# SOME CONTROVERSIES ON THE STATISTICAL METHODS USED IN THE EDUCATIONAL RESEARCH

*Olga Yanushkevichiene* [1*]
*Romanas Yanushkevichius* [2]
*Marilena Aura Din* [3]

## ABSTRACT

*This paper aims to draw attention to two aspects on the use of mathematical statistics for educational research purposes. Firstly, we need to take care about the adequacy of the application of a mathematical statistics method to examine various phenomena of teaching-learning process. As an example considered in our paper is the adequacy of the Student's t-test for testing hypotheses in the educational researches. Secondly, we must take care that research in the field of pedagogy were urgent and the experiment itself did not destroy the natural character of the phenomena. As an example, we try to understand how the computer affects the quality of mastering the material under study and the personal development of student. The results of our study on the two levels of interest are: Firstly, the use of Student's t-test is not adequacy with any type of educational data analyzed, if it does not meet the condition of normality. Secondly, two experiments in question to understand how the computer affects the learning process highlight different things in statistics: once the independence of the two variables and another once the existence of a relationship of dependency, but indirect one. Both highlights are controversial with the general perception on the positive effect of computer-aided learning on students' performances, so that a weak experimental design could alter the nature of a relationship between random variables.*

**KEYWORDS:** *educational researches, statistical methods, Student's t-test.*

## INTRODUCTION

For the consideration of various pedagogical phenomena we use here the methods of mathematical statistics in order to understand if the evaluation of the effectiveness of new methods of teaching are legitimate or not, and for this we take in consideration two aspects.

Firstly, we need to care about the adequacy of the application of a mathematical model to examine phenomena. Mathematical statistics are quite often used in the study of various pedagogical phenomena, but not always performed correctly. For example, A. Sidorkin (2013) wrote about the scandal broke out at the 2013 American Association of Colleges

---
[1*] corresponding author, Professor, Vilnius University Institute of Mathematics and Informatics, Lithuanian University of Educational Sciences, Akademijos str. 4, LT-08663 Vilnius, Lithuania, olgjan@zebra.lt
[2] Professor, Lithuanian University of Educational Sciences, romanas.januskevicius@leu.lt
[3] Associate Professor, Romanian-American University, Bucharest, Romania, din.marilena.aura@profesor.rau.ro

for Teacher Education (AACTE) conference, that "Cory Koedel, a respected researcher reported that all of their models are based on incorrect interpretation of statistics".

Koedel (2015) wrote: "Prior research has overstated differences in teacher performance across preparation programs for several reasons, most notably because some sampling variability in the data has been incorrectly attributed to the preparation programs". Another example is that very often the Student's t-test is used to test hypotheses in the educational researches, even though "normal distribution does not happen as often as people think" (A. Buthmann) and that Student's t-test couldn't be used for Non-Normally Distributed Data. He gave some examples from a technique with Non-Normally Distributed Data. But we never met in educational literature, that Student's t-test couldn't be used for educational researches. Otherwise, it's written (see http://simon.cs.vt.edu/SoSci/converted/T-Dist/), that in the case of Non-Normally Distributed Data we can still use the t-distributions.

But, the mathematical statistis theory prove that the t-distribution comes for sample distribution of size n,when the variable X of a population is *normally distributed* with mean μ and *σ is unknown*, i.e. when the variable under research $X \approx N\left(\mu, \sigma^2\right)$

Then the sample mean of X has a normal distribution:

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

and the standardized variable of the sample mean

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}} = \frac{\overline{x} - \mu}{s_x / \sqrt{n}}$$

has the t-distribution with n − 1 degrees of freedom, where $s_{\overline{X}}$ is the standard error of the sample mean.

Secondly, we must take care the research in the field of pedagogy and the experiment itself did not destroy the natural character of the phenomena. The experimenter himself should demonstrate the scientific honesty and did not send a phenomenon in the direction which is useful for him. As an example we consider the situation with the use of computers at school. Application of computers in the learning process is perceived as something that is clearly a positive one: e-Learning Gets Real, Making Knowledge Interesting, Bringing the World Closer, Presenting Creative Options and so on (see http://www.buzzle.com/articles/use-of-computers-in-education.html). Bialo E. R., Sivin-Kachala J. (1996) wrote that "comparisons of traditional mathematics instruction to its computer-assisted counterpart also yielded positive learning results related to the use of technology". On the other hand, there are some critics for using computer at schools. Mendels (2000) cite C. Stoll, who said that "giving a prominent place to technology in the classroom could end up doing a lot of real harm to students". Koblitz N. (1996) wrote: "If the best computers in the world are unable to translate from French into English, then they certainly cannot help my calculus students do what is the main point of the course: translating word problems into mathematics". It has not been studied enough how the

computer affects the quality of mastering the material under study and the personal development of student.

We will mention here two examples of this kind of research and their limitations. Each of the two experiments designed to assess the effects of using computer on the learning performance of students ignore that could be more other exogenous factors to influence in turn the time spent at the computer. The time spent with computer programs or applications specially tailored for the respective disciplines may change essentially the statistical results of such approach.

## METHODOLOGICAL APPROACH

We consider the use of various methods of mathematical statistics, focusing on two aspects of the problem described above.

Considering the first aspect of the problem we will use the SPSS (Statistical Package for the Social Sciences) and apply the Kolmogorov-Smirnov method for testing the hypothesis of normality of the sample, because we will check the adequacy of the Student's t-test which needs the normal distribution.

Further we will compare two methods of mathematical statistics for investigating educational problems: Student's t-test and the Wilcoxon signed rank test (we will use the SPSS also). We will apply these methods to the results of the experiments, which was done by one of the authors in Ohio University (Yanushkevichiene O., Kriegman S. & Phillips N. 2011).

Considering the second aspect of the problem we will use the Kruskal-Wallis test from the SPSS package, the descriptive statistics and the correlation coefficient.

## PARTICIPANTS IN THE EXPERIMENTS

Considering the first aspect of the problem we will investigate two examples of educational data. First of these will be the annual estimates in mathematics in the tenth grade in 2011, we will check the normality of it. The data were taken from the site of Lithuanian University of Education science (www.estudijos.vpu.lt). It was the annual scores of 581 tenth grade students of Vilnius and Vilnius District's schools.

The participants in the second study were undergraduate students enrolled in an introductory course in probability and statistics at Ohio University in Athens, OH. There were 27 total participants aged between 18 and 20 years of age. The aim was to compare the effectiveness of mathematics teaching by traditional method, and the method of group work.

At the beginning of group work days, the instructor presented the class with between 2 and 4 topics that the students were expected to have mastered by the end of the class period (approximately 50 minutes). To aid in the comprehension of these topics, students were given between 4 and 6 problems from their textbooks to work on. Next, students were organized into between 4 and 6 small groups, each containing between 4 and 6 students. Each group was then assigned a leader by the instructor whose task was to lead group work on assigned problems. Two of these leaders were students enrolled in higher

level statistics courses and who were not students in the research class. Leaders for the other groups were students in the class who either volunteered or were selected by the instructor to act as group leaders.

Students were then instructed to work reading their textbook and working on assigned problems. If students had a question about the course work, they were instructed to discuss the problem with other members of the group, and subsequently ask their group leader for help. The instructor followed the work of each group and tried to understand the weaknesses of students' understanding of material and fixed the attention of students on those points. If the group leader could not answer the question, the group leader directed the question to the course instructor. These group-work sessions lasted for approximately 40 minutes; during which time students completed between 3 and 6 problems from their textbooks.

X quizzes were assigned to students to assess their comprehension of course material. In each quiz, one problem was given whose content was discussed during group-work activities, in addition to one problem whose content was taught during a traditional lecture format. For each problem, answers were scored on discrete 0 to 4 scale, with 0 being the worst and 4 being the best score.

Considering the second aspect of the problem we surveyed seventh-graders of school named by Sofia Kovalevskaya (Vilnius, Lithuania). Students were asked to indicate the average time they spend at the computer and their annual scores in mathematics which changed from 2 to 10. It was not specifide what did they do: just playing games or play games specifically designed to stimulate thinking, creativity, understanding and self-taught learning.

Another experiment was conducted by student E. Poryadina under the guidance of the first co-author of this paper in Moscow. It was attended by 80 students from the second and the third grades. Asking students questionnaire included a question about the leisure of children, the nature and duration of their work at the computer, the mathematical scores, and a number of other issues. The jobs offered to children to assess the development of thinking needed to establish a connection between the words, to eliminate the unnecessary from the number of similar objects, to find patterns in the number series. Each task corresponds to a certain number of points that could be obtained for the right decision.

## THE RESULTS OF THE STATISTICAL METHODS

Let us consider first example of data from the educational researches: the annual estimates of Vilnius schools in mathematics in the tenth grade. The grades of 23 schools were investigated, it were the scores of 481 students. The Figure 1 presents the histograms of scores, which varies from 1 to 10.

If we check the normality of the samples by using the Kolmogorov-Smirnov criterion from SPSS (Fig. 2), we get that the hypothesis of normality of distribution can be rejected with a minimum significance level 0.0001. This means that we cannot assume that the data are normally distributed. Thus, to apply the Student's t-test would be not properly.
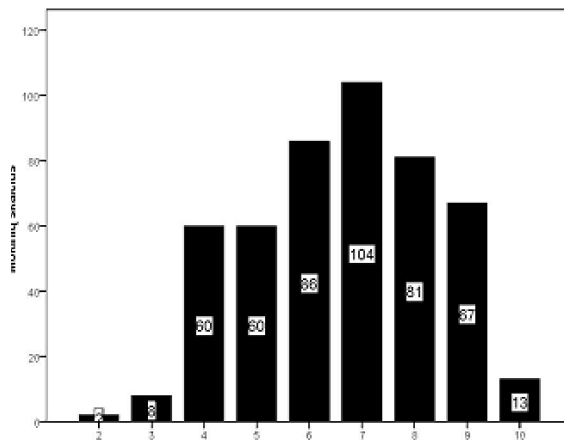
Figure 1. Histograms of scores from the state schools of Vilnius. Horizontal postponed the scores of students, vertical - the number of students with the appropriate scores

Let us investigate the second example. In the previous research of Yanushkevichiene O., Kriegman S. & Phillips N. (2011) the descriptive statistics were used for comparing students' scores for the solution of problem whose content was discussed during group-work activities with the scores for the problem whose content was taught during a traditional lecture format, henceforth referred to as "Traditional". The results of the research provided some evidence for the effectiveness of group work on the comprehension of course material in an undergraduate mathematics course setting. In the current research we will use stronger methods: Student's t-test and the Wilcoxon signed rank test. Tables 1 and 2 present the scores from the quizzes.

## NPar Tests

**One-Sample Kolmogorov-Smirnov Test**

| | | kl_8 |
|---|---|---|
| N | | 481 |
| Normal Parameters [a] | Mean | 6,64 |
| | Std. Deviation | 1,736 |
| Most Extreme Differences | Absolute | ,133 |
| | Positive | ,098 |
| | Negative | -,133 |
| Kolmogorov-Smirnov Z | | 2,927 |
| Asymp. Sig. (2-tailed) | | ,000 |

a. Test distribution is Normal.

## Explore

**Tests of Normality**

| | Kolmogorov-Smirnov [a] | | |
|---|---|---|---|
| | Statistic | df | Sig. |
| kl_8 | ,133 | 481 | ,000 |

a. Lilliefors Significance Correction

Figure 2. Testing the hypothesis of normality.

Let us apply the Kolmogorov-Smirnov criterion from SPSS for testing the hypothesis of normality of the samples in these cases.

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| q | ,232 | 27 | ,001 | ,795 | 27 | ,000 |

a. Lilliefors Significance Correction

Figure 3. The result of testing the hypothesis of normality using the Kolmogorov-Smirnov criterion (the scores from the first quiz, traditional method)

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| q | ,256 | 27 | ,000 | ,837 | 27 | ,001 |

a. Lilliefors Significance Correction

Figure 4. The result of testing the hypothesis of normality using the Kolmogorov-Smirnov criterion (the scores from the first quiz, group method)

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| w | ,372 | 25 | ,000 | ,688 | 25 | ,000 |

a. Lilliefors Significance Correction

Figure 5. The result of testing the hypothesis of normality using the Kolmogorov-Smirnov criterion (the scores from the second quiz, traditional method)

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| w | ,393 | 25 | ,000 | ,666 | 25 | ,000 |

a. Lilliefors Significance Correction

Figure 6. The result of testing the hypothesis of normality using the Kolmogorov-Smirnov criterion (the scores from the second quiz, group method)

We can see from Figures 1-4, that with significance levels less than 0.001 we should reject the hypothesis of normality in all cases and the application of Student's t-test will not be correct.

Let us check now the hypothesis H: there is no difference between the two comprehensions of course material (group and traditional). Let significance level be 0.05.

By using the Student's t-test statistic in SPSS, we have got the results, which are presented in the Figure 5.

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| trad | Equal variances assumed | 14,364 | ,000 | -1,318 | 52 | ,193 | -,48148 | ,36535 | -1,21461 | ,25165 |
| | Equal variances not assumed | | | -1,318 | 42,683 | ,195 | -,48148 | ,36535 | -1,21844 | ,25548 |

Figure 5. Student's t-test

The lowest significance level for Student's t-test is 0.19 and we should accept the null hypothesis H.

By using the Wilcoxon signed rank test, we have got the results, which are presented in the Figure 6.

**Test Statistics[b]**

| | gr - tr |
| --- | --- |
| Z | -2,586[a] |
| Asymp. Sig. (2-tailed) | ,010 |

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

Figure 6.  Wilcoxon signed ranks test

The lowest significance level of this test is 0.01 and we should reject the null hypothesis H. We can do the conclusion, that the result of Student's t-test is wrong, due to the fact, that the distribution of database is not normal.

Let us talk about the second aspect of considered problem. In recent years, education in many countries undergoes considerable reform, and the biggest changes are that the education is done largely computerized. Application of computers in the learning process is perceived as something that is clearly a positive, modern. On the other hand, it has not been studied yet how the computer affects the quality of mastering the material under study and the personal development of student. Such studies require a lot of time and effort. Today we will mention only two cases of this kind of research.

First we will study how the time spent at the computer influences the student's mathematical scores at the 7-th grade in Vilnius. Table 3 presents the results of the experiment.

The Kruskal-Wallis test from the SPSS package was used to test the hypothesis, that the random variable X is independent of the time spent at the computer. By using the Kruskal-Wallis test, we have got the results, which are presented in the Figure 7.
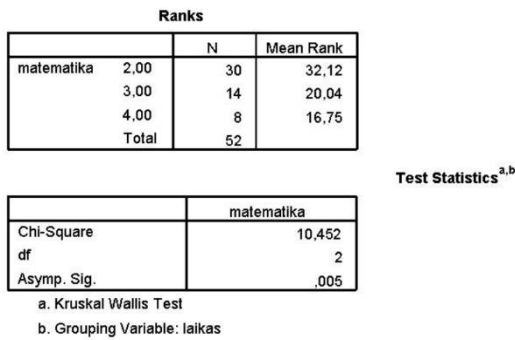
**NPar Tests**

**Kruskal-Wallis Test**

Ranks

|  |  | N | Mean Rank |
|---|---|---|---|
| matematika | 2,00 | 30 | 32,12 |
|  | 3,00 | 14 | 20,04 |
|  | 4,00 | 8 | 16,75 |
|  | Total | 52 |  |

Test Statistics[a,b]

|  | matematika |
|---|---|
| Chi-Square | 10,452 |
| df | 2 |
| Asymp. Sig. | ,005 |

a. Kruskal Wallis Test

b. Grouping Variable: laikas

Figure 7. Kruskal-Wallis test

We should reject the hypothesis with the significance level 0.005 and it is easy to see, that the more student spends time at a computer, the lower his score in mathematics.

Next experiment was conducted in Moscow. It was attended by 80 students from the second and the third grades. Students questionnaire (Poryadina E. 2015) included a question about the leisure of children, the nature and duration of their work at the computer, the mathematical scores, and a number of other issues. The jobs offered to children to assess the development of thinking needed to establish a connection between the words, to eliminate the unnecessary from the number of similar objects, to find patterns in the number series. Each task corresponds to a certain number of points that could be obtained for the right decision. Table 4 presents the results of the experiment.
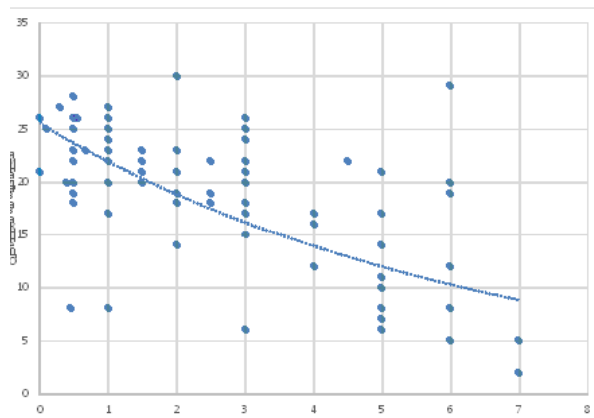


Figure 7. The relationship between the times spent at the computer and the assessment of the level of thinking. Horizontal postponed the times spent at the computer, vertical - the assessment of the level of thinking

In order to quantify the relationship between the time spent at the computer and the assessment of the level of thinking the correlation coefficient between these values was calculated. It was equal to - 0.61, which indicates the presence an inverse relationship between the number of solved pupil assignments and in the meantime, his daily spend at the computer.

## CONCLUSIONS AND FURTHER DEVELOPMENTS

As outlined previously, there were investigated two aspects of mathematical statistics' application to educational research.

In the first aspect - compliance with the pedagogical phenomena and mathematical methods used for his research - we looked at the possibility of using the Student's t-test for testing hypotheses in the educational researches. Application of the Student's t-test is possible if data have the normal distributions. In two presented examples the data are not normal. Attempting to use the Student's t-test knowingly gave wrong result. Of course the results of two experiments are not statistically significant, but the experience of authors indicates that in most cases the data of pedagogical experiments have bimodal distribution and are not normal. Perhaps the reason for this phenomenon lies in the fact that besides the natural ability, the distribution of which is close to normal, the great importance to the learning process has the presence or lack of motivation to learn. This is what can create the effect of bimodal distributions. In any case, Student's t-test for a study of pedagogical phenomena should be used with caution.

In the second aspect - the requirement that the research did not pursue conjectural goals - we looked at the use of the computers in the process of learning. Application of computers is perceived as something that is clearly a positive, modern. In contrast, in the two experiments discussed, the hole time spend at the computer is inversely proportional to the assignments of pupil. We can conclude that the computer' affects to the quality of mastering the material under study and the personal development of student should be studied very carefully.

Designing such an experiment should be carefully studied in advance the independent variable „time spends on computer". Therefore it is imperative required to delimitate this too general variable into the variable „time dedicated to those computer applications which are especially designed to develop the capacity of learning at the respective discipline". Only this approach would reveal the true relationship between variables, i.e. such kind of time spent on computers would positively affect on the variable under research that is student's scores at corresponding discipline. Perhaps such an experiment will highlight the limitations of previous studies that had conclusions which contrast with what it is of common sense: the computer may be a useful tool for the professional and personal developments of students, with caution on how they spend the time at computer.

## REFERENCES

[1]     Bialo E. R., Sivin-Kachala J. (1996) The Effectiveness of Technology in Schools: A Summary of Recent Research. SLMQ Vol. 25, No. 1.

[2]     Buthmann A. Dealing with Non-normal Data: Strategies and Tools http://www.isixsigma.com/tools-templates/normality/dealing-non-normal-data-strategies-and-tools/

[3]     Koblitz N. (1996) The Case Against Computers in K-13 Math Education (Kindergarten through Calculus) The Mathematical Intelligencer, Vol. 18, No. 1

[4]     Koedel C. (2015) Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs?  MIT press Journals. Vol. 10, No. 4, P. 508-534

[5]     Mendels P. (2000) Technology Critic Takes On Computers in Schools. http://partners.nytimes.com/library/tech/00/04/cyber/education/05education.html

[6]     Poryadina E. (2015). Analysis of the effect of computer games on the development of thinking of younger pupils. Coursework, the Faculty of Education, Saint Tikhon's Orthodox University, p. 1:32 (in Russian).

[7]     Sidorkin A. (2013) The Value-Added Scandal. http://sidorkin.blogspot.lt/2013/03/the-value-added-scandal.html

[8]     Yanushkevichiene O., Kriegman S. & Phillips N. (2011). An Investigation of the Activities of Undergraduate Students' Work by Using Statistical Methods. Pedagogika, 102, 88-92.