Экспериментальная апробация метода интервальной оценки результатов выполнения системы тестовых заданий с единственным верным ответом

Палкин Константин Сергеевич
адъюнкт кафедры кораблевождения,
Военный институт (военно-морской) ВУНЦ ВМФ “Военно-морская академия”,
Ушаковская наб., д. 17/1, г. Санкт-Петербург, Россия, 197045; тел. +79312027052;
palkinks@mail.ru

Печников Денис Андреевич
кандидат технических наук, доцент, доцент кафедры кораблевождения,
Военный институт (военно-морской) ВУНЦ ВМФ “Военно-морская академия”,
Ушаковская наб., д. 17/1, г. Санкт-Петербург, Россия, 197045;тел. +7921780580724;
19pda72@bk.ru

Аннотация

В статье представлены результаты практического применения метода интервальной оценки результатов выполнения системы одиночных критериально-ориентированных тестовых заданий с единственным верным ответом

The article presents the results of practical application of the method of interval estimation results of a system of single criterion-oriented test tasks with a single correct answer

Ключевые слова

критериально-ориентированный тест; тестовое задание; результаты тестирования; точечная оценка; интервальная оценка; доверительный интервал; заданная погрешность

criterion-oriented test; test task; results of testing; the point estimate; interval estimate; confidence interval; given error

Введение

Для оценки результатов выполнения систем тестовых заданий с единственным верным ответом обычно используется показатель частости успеха вида

$$B = \frac{\sum_{i=1}^{n} j_i}{n}, \qquad (1)$$

где: i ($i = \overline{1,n}$) – номер тестового задания, n - общее число тестовых заданий, j_i ($j_i = 0,1$) - результат выполнения отдельной попытки при наличии соответствия $j_i = 1$ – ошибок нет, $j_i = 0$ – ошибки есть.

Частость, определяемая по формуле (1), традиционно рассматривается как точечная эмпирическая оценка, достоверность которой может быть оценена только на основании закона больших чисел. В [1] было показано, что результат выполнения любого задания с единственным верным ответом всегда представляется дихотомической переменной вида “да – нет (правильно – неправильно, верно - неверно и т.п.)”, а процедура его определения в процессе решения выборки из n тестовых заданий соответствует схеме Бернулли. На этой основе был предложен

метод, который позволяет оценивать результаты тестирования не как частость, а как вероятность успеха, которая имеет вполне определенный доверительный интервал.

Реализация метода предполагает [1]:

1. Определение значения вероятности p путем решения задачи нелинейного программирования вида

$$\left.\begin{array}{c} C_n^k p^{n-k}(1-p)^{k+1}\sum\limits_{s=1}^{s=m} p^{h_s} \to \max\limits_{p}, \\ p \in (0,1),\ k=\overline{1,n}, \\ h=\overline{1,(n-k)},\ \sum\limits_{s=1}^{s=m} h_s = n-k \end{array}\right\}, \qquad (2)$$

где: p – вероятность безошибочного выполнения задания; k – число заданий, выполненных с ошибкой; n - общее число заданий; $n-k$ – число заданий, выполненных без ошибок; h ($h=\overline{1,(n-k)}$) – номер i последнего задания, выполненного без ошибки, определяющий величину соответствующего вектора $\vec{h}_s$ непрерывного успеха; m – число векторов $\vec{h}_s$ непрерывного успеха.

В (2) величина каждого из векторов $\vec{h}_s$ непрерывного успеха оценивается как

$$h_s = \sum_{\substack{j_i=1 \\ j_i \in \vec{h}_s}} j_i \ \ (s=\overline{1,m}), \qquad (3)$$

а корректность расчетов $\sum\limits_{s=1}^{s=m} h_s$ определяется выполнением условий

$$\sum_{s=1}^{s=m} h_s = n-k\,; \qquad (4)$$

$$m = \begin{cases} k & \text{при} \quad j_n = 0 \\ k+1 & \text{при} \quad j_n = 1 \end{cases}, \qquad (5)$$

где j_n - результат выполнения последнего тестового задания ($j_n = 0$ задание выполнено с ошибкой, $j_n = 1$ - задание выполнено без ошибок).

Для решения (2) целесообразно использовать стандартную функцию “*Solve* (Поиск решения)” *Microsoft Excel*.

2. Расчет доверительного интервала полученной оценки (2) вероятности p успеха по формулам оценки доверительных интервалов биномиального распределения:

1) точная оценка доверительного интервала:

$$\left(\frac{(n-k)}{(n-k)+(k+1)F_{2(k+1),2(n-k),1-\varepsilon/2}},\ \frac{(n-k+1)F_{2(n-k+1),2k,1-\varepsilon/2}}{k+(n-k+1)F_{2(n-k+1),2k,1-\varepsilon/2}} \right), \qquad (6)$$

где n - число испытаний, k - число ошибок, а $F_{f,g,\alpha}$ - квантиль порядка α распределения F с f,g степенями свободы;

2) приближенная оценка доверительного интервала:

$$\left(p - u_{1-\varepsilon/2}\sqrt{\frac{p(1-p)}{n}},\ p + u_{1-\varepsilon/2}\sqrt{\frac{p(1-p)}{n}} \right), \qquad (7)$$

где: $u_{1-\varepsilon/2}$ - квантили стандартного нормального распределения порядка $(1-\varepsilon/2)$.

Для оценки работоспособности метода и выработки рекомендаций по его применению метод следовало экспериментально апробировать.

Результаты апробации метода интервальной оценки результатов критериально-ориентированного тестирования

Проверка работоспособности и оценка целесообразности применения рассматриваемого метода была проведёна в процессе педагогического эксперимента.

Группе из 134 курсантов был предложен тест по знанию основных понятий навигации. Тест включал 50 закрытых заданий с выбором из предложенных 5 альтернатив единственного верного ответа.

Тестирование проводилось с применением компьютерной системы тестирования (КСТ) “Система автоматизированного контроля (САК)”, которая входит в состав автоматизированной системы обучения “Медиатор” [2], разработанной НИИ “Центрпрограммсистем”. После проведения тестирования результаты обрабатывались с помощью пакетов Microsoft Excel и SPSS 11.5.

Динамика изменения оценок успешности, определенных точечным методом по формуле (1) и интервальным методом по формулам (2-6), каждого испытуемого была представлена в виде 4 графиков: 1) частости успеха, рассчитанной по формуле (1), 2) вероятности успеха, определенной в результате решения задачи (2); 3) границ доверительного интервала, рассчитанных по формуле (6).

В качестве примеров типичного вида графиков ниже на рис. 1-3 представлены результаты тестирования “высоко”, “средне” и “низко” успешных испытуемых, которые были фактически получены в процессе эксперимента.

Представляется целесообразным отметить, что в приведенной ранжировке уровней успешности понятия “высоко”, “средне” и “низко” используются условно. Они соответствуют своему общепринятому значению только для случая, когда преподаватель как лицо, принимающее решение (ЛПР), рассматривает в качестве возможных оценок успешности весь интервал $1 > p > 0$ значений вероятности успеха. Обычно это не соответствует действительности. Так приведенные в [3,4] представления системы предпочтений преподавателей в традиционной 4-балльной шкалы свидетельствуют, что низший балл оценки “2 − неудовлетворительно” соответствует не $p = 0$, а $p = 0{,}5$.

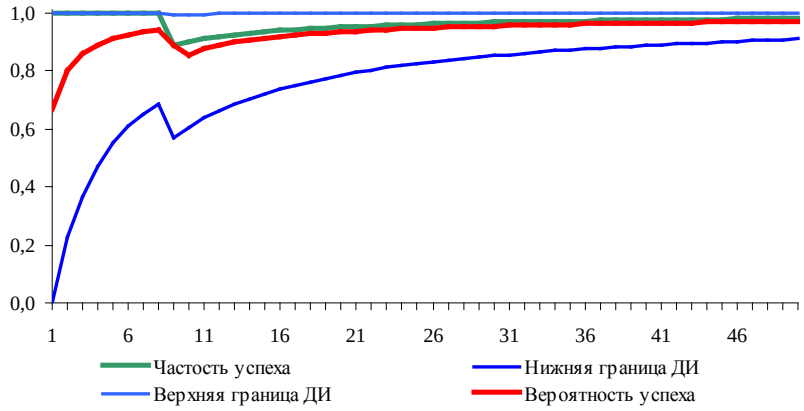

Рис. 1. Динамика изменения показателей “высоко” успешного обучаемого, закончившего тестирование с результатом $0{,}909 < p = 0{,}981 < 0{,}999$ при $\alpha = 0{,}1$

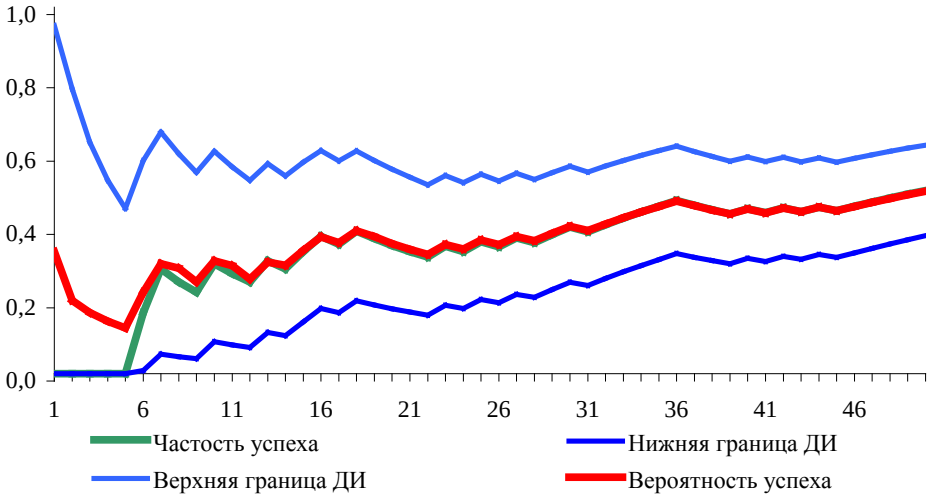

Рис. 2. Динамика изменения результатов “средне” успешного обучаемого, закончившего тестирование с результатом $0,376 < p = 0,500 < 0,624$ **при** $\alpha = 0,1$

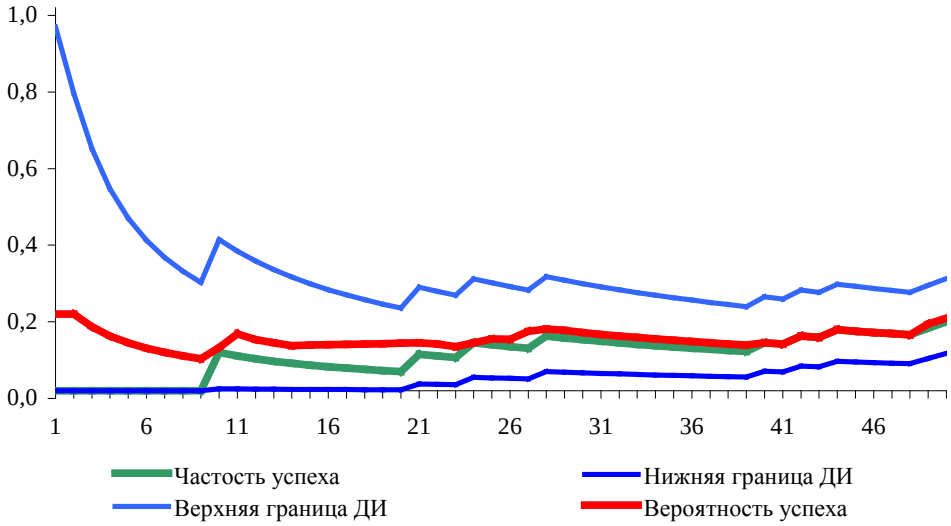

Рис. 3. Динамика изменения результатов “низко” успешного обучаемого, закончившего тестирование с результатом $0,097 < p = 0,179 < 0,293$ **при** $\alpha = 0,1$

Визуальный анализ всей совокупности 134 индивидуальных графиков позволил сделать следующие выводы:

1) значения интервальной и точечной оценок с ростом числа выполненных тестовых заданий имеют тенденцию к сходимости;

2) различия между интервальной и точечной оценками могут достигать существенных значений только для первых 10-15 тестовых заданий;

3) одним из отличий интервальной оценки от точечной является ее осторожность: в диапазоне оценок $p = 0,6 - 1,0$ интервальная оценка (2) всегда ниже точечной оценки (1), и, наоборот, в диапазоне $p = 0,0 - 0,4$, интервальная оценка (2) всегда выше оценки (1).

Количественный анализ экспериментальных данных производился с привлечением понятия и оценок p^* истинного балла.

Истинный балл (*true score*) обычно трактуют как “предел среднего значения наблюдаемых баллов, достигаемый при бесконечном увеличении числа выполнения учеником одного и того же теста. В целом можно считать, что истинный балл − это показатель испытуемого в гипотетической генеральной совокупности заданий бесконечного теста” [5]. При анализе данных эксперимента было принято допущение о корректности использования в качестве истинного балла p^* оценок (2) результатов выполнения всех 50 тестовых заданий.

Для каждой из попыток выполнения тестовых заданий всех испытуемых была произведена оценка модулей отклонения текущих оценок частости b_i $(i=\overline{1,50})$ и вероятности p_i $(i=\overline{1,50})$ успешного выполнения задания от соответствующих оценок p^* истинного балла

$$\left.\begin{aligned}\Delta b_i &= \left|b_i - p^*\right|;\\ \Delta p_i &= \left|p_i - p^*\right|\end{aligned}\right\}. \qquad (8)$$

Результаты статистической обработки оценок вида (8) в целом подтвердили итоги визуального анализа индивидуальных графиков:

1) погрешности Δp_i интервальных оценок (2) значимо меньше погрешностей Δb_i точечных оценок (1) только для первых 10-13 тестовых заданий;

2) с ростом числа выполненных заданий ряды значений погрешностей интервальной и точечной оценок имеют тенденцию к сходимости и оценка различий между ними теряет смысл.

На рис. 4 в обобщенном виде представлена динамика изменений отклонений вида (8) с ростом числа выполненных заданий. Эта динамика в явном виде демонстрирует справедливость сделанных выше выводов.

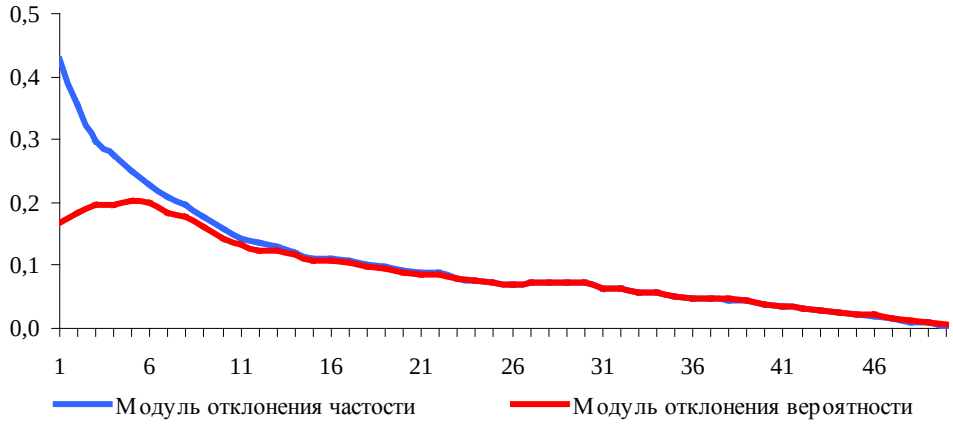

Рис. 4. Обобщенная динамика изменения отклонений исследуемых оценок от истинного балла испытуемого

Как отмечалось выше, принципиальное отличие между точечной и интервальной оценками состоит в том, что интервальная оценка характеризуется вполне определенной надежностью.

Параметром, непосредственно определяющим надежность (достоверность) оценок результатов тестирования, является ширина доверительного интервала (ДИ)

этих оценок. Графики изменения обобщенного значения этого параметра для уровней значимости $\alpha = 0,05$ и $\alpha = 0,1$ представлены на рис. 5.

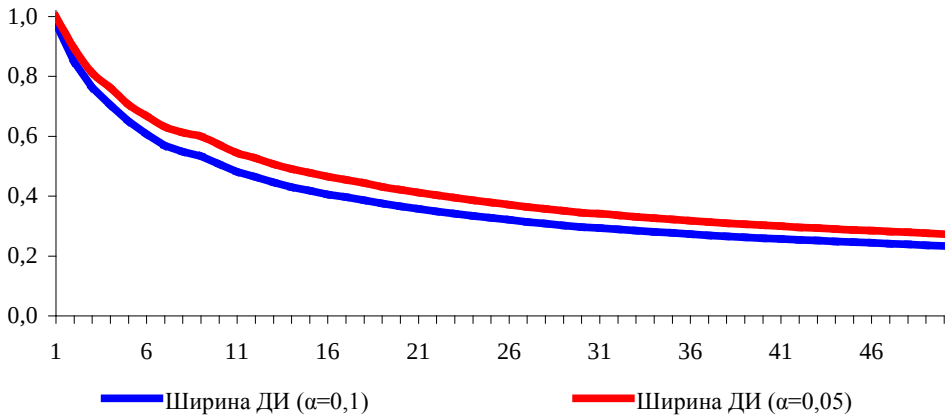

Рис. 5. Динамика изменения ширины доверительного интервала интервальных оценок результатов тестирования

Приведенные графики наглядно демонстрируют, что ширина ДИ L_i с ростом числа i $(i = \overline{1,50})$ выполненных заданий асимптотически снижается до некоторой величины. Средние значения минимальной ширины ДИ, достигнутой в имеемой группой испытуемых после выполнения всего теста, составляют $L_{\min}^{0,05} = 0,280$ для $\alpha = 0,05$ и $L_{\min}^{0,1} = 0,239$ для $\alpha = 0,1$. Величина этой минимальной ширины $L_{\min}$ ДИ для каждого испытуемого индивидуальна и закономерно обусловливается достигнутым значением $p*$ истинного балла оценок, определяемого по (2) после выполнения всех 50 тестовых заданий. Реализация зависимости $L_{\min}$ от вероятности $p*$ в рассматриваемой выборке представлена на рис. 6.

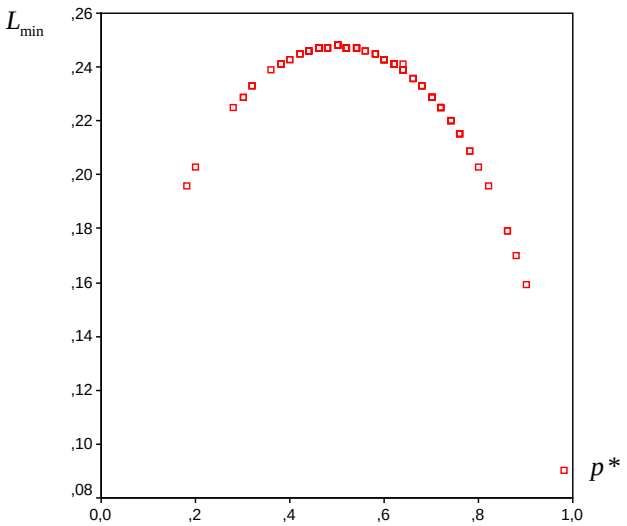

Рис. 6. Зависимость ширины ДИ от величины истинного балла результатов тестирования (экспериментальная оценка)

Графики на рис. 5 и рис. 6. в явном виде свидетельствуют, что возможность снизить погрешность оценки результатов тестирования за счет роста числа тестовых заданий имеет вполне определенный предел.

При этом влияние каждого из 50 заданий теста на изменение ширины ДИ определяется номером этого задания в той случайной последовательности заданий, которая предъявлялась испытуемому. Влияние выполненного числа тестовых заданий на ширину ДИ оценивалось следующими показателями:

1) показателем $l(i)$ относительной ресурсной результативности теста

$$l(i) = \frac{L_i - L_n}{L_1 - L_n}, \tag{9}$$

где: $(L_i - L_n)$ – оценка возможного уменьшения ширины ДИ после выполнения i-ого $(i = \overline{1,n})$ тестового задания; $(L_1 - L_n)$ − уменьшение ширины ДИ, достигаемое за счет выполнения всего теста (числа n тестовых заданий).

2) показателем r_i эффективности i-ого $(i = \overline{1,n})$ тестового задания

$$r_i = \frac{L_i - L_{i+1}}{L_1 - L_n}. \tag{10}$$

Графики показателя (9) относительной ресурсной результативности теста, которые представлены на рис. 6, позволяют оценить на сколько использована возможность уменьшения ширины ДИ за счет роста числа предъявленных тестовых заданий при рассматриваемом числе i выполненных тестовых заданий.

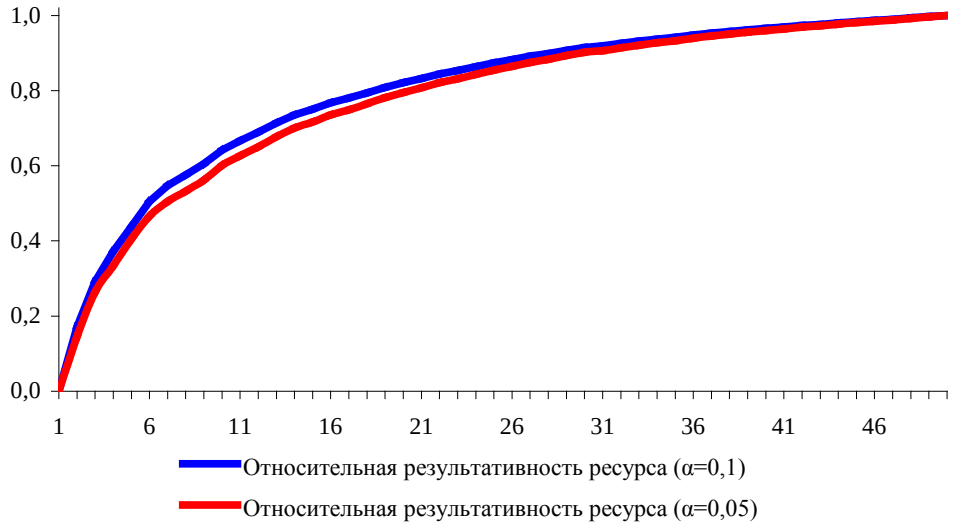

Рис. 6. Вид кривых относительной ресурсной результативности теста

Полученные в процессе эксперимента данные свидетельствуют, что увеличение числа тестовых заданий как способ повышения достоверности оценок результатов тестирования эффективно только для $i = \overline{1,20}$. Так для $\alpha = 0,1$ (см. рис. 6) $l(20) = 0,821$. Это означает, что после предъявления первых 20 тестовых заданий возможность сокращения ширины ДИ оказывается реализованной уже на $82,1\%$, а предъявление оставшихся 30 заданий сократит ширину ДИ только на оставшиеся $17,9\%$.

Еще нагляднее нецелесообразность разработки и предъявления тестов большой длины с позиций обеспечения достоверности оценок их выполнения демонстрирует показатель (10), кривые изменения которого представлены на рис. 7.

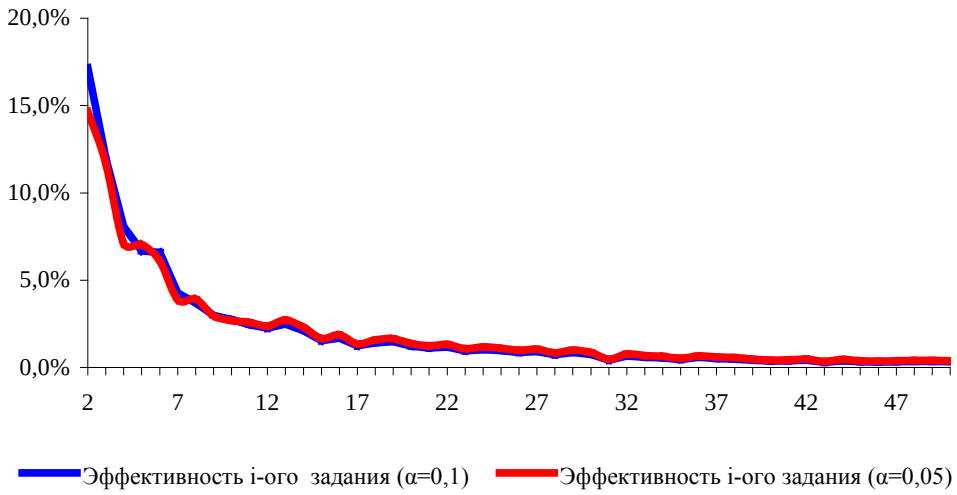

Рис. 7. Кривые эффективности тестовых заданий для уменьшения ширины доверительного интервала оценки результатов тестирования

Проведённые на рис. 7 кривые показывают, на какой процент каждое из последующих последовательно предъявляемых тестовых заданий сокращает исходную ширину L_1 ДИ, определяемую результатом выполнения первого ($i=1$) задания. Если второе ($i=2$) и третье ($i=3$) задания сокращают эту ширину (см. рис. 7) соответственно на 17,2% и 12,0%, то шестое ($i=6$) и седьмое ($i=3$) − только на 6,6% и 4,3%. Начиная с $i=23$, когда ресурсная результативность теста достигает значения $l(23)=0,854$ (см. рис. 6), каждое последующее тестовое задание уменьшает исходную ширину L_1 ДИ (см. рис. 7) менее чем на 1%.

Выводы

1. Метод интервальной оценки результатов критериально-ориентированного тестирования по моделям (2-7) работоспособен и обеспечивает определение не только значения оценки результатов тестирования, но и ее погрешности.

2. Применение метода интервальной оценки дает более точные оценки результатов тестирование при применении тестов небольшой длины (до 15-20 тестовых заданий).

3. Увеличение числа тестовых заданий в тесте свыше 25 заданий из соображений повышения надежности (снижения погрешности) получаемых оценок нецелесообразно.

3. Полученные значения ширины доверительного интервала оценок результатов тестирования свидетельствуют, что метод тестирования способен обеспечить достаточно надежную ($\alpha \le 0,1$) оценку знаний только “высоко” успешных ($p > 0,80$) и “низко” успешных ($p < 0,20$) обучаемых. Представляется возможным считать, что лежащие в этих интервалах оценки p имеют погрешность $\Delta p = \pm 0,1$.

4. Фактическая система предпочтений преподавателей, задаваемая в установках КСТ, формулируется в традиционной 4-балльной шкале, которая

соответствует не всей шкале вероятности успеха, а только ее части, обычно представленной интервалом $1 > p > 0,5$. Поэтому оценка погрешностей результатов критериально-ориентированного тестирования в традиционной 4-балльной шкале требует отдельного экспериментального исследования.

Литература:

1. Печников А.Н., Палкин К.С. Метод интервальной оценки результатов выполнения системы одиночных тестовых заданий закрытого типа с единственным верным ответом // Образовательные технологии и общество (Educational Technology & Society). 2014. Т. 17. № 2. С. 491-501. URL: http://ifets.ieee.org/russian/periodical/journal.html (дата обращения: 05.01.2015).
2. ЗАО “Научно-исследовательский институт “Центрпрограммсистем”. Автоматизированная система обучения “Медиатор”. РОСПАТЕНТ. Свидетельство об официальной регистрации программы для ЭВМ №2012614998 от 05.06.2012 г.
3. Печников А.Н. Теоретические основы психолого-педагогического проектирования автоматизированных обучающих систем. - Петродворец: ВВМУРЭ им. А.С. Попова, 1995. - 326с. URL: http://www.pedlib.ru/Books/1/0224/1_0224-8.shtml (дата обращения: 14.01.2015)
4. Печников А.Н., Шиков А.Н. Проектирование и применение компьютерных технологий обучения. - СПб.: Изд-во ВВМ, 2014. - 393с. URL: http://elibrary.ru/download/98535745.pdf (дата обращения: 14.01.2015)
5. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. - М: Логос, 2002. – 432 с. URL: http://www.twirpx.com/file/101903/ (дата обращения: 05.01.2015)