию (далее — ЦТ) таким показателем не является, ибо главная цель централизованного тестирования — не определение объективного уровня знаний абитуриентов, а обеспечение условий для соблюдения принципа социальной справедливости при заполнении бюджетных мест в вузах [3]. Поэтому сертификаты о прохождении ЦТ выдаются даже при результате в один балл. Вуз же заинтересован в том, чтобы на все его специальности, в том числе и на непопулярные, захотели прийти учиться именно те, кто способен учиться. Особенно остро этот вопрос стоит для регионального вуза, для которого абитуриент с определенным набором сертификатов ЦТ является редким ресурсом.

В 2006–2009 уч. гг. в Гродненском государственном университете им. Янки Купалы (далее — ГрГУ, Республика Беларусь) разрабатывалась методика определения целевой аудитории факультета на основе анализа данных о результатах ЦТ в Гродненском регионе [2]. При помощи этой методики из генеральной совокупности абитуриентов выбиралась целевая аудитория, которая затем подвергалась воздействию определенных инструментов из маркетингового комплекса средств продвижения.

По мере реализации этой деятельности актуализировалась задача количественной оценки

ее экономической эффективности. Подобная проблема не нова, так как затраты на стимулирование продаж, к сожалению, обычно относят на непроизводительные издержки. С течением времени на получение одинакового покупательского отклика необходимо тратить все больше средств на рекламу и при этом диверсифицировать каналы распространения рекламы. Отсюда возникает естественное желание, в нашем случае руководства вуза, оптимизировать затраты на рекламу на базе разумных расчетов, учитывающих реальность конкретной ситуации рекламирования.

При выявлении целевой аудитории конкретного факультета данные о результатах централизованного тестирования являются вторичными, так как они были собраны Республиканским институтом контроля знаний для своих целей. В настоящее время эти данные доступны вузу практически бесплатно и в имплицитной форме содержат информацию, необходимую для решения маркетинговых задач. Однако для перевода информации из имплицитной формы в явную необходимы методы, получившие название data-mining. Этим английским словосочетанием, которое пока не имеет устоявшегося перевода на русский язык, называется совокупность методов обнаружения в данных ранее неизвестной, нетривиальной, практически полезной и доступной интерпретации информации, необходимой для принятия решений. К их числу относят также и факторный анализ.

В качестве заказчика исследования результатов централизованного тестирования методами data-mining выступил факультет математики и информатики ГрГУ, испытывавший несколько лет подряд трудности с набором будущих учителей математики.

Шаг первый: выявление факторной модели данных

Анализ результатов прошедшего в Гродненской области централизованного тестирования на базе описательной статистики позволил получить тривиальные выводы. Например, о том, что плотность распределения абитуриентов, хорошо сдающих тест по иностранному языку, выше в областном центре, чем в глубинке. Естественно предполагать, что в областном центре, где расположено несколько вузов, количество репетиторов по английскому значительно выше, чем в районном центре с двумя средними школами.

После исчерпания информационного потенциала описательной статистики было принято решение применить к базе данных результатов ЦТ факторный анализ по методу принципиальных компонент [1]. Предполагалось выяснить, под влиянием скольких факторов находится совокупность этих данных. В результате процедуры факторного анализа были выявлены простые факторные модели тестов по предметам. Простая структура этих моделей образована одним или, значительно реже, двумя факторами. Однофакторная модель результатов ЦТ наиболее адекватна идеологии, заложенной в процедуру централизованного тестирования. Эта идеология построена на декларации, что тест измеряет только знания абитуриента, и ничего больше. Выявление иных значимых факторов фактически означает дефект теста и делает процедуру тестирования неадекватной цели, для которой она предназначена.

Факторное пространство многомерно, и поэтому индивидуальные достижения абитуриентов при тестировании располагаются в этом пространстве облаком. Простая же сумма баллов,

полученных абитуриентами на ЦТ, редуцирует это облако до одномерной прямой линии. Величина первого фактора не равна сумме баллов за ответы теста. Она дополнительно учитывает взаимную корреляцию правильных и неправильных ответов и демонстрирует индивидуальность достижений каждого абитуриента. Фактически первый фактор характеризует системность знаний абитуриента и дифференцирует абитуриентов с одинаковой суммой баллов между собой (рис. 1).

После обработки результатов тестов по всем предметам получена аналогичная картина. Это позволило предположить, что методом принципиальных компонент на основе данных централизованного тестирования можно рассчитывать новый количественный показатель качества абитуриента, который фактически является индикатором системности знаний абитуриента. Численное значение индикатора системности знаний абитуриента может быть равно его балльной оценке по тесту ЦТ либо быть выше или ниже этой оценки.

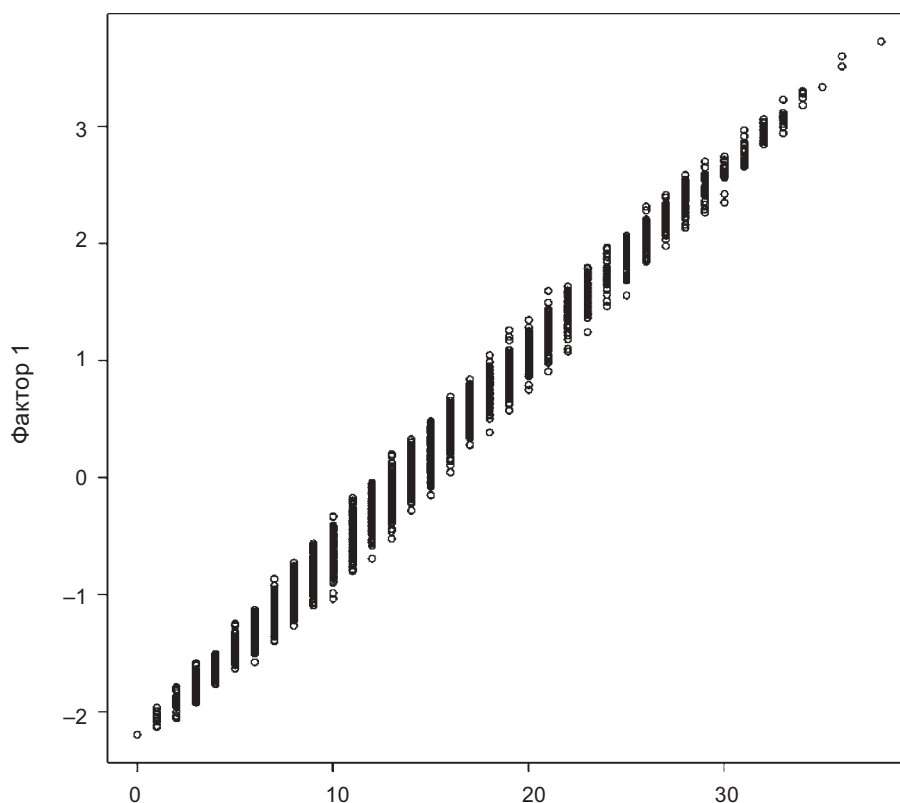


Рис. 1. Диаграмма разброса индивидуальных достижений абитуриентов при централизованном тестировании по белорусскому языку. Положение точек относительно горизонтальной оси отражает сумму правильных ответов на задания теста, по вертикальной оси — величину первого фактора

Шаг второй: проверка модели

С целью проверки полученной факторной модели на устойчивость и достоверность параметров была реализована процедура рандомизации данных. В модель, построенную по данным ЦТ по математике, были подставлены результаты тестирования гипотетических абитуриентов, которые отвечали на вопросы теста случайным образом.

Наглядно разница между реальным и рандомизированным распределениями количества правильных ответов видна, когда данные о количестве правильных ответов гипотетических абитуриентов, отвечавших случайным образом на вопросы теста (рис. 2), заменяются результатами реальных абитуриентов (рис. 3). Если в первом случае в явном виде выступает нормальное распределение, что полностью согласуется с теорией, то во втором случае распределение разворачивается и сжимается под воздействием первого фактора, который выше был интерпретирован как фактор, определяемый наличием системности знаний абитуриента. Именно этим фактором определяется разброс по вертикали результатов абитуриентов с одинаковой суммой положительных ответов на задания теста.

Шаг третий: проверка значимости полученного индикатора для принятия управленческих решений

Следующим шагом исследования стало выяснение, насколько важно для университетского менеджера знание индекса системы знаний абитуриентов, по сравнению с такими параметрами качества абитуриента, как сумма набранных баллов по тестированию и простая сумма правильных ответов на задания теста при тестировании. Для этого был осуществлен анализ связи всех вышеуказанных показателей с показателями успеваемости студентов в течение первого семестра. Выявление наличия и силы связи осуществлялось путем кластерного анализа объединения (построение дерева связей). В качестве объекта исследования были избраны студенты-первокурсники факультета экономики и управления. Для анализа были взяты данные об их промежуточной успеваемости (результаты промежуточной аттестации в семестре) и успеваемости по итогам сессии, которые после соответствующей очистки и подготовки были сопоставлены с суммой баллов по тестированию и индексом системы знаний на предмет возможной кластеризации.

По итогам кластерного анализа близкими друг другу оказались группы данных, содержащие результаты сессии (первого, второго и третьего

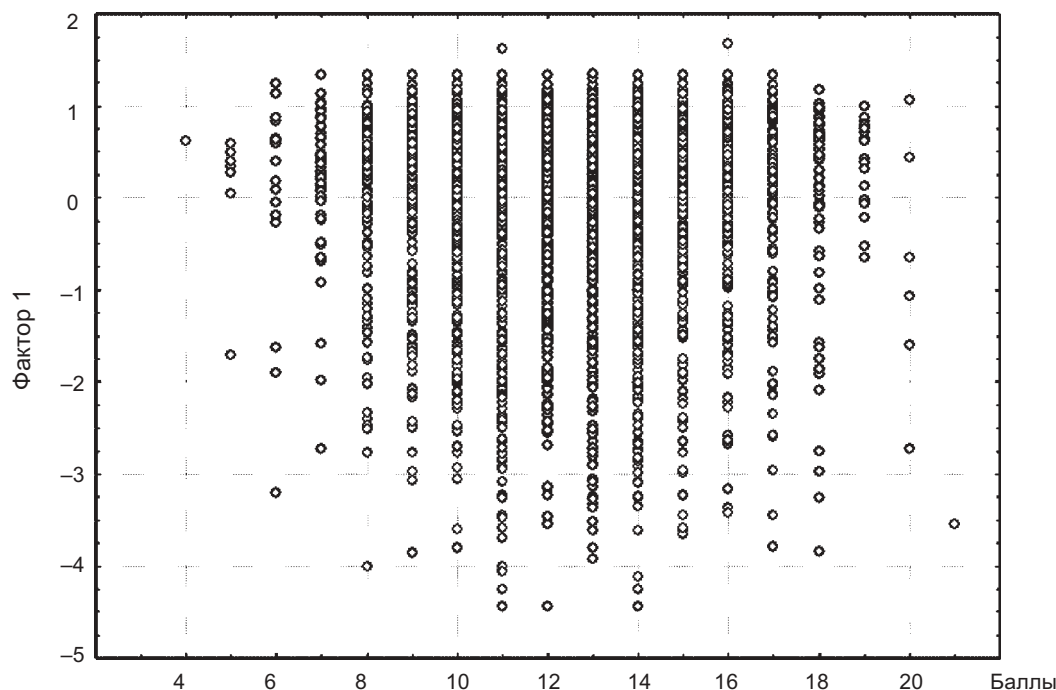


Рис. 2. Распределение количества правильных ответов на вопросы теста в зависимости от величины первого фактора, рассчитанного по рандомизированным данным

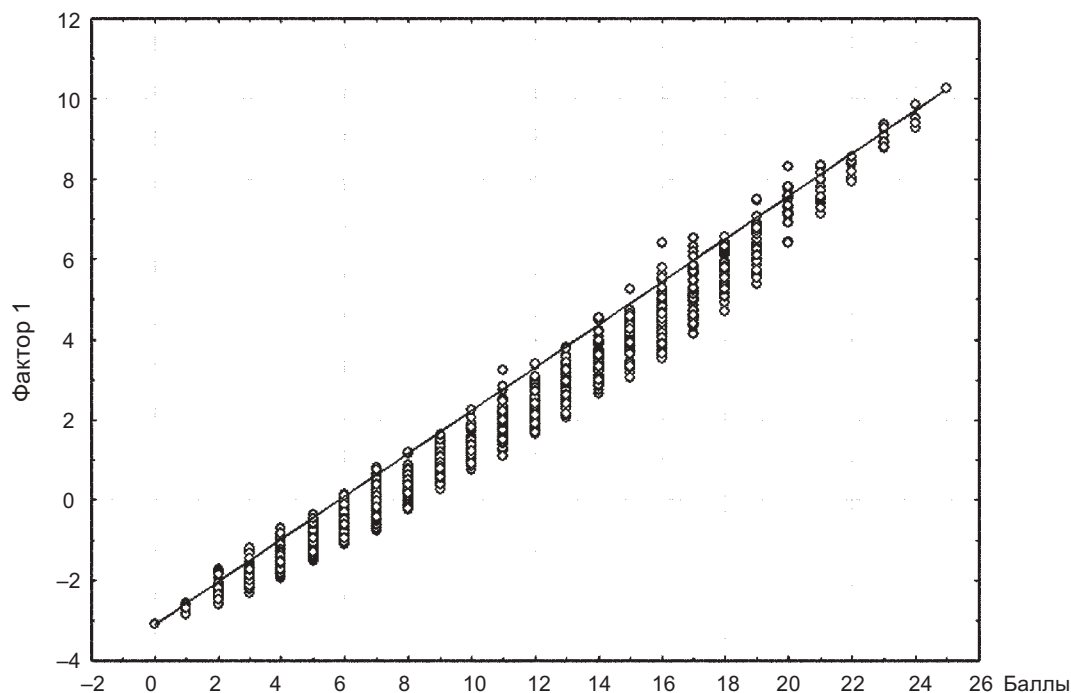


Рис. 3. Оценка централизованного тестирования по математике. Распределение количества правильных ответов на вопросы теста в зависимости от величины первого фактора

экзамена). Примерно в четыре раза дальше от этих групп расположилась группа данных, включающая индекс системы знаний, примерно в восемь раз дальше от них находятся значения простых сумм правильных ответов на вопросы теста. Самым неожиданным результатом процедуры кластеризации оказался тот факт, что группа данных, состоящая из сумм баллов абитуриентов по результатам централизованного тестирования, находится очень далеко (примерно в 362 (!) раза дальше) от данных, содержащих результаты сессии. Аналогичная кластеризация наблюдалась для средней оценки по результатам зимней сессии для студентов-экономистов и для результатов промежуточной аттестации знаний в течение семестра. Это позволило обоснованно говорить о том, что первого сентября первокурсники свою учебную жизнь начинают практически с нуля и их дальнейшие успехи определяются не столько объемом школьных знаний, сколько наличием целостной системы знаний и умением уложить в эту систему знания по новым для них дисциплинам.

Результаты кластерного анализа позволили сделать вывод о том, что «чистая» сумма баллов, полученная абитуриентом во время централизованного тестирования, содержит в себе очень мало прогностической информации о будущей успешности абитуриента во время обучения в вузе. Зато несравненно более точный прогноз о том,

будет ли способен абитуриент учиться в качестве студента, можно построить на основе индекса системы знаний, рассчитанного по результатам централизованного тестирования. Прогностическое значение индекса системы знаний увеличивается с уменьшением суммарного балла по тестированию. Поэтому вычисление индекса системы знаний абитуриентов особенно важно для специальностей с низким конкурсом и невысоким проходным баллом (рис. 4).

Пустое «окно» в диаграмме разброса (рис. 4, выделено пунктиром) демонстрирует, что практически все студенты, имевшие, будучи абитуриентами, рассчитанный по первому фактору индекс системы знаний выше, чем 0, по результатам первой сессии имеют суммарный балл от 10 и выше. Согласно принятой в Республике Беларусь 10-балльной оценке знаний, получение в сессию менее 10 баллов по трем экзаменам означает безусловную неуспеваемость и автоматически влечет отчисление. Рисунок 5 демонстрирует, что величина индекса системы знаний, рассчитанного по результатам ЦТ по математике, практически определяет максимальный средний балл студентов математического факультета во время первой сессии.

Из рис. 5 однозначно следует вывод, что для того, чтобы студент получил в первую сессию средний балл не ниже, чем 4, по десятибалльной шкале, его абитуриентский показатель рейтинга

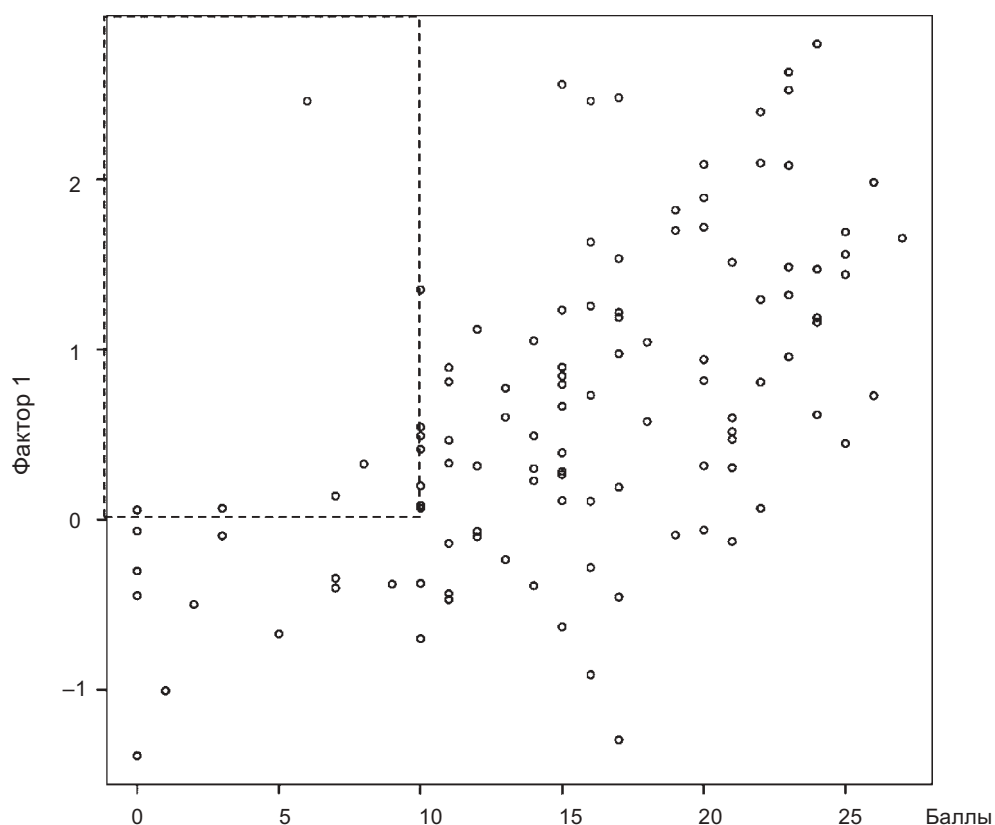


Рис. 4. Распределение результатов первой сессии студентов первого курса математического факультета в виде простой суммы баллов (горизонтальная ось) в зависимости от величины первого фактора (вертикальная ось), рассчитанного по результатам централизованного тестирования по математике

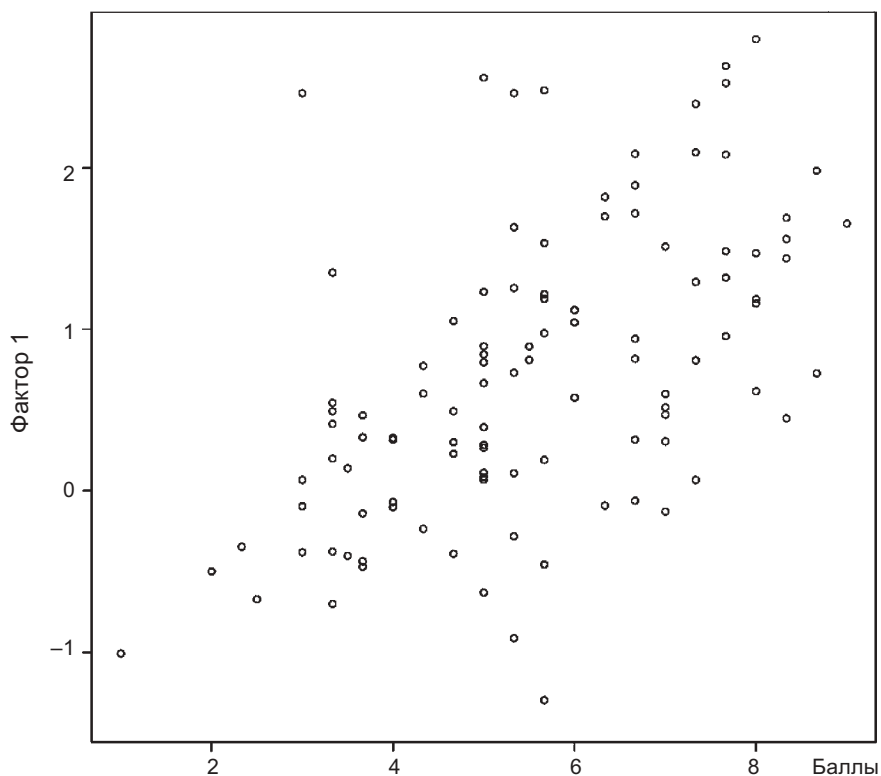


Рис. 5. Распределение средних баллов, полученных студентами математического факультета во время первой сессии (горизонтальная ось), в зависимости от величины первого фактора (вертикальная ось), рассчитанного по результатам централизованного тестирования по математике

по первому фактору должен быть не ниже 0,5. Если руководство факультета заинтересовано, чтобы первокурсники практически не имели неудовлетворительных оценок в сессию, то при зачислении необходимо брать абитуриентов с показателем рейтинга выше, чем 1. Данный вывод можно экстраполировать и на другие факультеты с учетом специфики их учебного процесса.

Следует заметить, что пять выброшенных вверх из общего распределения точек на рис. 5 — это низкие результаты конкретных студентов, перспективных с точки зрения величины их индекса системности знаний. После обсуждения результатов исследования с руководством факультета и выяснения причин уникальности этих результатов в каждом конкретном случае деканом были произведены соответствующие корректирующие действия и приняты предупреждающие меры.

По результатам данного исследования была разработана стандартная процедура определения целевой группы абитуриентов по итогам централизованного тестирования. Руководство заинтересованных факультетов получило возможность за две недели (времени от момента получения абитуриентом результата тестов до начала подачи заявлений) провести кампанию персонализированного профориентирования в этой целевой группе. Как правило, использовалась прямая почтовая рассылка, сравнительно дорогой и в то же время, при точном определении целевой аудитории, высокоэффективный способ стимулирования спроса. Крайне важно понимать, что эффект производит только конкретно адресованное письмо, которое получают далеко не все абитуриенты. Оно оказывает сильный психологический эффект на абитуриента и его родителей. Авторам известны случаи, когда письмо распространялось среди родственников и знакомых, служило предметом гордости и инициировало вторичную волну повышенного интереса к «непопулярным» в обычных условиях специальностям.

Шаг четвертый: бутстреп-процедура для доказательства эффективности целевого маркетинга

В 2009 г. в эксперимент по повышению качества абитуриентов «на входе» в образовательный процесс включился физико-технический факультет ГрГУ. Для определения целевой аудитории для рекламной рассылки при помощи методов многомерной статистики были одновременно проанализированы результаты тестов по фи-

зике и математике 4 116 абитуриентов. Из них для списка рассылки по указанной выше методике отобрано 598 потенциальных кандидатов к поступлению на факультет, которым были высланы персональные (т. е. с указанием фамилии, имени, отчества) приглашения к поступлению на физтех ГрГУ. Число 598 определялось ограничениями бюджета на рекламную кампанию.

По итогам приемной кампании 2009 г. на физико-технический факультет было принято 146 абитуриентов, из них 61 в свое время получил письмо-приглашение. Как оценить количественный эффект персонализированной рассылки?

Если известно количество откликов в результате проведенной кампании и количество контактов потенциальных студентов с определенным каналом распространения рекламы, то можно, построив статистическую модель случайного поведения абитуриентов, многократно (10 тыс. и более раз) сравнить реальную картину поведения абитуриентов с моделью. Таким образом, фактически сформулирована задача для бутстреп-анализа.

Предложенный в 1977 г. Бредли Эфроном бутстреп-метод [1] отличается от традиционных методов статистического анализа тем, что он предполагает многократную обработку различных частей одних и тех же данных, как бы поворот их «разными гранями», и сопоставление полученных таким образом результатов. Бутстреп-процедура не требует информации о виде закона распределения изучаемой случайной величины и в этом смысле может рассматриваться как метод непараметрической статистики, т. е. она работает без опоры на существенную часть априорной информации, чем, по-видимому, и обусловлен выбор термина [4].

Для обоснованности результатов бутстреп-процедуры в данном конкретном случае вначале необходимо было выяснить, насколько полученный результат приемной кампании близок к случайному. Для этого было рассчитано распределение вероятности количества зачисленных абитуриентов, получивших письмо-приглашение, в численном эксперименте по случайной рассылке 598 писем. Для этого случайным образом было выбрано 598 человек из 4 165 возможных кандидатов и проверено, сколько среди них находилось тех, кто действительно поступил на физический факультет. Данные 10 тыс. попыток позволили построить распределение числа поступивших абитуриентов, получивших письмо-приглашение случайным образом.

Из рис. 6 очевидно, что вероятность получить среди зачисленных абитуриентов более 37 человек с таким письмом-приглашением составила менее 0,0001.

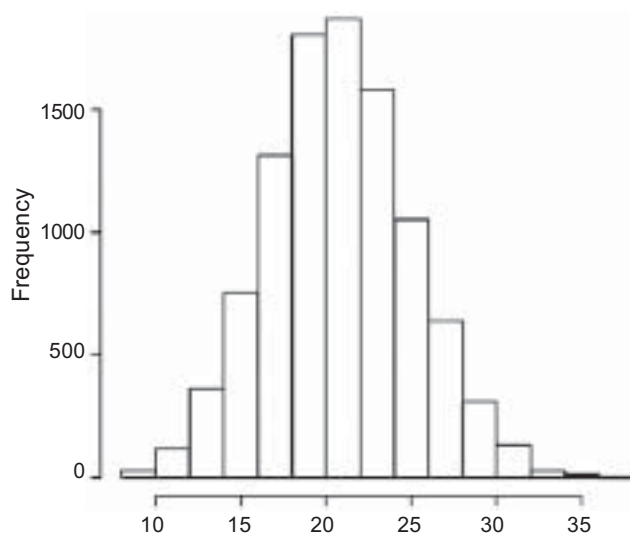


Рис. 6. Гистограмма, демонстрирующая распределение числа поступивших на факультет абитуриентов, получивших письма-приглашения, по результатам 10 тыс. попыток в ходе бутстрепа, в случае случайной рассылки 598 писем среди 4 165 человек

Однако рассылать письма-приглашения всем без исключения сдавшим тест в случае, когда таковыми считаются даже абитуриенты, получившие один балл из ста возможных, не имеет смысла. Поэтому следующим действием стало изучение распределения плотности вероятности количества поступивших абитуриентов в случае рассылки приглашений только абитуриентам с оптимальными для факультета результатами тестирования.

На рис. 7 построено 6 распределений для последовательно возрастающих (на рисунке — справа налево) уровней требований к результату тестирования исходя из предположения, что известна оптимальная нижняя граница требований. Форма графиков близка к нормальной форме распределения Гаусса. Очевидно, что в самом удачном случае вероятность получить в числе зачисленных 61 и более абитуриентов с письмом-приглашением составляет менее 0,0022 (22 случая на 10 тыс. попыток). Этот практически крайний случай задан оптимальной нижней границей требований. Все другие случаи отсеечения абитуриентов по нижней границе результатов тестирования дают еще меньшую вероятность получить на выходе рассылки 61 отклик. Но и распределение с оптимальной нижней

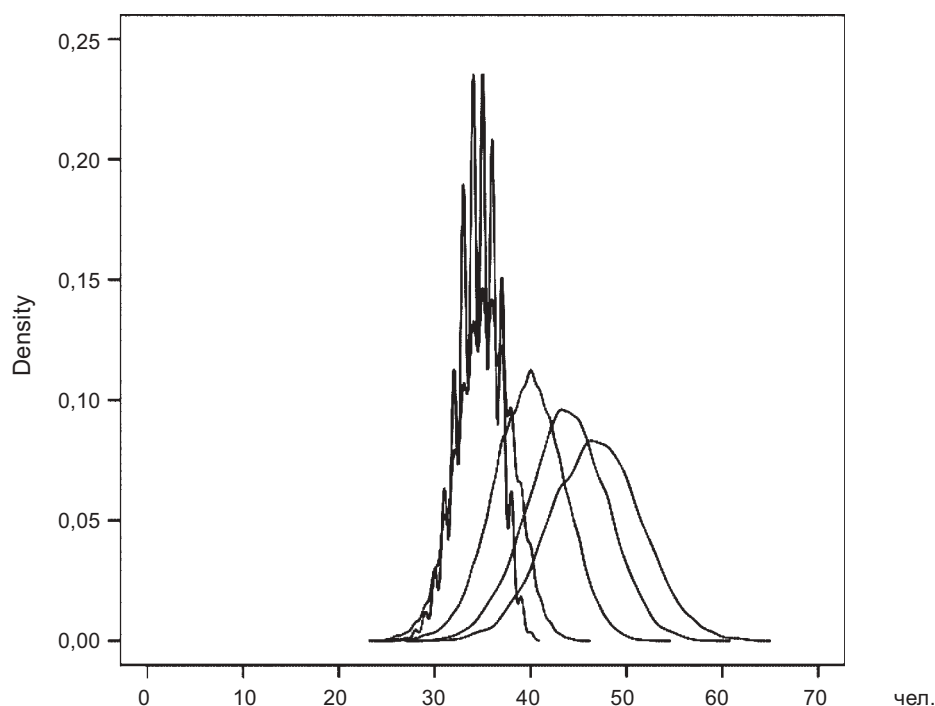


Рис. 7. Варианты распределения плотности вероятности количества поступивших на факультет абитуриентов, получивших письмо-приглашение: по горизонтали — количество поступивших, по вертикали — плотность вероятности. В численном эксперименте использовались реальные данные вступительной кампании 2009 г.

границей демонстрирует максимальную плотность вероятности в районе 46 человек, что говорит о крайне высокой достоверности наличия эффекта от рассылаемых писем-приглашений.

Для количественной оценки эффекта от целевой рассылки было построено распределение разницы между числом фактически принятых во время вступительной кампании абитуриентов, получивших приглашение, и эффектом от случайной рассылки приглашений.

На рис. 8 отражена картина, когда от реального результата приемной кампании (61 абитуриент с письмом-приглашением в числе 146 зачисленных) отняты гипотетические результаты, которые можно получить в результате случайной рассылки 598 приглашений всем, чьи результаты выше оптимальной границы требований. Несмотря на то, что на рисунке отражены только пять распределений из 10 тыс. просчитанных компьютером, видно, что он представляет собой практически зеркальное отражение рис. 7. При этом разница между реальным результатом и гипотетически наиболее вероятным равна 15. Это те пятнадцать человек, о которых можно сказать, что они пришли на факультет благодаря не случайной, а целенаправленной рассылке писем-приглашений. Для пятнадцати человек письмо-

приглашение сыграло роль последнего импульса к принятию решения. При этом практически невозможно определить этих 15 человек из 61, можно лишь утверждать, что именно 15 человек поступило на факультет благодаря использованной при рассылке методики целевой адресации писем-приглашений.

После учета всех издержек факультета на осуществление целевой прямой рассылки расходы на привлечение этих пятнадцати человек составили около 8 долл/чел. С учетом того, что выявленные на основе расчета индекса системности знаний целевые абитуриенты впоследствии способны успешно учиться пять лет, столь невысокие затраты можно смело признать высоко rentабельной инвестицией.

Выбор бутстреп-процедуры для оценки достоверности и определения величины эффекта рассылки был обусловлен невыраженностью эффекта от прямой почтовой рассылки на фоне эффектов от других факторов стимулирования. В результате оценен количественный эффект от примененной методики профориентации, составляющий не менее 30 %. Предложенная бутстреп-процедура применима для оценки эффективности и других управляющих воздействий на спрос в рамках маркетинговой деятельности.

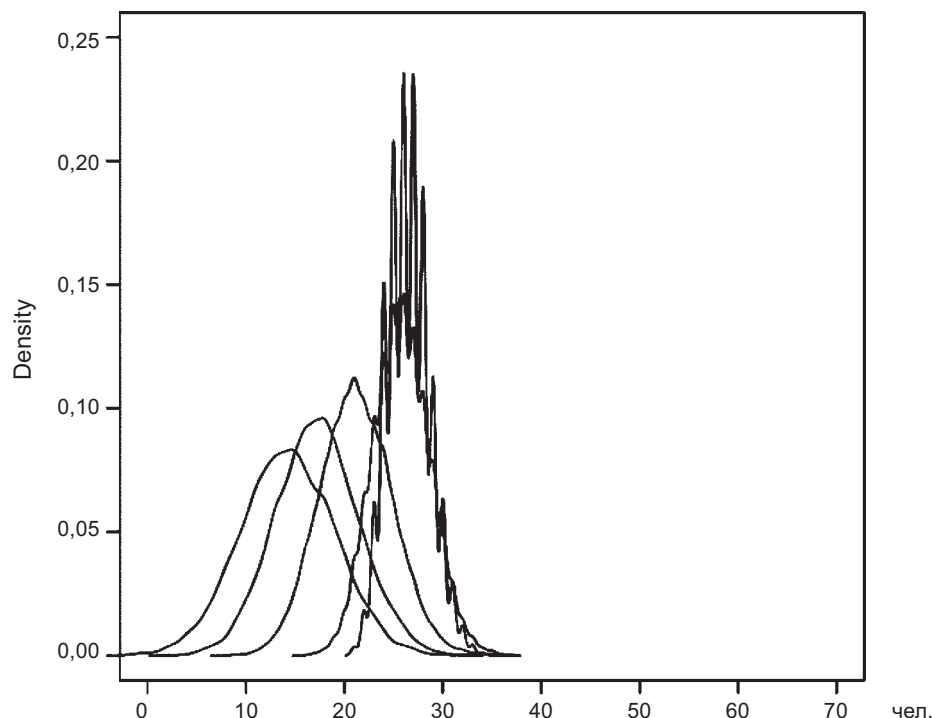


Рис. 8. Распределение плотности вероятности разницы между количеством абитуриентов, фактически поступивших на факультет с письмом-приглашением, и количеством абитуриентов, которое можно получить в результате случайной рассылки 598 писем-приглашений всем потенциальным кандидатам, чьи результаты тестирования находятся выше оптимального уровня требований

Выводы

По итогам проводимой в Гродненском государственном университете работы по повышению информативности результатов централизованного тестирования необходимо отметить, что выявленный авторами в результате факторного анализа феномен распределения абитуриентов позволяет повысить эффективность входного контроля качества абитуриентов и открывает, по меньшей мере, несколько возможностей его практического использования. Во-первых, рейтинг индивидуальных достижений абитуриентов по результатам централизованного тестирования, рассчитанный с учетом индекса системы знаний, позволяет более точно прогнозировать последующие успехи студентов и при желании установить обоснованный «отбраковочный» уровень качества абитуриента «на входе» в образовательный процесс. Во-вторых, будучи представлен в виде таблицы с персональными данными абитуриентов, их идентификационными признаками, суммой баллов по тестированию и индексом системы знаний, рейтинг позволяет дифференцировать абитуриентов с одной суммой баллов по централизованному тестированию. То есть во время процедуры зачисления при рассмотрении

спорных вопросов у деканов появляется еще один количественный критерий качества абитуриентов — индекс системы знаний. В-третьих, на основе рейтинга можно выявить школы и конкретных учителей, которые качественно готовят абитуриентов, и установить с ними долгосрочную прямую коммуникацию. Таким образом, рейтинг дает менеджменту регионального вуза возможность сконцентрировать усилия по профориентационной работе на более узких целевых сегментах, что позволяет с большей эффективностью расходовать ограниченные ресурсы.

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. М. : ЮНИТИ, 1998. С. 512–595.

2. Бойко В. К., Войтукевич Ю. А., Лявшук В. Е. и др. Повышение эффективности входного контроля качества абитуриентов при помощи методов многомерной статистики // Веснік ГрДУ імя Я. Купалы. Сер 2 : Матэматыка. 2007. № 1. С. 78–90.

3. Кириленко И., Козлович Н. Поступательное движение // Беларусь сегодня. 2009. № 130, 16 июля.

4. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. М. : Финансы и статистика, 1988. 263 с.

