Теория

АЛГОРИТМ УТОЧНЕНИЯ ОЦЕНКИ ПРИ КОМПЬЮТЕРНОМ ТЕСТИРОВАНИИ

Олег Деменчёнок
Восточно-Сибирский институт МВД России
AskSystem@yandex.ru

Средствами математической статистики проведён анализ случайной погрешности результата тестирования. Для компьютерного тестирования предложен алгоритм перевода доли правильных ответов в оценку. Этот алгоритм основан на расчёте разности значений вероятности оценок с использованием функции Лапласа.

Ключевые слова: *тест, компьютерное тестирование, доля правильных ответов, оценочная шкала, учёт случайной погрешности.*

Тестовый балл, как правило, не используется непосредственно. Часто он переводится в педагогическую оценку, для чего сравнивается с пороговыми значениями некой общепринятой шкалы оценок: зачтено — не зачтено; неудовлетворительно — удовлетворительно — хорошо — отлично и т.д. При этом некоторые считают, например, что за 75% правильных ответов может быть выставлена оценка «хорошо», а за 74,5% — «удовлетворительно». Т.е. за почти оди-

наковые результаты выполнения теста могут быть выставлены существенно различающиеся оценки, что представляется недостаточно обоснованным. Именно оценка является информаций об успехе или неуспехе, на основе оценки принимается решение о ходе процесса обучения. По оценке студенты судят об уровне своих знаний, а также об объективности педагога. Известно, что оценка приводит к благоприятному воспитательному эффекту только тогда, когда обучаемый внутренне согласен с ней. Ощущение несправедливости полученной оценки ослабляет мотивацию обучения, может привести к возникновению конфликтных ситуаций. Поэтому повышение обоснованности оценки представляется практически значимой задачей.

Важнейшей причиной неточности педагогической оценки является неоднозначность критериев оценивания. Например, оценка «неудовлетворительно» обычно рекомендуется в случае, если обучаемый не знает значительной части программного материала, допускает существенные ошибки, с большими затруднениями выполняет практические задания, задачи. Поскольку каждый из преподавателей имеет своё собственное представление о «значительной части», «существенных ошибках» и «больших затруднениях», то один и тот же ответ разными преподавателями совершенно добросовестно может быть оценен по-разному.

Результат тестового контроля определяется по заранее установлённым правилам и независим от личности преподавателя. Устраняя субъективизм процедуры оценивания, тестовый контроль знаний всё же не гарантирует точность оценки. Необходимо понимать, что тестирование позволяет достичь высокой степени объективности оценки, не гарантируя этого автоматически. Не затрагивая в данной работе вопросы соответствия теста целевому назначению (валидность) и другие характеристики теста, а также формирование шкалы оценок, сосредоточимся на анализе присущих тестированию случайных погрешностей.

Анализ случайной погрешности результатов тестирования

Погрешность — неизбежная часть любого измерения, и педагогические измерения не являются исключением. В статистике различают три основных вида ошибок: систематические, грубые и случайные.

Систематические ошибки однонаправлено либо преувеличивают, либо преуменьшают

результаты измерений. При тестировании причинами систематической погрешности могут стать ошибки в разработке и применении теста. Например, если использовать тест по высшей математике, разработанный для технической специальности, при тестировании студентов гуманитарного вуза, то получим систематическое занижение оценки. Случайное угадывание правильных ответов и недостаточный контроль за испытуемыми (соответственно, использование запрещенных справочных материалов, помощь других лиц и даже подмена тестируемого) увеличивают, по сравнению с истинным, значение тестового балла. Универсальных методов устранения систематических ошибок не существует, общая рекомендация — минимизировать влияние вызывающих систематические ошибки факторов.

Грубые ошибки возникают вследствие просчёта при вычислении тестового балла или неправильной регистрации результата (например, запись оценки в строку экзаменационной ведомости, не соответствующую фамилии тестируемого). Иногда грубые ошибки хорошо заметны и легко устранимы.

Случайными можно считать ошибки ввода данных; ошибки, вызванные неверным истолкованием условия задания и т.п. Единственно возможный способ объективного учёта случайных погрешностей состоит в определении их статистических закономерностей. Случайные ошибки происходят по различным случайным причинам, действующим при каждом из отдельных измерений непредвиденным образом, то в сторону уменьшения, то в сторону увеличения результатов.

Каковы источники случайных ошибок в случае тестового контроля? Основная причина — ограниченность числа заданий. Понятно, что чем больше заданий выполняет студент, тем полнее может быть представление о его знаниях. Проведение тестирования основано на формировании ограниченного набора тестовых заданий, что даёт возможность лучше организовать тестирование, обеспечивает быстроту проведения контроля знаний, приводит к экономии затрат труда на получение и обработку информации. Однако ограниченный набор заданий не всегда достаточен для полной проверки структуры и глубины знаний. Возникающие ошибки репрезентативности в сочетании с фрагментарностью знаний части обучаемых могут привести к зависимости тестового балла от того, какие именно задания предложены конкретному студенту («счастливый» и «несчастливый» билет).

Определённое влияние оказывает широкое распростра-

нение двоичной системы оценки правильности ответа на каждое задание (правильно или неправильно, 1 или 0). Ввиду малого объёма отдельного задания сложно различать степень правильности ответов. В результате неполные или неточные ответы квалифицируются как незнание ответа, что не всегда оправдано. Вместе с тем, правильный ответ, оцениваемый максимальным баллом, не всегда соответствует известным критериям оценки «отлично» — точное и прочное знание материала в заданном объёме; исчерпывающее и логически стройное его изложение; умение обосновывать принятые решения, обобщать материал[1] и др.

Педагогическое тестирование можно попытаться сравнить с определением площади некоторой фигуры по методу Монте-Карло[2]. Для применения этого метода фигуру вписывают в другую, известной площади (например, в квадрат), и случайным образом «бросают» точки, подсчитывая число попаданий в фигуру. При достаточно большом числе испытаний отношение числа точек, попавших внутрь фигуры, к общему числу точек стремится к отношению их площадей. Тогда квадрат — это область, в которой проверяются знания; фигура неизвестной формы и площади — структура и глубина знаний тестируемого, а точки — тестовые задания. При достаточно большом числе заданий доля правильных ответов p приближается к истинной величине относительного объёма знаний тестируемого.

В таком случае нужно рассматривать доверительный интервал доли правильных ответов — интервал, который с заданной вероятностью a накроет неизвестное значение. Например, доверительный интервал доли правильных ответов $p = 0{,}75 \pm 0{,}05$ при вероятности 0,9 означает, что с вероятностью 90% истинное значение p находится в интервале 0,7...0,8.

При обработке данных будем исходить из того, что погрешности имеют нормальное распределение. Если считать, что погрешность измерения определяется в результате совокупного действия многих малых факторов, действующих аддитивно и независимо друг от друга, то в силу Центральной Предельной Теоремы теории вероятностей погрешность измерения хорошо приближается (по распределению) нормальной случайной величиной[3].

Аналитически доверительный интервал доли правильных ответов записывается в виде $p \pm \Delta p$,

$$\Delta p = \varepsilon \cdot \sigma_{\bar{p}} = \varepsilon \frac{\sigma}{\sqrt{m}} = \varepsilon \frac{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2}}{\sqrt{m(m-1)}}, \quad (1)$$

1 — *Буланова-Топоркова М.В. и др.* Педагогика и психология высшей школы. Ростов-на-Дону: Феникс, 2002. 544 с.

2 — *Секей Г.* Парадоксы в теории вероятностей и математической статистике: Пер. с англ. М.: Мир, 1990. 240 с.

3 — *Орлов А.И.* Прикладная статистика. М.: Экзамен, 2006. 669 с.

где Δp — погрешность определения доли правильных ответов, вызванная действием случайных факторов; σ — среднее квадратичное отклонение результатов выполнения i-го задания x_i от среднего значения ; m — число заданий; $\sigma_{\bar{p}}$ — среднее квадратичное отклонение доли правильных ответов от истинного значения; ε — табличный коэффициент для заданного значения вероятности α ($\alpha = 0{,}68$ соответствует $\varepsilon = 1{,}0$; $\alpha = 0{,}90$ соответствует $\varepsilon = 1{,}65$; $\alpha = 0{,}997$ соответствует $\varepsilon = 3{,}0$ и т.д.).

Простой анализ выражения (1) показывает, что случайная погрешность зависит от однородности результатов выполнения отдельных заданий и количества заданий m. Нетрудно заметить, что если все ответы правильны ($x_i = \bar{x}$, $\sigma = 0$), то случайная погрешность равна нулю. Аналогично, $\Delta p = 0$ в случае, когда ответы полностью неверны ($x = \bar{x} = \sigma = 0$). Это означает, что случайная погрешность отсутствует только в этих двух крайних случаях.

Очевидно, что максимальное значение Δp принимает в случае использования двоичной системы оценивания правильности ответа (правильно или неправильно) при условии равенства количества правильных и неправильных ответов ($x_i = 0; 1; 0; 1; 0; 1...$):

$$\Delta p_{\max} = \varepsilon \frac{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}}{\sqrt{m(m-1)}} = \varepsilon \frac{\sqrt{0{,}5^2 m}}{\sqrt{m(m-1)}} = \frac{\varepsilon}{2\sqrt{m-1}}. \quad (2)$$

Например, случайная погрешность доли правильных ответов для теста с m=20 заданиями при доверительной вероятности 0,68 не превышает $\Delta p_{\max} = 0{,}11$ (или 11%), при $m = 50$ $\Delta p_{\max} = 0{,}07$, при $m = 200$ – 0,035. Из этого уравнения легко получить зависимость для расчёта количества заданий, гарантирующего, что случайная погрешность не превысит заданного значения:

$$m = \frac{\varepsilon^2}{4\Delta p_{\max}^2} + 1. \quad (3)$$

Так, для обеспечения случайной погрешности не более 0,05 (5%) при указанной доверительной вероятности требуется 101 задание. Графически зависимость $m = f(\Delta p, e)$ представлена на рис. 1.

Алгоритм уточнения оценки при компьютерном тестировании

Вернёмся к задаче перевода результата тестирования в качественные показатели типа «хорошо», «удовлетворительно» и т.п. При таком переводе статистически неразличимые результаты могут привести к разным оценкам. Так, например, доли правильных ответов p = 0,59 и p = 0,61 при ошибке Δp = 0,05

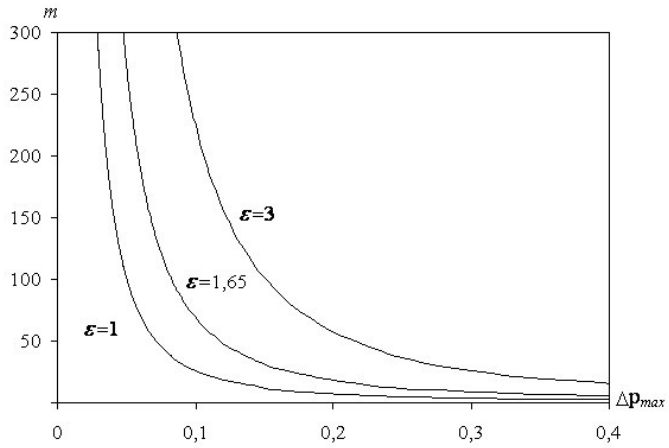

Рис. 1. Зависимость количества заданий теста от максимально допустимой величины случайной погрешности

соответствуют практически одинаковым интервалам 0,54...0,64 и 0,56...0,66. Однако при пороговом значении для удовлетворительной оценки $R_3 = 0{,}6$ оценки будут кардинально отличаться: первый обучаемый получит «неудовлетворительно», а второй — «удовлетворительно». Реально же данная ситуация означает, что оценка лежит в пределах от «неудовлетворительно» до «удовлетворительно». Что делает в таких случаях опытный преподаватель? Для уточнения оценки задаёт дополнительные задания. Если при бланковом тестировании организовать подобное сложно, то при компьютерном тестировании можно смоделировать подобный алгоритм работы преподавателя и в спорных случаях автоматически выдавать дополнительные задания.

На рис. 2 схематично показано сопоставление результата тестирования в виде доверительного интервала p со шкалой оценивания (0...R_3 — «неудовлетворительно», R_3...R_4 — «удовлетворительно», R_4...R_5 — «хорошо», свыше R_5 — «отлично»).

В случае, когда доверительный интервал полностью помещается между двумя соседними значениями шкалы оценивания, можно утверждать, что с вероятностью не меньшей α результат соответствует оценке R_i. Так, на рис. 2а доверительный интервал доли правильных ответов располагается между значениями R_3 и R_4. Следовательно, результат выполнения теста оценивается на «удовлетворительно».

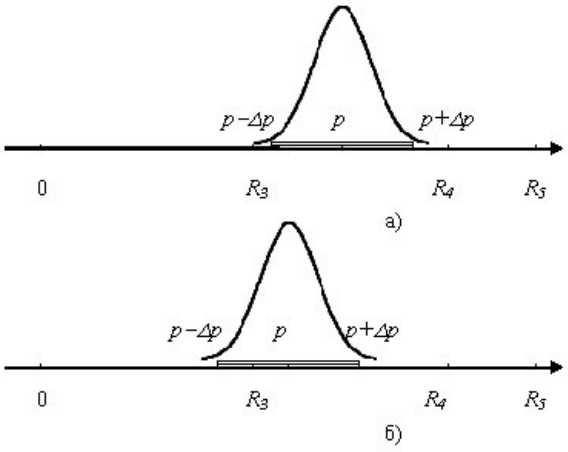

Рис. 2. Сравнение доверительного интервала доли правильных ответов с пороговыми значениями шкалы оценивания

Возможен также вариант, когда значение шкалы оценивания окажется внутри доверительного интервала (рис. 2б). Возникает неоднозначность: с вероятностью $P_1 = P(p - \Delta p < p < R_3)$ результат соответствует оценке «удовлетворительно», а с вероятностью $P_2 = P(R_3 \le p < p + \Delta p)$ результат соответствует оценке «хорошо».

Очевидно, что при близких значениях p и R_i вероятности примерно равны $P_1 \approx P_2$. Следовательно, в таком случае равновероятны две разные оценки, что существенно затрудняет оценивание ответа.

Вероятность попадания результата тестирования p в промежуток $[p_1, p_2]$ в предположении нормального распределения величины p[4]:

$$P = \int_{\rho_1}^{\rho_2} \frac{1}{\sqrt{2\pi}\sigma} e^{-(p-\bar{p})^2/(2\sigma^2)} dp = \\ = F\left(\frac{p_2 - p}{\sigma}\right) - F\left(\frac{p_1 - p}{\sigma}\right), \quad (4)$$

где F — функция нормированного и централизованного нормального распределения (функция Лапласа).

Функция Лапласа обычно задаётся в табличном виде, неудобном для автоматизированного вычисления. Проблема устраняется путем разложения функции в ряд:

$$F(t) = \int_0^t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \approx \\ \approx \frac{1}{\sqrt{2\pi}} \sum_{s=1}^{z} \frac{(-1)^s t^{2s+1}}{(2s+1) \cdot 2^s \cdot s!}, \quad (5)$$

где $t = \frac{p_i - p}{\sigma}$, s — число используемых в расчёте членов ряда

4

Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся втузов. М.: Наука, 1981. 720 с.

(исходя из точности расчёта $\delta \le 0{,}0001$ принято $s = 36$).

Принимая доверительную вероятность равной 0,68, получим:

$$\Delta p = \varepsilon \cdot \sigma_{\bar{p}} = \sigma_{\bar{p}}, \qquad (6)$$

$$P_1 = P(p - \Delta p < p < R_i) =$$
$$= F\left(\frac{R_i - p}{\sigma_{\bar{p}}}\right) - F\left(\frac{(p - \Delta p) - p}{\sigma_{\bar{p}}}\right) =$$
$$= F\left(\frac{R_i - p}{\sigma_{\bar{p}}}\right) + F(1), \qquad (7)$$

$$P_2 = P(R_i < p < p + \Delta p) =$$
$$= F\left(\frac{p - (p + \Delta p)}{\sigma_{\bar{p}}}\right) - F\left(\frac{R_i - p}{\sigma_{\bar{p}}}\right) =$$
$$= F(1) - F\left(\frac{R_i - p}{\sigma_{\bar{p}}}\right) \qquad (8)$$

Рассмотрим пример. В таблице представлены результаты тестирования трёх студентов (с расчёты проведены для $\alpha = 0{,}68$; результаты выполнения заданий x_i для простоты представлены целыми числами).

Предположим, заданы пороговые значения для четырёхбалльной шкалы $R_3 = 0{,}6$; $R_4 = 0{,}75$ и $R_5 = 0{,}9$ (иными словами до 60% правильных ответов — «неуд.», 60–75% — «удовлетворительно», 75–90% — «хорошо» и свыше 90% — «отлично»). Тогда доверительный интервал результата тестирования испытуемого Иванова (см. табл. 1) $p \pm \Delta p = 0{,}9...\ 1{,}0$ с вероятностью 0,68 превышает $R_5 = 0{,}9$ и, следовательно, соответствует оценке «отлично».

Результаты тестирования

В таблице приводится пример оценок, полученных тремя студентами.

В доверительный интервал результата студента Петрова 0,708...0,892 попадает пороговое значение $R_4 = 0{,}75$. Вероятности P_1 и P_2 в этом случае:

$$P_1 = P(0{,}708 < p < 0{,}75) =$$
$$= F\left(\frac{0{,}75 - 0{,}8}{0{,}092}\right) + F(1) = 0{,}13,$$

$$P_2 = P(0{,}75 < p < 0{,}892) =$$

Фамилия	Результаты выполнения заданий x_i	p			
Иванов	1;1;1;1;1;1;1;1;1;0;1;1;1;1; 1;1; 1;1;1;1	0,95	0,224	0,050	0,9...1,0
Петров	1;1;0;0;1;1;1;1;1;0;1;1;1;1; 1;1; 1;1;1;0	0,80	0,410	0,092	0,708...0,892
Сидоров	1;1;1;0;0;1;0;0;0;0;1;1;1;1;1; 0; 1;1;0;1	0,60	0,503	0,112	0,488...0,712

$$= F(1) - F\left(\frac{0{,}75 - 0{,}8}{0{,}092}\right) = 0{,}55.$$

Это означает, что с вероятностью 0,55 результат тестирования соответствует оценке «хорошо», а с вероятностью 0,13 — «удовлетворительно».

Ещё менее точно определена оценка Сидорова. На середину доверительного интервала его результата приходится пороговое значение R_3 = 0,6. Поэтому практически с равной вероятностью результат тестирования может быть интерпретирован как «удовлетворительно», так и «неудовлетворительно», что не позволяет определить оценку с приемлемой точностью.

В этих условиях разумным представляется в качестве критерия обоснованности оценки принять вероятность её соответствия данному результату выполнения теста. Если значение шкалы оценивания окажется внутри доверительного интервала, рекомендуется найти разность вероятностей двух оценок. При значении разности меньшей заданной следует выдавать дополнительные задания до тех пор, пока вероятность одной из оценок не окажется существенно большей.

Тогда в основу процедуры оценивания при компьютерном тестировании может быть положен предлагаемый ниже алгоритм.

1. Задаём количество заданий m, максимальное количество заданий с учётом дополнительных $m_{\max}$ и минимальную величину разности вероятностей оценок, при которой оценка считается найденной $\Delta P_{\min} = |P_2 - P_1| \approx 0{,}3...0{,}4$. Значение доверительной вероятности можно принять постоянным и в настройках теста не задавать.

2. Предлагаем тест из m заданий.

3. После получения ответов на задания теста определяем доверительный интервал доли правильных ответов, который сравниваем с пороговыми значениями R_i шкалы оценок:

а) если доверительный интервал полностью помещается между двумя соседними значениями шкалы оценивания, то оценку можно считать найденной, тестирование окончено;

б) если значение шкалы оценивания окажется внутри доверительного интервала и $|P_2 - P_1| \geq \Delta P_{\min}$, то в качестве итоговой принимаем оценку, вероятность которой больше; тестирование окончено;

в) если значение шкалы оценивания окажется внутри доверительного интервала и $|P_2 - P_1| < \Delta P_{\min}$, то оценка определена с недостаточной точностью, переходим к пункту 4.

4. Если общее количество выполненных заданий меньше $m_{\max}$, то предлагаем дополнительное задание и переходим

к пункту 3. В противном случае в качестве итоговой принимаем оценку, вероятность которой больше; тестирование окончено.

Если тестовые задания оцениваются разным количеством баллов, то долю правильных ответов можно заменить отношением индивидуального тестового балла к сумме баллов за все задания. Особенностью изложенного алгоритма является учёт не столько погрешности результата тестирования, сколько вероятности его соответствия той или иной оценке. Кроме того, методы математической статистики применены для обработки результатов ответа одного обучаемого, а не группы. Такой подход даёт возможность повысить обоснованность каждой конкретной оценки, что соответствует целям учебного процесса.

Алгоритм реализован в рамках авторской системы автоматизированного обучения и контроля знаний «Assistent» и используется в учебном процессе ряда вузов. Ознакомиться с программой и получить бесплатно дистрибутив можно на сайте www.asksystem.narod.ru.