Ученые записки Таврического национального университета им. В.И. Вернадского

Серия "Математика. Механика. Информатика и Кибернетика" N 1 (2005) 95 – 104

УДК 51.77: 37.015.42

В.А. ТЕМНЕНКО

МАТЕМАТИЧЕСКАЯ ТЕОРИЯ СОВЕРШЕННОГО ТЕСТА

Введение

Теоретическая физика содержит немалое число идеализированных моделей: "совершенный газ", "идеальная жидкость", "абсолютно твердое тело" и т.д. Этим моделям ничего не соответствует в природе, но изучение их помогает понять некоторые фундаментальные черты реальности, не осложненные множеством важных, но второстепенных деталей.

В тестологии, как и в теоретической физике, такие идеализированные модели могли бы приносить пользу. Важнейшая из таких моделей – ***модель совершенного теста*** – может быть описана следующим образом.

1. Айтем-Банк I

Пусть нам дано некоторое множество $I = \{I_i\}$, настолько большое, что для практических целей может считаться бесконечным. Элементы этого множества I_i – тестовые задачи ("айтемы") с заранее известными индексами трудности p_i : $0 < p_i < 1$. Эти индексы определены из предшествующих выборочных претестов на достаточно больших и репрезентативных выборках испытуемых. Если вероятность угадать ответ ничтожно мала, то p_i – это доля испытуемых, не решивших задачу номер i в этих испытаниях.

Так можно приписывать эмпирическую меру трудности при решении многих математических задач, в которых вероятность угадать конкретный ответ вида, например, $x = (\pi + \sqrt{2})/(\sqrt{13} + \sqrt{5})$ – действительно, не заслуживает быть учитываемой.

Если же испытуемый должен выбрать ответ среди предложенных ему четырехпяти опций, то эмпирически определенная мера трудности задачи нуждается в корректировке на "угаданный ответ". Пусть q' – эмпирически определенная доля испытуемых, выбравших правильный ответ, а l – число предложенных опций ответа. Тогда выражение

$$s' = \frac{1 - q'}{l - 1} \tag{1}$$

описывает среднюю, в расчете на один дистрактор (т.е. опцию с неправильным ответом) вероятность выбора неправильного ответа. Предполагая, что вероятность выбора ***наугад*** правильного ответа такая же, как и неправильного, для определения доли q испытуемых, правильно (и не случайно) выбравших ответ в этой задаче, из эмпирически определенной доли q' необходимо вычесть вероятность s' :

$$q = q' - s',$$

или с учетом (1):

$$q = \frac{q'l - 1}{l - 1}.$$

Если определенная таким образом величина q оказывается отрицательной, то соответствующая задача не должна включаться в Айтем-банк, т.е. упомянутое выше множество I , т.к. она некорректно сформулирована и провоцирует выбор неправильного ответа. Если же $0 < q < 1$, то величина:

$$p = 1 - q,$$

будет являться индексом трудности задачи.

Пополняя множество I , мы не должны включать в него элементы с $p = 0$ (задачи, которые решают все) и $p = 1$ (задачи, которые не смог решить ни один испытуемый). Обширность множества I означает, что мы можем найти в нем элемент с любым значением $p : 0 < p < 1$.

2. Совершенный тест T

Определим ***совершенный тест T*** как конечную (но, вообще говоря, достаточно большую) выборку из множества I , сконструированную в соответствии с заранее заданной ***функцией трудности теста*** $f(p)$, задающей распределение задач в тесте по индексу трудности, так что $f(p)dp$ – это доля задач трудности p в малом интервале шириной dp .

Функция $f(p)$ равняется нулю на краях интервала трудности:

$$f(0) = f(1) = 0, \tag{2}$$

положительна внутри интервала

$$f(p) > 0 \quad \text{при} \quad 0 < p < 1, \tag{3}$$

и нормирована на единицу:

$$\int_0^1 f(p)dp = 1. \tag{4}$$

Кроме обязательных условий (2), (3) и (4) на функцию трудности теста $f(p)$ можно наложить некоторые дополнительные, диктуемые математическим удобством

оперирования с этой функцией. Будем считать ее непрерывной и дифференцируемой. Кроме того, удобно считать ее ***унимодулярной*** -- т.е. имеющей единственный максимум в некоторой точке p_m между $p = 0$ и $p = 1$.

Составители тестов должны указывать функцию теста $f(p)$, описывающую распределение задач по трудности. (Насколько нам известно, на самом деле они этого не делают).

Если известна функция $f(p)$, несущая полную статистическую информацию о тесте, то могут быть посчитаны некоторые интегральные статистические характеристики теста; например, средняя по тесту трудность задач, которую мы обозначим P или $\langle p \rangle_f$:

$$P = \int_0^1 p \cdot f(p) dp \tag{5}$$

и дисперсию σ_p:

$$\sigma_p^2 = \int_0^1 (p - P)^2 f(p) dp = \langle p^2 \rangle_f - P^2. \tag{6}$$

Тесты можно классифицировать в зависимости от значений этих интегральных характеристик. Тесты с $P > \frac{1}{2}$ назовем тестами высоких достижений, или ***силинговыми*** тестами. (Термин ceiling – верхний уровень, потолок – встречается в тестологических глоссариях). В таких тестах трудные задачи представлены с большим относительным весом, чем легкие. Тесты с $P = \frac{1}{2}$ назовем ***нейтральными***. В них легкие и трудные задачи присутствуют в равных пропорциях. Тесты с $P < \frac{1}{2}$ назовем ***утешительными*** – в них больше вес нетрудных задач с низким индексом p.

Как нам представляется, тесты, служащие для определения общих способностей человека, - например, такие, как абитуриентский тест SAT I в США и его аналоги в Восточной Европе и странах бывшего СССР – должны быть нейтральными (ориентация на общее ранжирование испытуемых), а тесты, оценивающие готовность к определенному виду деятельности, например, "предметные" тесты SAT II в США - должны быть силинговыми (ориентация на отбор лучших среди испытуемых). Утешительные тесты, вероятно, уместны в сфере развлечений.

Если тест сконструирован так, что его дисперсия σ_p меньше некоторого наперед заданного числа ε , то такой тест будем называть ε-сфокусированным. При достаточно малом ε можно говорить о "хорошо сфокусированном" тесте.

3. Эйлеровы тесты

Учитывая условия (2) и (3), налагаемые на $f(p)$, общий вид функции трудности удобно представить так:

$$f(p) = p^{\alpha}(1-p)^{\beta} \cdot \varphi(p), \quad (\alpha > 0, \beta > 0) \tag{7}$$

где $\varphi(p)$ – гладкая функция, не обращающаяся в нуль при $0 \le p \le 1$. Если $\varphi(p) = A = const$, то такую функцию $f(p)$ назовем Эйлеровой $f_E(p;\alpha;\beta)$, а тест, описанный этой функцией - эйлеровым тестом:

$$f_E(p;\alpha;\beta) = A_{\alpha\beta} \cdot p^{\alpha}(1-p)^{\beta}. \tag{8}$$

Эйлеровы функции задают простейшее двухпараметрическое семейство тестов. Постоянная $A_{\alpha\beta}$ определяется из условия нормировки (4).

$$A_{\alpha\beta} \cdot \int_0^1 p^{\alpha}(1-p)^{\beta} dp = 1. \tag{9}$$

Эта постоянная не меняется при перестановке индексов α и β. Интеграл в формуле (9) выражается через Бета-функцию Эйлера [1, (6.2.1)]:

$$\int_0^1 p^{\alpha}(1-p)^{\beta} dp = B(\alpha+1, \beta+1), \tag{10}$$

которая, в свою очередь, выражается через Гамма-функцию Эйлера Г:

$$B(\alpha+1, \beta+1) = \frac{\Gamma(\alpha+1) \cdot \Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}. \tag{11}$$

Подстановка формул (10) и (11) в условие нормировки (9) дает выражение нормировочной константы $A_{\alpha\beta}$:

$$A_{\alpha\beta} = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1) \cdot \Gamma(\beta+1)}. \tag{12}$$

Если эйлеровы коэффициенты α и β целые числа, $\alpha = n$ и $\beta = m$, то Гамма-функция сводится к факториалам, и выражение (12) упрощается:

$$A_{nm} = \frac{(n+m+1)!}{n!m!}. \tag{13}$$

Формула (12) упрощается и в том случае, если один или оба Эйлеровы коэффициента полуцелые (см. [1, (6.1.12)]):

$$A_{n+\frac{1}{2},m} = \frac{(2 \cdot (n+m)+3)!!}{2^{m+1} \cdot m! \cdot (2n+1)!!}, \tag{14}$$

где выражение $(2n+1)!!$ означает произведение всех целых нечетных чисел до числа $2n+1$ включительно;

$$A_{n+\frac{1}{2},m+\frac{1}{2}} = \frac{2^{n+m+2} \cdot (n+m+2)!}{\pi \cdot (2m+1)!! \cdot (2n+1)!!}. \tag{15}$$

В таблице (1) представлены значения эйлеровой нормировочной константы $A_{\alpha\beta}$ для целых и полуцелых значений индексов α и β ; посчитанные по формулам (13), (14), (15):

Таблица 1. Эйлерова нормировочная константа $A_{\alpha\beta}$ для целых и полуцелых значений индексов α и β.

α \ β	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3
$\frac{1}{2}$	$\frac{8}{\pi}$	$\frac{15}{8}$	$\frac{16}{\pi}$	$\frac{105}{16}$	$\frac{128}{5\pi}$	$\frac{315}{32}$
1	$\frac{15}{8}$	6	$\frac{35}{4}$	12	$\frac{63}{4}$	20
$1\frac{1}{2}$	$\frac{16}{\pi}$	$\frac{35}{4}$	$\frac{128}{3\pi}$	$\frac{315}{16}$	$\frac{256}{3\pi}$	$\frac{1155}{32}$
2	$\frac{105}{16}$	12	$\frac{315}{16}$	30	$\frac{693}{16}$	60
$2\frac{1}{2}$	$\frac{128}{5\pi}$	$\frac{63}{4}$	$\frac{256}{3\pi}$	$\frac{693}{16}$	$\frac{1024}{5\pi}$	$\frac{3003}{32}$
3	$\frac{315}{32}$	20	$\frac{1155}{32}$	60	$\frac{3003}{32}$	140

Эйлеровы тесты с $\alpha = \beta$ являются нейтральными; при $\alpha > \beta$ – тест силинговый; при $\alpha < \beta$ – это утешительный тест.

Средняя трудность эйлеровских тестов $P_{\alpha\beta}$ может быть посчитана по формуле (5) с учетом (8) и (12); что приводит к весьма простому выражению:

$$P_{\alpha\beta} = \frac{\alpha+1}{\alpha+\beta+2}. \tag{16}$$

Подсчитанная по формуле (16) средняя трудность представлена в таблице 2:

Таблица 2. Средняя трудность эйлеровских тестов $P_{\alpha\beta}$.

α \ β	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{8}$	$\frac{1}{3}$	$\frac{3}{10}$	$\frac{3}{11}$
1	$\frac{4}{7}$	$\frac{1}{2}$	$\frac{4}{9}$	$\frac{2}{5}$	$\frac{4}{11}$	$\frac{1}{3}$
$1\frac{1}{2}$	$\frac{5}{8}$	$\frac{5}{9}$	$\frac{1}{2}$	$\frac{5}{11}$	$\frac{5}{12}$	$\frac{5}{13}$
2	$\frac{2}{3}$	$\frac{3}{5}$	$\frac{6}{11}$	$\frac{1}{2}$	$\frac{6}{13}$	$\frac{3}{7}$
$2\frac{1}{2}$	$\frac{7}{10}$	$\frac{7}{11}$	$\frac{7}{12}$	$\frac{7}{13}$	$\frac{1}{2}$	$\frac{7}{15}$
3	$\frac{8}{11}$	$\frac{2}{3}$	$\frac{8}{13}$	$\frac{4}{7}$	$\frac{8}{15}$	$\frac{1}{2}$

Дисперсия эйлеровских тестов $\sigma_{\alpha\beta}$ в сочетании с формулами (6), (8) и (12) имеет следующий вид:

$$\sigma_{\alpha\beta} = \sqrt{\frac{(\alpha+1)\cdot(\beta+1)}{(\alpha+\beta+2)^2 \cdot (\alpha+\beta+3)}}. \tag{17}$$

Подсчитанная по формуле (17) дисперсия эйлеровских тестов $\sigma_{\alpha\beta}$ представлена в таблице 3.

Таблица 3. Дисперсия эйлеровских тестов $\sigma_{\alpha\beta}$.

α \ β	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3
$\frac{1}{2}$	0.25	0.233285	0.216506	0.201008	0.187083	0.174685
1	0.233285	0.223607	0.211880	0.2	0.188682	0.178174
$1\frac{1}{2}$	0.216506	0.211880	0.204124	0.195304	0.186339	0.177646
2	0.201008	0.2	0.195304	0.188982	0.182033	0.174964
$2\frac{1}{2}$	0.187083	0.188682	0.186339	0.182033	0.176777	0.171117
3	0.174685	0.178174	0.177646	0.174964	0.171117	0.166667

Чем меньше дисперсия σ, тем больше в тесте удельный вес задач с трудностью, близкой к средней трудности теста и тем беднее представлены как очень трудные, так и очень легкие задачи. Трудно строго мотивировать выбор дисперсии σ, приемлемый для стандартизованного массового тестирования, но представляется, что значение $\sigma = 0.2$ могло бы рассматриваться в качестве верхнего предела; однако реально используемое значение не должно быть ***существенно*** меньше, чем это предельное значение.

Исходя из этих соображений, можно рекомендовать для нейтрального теста в качестве стандартной функции трудности Эйлерову функцию с индексом $\alpha = 2$ и $\beta = 2$:

$$f_{st,n} = 30p^2(1-p)^2. \tag{18}$$

Индексы у функции призваны напомнить о том, что эта функция задает нейтральный тест (индекс n) и рекомендуется в качестве стандартизованной для массового тестирования. Для этой функции дисперсия $\sigma = 0.188982$.

При конструировании силинговых тестов дополнительная проблема связана с выбором среднего по тесту значения трудности задач P. Представляется, что для массового тестирования разумно принять $P = \frac{2}{3}$. Большее значение P уместно в условиях олимпиадных состязаний, задача которых – выявление небольшого числа победителей. Уменьшение P приближает тест к нейтральному и ослабляет селективные свойства теста, необходимые, например, при конкурсном отборе абитуриентов. С учетом этого обстоятельства, в качестве силинговой функции распределения трудности задач для массовых тестов можно рекомендовать эйлерову функцию с индексом $\alpha = 3$ и $\beta = 1$:

$$f_{st,c} = 20p^3(1-p). \tag{19}$$

Индексы у функции f призваны напомнить о том, что эта функция рекомендована в качестве стандартной (индекс st) функции трудности для массовых силинговых (индекс c) тестов.

Функция трудности (19) имеет среднее по тесту значение трудности $P = \frac{2}{3}$ и дисперсию $\sigma = 0.178174$.

Целочисленность индексов α и β в стандартных функциях (18) и (19), сама по себе не принципиальная, облегчает выполнение различных вычислений с этими функциями.

4. Ценовая функция теста

Конструирование теста требует, кроме выбора фундаментальной функции $f(p)$, задающей распределение задач по трудности, еще одного произвольного выбора. В конструкцию каждого теста заложена еще ценовая функция $c(p)$, определяющая ценность каждой решенной задачи. Произведение ценовой функции $c(p)$ на функцию трудности $f(p)$ задает распределение задач по ценности $\Psi(p)$:

$$\Psi(p) = c(p) \cdot f(p), \tag{20}$$

которое также должно быть нормировано:

$$\int_0^1 \Psi(p)dp = 1. \tag{21}$$

В правой части нормировочного соотношения (21) может стоять не единица, а любое другое положительное число, задающее общую цену (например, 100 или 1000 баллов) всех задач теста.

Каким требованиям должна удовлетворять ценовая функция $c(p)$?

Она должна быть положительной и монотонно растущей с ростом трудности задач. Представляется, что на эту функцию можно наложить такие дополнительные условия:

- $c(p)$ должна быть отлична от нуля даже при $p \to 0$ (поощряться должно решение даже простейших задач);
- $p \to 1$ при ценовая функция $c(p)$ должна расти неограниченно (растущее поощрение за решение самых трудных задач). Этот рост лимитирован только требованием сходимости нормировочного интеграла (21).

Простейшая ценовая функция, удовлетворяющая этим условиям, имеет вид:

$$c(p) = \frac{b}{1-p}. \tag{22}$$

Нормировочная постоянная b определяется из нормировочного условия (21) с учетом определения (20).

Стандартную функцию (22) можно использовать для стандартного нейтрального теста (18) и для силингового теста (19). В этих двух случаях ценовые функции (22) отличаются только значением нормировочной константы b. Для стандартного нейтрального теста $b = \frac{2}{5}$ и, соответственно

$$c_{st,n} = \frac{2/5}{1-p},$$

$$\Psi_{st,n} = 12p^2(1-p). \tag{23}$$

Эта функция $\Psi_{st,n}$ имеет максимум при $p = \frac{2}{3}$. С учетом вида распределения (23) испытуемому в нейтральном тесте можно рекомендовать следующую стратегию: сначала решать задачи с уровнем трудности $p \cong \frac{2}{3}$, а затем чередовать решение более трудных и более легких задач, отступая все дальше от точки экстремума функции Ψ.

Для силингового теста (19) нормировочная постоянная ценовой функции $b = \frac{1}{5}$. Соответственно,

$$c_{st,n} = \frac{1/5}{1-p},$$
$$\Psi_{st,n} = 4p^3. \tag{24}$$

Ценовое распределение (24) растет вплоть до $p = 1$. Это позволяет рекомендовать испытуемому в силинговом тесте следующую стратегию: начинать тест с самых трудных задач, постепенно смещаясь к более легким.

С учетом всего изложенного, можно дать такое формальное определение совершенного теста:

Совершенный тест – это тройка $\{I, f, c\}$, ***где*** I ***– бесконечный айтем-банк,*** a f u c ***– произвольно выбираемые функция трудности теста и ценовая функция теста, удовлетворяющие условиям нормировки и другим условиям, описанным в тексте.***

Конкретные выражения этих функций, рекомендованные в формулах (18), (19), (23) и (24) диктуются не математическими соображениями, а, скорей, социальными. Однако и само массовое тестирование принадлежит социальной среде.

5. Учет конечности и дискретности теста T

Генерируя тест T, т.е. конечное множество в N айтемов из бесконечного банка айтемов I в соответствии с функцией трудности теста $f(p)$, можно поступить следующим образом. Разобьем весь интервал трудностей $0 \le p \le 1$ на последовательные отрезки $p_{i-1} \le p \le p_i$; $0 \le i \le k$. $p_0 = 0; p_k = 1$. Эти отрезки могут быть равными по длине, но это условие не является обязательным. Общее число отрезков k должно быть существенно меньше числа задач в тесте N, но и не быть слишком малым. Например, при $N \cong 100$, $k \cong 10$.

Интегрируя функцию трудности $f(p)$ на i-ом отрезке, определим площадь S_i под графиком функции трудности:

$$S_i = \int_{p_{i-1}}^{p_i} f(p)dp.$$

Число задач N_i с индексом трудности p, принадлежащем интервалу $p_{i-1} \le p \le p_i$, которые нужно выбрать из банка I для теста T, определяется выражением:

$$N_i = [NS_i],$$

в котором квадратные скобки означают целое число, ближайшее к числу NS_i

Сумма определенных таким образом целых чисел N_i будет отличаться от N менее, чем на k . (А в большинстве случаев не больше, чем на $\sqrt{k}$).

Выстроенное таким образом распределение чисел N_i корректируется вручную добавлением или уменьшением задач, начиная с области максимума функции $f(p)$ равномерно в сторону возрастания и убывания p до тех пор, пока сумма чисел N_i не станет равной заданному размеру теста N .

Расчет по этой схеме для стандартной функции нейтрального теста (18) при разбиении единичного интервала изменения p на 10 равных отрезков ($k = 10$) при размере теста $N = 100$ дает следующий результат:

i	1	2	3	4	5	6	7	8	9	10
N_i	1	5	9	15	18	18	15	9	5	1

Сумма N_i равна 98, т.е. в этом распределении не хватает двух задач. Добавим по одной задаче на отрезках номер 4 и 7 по обе стороны от максимума функции распределения, что дает окончательную версию теста, совместимую в пределах погрешности дискретизации с функцией плотности (18); см. рис.1:

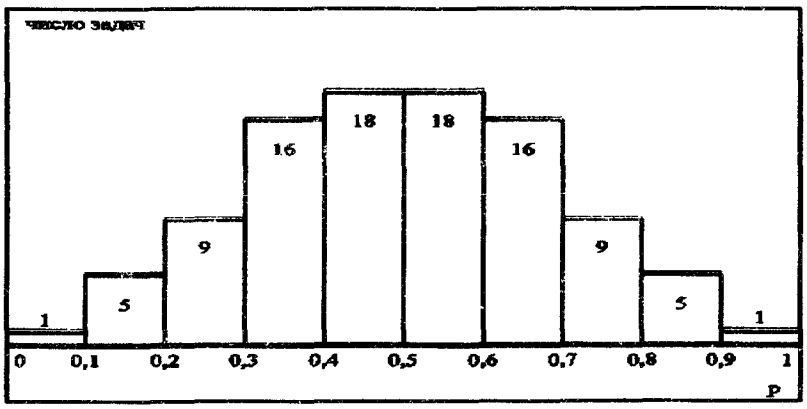

Рис.1. Распределение задач в стандартном нейтральном тесте (100 задач, 10 равных интервалов трудности).

Расчёт по этой схеме для стандартной функции силингового теста (19) при разбиении интервала трудностей на 10 равных отрезков ($k = 10$) и том же размере теста ($N = 100$) дает следующий результат:

i	1	2	3	4	5	6	7	8	9	10
N_i	0	1	2	6	10	15	19	21	18	8

В этом случае в "ручной" корректировке нет необходимости: сумма N_i равна размеру теста.

Результат c_{ind} выполнения теста для каждого испытуемого определяется суммой значений ценовой функции $c(p)$ по всем решенным задачам r: $c_{ind} = \sum_r c(p_r)$.

Последнее утверждение верно, если вероятность угадывания правильного ответа пренебрежимо мала. Для теста с фиксированным набором ответов в обработку результатов теста нужно вносить штрафные поправки, связанные с возможностью случайного угадывания правильного ответа. Однако обсуждение этой техники поправки уместно провести в отдельной статье.

Список литературы

[1] Справочник по специальным функциям. С формулами, графиками и математическими таблицами. /Под редакцией М.Абрамовица и И.Стигана. Пер. с англ. под ред. В.А.Диткина и Л.Н.Кармазиной. - М.: Наука, гл. ред. физ.-мат. лит., 1979. - 832 с.