Description of a Natural Experiment for Collecting Students' E-test Data

Zlatan Iliev, Todor Todorov, Adriana Borodzhieva, Irina Zheliazkova

Description of a Natural Experiment for Collecting Students' E-Test Data: *The paper describes a natural experiment for collecting students' e-test data: correct, missing, and wrong knowledge, performance time and mark measured within an intelligent and adaptive e-testing environment. The description itself includes: participants, materials, and applied statistical methods.*

Key words: *Experiment Description, Collecting, E-testing Environment, Students' Data.*

INTRODUCTION

Some educational researchers [4] envisage that consequently e-testing will become the main teaching activity at universities and the lecture information could be self learned from other sources, for example textbooks, television, and *INTERNET*. So the need of effective and efficient e-testing environments, as well as good authors and instructors for their support and test results analysis will be increasing. In the classical multiple choice test theory a lot of metrics for the students' test assessment exist and recently the wide usage of e-testing environments lead to new metrics. More often the experimental studies concern tests with multiple choice questions covering a single lecture topic or all lecture topics. Usually they are collected by means of a subject pretest and posttest for experimental and control groups.

Many researchers in the pedagogy diagnostics argue that most of the educational data concerning the tested students, test questions, and lecture topics can be treated as statistical variables. The data more often analyzed are the student's mark, scores, errors, performance time, etc. The dispersion, correlation, and regression analysis are more often applied to study the impact of different factors on a given dependent variable, correlation among them [15], and checking hypotheses about their distribution. The interpretation, computation, table and/or graphical visualization of such types of analysis are implemented within the learning or testing environments [3] or using some commercial software products such as *EXCEL*, *MATLAB*, and *STATGRAPHICS*.

The problem of the test results processing includes solving some subproblems, more important of which are: representation of different types of questions; measuring the student's knowledge and performance time, support the activities of the test author, development of an appropriate assessment scale; verification of the test validity and reliability, etc. [4, 7, 8, 9, 12, 14]. Another subproblem is verification of hypotheses, e.g. assumptions related to the measurements of different statistical variables. If the testing technology and organization are fixed, and the number of measurements is large, the theoretical normal distribution (known also as Gaussian) can be applied to describe the experimental data set.

In some previous papers of Zheliazkova's group [16, 17] a computer-based technology for statistical processing and visualization of both test and exercise data sets covering a lecture topic can be found. Different types of statistical analysis, such as dispersion, profile, correlation, regression and clustering, are applied on different data sets, measured by means of an intelligent and adaptive e-testing environment, put into practice more than a decade ago.

The present study continues the group's research for efforts focusing on verification of some hypotheses about normal distribution of a new experimental data set. The natural experiment description is presented further as follows: students-participants, used materials, and applied statistical methods. The conclusion summarizes the distinctive properties of the experiment in comparison with previous ones.

EXPERIMENT ORGANIZATION

The participants involved in the knowledge testing were 57 bachelor degree students from the specialty "Computer Systems and Technologies" at the University of Ruse. The test session was carried out within the framework of the course "Discrete Structures and Modeling", taught in the second semester of the academic 2012/2013 year. The test covered 30 hours taught lecture material and was created as an intelligent posttest. In order to measure the Correct Knowledge (*CK*), Missing Knowledge (*MK*), Wrong Knowledge (*WK*), performance *Time*, and student's *Mark*, an intelligent and adaptive e-testing environment was put into the teacher's team practice for five subjects. Three computer laboratories, each one with 15 computers, i.e. in total 45 students, were engaged in this natural experiment during the first pass. The other 12 students were tested at the second pass in one computer hall. The participants used this environment for the first time. As its interface is very intuitive for these students, short 5-10 minutes instructions were given to them before starting their registration in the environment.

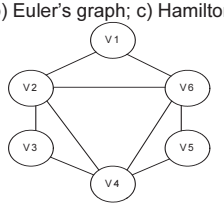
Depending on the type, each question brought a different number of scores p_{max}. The student's answer was reduced to: filling in an empty edit field, copying and pasting keywords from the test dictionary embedded in the environment. The special symbols ";" and ">" were used as separators of non-ordered and ordered subanswers respectively. Using "No" in an answer or subanswer was recommended in order to make difference between *MK* and *WK*. The student's final *Mark* was computed as a real number in the range from 2.00 to 6.00 depending only on the student's test *CK* scores in a traditional for Bulgarian assessment scale: $0 \le CK \le 0.4*P_{max}$ – "Poor (2)"; $0.4*P_{max} < CK \le 0.55*P_{max}$ – "Satisfactory (3)"; $0.55*P_{max} < CK \le 0.70*P_{max}$ – "Good (4)"; $0.70*P_{max} < CK \le 0.85*P_{max}$ – "Very good (5)"; $0.85*P_{max} < CK \le 1.0*P_{max}$ – "Excellent (6)". The experience accumulated during the last decade by Zheliazkova's research group has pointed out that such a non-linear scale leads to marks, acceptable by both teachers and students. An algorithm for computation of the continuous *Mark* is embedded with accuracy of two digits after the decimal point. One of the advantages of the e-testing environment is that it registers the test performance by means of the system time in seconds. That allows the teacher if he/she wants to use performance *Time* as additional variable for the student's assessment.

USED MATERIALS

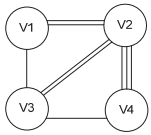
The number of questions included in the test was 27 with total scores $P_{max} = 241$ and time planned for the test performance is $T_{max} = 120$ min. The assessment scale also calculated automatically for the given test mark was in the following ranges: 0 ÷ 96 –"Poor (2)"; 97 ÷ 132 – "Satisfactory (3)"; 133 ÷ 168 – "Good (4)"; 169 ÷ 204 – "Very good (5)" ; 205 ÷ 241 – "Excellent (6)". The students were told that performance Time would be actually unlimited and that this parameter and also *MK* and *WK* would be used as assessment factors only for research purpose. The parameter *CK* is defined as a part of the scored student's answer that coincides with the teacher's one and the parameter *MK* as a part of the scored teacher's answer that is missing in the student's one. Actually from all 9 "question-answer" types in this intelligent test, only five types, namely: unordered pairs, multiple choice, ordered keywords, numbered keywords, and numbers, were used (Table 1). Here the question parameters automatically calculated have the following meanings: W – the weight of a subanswer; Q_t – maximum question scores; C_p – the degree of the system's prompt. Their values from the teacher's and student's sides are based on the ontology of different types of question answers. It is clearly seen that although intelligent, the test questions are related to the lower cognitive levels (understanding, memorizing, comparison, contradiction), e.g. in concordance with Bloom's taxonomy [1]. Due to the pages limitation of the conference papers, the raw data set with students' test results are given in the next paper [5].

Table 1 – Example types of intelligent questions

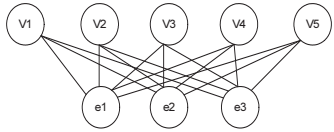
***QUESTION 1:* Unordered pairs**
Point out the correspondence between the graphs in figures 1), 2), 3) and their types: a) multigraph; b) Euler's graph; c) Hamilton's graph; d) Kenning's graph; e) full graph.



Question figure1



Question figure2

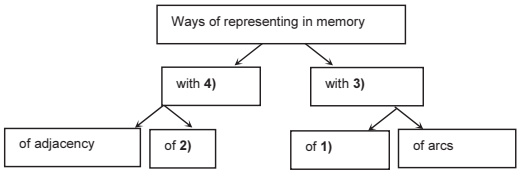


Question figure3

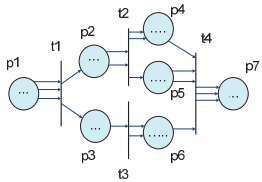
Answer: 1>b;1>c;2>a;3>d;3>e
Parameters: W=3; Qt=15; Cp=0.66

***QUESTION 2:* Formula**
Write down the formula for calculation of the graph cyclamate number (g), depending on the number of its nodes (n) and the number of its connections (m).
Answer: g=m-n+1
Parameters: W=2; Qt=14; Cp=0.50

***QUESTION 3:* Multiple choice**
Point out the type of the tree in figure4: a) root tree; b) oriented tree; c) binary; d) ordered tree.



Question figure4



Question figure5

Answer: a;b;c
Parameters: W=1; Qt=3; Cp=0.70

***QUESTION 4:* Ordered keywords**
Enter the missing words in figure4 in accordance with their numbers.
Answer: nodes>incidence>lists>matrix
Параметри: W=2; Qt=8; Cp=0.70

***QUESTION 5:* Numbered words**
Order the numbered actions in the right sequence: 1) Enter a recognized pattern, 2) Determination of the pattern's attributes, 3) Choosing the more informative attributes, 4) Learning or self-learning, 5) Pattern recognition.
Answer: 2>3>4>1>5
Parameters: W=2; Qt=10; Cp=0.50

***QUESTION 6:* Number**
Enter the vector of the new marking of Petri net in figure5 after transition t4 activation.
Answer: 333246
Parameters: W=1; Qt=6; Cp=0.00

APPLIED METHODS

In accordance of Bloom's taxonomy [1] the subproblems of the test results processing are related to the higher cognitive levels (analysis, synthesis, and evaluation). In this section the applied statistical methods with their properties are reminded.

Dispersion analysis: A given statistical variable can be presented as $X = x_1, x_2, \ldots, x_i, \ldots, x_n$, where n is the number of observations/measurements. This method is fundamental for other more complex statistical methods for analysis [2, 10, 11, 13] including verification of a hypothesis about a given distribution. The parameters of the dispersion analysis (also called descriptive statistics) have a simple calculation. More

precisely, the formulas of its parameters are: mean $\bar{x} = \left(\sum_{i=1}^{n} x_i \right) / n$; dispersion $D(X) = \left(\sum_{i=1}^{n} |x_i - \bar{x}|^2 p_i \right) / \sum_{i=1}^{n} p_i$ and standard deviation $\sigma(X) = \sqrt{D(X)}$, where n is the number of the students, xi is the value of the parameter for the ith student; pi is the additionally calculated probability of this parameter. The value of $\sigma(X)$ is larger if the scatter of the experimental data set *X* is bigger.

Normal distribution: Up to now about 30 continuous and discrete distributions are known from the applied mathematical statistics [2, 10, 11, 13]. The most popular and with wide use is the normal distribution. Its probability density function $f(x;\bar{x};\sigma)$ presents a theoretical continuous function defined as $f(x;\bar{x};\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$. The concrete shape of $f(x;\bar{x};\sigma)$ (the well-known "bell" curve) this function depends on the concrete values of the dispersion analysis and the maximum of $p(x;a;\sigma)$ will occur at $\bar{x} = a$, where a is some central tendency (Fig. 1). The geometrical interpretation of the probability density function predicts that the variable's scatter will be distributed symmetrically around a. The most expected value from any single measurement, i.e. the most probable value, would be *a* > 0 and can be more than 1. Under the assumption that the sum of *X* values gives the probability of 1.00, 68.27% of the values lie within 1.σ interval, 95.45% within 2.σ interval; and 99.73% within 3.σ interval (Fig. 2). If the distribution is different from the normal one it means that some factors concerning the test itself, e-testing environment and/or e-testing technology had not been taken into consideration.

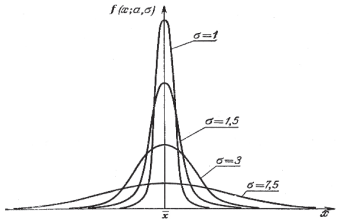


Fig. 1 – Normal distribution curves with σ = 1; 1,5; 3; 7,5

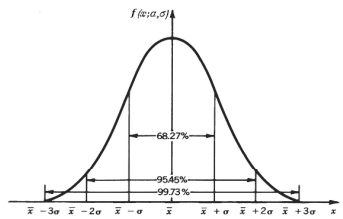


Fig. 2 – Three intervals with the percentage of the values

Pearson criterion: From the applied mathematical statistics [10, 11] a lot of criteria (Fisher's, Kolmogorov-Smirnov's, Anderson-Darling's, Bernoulli's, and so on) were found, each one with its own properties and applications. In this experimental study Pearson's criterion (also known as χ^2 criterion) was used to check if the null hypothesis H_0 "the experimental data set *X* is in concordance with the normal distribution". Recently it is the most widely used among other known chi-squared tests (Yates, likelihood ratio, portmanteau test in time series, etc.). The results of these statistical criteria are evaluated by reference to the chi-squared distribution. χ^2 criterion is suitable for unpaired data from large samples. For all students' data sets this criteria tests H_0 that their frequency distribution of the experimental sample (also called histogram) is consistent with a given theoretical distribution.

The value of Pearson's criterion is calculated as: $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, where: χ^2 –

Pearson's criterion, which asymptotically approaches a χ^2 distribution; O_i – an observed frequency; E_i – an expected (theoretical) frequency, asserted by the null hypothesis. Chi-squared distribution, shows χ^2 on the *x*-axis and *p*-value on the *y*-axis (Fig. 3). The chi-squared statistics can then be used to calculate a probability *p*-value by comparing the value of the statistic to a chi-squared distribution. The number of degree of freedom (*k*) is equal to the number of the students *n*, minus the reduction in degrees of freedom. The result about the numbers of *k* is valid when the original data are multinomial and hence the estimated parameters are efficient for minimizing the chi-squared statistic. More generally however, when maximum likelihood estimation does not coincide with minimum Chi-squared estimation, the distribution will lie somewhere between a chi-squared distribution the degree of freedom with $k = (n - 1 - p)$ and $k = (n - 1)$. As the value of χ^2 criterion is smaller so the histogram of the experimental data set is closer to the normal distribution.

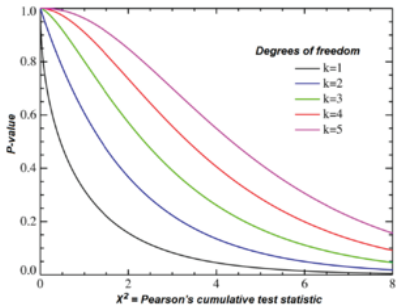


Fig. 3 – Family of curves P-value ( ) for k = 1, 2, 3, 4, 5

Representative sample size: Another frequently asked question concerning an experimental sample is "Is this sample representative enough?" The answer of this question is influenced by three main factors, e. g. the level of precision, level of confidence or risk, and degree of variability. In [6] tables and four formulas, namely for mean, proportions, correction for proportions, and simplified for proportion, were found for calculation of the representative sample size. The simplified formula for proportions proposed by Yamane [6] was chosen for this study, i.e. $n = \frac{N}{1 + N(e)^2}$, where *n* is the representative sample size; *N* is the experimental data size; e is the desired level of precision. If $N > n$ then the null hypothesis H_0 is accepted, else it is rejected, e.g. the alternative hypothesis H_1 is accepted.

CONCLUSION

The described experiment is natural, i.e. without using experimental and control groups. It was carried out as posttest, covering all lecture topics of a subject. The test includes five types of questions with three parameters: the subanswer's weight, maximal scores, and degree of prompt automatically calculated on the base of the question answer ontology. The used non-commercial intelligent and adaptive e-testing environment measures student's correct, missing, and wrong knowledge, time of the test performance, and student's mark too. Recently the assessment scale also automatically calculated takes into account only one factor, the student's correct knowledge. Nowadays the author's team is intended to embed several multi-factor models for the student's Mark, including also missing knowledge, wrong knowledge, and performance time or different combinations of them in order adapt the environment to the teacher's preferences.

REFERENCES

[1] Bloom, B. S.Taxonomy of Educational Objectives, Vol. 1, New York: McKay, 1956.

[2] Figliola, R., D. Beasley. Theory and Design for Mechanical Measurements, 1991.

[3] Gnatienko, G., Software for Expert Information Processing During Carrying Out the Examinations, International Journal "Information Technologies and Knowledge", 2007, Vol. 1, Number 3, pp. 272 – 278.

[4] Hristova, M. Algorithmization of a Multi-factor Model for Assessment of the Quality of Teaching in Higher Schools, Journal of Automatics and Informatics, 2008, Vol. 2, pp. 37 – 40 (in Bulgarian).

[5] Iliev, Zl., T. Todorov, A. Borodzheva, I. Zheliazkova. Analysis of the Students' Test Data for Verification of Some Hypotheses, Annual Conference, University of Ruse, 2014.

[6] Izrael G. D. Determining Sample Size, http://edis.ifas.ufl.edu/pdffiles/pd/pd00600.pdf

[7] Jelev, G., D. Minkovska. Approaches for Definition the Validity of the Results of the Test for Knowledge Mastering, Proceedings of the International Conference "Computer Science', 2004, Sofia, pp. 268 – 273.

[8] Jelev, G., Y. Minkova. Determination of Representative Sample Size and Knowledge Assimilation Test Results Processing. Problems & Discussion, Proceedings of the International Conference "Computer Science', 2004, Sofia, pp. 274 – 279.

[9] Karasev, V.A., S. S. Malomuzh, M. Yu. Sternin. The Conceptual Model of an Intelligent System for Training the Users of Lazer Technological Complexes, Proceedings of the ISA RAS, 2005, pp. 131 – 145 (in Russian)..

[10] Kobzar, A.I. Application Mathematical Statistics for Engineers and Scientists, Moscow, "FIZMATLIT", 2012 (in Russian).

[11] Kostova, V. Theory of Probability and Mathematical Statistics, Rousse University, 2005 (in Bulgarian).

[12] Radimova, D. System "Testing Students' Knowledge", Proceedings of the TAAC Conference, Kiev, 2012, pp. 76 – 78 (in Russian)..

[13] Smirnov, N., I. Dunin-Barkovskii. Short Course on Mathematical Statistics for Technical Applications, Moscow, "Mir", 1959 (in Russian).

[14] Sokolova, M., G. Totkov. About Test Classification in E-Learning Environment, Proceedings of the International Conference CompSysTech'05, 2005, pp. II.21-1 – II.21-6.

[15] Vasilev, Yu. Extracting of Knowledge (Data Mining) from a Database of Tests, Journal "Automatics and Informatics", Number 2, 2011, pp. 27 – 30 (in Bulgarian).

[16] Zheliazkova, I.I., R.T. Kolev. Task Results Processing for the Needs of Task-Oriented Design Environments, Int. Journal Computers & Education, 2008, No 51, pp. 86 –96.

[17] Zheliazkova, I.I., P.L. Valkova, G.T. Georgiev. A Computer-Based Technology for Processing and Visualization of Session's Data, Int. Journal of Information Technologies and Control, 2011, No 1, pp. 10 – 18.

Authors' Information:

Zlatan Iliev – fourth year student, Specialty Computer Systems and Technologies, University of Ruse, 8 Studentska Str., 7017 Ruse, Bulgaria; e-mail: zlatko92@mail.bg.

Assoc. Prof. Todor Todorov, PhD, Department of Applied Mathematics and Statistics, University of Ruse, 8 Studentska Str., 7017 Ruse, Bulgaria, ttodorov@uni-ruse.bg.

Head Assistant, Adriana Borodzhieva, PhD, Department of Telecommunications, University of Ruse, 8 Studentska Str., 7017 Ruse, Bulgaria, aborodjieva@ecs.uni-ruse.bg.

Assoc. Prof. Irina Zheliazkova, PhD, Department of Computer Systems and Technologies, University of Ruse, 8 Studentska Str., 7017 Ruse, Bulgaria, Irina@ecs.uni-ruse.bg.

The paper has been reviewed.