Method of user authentication on the basis of recognition of computer handwriting peculiarities

Leonid S. Kryzhevich[a]

[a] *Kursk state university, 33 Radisheva str., Kursk, 305000, Russian Federation*

Abstract
This article deals with the following hypothesis: each person has unique peculiarities of text typing. The process of typing can be expressed in the form of various metrics and analyzed with the help of statistical methods.

Keywords
normal distribution, de Moivre–Laplace integral theorem, Pearson's nonparametric test χ^2

1. Introduction

Nowadays people keep almost all sorts of data in digital forms, databases or cloud storage services, which can be accessed online. It is possible to keep important documents, treaties, banking data, passwords. If these forms of data are stolen, people can lose their personal or business information, their bank accounts can be wasted. Therefore, the number of evil-doers, who want to steal various forms of information, is increasing.

There are different ways to protect information. However, they are constantly getting out of date. To detect a transgressor, it is necessary to find out if this person has system access rights. This fact has led to ideas to authenticate users with the help of digital handwriting.

Each person has unique peculiarities of text typing. People type texts at a definite speed. The amount of time of keystrokes can vary as well. We decided to measure these characteristics and analyze them.

2. Conditions of the experiment

An experiment was carried out to get test results. About one hundred students of the faculty of mathematics, physics and information science of Kursk State University participated in the experiment [1]. Their aim was to type a text which included at least four sentences. At the same time, a special program measured the following characteristics for each symbol: the amount of time of a keystroke from the moment when the program was run (in milliseconds); ASCII of a pressed key; whether a key was pressed (1) or released (0).

In **Figure 1:** data fileFigure 1 you can see the file which includes statistical data for the further analysis.

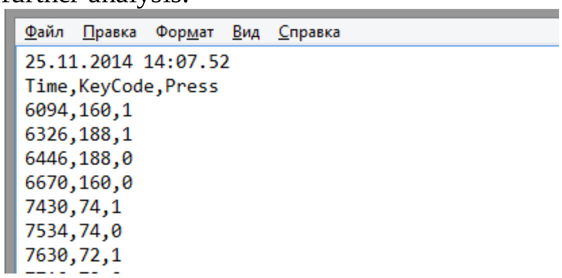

Figure 1: data file

The purpose of the experiment is to determine individual features of one typing session in order to find out in what way it differs from some other test patterns of other users.

3. Data analysis

Let us examine the analysis of statistics of the first feature noted – the amount of time of a keystroke. If we take all the consecutive measurements in pairs for the same symbol (when it was pressed and when it was released) from the test pattern and subtract the press time from the release time, we can see

Models and Methods for Researching Information Systems in Transport, Dec. 11-12, 2020, St. Peterburg, Russia
EMAIL: Leonid@programist.ru (L.S. Kryzhevich);
ORCID: 0000-0002-6736-498X (L.S. Kryzhevich);

the duration of press for each of the symbols. Let us depict test durations for all the symbols in a two-dimensional chart. The horizontal axis of the graph denominates time of a keystroke in milliseconds and the vertical axis denominates frequency of a keystroke (it is the ratio of the number of keystrokes of the definite duration to the total number of keystrokes). If the data are sorted according to the press time, the chart can be depicted in the following way (Figure 2).

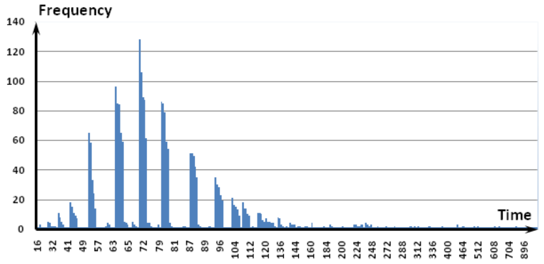

Figure 2: the time/frequency bar chart for the first typing session of a test person

3.1. Checking for normal distribution

Let us make a suggestion that this distribution is normal. To check it, we should analyze the received data with the help of Pearson's nonparametric test χ^2.

Let us divide our series into fourteen disjoint intervals. For each of the intervals we should count the number of test values which are included in it. It is obligatory to include at least five results of each key pressed into each of the intervals [2]. If we follow this rule, we can average out the values of these intervals according to the arithmetic mean and we can create a new chart (Figure 3).

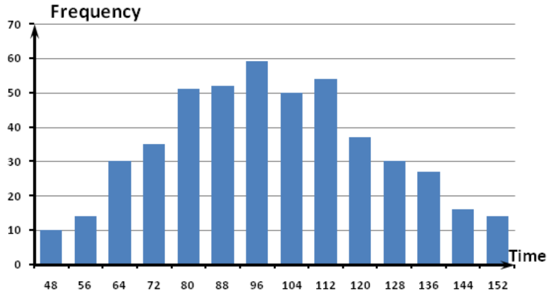

Figure 3: the averaged time/frequency bar chart for the first typing session of a test person

In order to find out if the distribution is normal, we should use Pearson's test χ^2 [3].

We should use the following indices:

x_1 – abscissa axis or time;

f – frequency,

(x_1*f) which should be used to calculate the weighted arithmetic mean;

S – cumulative frequency, which is calculated by adding each previous frequency to the following one; ($|x_i - x_{cp}|*f_i$) value, which is the difference between the current xi and the weighted arithmetic mean multiplied by the current frequency;

($(x_i - x_{cp})^2*f_i$) value, which is the difference between the current x_i and the weighted arithmetic mean which is raised to the second power and multiplied by the current frequency;

(f_i/f) – the ratio of the relative frequency to the total sum.

We should calculate the weighted arithmetic mean:

$$\overline{x} = \frac{\sum x_i * f_i}{\sum f_i} = \frac{47656}{479} = 99{,}49$$

These values are necessary for further calculations. Let us create a Table 1 that includes them.

The dispersion shows the measure of scatter of all the values in the series around the average value.

Let us calculate the mean square deviation: $\sigma = \sqrt{D} = \sqrt{626{,}079} = 25{,}022$

Let us check the suggestion that X is normally distributed with the help of Pearson's chi-squared test $K=\sum\frac{(ni-ni*)^2}{ni*}$, where n*i – theoretical frequencies, which are calculated according to the formula $n_i = \frac{n*h}{\sigma} * \varphi_i$.

Let us choose the mode for the following distribution. The mode is the most frequent value among the examined indices. In our case, we can choose the mode as x_i = 96 (the value of frequency is 59).

The median is also x_i= 96 because it is the first index where the value of the cumulative frequency is higher 479/2≈240.

In symmetrical distribution series the values of the mode and the median are similar to the average value (x_{cp}=Me=Mo), and in moderately asymmetrical series they can be calculated in the following way:

$$3*(x_{av}\text{-Me}) \approx x_{av}\text{-Mo}.$$

Table 1
The calculation table for empirical frequencies of the first typing session

x_i	The number, f_i	Relative frequency, $p_i=f_i/f$	$x_i * p_i$	Cumulative frequency, S
48	10	0.0209	480	0.0209
56	14	0.0292	784	0,0501
64	30	0.0626	1920	0,1127
72	35	0.0731	2520	0,1858
80	51	0.106	4080	0,2918
88	52	0.109	4576	0,4008
96	59	0.123	5664	0,5238
104	50	0.104	5200	0,6278
112	54	0.113	6048	0,7408
120	37	0.0772	4440	0,818
128	30	0.0626	3840	0,8806
136	27	0.0564	3672	0,937
144	16	0.0334	2304	0,9704
152	14	0.0292	2128	0,9996
Total	479	1	47656	

The range of deviation, which is the difference between the minimum and maximum values of x, is R = 152 - 48 = 104.

We can calculate the mean deviation:

$$d = \frac{\sum |x_i - \bar{x}| * f_i}{\sum f_i} = \frac{9896,284}{479} = 20,66.$$

Let us calculate the dispersion D

$$= \frac{\sum(|x_i - \bar{x}|)^2 * fi}{\sum f_i} = \frac{299891,708}{479} = 626,079.$$

x_i	$\lvert x - x_{av}\rvert * p_i$	$(x - x_{av})^2 * p_i$	Cumulative frequency, S
48	514.906	26512.824	10
56	608.868	26480.059	24
64	1064.718	37787.492	54
72	962.171	26450.669	89
80	994.021	19374.069	140
88	597.511	6865.769	192
96	205.946	718.875	251
104	225.47	1016.732	301
112	675.507	8450.187	355
120	758.848	15563.505	392
128	855.282	24383.567	422
136	985.754	35989.269	449
144	712.15	31697.379	465
152	735.132	38601.311	479
Total	9896.284	299891.708	

The following indices are used in the formula: n = 479, h=8 (the interval width), σ = 25.022, x_{cp} = 99.49, φ_i – the appropriate value from Laplace's table.

We can calculate the theoretical frequencies in Table 2.

Now we should compare the empirical and theoretical frequencies.

Table 2
The calculation table for theoretical frequencies of the first typing session

i	x_i	u_i	φ_i	n*i
1	48	-2.0578	0,0478	7.32
2	56	-1.7381	0,0878	13.446
3	64	-1.4184	0,1456	22.298
4	72	-1.0987	0,2179	33.371
5	80	-0.779	0,2943	45.071
6	88	-0.4592	0,3589	54.965
7	96	-0.1395	0,3951	60.509
8	104	0.1802	0,3918	60.003
9	112	0.4999	0,3521	53.923
10	120	0.8197	0,285	43.647
11	128	1.1394	0,2083	31.901
12	136	1.4591	0,1374	21.043
13	144	1.7788	0,0818	12.527
14	152	2.0986	0,044	6.739

We can create one more Table 3, with the help of which we are going to find the observed value of Pearson's test $\chi^2 = \sum \frac{(n_i - n_i^*)^2}{n_i^*}$.

We should include the following indices in the Table 3: i- the sequence number, n_i – the observed frequencies, n_i^* – theoretical frequencies, (n_i - n_i^*) – the difference between the observed and theoretical frequencies, $(n_i - n_i^*)^2 / n_i^*$– the difference, which is raised to the second power and divided by the current value of the theoretical frequency.

Later we should calculate the following indices: K_{emp} – the observed value of the bound of the critical region and K_{cr} - the theoretical value of the bound of the critical region.

Table 3
The calculation table for comparison of theoretical and empirical frequencies of the first typing session

i	x_i	u_i	φ_i	n*i
1	48	-2.0578	0,0478	7.32
2	56	-1.7381	0,0878	13.446
3	64	-1.4184	0,1456	22.298
4	72	-1.0987	0,2179	33.371
5	80	-0.779	0,2943	45.071
6	88	-0.4592	0,3589	54.965
7	96	-0.1395	0,3951	60.509
8	104	0.1802	0,3918	60.003
9	112	0.4999	0,3521	53.923
10	120	0.8197	0,285	43.647
11	128	1.1394	0,2083	31.901
12	136	1.4591	0,1374	21.043
13	144	1.7788	0,0818	12.527
14	152	2.0986	0,044	6.739

The higher K_{emp} value differs from K_{cr}, the more convincing arguments against our main hypothesis can be provided [3].

Its bound K_{cr} = χ2(k-r-1;α) can be calculated according to the distribution tables χ^2 and the set values x_{av} and σ(determined according to the series), k = 14, r=2,the significance level α is determined as 0,05.

$K_{cr}(0.05;11) = 19.67514$; $K_{emp} = 17.99$.

The observed value of Pearson's statistics does not touch the critical region: (K_{emp}<K_{cr}.) It can be fair to say that the data from the series follow the rules of normal distribution.

Paying attention to the same ideas, we can check the second set of data series (Figure 4) of the same person but for different text extracts with the help of Pearson's test.

Table 4
The calculation table for empirical frequencies of the second typing session

xi	The number, fi	Relative frequency, pi=fi/f	xi * pi	Cumulative frequency, S
56	5	0.0116	280	0.0116
64	15	0.0349	960	0,0465
72	35	0.0814	2520	0,1279
80	38	0.0884	3040	0,2163
86	45	0.105	3870	0,3213
88	53	0.123	4664	0,4443
96	56	0.13	5376	0,5743
104	51	0.119	5304	0,6933
112	41	0.0953	4592	0,7886
120	35	0.0814	4200	0,87
128	30	0.0698	3840	0,9398
136	15	0.0349	2040	0,9747
144	8	0.0186	1152	0,9933
152	3	0.00698	456	1,00028
Tota	430	1	4229	

xi	|x - xcp|*pi	(x - xcp)2*pi	Cumulative frequency, S
56	211.791	8971.06	10
64	515.372	17707.226	24
72	922.535	24316.303	54
80	697.609	12806.809	89
86	556.116	6872.563	140
88	548.981	5686.426	192
96	132.056	311.406	251
104	287.735	1623.36	301
112	559.316	7630.115	355
120	757.465	16392.954	392
128	889.256	26359.197	422
136	564.628	21253.645	449
144	365.135	16665.435	465
152	160.926	8632.348	479
Total	7168.921	175228.847	

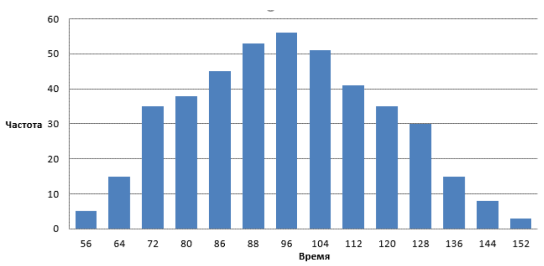

Figure 4: the averaged time/frequency bar chart for the second typing session of a test person

Let us create a Table 4 for the second distribution according to the described above.

We have the following values of the indices:

The weighted arithmetic mean (sample mean)$\bar{x} = \frac{\sum xi*fi}{\sum fi} = \frac{42294}{430} = 98,36$

The maximum value of repeat counts if x = 96 (f = 56) => the mode is 96.

Half of the sum of the cumulative frequency is 216. It is x_i = 96. Thus, the median is 96.

The range of deviation is 152 - 56 = 96.

The mean deviation is

$$d = \frac{\sum |xi - \bar{x}| * fi}{\sum fi} = \frac{7168{,}921}{430} = 16{,}67.$$

Table 5

The calculation table for theoretical frequencies of the second typing session

i	x_i	u_i	φ_i	n_i^*
1	56	-2.0983	0,044	7.498
2	64	-1.702	0,0925	15.763
3	72	-1.3057	0,1691	28.816
4	80	-0.9094	0,2637	44.937
5	86	-0.6122	0,3292	56.098
6	88	-0.5131	0,3485	59.387
7	96	-0.1168	0,3961	67.499
8	104	0.2795	0,3825	65.181
9	112	0.6758	0,3166	53.951
10	120	1.0721	0,2227	37.95
11	128	1.4684	0,1354	23.073
12	136	1.8647	0,0694	11.826
13	144	2.261	0,0303	5.163
14	152	2.6573	0,0116	1.977

Each value of the range differs from another index by 16.67

Let us calculate the dispersion:

$$D = \frac{\sum(|xi - \bar{x}|)^2 * fi}{\sum fi} = \frac{175228{,}847}{430} = 407{,}509$$

The mean square deviation is $\sigma = \sqrt{D} = \sqrt{407{,}509} = 20{,}187$

We can check the suggestion that X is normally distributed with the help of Pearson's chi-squared test [3]. We should calculate the theoretical frequencies, paying attention to the fact that: n = 430, h=8 (the interval width), σ = 20.187, xcp= 98.36.

$$n_i = \frac{n * h}{\sigma} * \varphi_i \Rightarrow n_i = \frac{430 * 8}{20{,}187} * \varphi_i = 170{,}41\ \varphi_i.$$

Table 6

The calculation table for comparison of theoretical and empirical frequencies of the second typing session

i	n_i	n_i^*	n_i-n_i^*	$(n_i$-$n_i^*)^2$	$(n_i$-$n_i^*)2/n_i^*$
1	5	7.498	2.498	6.2398	0.832
2	15	15.7627	0.7627	0.5818	0.0369
3	35	28.816	-6.184	38.242	1.327
4	38	44.9366	6.9366	48.1161	1.071
5	45	56.0983	11.0983	123.1722	2.196
6	53	59.3872	6.3872	40.796	0.687
7	56	67.4986	11.4986	132.2176	1.959
8	51	65.181	14.181	201.102	3.085
9	41	53.9512	12.9512	167.7325	3.109
10	35	37.9499	2.9499	8.7016	0.229
11	30	23.0732	-6.9268	47.98	2.079
12	15	11.8263	-3.1737	10.0723	0.852
13	8	5.1634	-2.8366	8.0465	1.558
14	3	1.9767	-1.0233	1.0471	0.53
∑	430	430			19.551

Let us calculate the theoretical frequencies (Table 5), paying attention to the appropriate values from Laplace's table.

Let us compare the empirical and theoretical frequencies. We can create a calculation Table 6 for the second typing session where the above mentioned values should be included. The table helps us to determine the observed value of the test: $\chi^2 = \sum \frac{(n_i - n_i^*)^2}{n_i^*}$.

According to the described above principle, we can see that: $K_{cr}(0.05;11) = 19.67514$; $K_{emp} = 19.55$. Thus, $(K_{emp} < K_{cr}) \Rightarrow$ the distribution is normal.

3.2. Comparison of series

Two sets of samples for one person are portrayed in the next Figure 5.

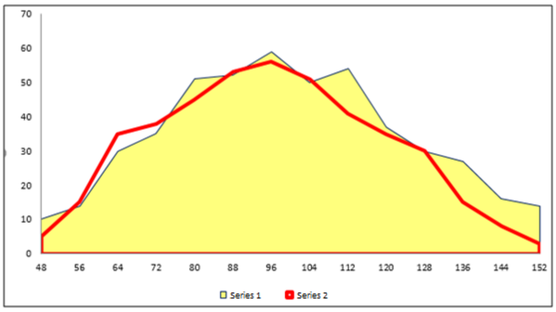

Figure 5: joint graphs for the sets of samples of the first and the second typing sessions

To show everything better, we can depict the graphs in the form of bar charts (Figure 6). The red bars denote the averaged chart of the first typing session, the blue bars are related to the second typing session.

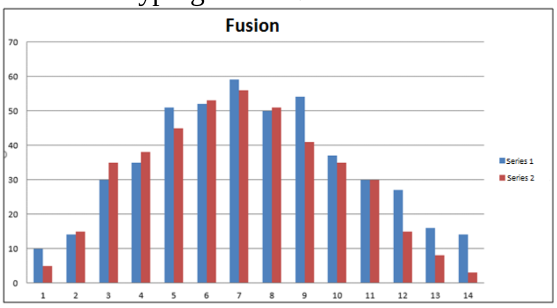

Figure 6: the graphs of the sample sets

To determine how much the typing style of one test person differs from his own, we should examine the crossing area of the graphs[4]. The first set of samples crosses the second set completely. Therefore, we should consider the second set to be the crossing area, whereas the first set of samples is the joining area.

We should use the following formula:

$\sum_{i=1}^{n} h * l_{max} - \sum_{i=1}^{n} h * l_{min}$, where:

h – width of the bars;

l_{max}= $\max(l_{i1}, l_{i2})$ – the maximum value out of the bar heights, which are grouped in pairs, from the two graphs;

l_{min}= $\min(l_{i1}, l_{i2})$ – the minimum value, respectively.

According to the described formula, for the first typing session we can see $\sum_{i=1}^{n} h * l_{max}$=0,010438+ 0,029228+ 0,06263+ 0,073069+ 0,093946+ 0,108559+ 0,11691+ 0,104384+ 0,085595+ 0,073069+ 0,06263+ 0,031315+ 0,016701+ 0,006263=2,0459.

For the second typing session we can see $\sum_{i=1}^{n} h * l_{min}$=0,020876827 +0,03131524 +0,073068894 +0,079331942 +0,106471816 + 0,110647182+ 0,123173278+ 0,106471816+ 0,112734864+ 0,077244259 +0,06263048 + 0,056367432 + 0,033402923+0,029227557 =1,749.

The hit rate is K1= 1,749 /2,0459=0,85510≈86% is the level of coincidence between the two results of the same user.

We can check the hit rate between the values of normal distributions, which are corresponding to the sets noted [5]. We should use de Moivre–Laplace integral formula for normal distribution.

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{\frac{-(t-m)^2}{2\sigma^2}} dt,$$

where

σ – standard deviation;

t – the amount of time of a keystroke in milliseconds;

m – expected value.

According to that function, we can create the graphs of the two cases of the normal distribution, which are shown in Figure 7.

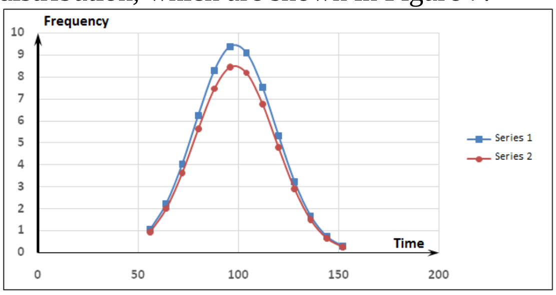

Figure 7: graph of normal distributions, corresponding to both samples

where

S1 – the area, which is limited to the first graph,

S2 – the area, which is limited to the second graph.

The hit rate of the theoretical graphs is K2=$\frac{S_1 1\cap S_2}{S_1 \cup S_2}$=$\frac{53,082}{59,075}$=0,899582≈90% - is the level of the coincidence.

Even taking into consideration the high error level, we have 86% of coincidence for the empirical and 90% of coincidence the theoretical values. Therefore, we can conclude that each person has individual peculiarities connected with the duration of pressing keys he or she follows while typing texts.

4. Scaling by multiple series

In Figure 8 we can see a range of the expected value for the amount of time of folding different keys pressed [4] in the sessions of the same user during different days (the days are marked in different colours).

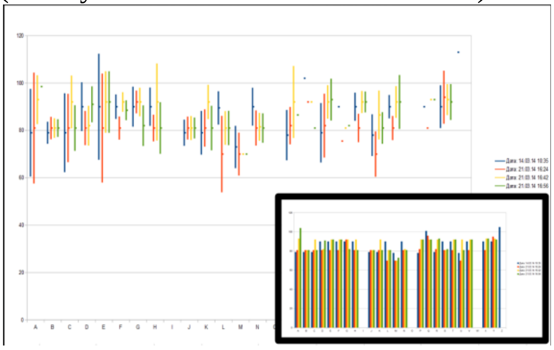

Figure 8: the graph for the cases of normal distribution

In the bottom right corner on the axis of the ordinates, we can see the average amount of time of holding the keys pressed.

In Figure 9, the similar characteristics are illustrated to show the typing sessions of different users.

The comparative analysis of the received results gives an opportunity to conclude that the amount of time of holding different keys pressed is a very informative value that shows a user’s typing technique[6]. Despite partly random scatter of averaged amounts of time of holding keys pressed, the statistical analysis of the differences lets identify various versions of keyboard typing of the same user and distinguish typing variants of different users[7].

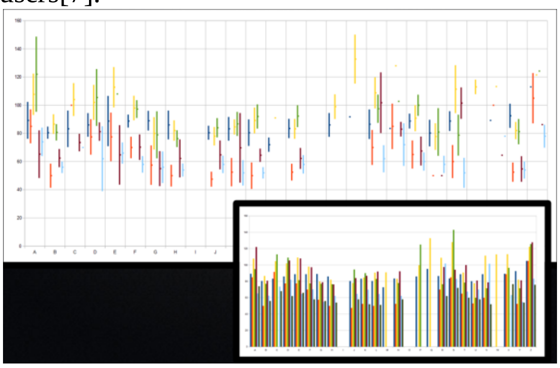

Figure 9: the graph for the distribution of typing sessions of different users

The results of the experiment show that, in most cases, periods of time of holding keys pressed are random sets of samples, which are normally distributed.

5. Summary

This method of identification during the process of the user’s authorization can be used in samplings of various volumes. K value of each user can differ a bit in different typing sessions. The fact that K value can be close or not so close to 1 depends on the level of development of the user’s keyboard handwriting. If a user has weak typing skills, the critical value K for his authorization can be determined according to the results of the comparative analysis of his several typing sessions[8]. The further analysis of typing sessions of such users can be made more accurate if we do not take into consideration those keys, the amounts of time of holding which pressed have a high level of standard deviation (for example, far higher than the standard deviation of the whole typing session).

References

[1] AragÃ³n-MendizÃ¡bal E., Delgado-Casas C., Romero-Oliva M. F., A comparative study of handwriting and computer typing in note-taking by university students, Comunicar (2016). doi: 10.3916/C48-2016-10

[2] Summary and classification of statistics, 2018, URL: http://www.grandars.ru/student/statistika/gruppirovka-statisticheskih-dannyh.html

[3] Shulenin, V.P., Mathematical statistics, NTL Publishing House, Tomsk, 2012.

[4] Kryzhevich L.S., Rakov A.S. Kostenko I.V., Arkhipova V.V. Lukin D.E., Testing statistical hypotheses about the time parameters of keytyping, "Problems of cybersecurity, modeling and information processing in modern sociotechnical systems", KSU, Kursk, 2017.

[5] Gmurman V.E., Probability theory and mathematical statistics, 9th edition, Vysshaya shkola, Moscow, 2003.

[6] Fedorowich L. M., Côté J. N., Effects of standing on typing task performance and upper limb discomfort, vascular and muscular indicators, Applied Ergonomics (2018). doi: 10.1016/j.apergo.2018.05. 009.

[7] Kryzhevich L. S., Matyushina S. N., Kostenko I. V., Providing access to electronic equipment based on computer

handwriting recognition, "Current research in the field of exact sciences and their study in secondary and higher educational institutions", KSU, Kursk, 2015.

[8] Yoo W. G., Effects of different computer typing speeds on acceleration and peak contact pressure of the fingertips during computer typing, Journal of Physical Therapy Science (2015). doi: 10.1589/jpts.27.57