Reliability of Decision in Testing Problems

Mikhail M. Lutsenko
Emperor Alexander I St. Petersburg State Transport University
Saint Petersburg, Russia
ml4116@mail.ru

Dzhemil A. Seytmanbitov
Emperor Alexander I St. Petersburg State Transport University
Saint Petersburg, Russia
dzhem93@gmail.com

Anatoly M. Baranovskiy
Emperor Alexander I St. Petersburg State Transport University
Saint Petersburg, Russia
bamvka@mail.ru

Abstract

In this paper, a statistical game was defined and solved. Its solution is: the optimal randomized decision rule, the probability of a correct decision on this rule, and the worst a priori distribution of the test subjects knowledge levels. We have developed a method for assessment the accuracy and reliability of decision making by on test results. The proposed program allows you to assessment the reliability of the solution for a test containing 10 items with different levels of difficulty, and 11 different levels of knowledge level.

Introduction

The main purpose of any testing is to assessment of test-takers knowledge level and make a decision by the result. Unfortunately, the test result (the number of completed test items) depends not only on the test-takers knowledge levels, but also on many other factors that are hardly predicted. So, an adequate model of the decision making problem must include probabilistic components. We need a flexible model for building optimal randomized solutions that takes into account various types of solutions, a priori distributions at different knowledge levels, and different item difficulties. This model can be executed in the scope of statistical game theory [Lin97].

1 Methods and Algorithms

Let's $\Theta = \{\theta_1, \theta_2, \dots, \theta_m\}$ be the set of possible test-takers knowledge levels, $X(\theta)$ – random variable (the number of points tested with a knowledge level θ when take test T). Let's $D = \{d_1, d_2, \dots, d_n\}$ – be the set of possible decisions that the decision maker can make by the test results.

Examples of such solutions are the following sets of solutions.

Accurate assessment of the test-taker knowledge level:

$$d_1 = \theta_1, d_2 = \theta_2, \dots, d_n = \theta_n,$$

(The solution d_j is that the test-taker knowledge level is θ_j).

Interval assessment of the test-taker knowledge level:

$$d_1 = \Delta_1, d_2 = \Delta_2, \dots, d_m = \Delta_m,$$

where $\Delta_1, \Delta_2, \dots, \Delta_m \subseteq \Theta$ is a set of partially intersecting intervals. Their elements can be interpreted as poorly prepared test-takers, satisfactorily prepared, etc., or as test-takers ready to execute task 1, task 2, etc. (Solution d_j is that the interval Δ_j include the test-taker knowledge level θ).

Let's $h(d, \theta)$ the benefit of the decision maker when it made the decision d, and the knowledge level was θ. Using the benefit function, you can modeling the many other sets of solutions.

For example, the decision maker benefit function

$$h(d, \theta) = \begin{cases} 1, \text{if } \theta \in \Delta(d) \\ 0, \text{otherwise} \end{cases}$$

is built on any set of confidence intervals $\Delta_1, \Delta_2, \dots, \Delta_m \subseteq \Theta$ [Lut03].

Let's $X = \{1,2, \dots, N\}$ the set of possible values of a random variable $X(\theta)$, and, $P_\theta(x) = P(X(\theta) = x)$ the probability of the corresponding event. These probabilities can be calculated as probability for Bernoulli trials$P_\theta(x) = C_N^x p^x (1-p)^{N-x}$, if the difficulties of all test items are the same. In the case when the difficulties of the items are different, we will define a random variable $X_k(\theta)$. Its value is set to one if the test-taker with the knowledge level θ completed the j-th item of the test and zero otherwise. Let's $p(k, \theta) = P(X_k(\theta) = 1)$ the probability of the corresponding event. Then the random variable $X(\theta)$ is equal to the sum of the corresponding random variables:

$$X(\theta) = X_1(\theta) + X_2(\theta) + \cdots + X_N(\theta)$$

In this case, the probability $P_\theta(x)$ is calculated using the known formulas [Ney00].
Let's δ is a function that gives each observed point x a solution from the set D, that is $\delta: X \to D$.
We denote the set of all solving functions by $D = D^X$.

Let's

$$H(\delta, \theta) = \sum_{k=1}^{N} P_\theta(x_k) h(\theta, \delta(x_k)) \qquad (1)$$

is the expected value of the decision maker benefit if it uses the solving function δ, and the test-takers knowledge level is θ. If the function $h(d, \theta)$ is defined through a set of confidence intervals, then the function $H(\delta, \theta)$ is equal to the probability that the accurate value of the test-taker knowledge level is in the required confidence interval. This interval is built on the observed point x according to the solving function δ.
Note that the lowest probability that a set of $\Delta_1, \Delta_2, \dots, \Delta_m$, generated by the solving function δ, will include the unknown parameter θ, is called the confidence probability for this set (for this solving function), that is

$$\gamma = \gamma(\delta) = \min_{\theta \in \Theta} P(\theta \in \Delta(\delta(X_\theta))).$$

If the a priori distribution of knowledge levels ν is known, then the best solving function δ_ν can be builded according to this distribution

$$H(\delta_\nu, \nu) = \max_{\delta} H(\delta, \nu),$$

this function is called the Bayesian solving function [Lut00].
If the a priori distribution is unknown, then the best solving function should be found from the solution of the statistical game $\Gamma = \langle D, \Theta, H \rangle$. Where D – is the set of solving functions (the set of decision maker strategies), Θ – is the set of possible test-taker knowledge levels (the set of condition of nature), and the decision maker benefit function in a statistical game whose values are found by the formula (1).
To solve the matrix game, let's make a pair of mutually dual problems. From the first problem, we find: the best randomized solving function $\mu = (\mu_1, \mu_2, \dots, \mu_N)$, from the second, the worst a priori distribution ν, and the total value of these games is the value of the game Γ.

Direct problem:

$$\nu \to \max,$$

$$\sum_{k=1}^{N} \Lambda^k B \mu_k \geq \nu 1_m$$

$$\sum_{j=1}^{n} \mu_k^j = 1; \qquad \mu_k^j \geq 0; \qquad k = \overline{1, N}; \qquad j = \overline{1, n}.$$

Dual problem:

$$\nu = \sum_{k=1}^{N} u_k \to \min,$$

$$\nu^t \Lambda^k B \leq u_k 1_n^t; \qquad k = \overline{1, N}; \qquad \sum_{i=1}^{m} \nu_i = 1.$$

There are many ways to solve linear programming problems. The most appropriate method here would be the dynamic method [Lut90], specially developed by the author for statistical games with threshold benefit functions. However, in the simplest cases, the statistical game can be solved using MS Excel. Although these methods often do not provide an exact solution, they always indicate valid solutions to problems and, consequently, the upper and lower bounds of the matrix game.

2 Approbation

Let's assume that the test includes 10 questions, and the Statistician makes a decision by the results of this test. The set of observations X includes 11 numbers: from zero to 10. The probability of a correct answer to one test question is equal to the test-taker knowledge level. The possible values of test-taker knowledge level are set Θ = {0,95; 0,85; 0,75; 0,65; 0,55; 0,45; 0,35; 0,25; 0,15; 0,05}.
Then the probability of correctly answering x test items is calculated as probability for Bernoulli trials:

$$P_\theta(x) = C_{10}^x \cdot \theta^x \cdot (1 - \theta)^{10-x}, \quad x = \overline{0; 10}.$$

Statistics assess the knowledge level in the subject. It puts one of the following four grades: D = {A, B, C, D}. An excellent grade is given to test-takers with 95% and 85% knowledge, good from 75% to 55%, satisfactory from 45% to 35%, and unsatisfactory to the rest.
Let's make a statistical game $\Gamma = \langle D, \Theta, H \rangle$ and solve it in mixed strategies. The benefit matrix in the game has a size of 44×10. Unfortunately, MS Excel tools do not allow you to accurately solve two mutually dual problems. But we get upper and lower assessment of the game value, a randomized decision function, and the worst a priori distribution of the parameter θ [Lut11].
As a result, we get the lower (0.519) and upper (0.562) assessments of the game value.

Table 1: Randomized decision function μ.

		μ_{10}	μ_9	μ_8	μ_7	μ_6	μ_5	μ_4	μ_3	μ_2	μ_1	μ_0
decision	*A*	1,00	0,49	0,75	0	0	0	0	0	0	0	0
	B	0	0,51	0,24	0,95	0,75	0,70	0	0	0	0	0
	C	0	0	0	0,05	0,25	0,30	1,00	1,00	0	0	0
	D	0	0	0	0	0	0	0	0	1,00	1,00	1,00

The columns in this table indicate the probabilities with

which the Statistician indicates a particular solution depending on the observation.
So, the probability of correct decision of the statistics about the test-taker knowledge level by the test results is in the range from 0.52 to 0.56. Thus, in about 50% of cases, the Statistician will make an incorrect decision about the test-taker knowledge level [Sha13].

Table 2: Worst a priori distribution of the parameter θ.

θ_i	0,95	0,85	0,75	0,65	0,55	0,45	0,35	0,25	0,15	0,05
ν_i	0	0,11	0,01	0,04	0,25	0,19	0,15	0,26	0	0

The resulting game values are a lower assessment and can be improved with a known a priori distribution. In addition, it seems unlikely that the a priori distribution of knowledge levels coincides with the worst a priori distribution [Sha14].
Although the above statement does not take into account all the features of the testing organization, it can be clarified if necessary. However, the value of the game will not improve much if you enter more items into the test. Similar examples are considered in [Lut14].

3 Rasch model

The modern method of assessing the test-takers knowledge level is based on the Item Response Theory (IRT) [Lin97]. Let's enumeration the main assumptions of this theory.

- Each test-takers has a certain knowledge level θ from the set of possible (acceptable) levels $\Theta \subseteq \mathbb{R}$.
- Each item of the test τ is assigned a characteristic function of the satisfiability of this item $p_\tau(\theta)$. Its value is the probability of the item completed by the test-taker with the knowledge level θ. It is obvious that $0 \leq p_\tau(\theta) \leq 1$ when $\theta \in \Theta$.
- The assessment of the test-taker knowledge level is based on the result of performing N items $\tau_1, \tau_2, .., \tau_N$, the characteristic functions $p_{\tau_1}(\theta), p_{\tau_2}(\theta), \ldots, p_{\tau_N}(\theta)$.
- Difficulty of the item τ, and the knowledge level of the test θ can be measured in the same units, so the difference $\tau - \theta$ shows the extent of exceeding the difficulty of the item over the test-taker knowledge level [Lut15].

In the Item Response Theory it is assumed that the probability of correctly take an item of difficulty τ by a test-taker with knowledge level θ is equal to

$$p_\tau(\theta) = p(\theta - \tau) = (1 + exp\,(-(\theta - \tau)))^{-1} \text{ (Rasch model).}$$

We now turn to the general case of parameter assessment in the rush model. Suppose that n test-takers take a test T containing N items of difficulty: $\tau_1 < \tau_2 < \ldots < \tau_N$. Then the probability that the i-th test-takers performed j-th item of the test is equal to

$$p_{i,j} = (1 + exp\,(\theta_i - \tau_j))^{-1}, \tau_j \in \mathbb{R},$$

Let's c_j – is the number of participants who correctly performed the item with the number j (the number of initial points j-th item); b_i – is the number of correctly completed items participant number i. (As a rule, these are all integers from 0 to N inclusive). Assessment $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_N$; $\hat{\tau}_0, \hat{\tau}_1, \ldots, \hat{\tau}_N$ of the corresponding parameters can be obtained by the method of moments or by the method of greatest likelihood. To do this, need to solve a system of equations.

$$\begin{cases} \sum_{j=1}^{N} p_{i,j} = b_i, i = \overline{1,n}; \\ \sum_{i=1}^{n} p_{i,j} = c_j, j = \overline{1,N}. \end{cases} \quad (2)$$

The possible values of the right parts of this system (numbers b_i) are integers from 0 to N. So system (2) consists of $2N+1$ equations and contains $2N+1$ unknowns.

Conclusion

In this paper, the problem of calculating the reliability of decisions made based on the results of testing was set and solved. The solution of the statistical games found: the optimal randomized decision rule (the best assessment of the test-taker knowledge level), the probability of a correct decision on this rule, worst the a priori distribution of the levels of knowledge tested. The advantage of this approach is that we do not impose any restrictions on the distribution of test-takers types and that the solution of these statistical games is obtained by standard methods. In addition, the resulting solution is quite resistant to small changes in the problem conditions.

Reference

[Lin97] van der Linden, Win. J., R.K. Hambleton, Handbook of Modern Item Response Theory. Edition. 1997, Springer – Verlag, New York, P.510.

[Lut90] Lutsenko M.M. Game theoretic method for assessment the parameter of the binomial distribution, Probability theory and its applications. 1990, №3. Pp. 471-481.

[Lut00] Lutsenko M.M., Ivanov M.A. Minimax confidence intervals for the parameter of a hypergeometric distribution, Automation and remote control. 2000, №7. Pp. 1125-1132.

[Lut03] Lutsenko M.M., Maloshevskii S.G. Minimax confidence intervals for the binomial parameter, Journal of statistical planning and inference. 2003, №1. Pp. 67-77.

[Lut11] Lutsenko M.M., Shadrinceva N.V. Educational Testing Accuracy, News of St. Petersburg State

Transport University. 2011, №4(29). Pp. 250-258.

[Lut14]. Lutsenko M.M., Seytmanbitov D.A. Test explicitly in Rasch model, Proceedings of the international banking institute. 2014. Pp. 114-116.

[Lut15] Lutsenko M., Seytmanbitov D., Game-theory Method for Knowledge Assessment, SING11-GTM2015 European Meeting on Game Theory. 2015. Pp. 125-126.

[Ney00] Neyman Yu.M., Khlebnikov V.A.: Introduction to the theory of modeling and parameterization of pedagogical tests. 2000. 168 p.

[Sha13] Shadrinceva N.V., Seytmanbitov D.A. About the reliability of testing in the rush model, Mathematical modeling in education, science, and manufacturing. 2013. Pp. 156-157.

[Sha14] Shadrinceva N.V., Seytmanbitov D.A. Reliability of testing in the rush model, Institute of information technology and management SPBSTU. 2014.