Исследование достоверности педагогических измерений с использованием аналитических свойств модели G.RASCH

Сергей Бутенков

Михаил Бойченко

Виталий Кривша

Южный федеральный университет, Таганрог

saab@tsure.ru

Опубликовано в ж. «Педагогические Измерения» № 1, 2007 г.

В работе вводятся методы получения оценок достоверности тестирования и шкалирования тестовых заданий, основанные на использовании аналитических свойств модели педагогических измерений по G.Rasch. В классической математической статистике для этих целей используются методы, основанные на применении функций Лапласа, не имеющих аналитического выражения и требующих применения численных методов расчета. Использование предложенных аналитических формул вычисления основных показателей качества измерений позволяет, значительно упростить расчеты, сводя их к вычислению элементарных функций. Наличие аналитических зависимостей позволяет теоретически исследовать зависимость показателей тестов от параметров тестовых заданий. На основе полученных формул даются практические рекомендации по выбору некоторых параметров тестов.

Ключевые слова: проектирование тестов, теория педагогических измерений, Rasch Measurement, логистические функции, доверительные границы, оценки достоверности.

Введение

Любое измерение, особенно педагогическое, от результатов которого зависит личная судьба больших масс обучаемых как в школах, так и в других учебных заведениях, должно иметь математически обоснованные доказательства того, что используемая методика тестирования является корректной. В классической теории тестирования [1] используются методы математической статистики, основанные на теоремах *закона больших чисел.* Эти методы успешно применяются также для контроля качества массовой продукции и для аналогичных задач в других отраслях науки и техники [2].

Математической основой многих статистических методов является применение функций Лапласа [3]. Этот класс функций не имеет аналитического выражения и требует для своего вычисления специальных методов расчета [4]. Это обстоятельство

заметно усложняет вычисления, поскольку при ручных вычислениях требуются таблицы основных статистических распределений, а при использовании компьютеров требуются значительные затраты машинного времени.

Оценки достоверности получаемых при этом результатов связаны с широким использованием *нормального закона распределения*, согласно которому, независимо от частных распределений случайных величин, их сумма будет распределена по *нормальному закону*. Для практического применения закона больших чисел центральную роль играет интегральная функция нормального распределения

$$N(x;m;\sigma)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{u-m}{\sigma}\right)^2}du, \qquad (1)$$

которая широко использовалась в ранних моделях педагогических измерений L.L.Thurstone, M.W.Richardson, W.A.Fergusson, D.M.Lawly, F.M.Lord и других [3]. Однако подобные модели имеют недостатки:

1. Функция (1) записана в виде интеграла с переменным верхним пределом и не может быть выражена через элементарные функции. Поэтому её значения приходится вычислять путем разложения в ряд или численного интегрирования [4], что значительно увеличивает объем вычислений;

2. Невозможны аналитические исследования формул, использующих функцию (1) и вывод на их основе простых рабочих формул, с использованием всех полезных свойств функции (1).

В процессе развития теорий педагогических измерений G.Rasch предложил *априорно модельный* подход к решению задач определения вероятности правильного ответа, который основан на использовании вместо функции (1) произвольных функций, имеющих график, сходный с графиком (1) и нормируемых с помощью искусственно вводимых параметров [5, 6]. Наиболее простым и удобным для аналитических расчетов является семейство *логистических функций* вида

$$L(x;D)=\frac{1}{1+e^{-D\cdot x}}, \qquad (2)$$

в котором параметр D играет роль *нормирующего множителя*. При правильном выборе D, вместо значений функции (1) можно вычислять значения её модели (2), причем такая замена дает среднеквадратическую погрешность вычисления не выше 10^{-5}, что вполне достаточно для педагогических измерений [9].

Априорно модельный подход, предложенный G.Rasch, неоднократно доказывал свою актуальность и практическую полезность, однако не все его потенциальные возможности раскрыты до настоящего времени. Например, при оценке достоверности результатов педагогических измерений (в том числе, и по модели G.Rasch) сейчас используются *численные методы* математической статистики [2, 7]. Между тем, априорно-модельный подход, благодаря использованию аналитических функций типа (2), позволяет получить ряд *аналитических решений* для задач оценки достоверности результатов тестирования, приводящих к сравнительно простым и удобным для практического использования формулам.

Задачи исследования

В работе предлагается модельный подход к получению оценок достоверности результатов педагогических измерений с помощью математической модели (2). Для уточнения математической формулировки задачи введем формальное описание классического подхода к получению статистических оценок результатов тестирования.

Статистическим материалом для вычислений при проведении педагогических измерений является матрица исходных тестовых баллов, или тестовая матрица. Формально представим тестовую матрицу T размером $m \times n$ в виде $T = \left\{ t_{ij} \in (0,1) \right\}, i = 1,2,\ldots,m; \quad j = 1,2,\ldots,n$. Здесь m – число испытуемых, n – число заданий теста. Суммированием по строкам матрицы T можно найти статистику, называемую *относительной частотой правильных ответов* i-го испытуемого на n заданий теста

$$P_n^*(A) = \frac{1}{n} \sum_{k=1}^{n} t_{i,k} \,. \tag{3}$$

Относительная частота (3) является эмпирической статистической оценкой вероятности события A (i*-й испытуемый имеет высокий уровень знаний по теме теста*), $P(A) = \lim\limits_{n \to \infty} P_n^*(A)$.

Другой важной в теории педагогических измерений статистикой является сумма по столбцам тестовой матрицы T, называемая *относительной частотой правильных ответов* m испытуемых на j задание теста

$$P_m^*(B) = \frac{1}{m}\sum_{k=1}^{m} t_{k,j} \,. \qquad (4)$$

Относительная частота (4) является статистической оценкой вероятности события B (j *-е задание имеет низкий уровень трудности*), $P(B) = \lim_{m\to\infty} P_m^*(B)$.

Вычисление вероятностей этих событий $P(A)$ и $P(B)$ по матрице исходных тестовых результатов T является ключевым моментом, определяющим достоверность педагогического измерения [3], поскольку оценки (3) и (4) *сходятся по вероятности* (т.е. достаточно медленно) при значительном увеличении объема выборки. Для (3) объемом выборки является количество n заданий теста, которое не может быть очень большим. Для (4) объем выборки - это количество m тестируемых по каждому заданию, которое в процессе пробного тестирования для шкалирования заданий может быть сделано весьма большим. Поскольку методики определения достоверности для (3) и (4) одинаковы, в дальнейшем будем описывать исследование только одной из них.

В инженерных приложениях для обеспечения достоверности полученных значений вероятностей по (3) и (4) используется следующий критерий [8]

$$\begin{cases} \sqrt{n \cdot p \cdot (1-p)} >> 1, \textit{при } p \neq 0,5; \\ n \cdot p \cdot (1-p) >> 1, \textit{при } p = 0,5, \end{cases} \qquad (5)$$

где n – объем выборки, p – оценка вероятности. Эмпирическим здесь является *отношение сравнения величин* $x >> y$, означающее «x *намного больше, чем* y». В разных источниках предлагаются разные численные оценки этого отношения для различных практических приложений методов математической статистики [2].

Задачей настоящей работы является объективизация такого отношения (с помощью возможностей модели G.Rasch) применительно к задаче оценки достоверности результатов тестирования, т.е. получение рабочих формул для вычисления необходимого значения объема выборки в процессе проектирования тестов.

Концептуальная основа работы.

В основу метода повышения достоверности статистических оценок (3) и (4) положены идеи метода интервальных оценок случайных величин [2]. Пусть мы имеем относительную частоту правильных ответов по (3). В статистике (и в теории

педагогических измерений по модели G.Rasch) предполагается, что эта оценка распределена асимптотически нормально (по закону (1)) с параметрами

$$M_{P_n^*(A)} = p,\ S_{P_n^*(A)} = \sqrt{D_{P_n^*(A)}} = \sqrt{\frac{p \cdot (1-p)}{n}}, \qquad (6)$$

где $p = P_n^*(A)$ согласно (3). Такие оценки в статистике называются *точечными оценками* случайной величины, однако достоверность таких оценок весьма невысока [2]. Для улучшения точечных оценок переходят к *интервальным оценкам* случайной величины, которые определяются заданным параметром *уровня достоверности оценки* $\alpha \in (0,1)$. Обычно в статистике выбирают стандартные значения уровня достоверности $\alpha = 0{,}1;\ 0{,}05;\ 0{,}01$. Заданное значение уровня α определяет *доверительные границы*, в которых лежит значение точечной оценки.

Пусть $(X_1, X_2, \ldots, X_n)$ есть выборка из генеральной совокупности с признаком X, распределение которой зависит от параметра γ. Пусть $\underline{\Gamma}(X_1, X_2, \ldots, X_n)$ и $\overline{\Gamma}(X_1, X_2, \ldots, X_n)$ – такие функции выборки, что при произвольном значении параметра γ выполняется условие

$$P\left(\underline{\Gamma}(X_1, X_2, \ldots, X_n) < \gamma < \overline{\Gamma}(X_1, X_2, \ldots, X_n)\right) = 1-\alpha. \qquad (7)$$

Тогда случайный интервал $\left(\underline{\Gamma}, \overline{\Gamma}\right)$ называется *доверительной оценкой параметра* γ с мерой достоверности $1-\alpha$.

Если имеется реализация $(x_1, x_2, \ldots, x_n)$ выборки $(X_1, X_2, \ldots, X_n)$, то реализация доверительной оценки дает интервал $\left(\underline{\gamma}, \overline{\gamma}\right)$, и в большом ряду выборок истинное значение γ лежит примерно в $(1-\alpha)\cdot 100\%$ случаев внутри вычисленных доверительных границ. Равенство (6) можно интерпретировать так: случайный интервал $\left(\underline{\Gamma}, \overline{\Gamma}\right)$ покрывает истинный параметр с вероятностью $1-\alpha$. Такая интервальная интерпретация дает возможность делать выводы о достоверности полученных оценок случайных величин.

Основные теоретические положения.

Для получения интервальной оценки величин (3) и (4) используются свойства интегральной функции распределения случайной величины [2]. При этом, чтобы

свести вычисления к типовым, значения случайной величины обычно нормируют. Так, если исходная величина X распределена по нормальному закону (1), то нормированная величина

$$Z = \frac{X - m_x}{\sigma_x} \tag{8}$$

имеет *стандартный нормальный закон* распределения с параметрами $m = 0$, $\sigma = 1$, что записывается как $Z \in N(z;0,1)$.

С учетом (8) можно записать условие попадания оценки в доверительный интервал с заданным уровнем достоверности в виде

$$P\left(-z_\alpha < Z < z_\alpha\right) = \frac{1}{\sqrt{2\pi}} \int_{-z_\alpha}^{z_\alpha} e^{-\frac{x^2}{2}} dx = 2\Phi_0(z_\alpha) = 1 - \alpha, \tag{9}$$

где z_α – пороговое значение оценки, которое и определяет доверительные границы (7) для случая распределения (9). В стандартных методах статистики значения z_α находятся численным решением уравнения относительно функции Лапласа $\Phi_0(x)$ [2]:

$$2\Phi_0(z_\alpha) = 1 - \alpha. \tag{10}$$

Введем, согласно идее метода G.Rasch, *модель интегральной функции распределения* типа (2), и относительно неё запишем условие (10) в виде

$$\left| \frac{1}{1 + e^{-D \cdot z_\alpha}} - 0.5 \right| = \frac{1 - \alpha}{2}, \tag{11}$$

где D – нормирующий коэффициент модели, $D \approx 1,7$ согласно [7].

В отличие от уравнения (10), из уравнения (11) можно аналитически найти пороговые значения z_α как

$$z_\alpha = \pm \frac{1}{D} \ln\left(\frac{2 - \alpha}{\alpha}\right). \tag{12}$$

Теперь мы можем аналитически найти границы доверительного интервала оценки вероятности исходной случайной величины $x_\alpha = m_x \pm \sigma_x \cdot \left| z_\alpha \right|$ или с учетом (6), (12)

$$p_\alpha = p \pm \left| z_\alpha \right| \cdot \sqrt{\frac{p \cdot (1 - p)}{n}}, \tag{13}$$

где значение p определяется выражением (6).

Для оценки точности полученной оценки вероятности (3) или (4) важную роль играет *длина доверительного интервала* $l = \overline{p} - \underline{p}$, которая с учетом (12) может быть найдена аналитически в виде функции основных параметров теста:

$$l(p,\alpha,n) = \frac{2}{D}\ln\left(\frac{2-\alpha}{\alpha}\right)\cdot\sqrt{\frac{p\cdot(1-p)}{n}}\,. \qquad (14)$$

На основе функции (14), в силу ее аналитичности, можно получить функцию, явно выражающую объем выборки n, при котором обеспечивается заданный уровень достоверности α в зависимости от желаемого значения длины доверительного интервала l^* и оцениваемой вероятности p:

$$n(p,\alpha,l^*) = \left\lfloor \frac{4\cdot p\cdot(1-p)\cdot\left(\ln\left(\frac{2-\alpha}{\alpha}\right)\right)^2}{\left(D\cdot l^*\right)^2} \right\rfloor, \qquad (15)$$

где стандартная целочисленная функция вещественного аргумента означает выделение его целой части.

С помощью полученных рабочих формул (11) - (15) можно полностью исследовать параметры проектируемого теста и выбрать те, которые обеспечивают необходимую достоверность результатов тестирования. Примеры применения рабочих формул приведены в следующем разделе.

Исследование полученных результатов.

В предыдущих разделах работы были выполнены преобразования, основанные на замене функций, порождаемых нормальным распределением (1) на более простые модельные функции типа (2). В результате получены достаточно простые аналитические формулы для исследования достоверности как оценки вероятности правильного ответа на задания теста (3), зависящей от уровня знаний каждого тестируемого. С помощью тех же формул возможна также и оценка вероятности правильного выполнения каждого задания (4), которая зависит от уровня трудности каждого задания [5].

Рисунок 1 демонстрирует интервальную оценку вероятности с помощью аналитических выражений (2) и (13).

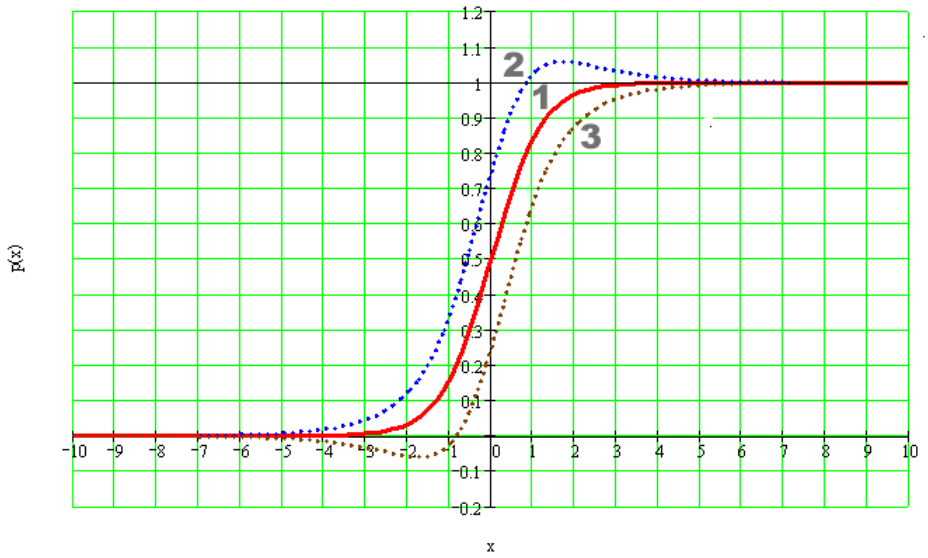

Рис.1. Кривые точной интегральной функции распределения вероятности p и её верхнего и нижнего значений, получаемых при обработке результатов тестирования.

Кривая 1 соответствует точному значению вероятности p правильного ответа i -го испытуемого на $n = 20$ заданий теста на уровне достоверности $\alpha = 0.1$. Кривая 2 соответствует верхней границе $\overline{p}$ доверительного интервала, а кривая 3 – нижней границе $\underline{p}$ доверительного интервала для того же теста. Рисунок подтверждает, что с изменением значения вероятности p изменяется и доверительный интервал $\left(\underline{p}, \overline{p}\right)$ для этого значения согласно (13).

Аналитическое исследование полученной в работе функции (14) на максимум по переменной p показывает, что, к сожалению, наибольшая длина доверительного интервала $\left(\underline{p}, \overline{p}\right)$ и, соответственно, наименьшая точность вычисления вероятности имеет место как раз в точке наибольшей разрешающей способности теста при $x = 0,\ p(x) = 0{,}5$.

Аналитическое исследование функции (14) показывает также, что влияние уровня достоверности α на длину доверительного интервала значительно меньше, чем влияние других параметров, поскольку α входит в (14) под знаком логарифма. Поэтому в дальнейшем рассмотрим влияние на точность оценок вероятностей объема выборки n.

Рисунок 2 изображает зависимость максимального значения длины доверительного интервала $\left(\underline{p}, \overline{p}\right)$ по формуле (14) от объема n выборки для (3) или (4) при одинаковом уровне достоверности $\alpha = 0.1$.

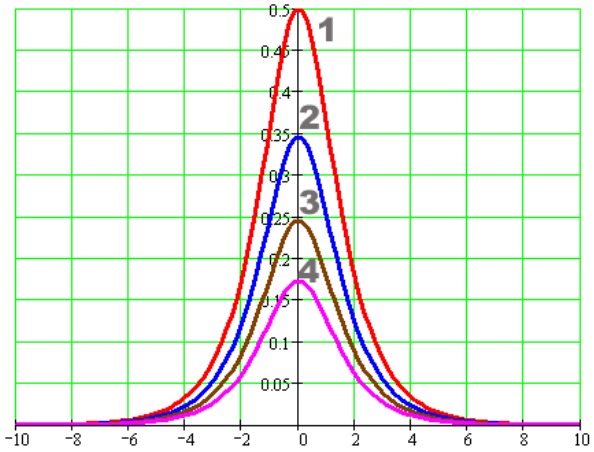

Рис.2. Зависимость длины доверительного интервала $\left(\underline{p}, \overline{p}\right)$ для вероятности $p = 0,5$ от значения x для различного числа заданий теста.

Кривая 1 – тест из 12 заданий, 2 – из 25 заданий, 3 – из 50 заданий, 4 – из 100 заданий. Очевидно, что даже при числе заданий теста $n = 100$ не достигается приемлемое значение длины доверительного интервала $l = \left(\underline{p}, \overline{p}\right)$, например, $l = 0.1$ (десятипроцентная точность).

Рисунок 3 демонстрирует аналитическую зависимость длины доверительного интервала $\left(\underline{p}, \overline{p}\right)$ от объема выборки n (количество заданий теста при оценке знаний и количества тестируемых, ответивших на данное задание). Здесь также выбрано значение $p = 0,5$ для получения верхней оценки точности.

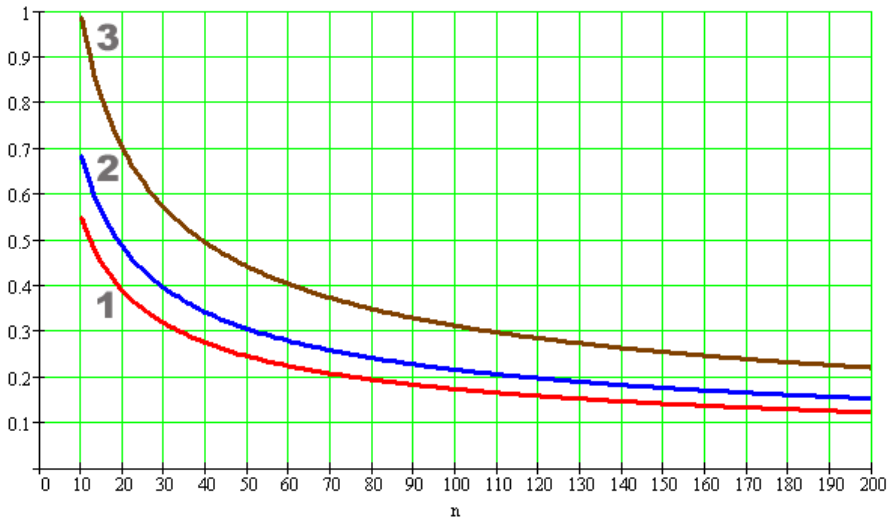

Рис.3. Зависимость максимальной длины доверительного интервала $\left(\underline{p}, \overline{p}\right)$ по вероятности от доверительного уровня α и числа заданий теста.

Кривая 1 соответствует значению доверительного уровня $\alpha = 0.1$, 2 – уровню $\alpha = 0.05$, 3 – уровню $\alpha = 0.01$. Очевидно, что в случае, когда n – число заданий, достижение высокой точности вычисления вероятности правильного ответа по (3) невозможно при разумном числе заданий ($n < 100$), в том числе и при большом значении доверительного уровня $\alpha = 0.1$. Зато для оценки уровня трудности задания по (4) малая длина интервала $\left(\underline{p}, \overline{p}\right)$ может быть достигнута при $n > 2000$, в том числе и при достаточно малом доверительном уровне $\alpha = 0.01$.

Рассмотрим теперь решение обратной задачи с помощью рабочей формулы (15). Задавшись желаемыми значениями основных параметров теста, можно оценить необходимые объемы выборок n как для вероятности правильного ответа (3), так и для вероятности правильного выполнения задания (4).

Следующий рисунок демонстрирует зависимости, показывающие, при каких объемах выборки достигается желаемая длина доверительного интервала $\left(\underline{p}, \overline{p}\right)$.

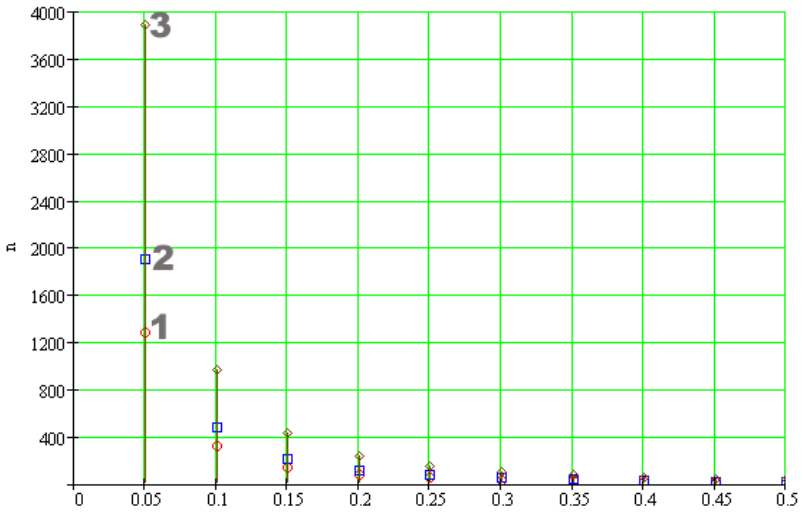

Рис. 4. Зависимость необходимого объема выборки n от доверительного уровня α и желаемой длины доверительного интервала $l = \left(\underline{p}, \overline{p} \right)$.

Точки 1 соответствуют доверительному уровню $\alpha = 0.1$, 2 – уровню $\alpha = 0.05$, 3 – уровню $\alpha = 0.01$. Очевидно, что в силу свойств выведенной в работе аналитической зависимости (15) требуемые объемы выборок значительно возрастают при уменьшении желаемой длины доверительного интервала $l = \left(\underline{p}, \overline{p} \right)$ для вероятности. Тем не менее, при использовании контингента около 4000 человек для шкалирования тестовых заданий, полученные оценки вероятности правильного выполнения задания (4) могут быть вполне точными.

Следующий рисунок изображает эту же зависимость в области малых объемов выборок n заданий, что имеет место при вычислении вероятности правильного ответа (3).

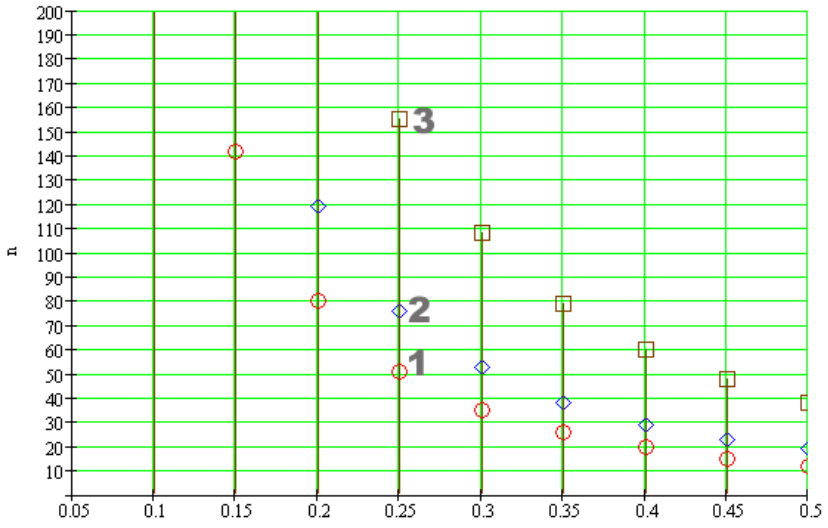

Рис. 4. Зависимость необходимого количества заданий теста n от доверительного уровня α и желаемой длины доверительного интервала $l = \left(\underline{p}, \overline{p} \right)$.

Точки 1 соответствуют значению $\alpha = 0.1$, 2 – значению $\alpha = 0.05$, 3 – значению $\alpha = 0.01$. Перспективы получения точного значения вероятности (3) по тестам с числом заданий до $n = 40$ представляются печальными, так как длину доверительного интервала не удается сделать меньше, чем 0,25 даже при низком доверительном уровне $\alpha = 0.1$.

Выводы

Полученные результаты позволяют аналитически исследовать точность получаемых при тестировании вероятностей (3) и (4) и проектировать тесты с учетом требуемой достоверности, задаваемой уровнем α.

При выборе числа заданий теста n можно использовать оценку (3) а при шкалировании трудности заданий теста – оценку (4).

Полученные в работе аналитические зависимости можно использовать для выбора основных параметров теста по заданным значениям уровня достоверности и длины доверительного интервала, которые являются объективными оценками точности педагогических измерений и позволяют определять величину n в (5).

Еще более полезными будут полученные формулы при использовании в автоматических системах тестирования [9], позволяющих оперативно менять показатели тестов в процессе работы с тестируемым.

На основании полученных в работе результатов возможно дальнейшее развитие методов построения статистик, используемых в модели G.Rasch, которые основаны на функциях, обратных к модельным функциям (2). Это приведет к уточнению и улучшению оценок не только в начальной форме (вероятностей) но и в логитах.

Литература

1. Аванесов В.С. Определение исходных понятий теории педагогических измерений. «Педагогические измерения» № 3, 2005г
2. Смирнов Н.В., Дунин-Барковский И.В. Краткий курс математической статистики для технических приложений. - М.: Физматгиз, 1959.

3. Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дис. докт. пед. наук. С-Петербургский гос. университет, 1994. –339с.
4. Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. - М.: Физматлит, 1969.
5. Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. With a Foreword and Afterword by B.D. Wright. The Univ. of Chicago Press. Chicago & London, 1980.
6. Karabatsos G. Axiomatic measurement theory as a basis for model selection in item response theory. Paper presented at 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA. 1999, July.
7. Булыгин В.Г. Основы автоматизации процесса обучения. - Йошкар-Ола, 2003.
8. Бендат Дж., Пирсол А. Измерение и анализ случайных процессов. - М.:Мир, 1974.
9. Бойченко М.М. Современная система контроля знаний обучаемых. - В кн.: Материалы международной конференции “Динамика процессов в природе, обществе и технике: информационные аспекты”, часть I, Таганрог:ТРТУ, 2003, с. 12-13.