Lietuvos matematikos rinkinys
Proc. of the Lithuanian Mathematical Society, Ser. A
Vol. **54**, 2013, 37–42

ISSN 0132-2818
www.mii.lt/LMR/

The construction of test reliability statistical criteria by a computer simulation

Olga Navickienė[1], Aleksandras Krylovas[1], Natalja Kosareva[2]

[1] *Mykolas Romeris University, Faculty of Economics and Finance Management*
Ateities 20, LT-08303

[2] *Vilnius Gediminas Technical University, Faculty of Fundamental Sciences*
Saulėtekio al. 11, LT10223 Vilnius

E-mail: navickiene@mruni.eu, krylovas@mruni.eu, natalja.kosareva@vgtu.lt

Abstract. The paper describes a computer simulation experiment of knowledge assessment test when item's characteristic functions are power functions and knowledge level of the population has uniform or normal distribution. The inversions of the conditional means were calculated and probability distributions of the numbers of inversions were simulated. In practice of construction of knowledge tests for checking their suitability (validity) observed values of inversions are compared with reference values obtained by Monte Carlo experiments.

Keywords: knowledge assessment tests, Monte Carlo experiments, mathematical modelling, statistical criteria.

1 Introduction

In the works [3, 2, 4] of this article co-authors a mathematical model of closed type knowledge assessment test (with multiple-choice answers) was proposed. The model was based on the assumptions that each test taker has a certain level of knowledge $p \in (0,1)$ and probabilities of the correct answer to j-th test item could be expressed with nondecreasing functions of variable p – the so-called characteristic functions of test items $k_j : [0,1] \to [0,1]$. Let's us denote $M_n(p)$ the average number of correct answers to n items test of a test taker whose knowledge level is p: $M_n \in \{0,1,\ldots,n\}$. It's clear that $M_n(p)$ must be nondecreasing function of the variable p, i.e. the test takers with a higher knowledge level on average are receiving more test points (numbers of correct items). In theory, this allows to find the inverse function M_n^{-1}, to determine the knowledge level of test takers according to their received test points. However it is difficult in practice and it requires the further study. The paper [6] presents an empirical study of such a task when the knowledge level of test takers was considered as a priori known information. Student's knowledge was assessed in various ways, then the average value of the estimates was calculated and depending on the received value the test taker's grade point was restored. With regard to the stochastic nature of phenomenon sometimes the stronger student (having higher knowledge level p_i) is receiving lower test scores and vice versa, the student receiving higher test score can have a lower level of knowledge. If such discrepancies are occurring frequently, the test isn't qualitative; it unreliably measures student's knowledge, so the test hasn't *validity* property. In this article we offer the design scheme of an evaluation criteria for such discrepancies which is realized by Monte Carlo experiments.

2 The experiment

Some methodological assumptions were done for the experiment. There are m students having knowledge levels distributed according to $m_1+m_2+\cdots+m_k=m$ where m_i is the number of students having knowledge level p_i. Theoretical assumption of experiment – characteristic functions k_j of test items are known and they are power functions: $k_j(p)=p^\alpha$. Depending on the parameter $\alpha\in(0,1)$ values these functions simulate the difficulty of the questions (test items): the greater is α the harder the question being asked (student with the level of knowledge p answers to the question correctly with a lower probability p^α with the higher α value). The difficulties of test questions considered to be known, so test consists of questions with known n characteristic functions $k_1,k_2,\ldots,k_n$. The result of experiment is the computer simulation of student's responses to n test items – random vectors $r=(r_1,r_2,\ldots,r_n)$ where $r_j\in\{0,1\}$ consisting of zeros and ones generated by a priori distributions $P\{r_j=1\}=k_j$, Note that the characteristic function can have another form (see [4, 7]), and it would not change design of the study.

Calculations were limited with the test length of $n=15$ items. There were 5 "easy" items ($\alpha=0.25$), 5 items of "medium difficulty" ($\alpha=0.5$) and 5 "hard" items ($\alpha=0.75$). Each of them could be responded correctly ($r_j=1$) or not ($r_j=0$). By computer simulation ($m=50$) students' answers to such test questions have been generated. Students' knowledge levels were set in advance. The example of simulation result is random zeros and ones matrix shown in Fig. 1.

After that the numbers of students with a certain knowledge level (from 0.1 to 1) correctly responded to j ($j=0,1,\ldots,15$) test items were calculated (see Fig. 2). The element of this matrix

$$F=(f_{ij})_{10\times(n+1)},$$

f_{ij} – is the number of test takers having knowledge level $i=1,2,\ldots,10$ ($p_i=0.1,0.2,\ldots,1$) and correctly responded to $j=0,1,\ldots,15$ test items. Note that

Students	Items															Correct
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	1	1	0	1	1	1	0	0	1	1	1	0	0	0	1	9
2	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	7
3	1	1	1	1	0	1	0	0	1	1	1	0	0	0	0	8
4	0	1	1	1	0	1	0	0	1	1	0	0	0	0	0	6
49	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	14
50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15

Fig. 1. The model matrix of test results.

Knowledge level	Correctly answered to j test items															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0,10	0	0	0	2	0	1	0	0	1	1	0	0	0	0	0	0
0,20	0	0	0	0	0	0	1	1	2	0	1	0	0	0	0	0
0,30	0	0	0	0	0	0	0	0	0	3	1	1	0	0	0	0
0,40	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0
0,50	0	0	0	0	0	0	0	0	0	0	2	1	0	1	0	1
0,60	0	0	0	0	0	0	0	0	0	0	3	0	0	2	0	0
0,70	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	1
0,80	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	0
0,90	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2
1,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

Fig. 2. The matrix of processed results.

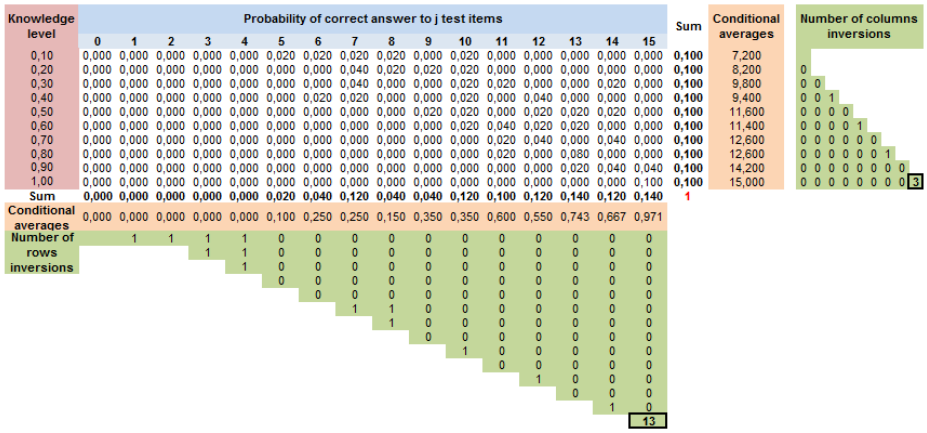

Fig. 3. The test results probability matrix with numbers of inversions.

$\sum_{i=1}^{10}\sum_{j=0}^{15} f_{ij} = m$. Dividing each element of the frequency results matrix by the number of tested students m we get the probability results matrix (see Fig. 3):

$$F = (p_{ij})_{10\times(n+1)},$$

where $p_{ij} = P(X = j, p = p_i)$ – is the probability of test takers having knowledge level $i = 1, 2, \ldots, 10$ ($p_i = 0.1, 0.2, \ldots, 1$) and correctly responded to $j = 0, 1, \ldots, 15$ test items.

Then conditional averages [1, 5] were calculated for each row and column of probabilities matrix:

$$m_j = \frac{\sum_{i=1}^{10} i \cdot p_{ij}}{\sum_{i=1}^{10} p_{ij}}, \qquad n_i = \frac{\sum_{j=1}^{15} j \cdot p_{ij}}{\sum_{j=1}^{15} p_{ij}}, \quad i = 1, 2, \ldots, 10,\ j = 0, 1, \ldots, 15,$$

m_j is the average knowledge level of students correctly responded to j test items. n_i is the average amount of correct answers for students having knowledge level i. It is obvious that the student with the higher knowledge level at an average get more test points (the number of correct responses). Then rows and columns inversions were calculated. For example, number of column inversions is the number of pairs (i_1, i_2) with $i_1 < i_2$ and $n_{i_1} \geqslant n_{i_2}$.

3 The results of experiment

Therefore 2 series of experiments described in previous section were conducted. In the first experiment probability distribution of population knowledge level was considered to be uniform and in the second – Gaussian (normal). During each experiment 100 random matrices of items responses were generated. Each of experiments was repeated for 5 times. One of the resulting probabilities matrices is shown in Fig. 3. Numbers of columns inversions are listed in the tables in Fig. 4. In the first table of Fig. 4 knowledge level in the population has uniform distribution (there are chosen

Columns inversions number (uniform distribution)	Experiment				
	1	2	3	4	5
	Frequency				
0	2	2	4	3	1
1	26	24	19	20	21
2	32	38	28	36	39
3	28	23	32	27	25
4	8	11	14	14	13
5	3	2	3	0	1
6	1	0	0	0	0

Columns inversions number (normal distribution)	Experiment				
	1	2	3	4	5
	Frequency				
0	8	6	12	6	6
1	19	30	27	24	32
2	38	30	35	45	38
3	23	27	21	15	16
4	12	6	5	10	8
5	0	1	0	0	0

Fig. 4. Numbers of columns inversions when knowledge level in the population has uniform and normal distributions.

Knowledge level	Number of students
0,10	1
0,20	3
0,30	5
0,40	7
0,50	9
0,60	9
0,70	7
0,80	5
0,90	3
1,00	1

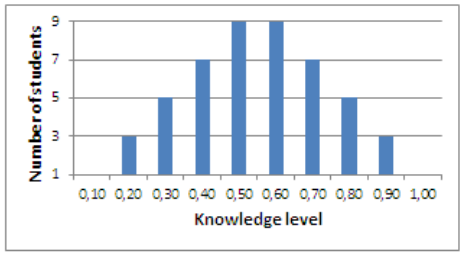

Fig. 5. Normally distributed data.

50 students, each 5 students have the same knowledge level), in the second – normal distribution (Fig. 5).

Statistical hypotheses about probability distribution of numbers of column inversions were tested. According to Kolmogorov–Smirnov goodness of fit criteria for all 5 cases of uniform distribution and 5 cases of Gaussian distribution of population knowledge level hypotheses of Poisson distribution of the number of columns inversions weren't rejected. Point estimates of parameter λ values (the average rate of column inversions occurrences) and p-values of two tailed goodness of fit criteria are listed in Table 1.

At an average for uniform distribution $\hat{\lambda} = 2.303$, for Gaussian distribution $\hat{\lambda} = 1.958$. As the number of column inversions has Poisson distribution we can calculate values of cumulative distribution function which are presented in Table 2.

So, number of column inversions for uniform distribution of population knowledge level in the case of $m = 50$, $n = 15$ (items are described in Section 2) couldn't

Table 1. Point estimates of average rate of column inversions occurrences for uniform and Gaussian distributions of knowledge level and p-values of goodness of fit criteria.

Uniform distribution			Gaussian distribution		
Experiment	$\hat{\lambda}$	p-value	Experiment	$\hat{\lambda}$	p-value
1	2.27	0.491	1	2.12	0.225
2	2.23	0.428	2	2	0.622
3	2.42	0.642	3	1.8	0.664
4	2.29	0.238	4	1.99	0.188
5	2.307	0.189	5	1.88	0.358

Table 2. Cumulative distribution function for number of column inversions occurrences (X) for uniform and Gaussian distributions of knowledge level.

Uniform distribution		Gaussian distribution	
x	$P(X \leqslant x)$	x	$P(X \leqslant x)$
3	0.796705	2	0.877311
4	0.928189	3	0.965357
5	0.979074	4	0.991991

exceed 5 and for Gaussian distribution couldn't exceed 4, if we choose the significance level $\alpha = 0.05$. At the same manner criteria for the critical values of numbers of row inversions could be constructed.

4 Conclusions

The purpose of this paper – to show the methodology for construction of test validity criteria. In this study the simulated test results were presented when having a priori information about: (1) the difficulty of the test items which are power functions, (2) the number of test takers, (3) the number of test items, (4) probability distribution of knowledge level in the population. If the number of experiments is sufficient, it is possible to construct probability distributions of the number of inversions. This allows evaluating the suitability (validity) of the test for assessment of student's knowledge levels. In practice, this would mean that if we have information about knowledge level distribution in the population and for any test we can calculate the numbers of inversions of the resulting test matrix, it is possible to compare distribution of number of inversions with the referenced distributions obtained by Monte Carlo experiments. If the number of inversions didn't exceed referenced values we can say with a certain confidence level that the test is reliable (valid). Note that the reference number of inversion's distributions depends on the above mentioned information and practical application of criteria requires a number of additional studies.

References

[1] W. Barth, P. Mutzel and M. Jünger. Simple and efficient bilayer cross counting. *J. Graph Algorithms Appl.*, **8**(2):179–194, 2004.

[2] A. Krylovas ir N. Kosareva. Mathematical modelling of forecasting the results of knowledge testing. *Techn. Econ. Devel. Economy*, **14**(3):388–401, 2008.

[3] A. Krylovas ir N. Kosareva. Žinių tikrinimo matematinis modelis. *Liet. mat. rink. LMD darbai*, **48**(49):217–221, 2008.

[4] N. Kosareva and A. Krylovas. A numerical experiment on mathematical model of forecasting the results of knowledge testing. *Techn. Econ. Devel. Economy*, **17**(1):42–61, 2011.

[5] A. Krylovas. *Diskrečioji matematika*. Technika, Vilnius, 2009.

[6] O. Navickienė and A. Krylovas. Studentų žinių vertinimo kokybės kriterijų modeliavimas. *Studijos šiuolaikinėje visuomenėje/Studies in Modern Society*, **3**(1):177–184, 2012.

[7] G. Rasch. *Probabilistic Models for some intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen. Expanded edition, The University of Chicago Press, 1980.

REZIUMĖ

Testų patikimumo statistinių kriterijų konstravimas kompiuterine simuliacija
O. Navickienė, A. Krylovas, N. Kosareva

Straipsnyje aprašomas kompiuterinis žinių tikrinimo testo modeliavimas, kai klausimų charakteringosios funkcijos turi laipsninį pavidalą, o žinių lygis populiacijoje turi tolygųjį arba normalųjį skirstinį. Apskaičiuojamos sąlyginių vidurkių inversijos ir modeliuojami inversijų skaičių tikimybiniai skirstiniai. Praktiškai konstruojant žinių tikrinimo testus, jų tinkamumui (validumui) tikrinti stebimos inversijų reikšmės lyginamos su etaloninėmis reikšmėmis, gautomis Monte Carlo eksperimentais.

Raktiniai žodžiai: žinių tikrinimo testai, Monte Carlo eksperimentas, matematinis modeliavimas, statistiniai kriterijai.