

**RPTSS 2017**  
**International Conference on Research Paradigm Transformation  
in Social Sciences**

**STEN SCORE METHOD AND CLUSTER ANALYSIS:  
IDENTIFYING RESPONDENTS VULNERABLE TO DRUG ABUSE**

Natalia Lukianova (a,b)\*, Yulia Burkatovskaya (a,b,c), Elena Fell (a)

\*Corresponding author

(a) Tomsk Polytechnic University, Lenin Av., 30, Tomsk, Russian Federation

(b) Tomsk State University, Lenin Av., 36, Tomsk, Russian Federation

(c) International Laboratory of Statistics of Stochastic Processes and Quantitative Finance, Lenin Av., 36, Tomsk,  
Russian Federation

*Abstract*

In this article, the authors assess the methodological reliability of big data processing in sociological research. The authors compare sten score method and cluster analysis as methods of processing the results of socio-psychological tests aimed at identifying groups of young people potentially vulnerable to drug addiction. The survey was conducted in eight universities in a city in Siberia with a large student population where 22884 students aged from 18 to 25 were questioned. First, the obtained results were processed by using the sten score method. Then, cluster analysis was conducted to define a high-risk group of students having a propensity for drug consumption. Advantages and disadvantages of the two methods for processing a large sample of data are compared. The results of this comparison demonstrate that the cluster analysis method is the most appropriate method for this type of research as it produces statistically correct data. The use of cluster analysis makes it possible to work with any type of information, both qualitative and quantitative data. On the other hand, the sten scores method can only be applied in certain conditions, i.e. where the original distribution resembles a normal distribution; where some theoretical basis to expect normal distribution exists, and where there is certainty that the normalization group is sufficiently large and representative to be a true reflection of the population.

© 2018 Published by Future Academy [www.FutureAcademy.org.UK](http://www.FutureAcademy.org.UK)

**Keywords:** Sten score method, cluster analysis, drug use, drug addiction, sociological research, statistical information.



This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1. Introduction

The paper compares two methods of psychological test results processing, sten score method and cluster analysis. The object of the investigation is results obtained during the work on the project “Social-psychological testing of students, aimed at early detection of non-medical consumption of narcotic and psychotropic substances”. The study was conducted in eight universities of a Siberian city with a large student population; 22884 students aged from 18 to 25 were questioned. First, the obtained results were processed by using sten score method. Then, cluster analysis was conducted to define a high-risk group of students having a propensity for drug consumption.

## 2. Problem Statement

The paper considers the methodological reliability of big data processing in sociological research. One of the tasks of the study, which was solved by statistical methods, was to identify amongst students a group of individuals who (Newcomb & Bentler, 1988), according to their personal characteristics and circumstances, are vulnerable to drug use and addiction (Restivo & Loughlin, 2016).

## 3. Research Questions

The research, presented in this article, focuses on the psychological predisposition to drug use (Dembo R, et al., 1988) and addiction, whereby a combination of certain character traits, education, attitude to life (Drugs, Brains, and Behavior: The Science of Addiction, 2014), personal and environmental coping mechanisms (Sinha, 2008), family relationships (Barrett & Turner, 2006), sense of loneliness, negative emotional ambiance (Hayaki et al., 2005) and other factors may form a psycho-protective coping behavior (Spooner & Hetherington, 2004) that makes a person vulnerable to drug addiction (Patrick, Wightman, Schoeni & Schulenberg, 2012). It is assumed that in those situations, bans, high prices or other obstacles in obtaining drugs do not play a significant role (Blum et.al., 2012; Sinha, 2008; Lukianova, Fell, Sibers, 2015)

To reveal the presence of individuals with this predisposition, socio-psychological testing was conducted in a group. The respondents were asked to fill in a questionnaire, which included 35 statements. 27 of them were direct markers, and 8 were inverse markers of drug-addict behavior (Spooner & Hetherington, 2004). Filling in the "passport part" of the questionnaire helped the subject to familiarize himself or herself with the theme of the questionnaire and prepared them for the work with the main block of questions. Each respondent was given the following instruction: "Read carefully each statement and choose from the proposed answers the one that, in your opinion, is the most accurate. The selected variant of the answer should be marked with a "+" sign in the corresponding table cell ". The possible answers were “yes”, “rather yes”, “rather no” and “no”.

The average duration of the procedure was no more than 30 minutes.

The test was applied in the study of Siberian city students aimed at identifying a high-risk group prone to drug-addict behavior. To process the results, sten scores method (Kline, 2000) was used. The test and the specific method were specifically designed for this research. However, the sten scores method has its limitations. First, the sten scores method requires information in numerical form, whereas in the

test respondents should answer 'yes', 'rather yes', 'rather no' and 'no', so, the information is given in an order scale. Hence, answers are translated into numbers; consequently, the order scale is transformed into the numerical scale. This leads to the loss of information and its distortion. Since there are no correct procedures for transferring qualitative information to quantitative information, there is a doubt about the correctness of the results obtained. Consequently, the sten scores method can only be applied in certain conditions (Kline, 2000): (1) where the original distribution resembles normal distribution; (2) where the authors have some theoretical basis to expect normal distribution, and (3) where the authors are confident that their normalization group is sufficiently large and representative to be a true reflection of the population. However, when using the sten scores method, a large number of respondents taking part cannot be guaranteed.

The authors propose to use cluster analysis to process such statistical information. The advantage of this method is that it works with any type of information, both qualitative and quantitative, as well as with data of various types. Cluster analysis is applicable to working with small samples and its purpose is to divide objects into a predetermined number of classes in such a way that the objects inside the class are similar to each other, whereas the objects from different classes are different. Applying this method makes it possible to isolate the risk group according to the criteria specified in the test, and also to estimate statistical reliability of the results.

#### **4. Purpose of the Study**

The purpose of the study is to compare the two methods of statistical analysis for the correctness of their application for identifying respondents who are at risk of drug-dependent behavior.

#### **5. Research Methods**

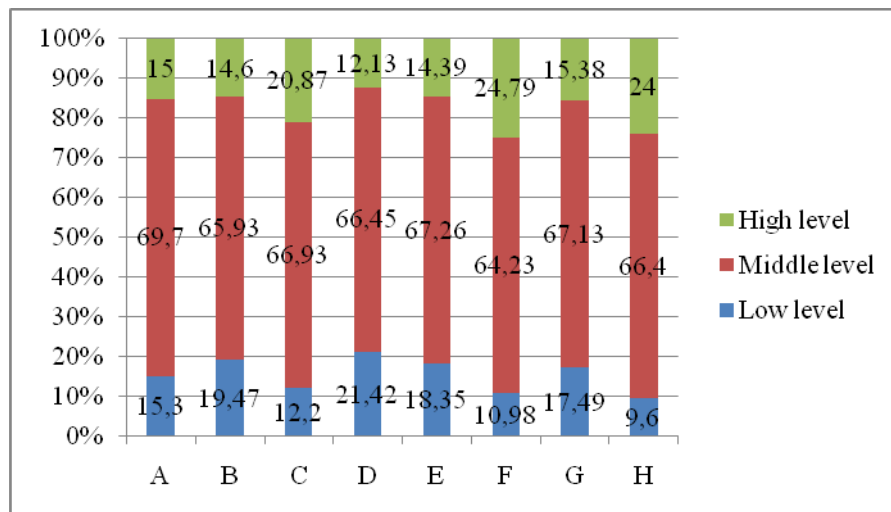
##### **5.1. Sten scores**

First, the authors interpreted the results by using sten scores, sten being an abbreviation for 'Standard Ten'. The procedure is described as follows.

1. Answers for direct markers of DAB (drag-addict behaviour) are translated into 'response scores': answers 'yes' is interpreted as '4', answers 'rather yes' – as '3', answers 'rather no' – as '2', and answers 'no' – as '1'.
2. Answers for inverse markers of DAB (drag-addict behaviour) are translated into 'response scores': answers 'yes' are interpreted as '4', answers 'rather yes' – as '3', answers 'rather no' – as '2', and answers 'no' – as '1'.
3. The 'raw score' for every  $i$ -th respondent is calculated as the sum  $x_i$  of all 35 response scores.
4. The mean  $M$  and the standard deviation  $s$  is calculated by the raw scores of all respondents.
5. By using the formula  $z_i = (x_i - M) / s$ , the raw scores are translated into 'standard scores'.

6. By using the sten formula  $Sten_i = 2z_i + 5,5$ , the obtained 'standard scores' are translated into 'standard tens' or 'stens'.
7. Calculating the percentage of respondents whose results in stens are above 7.5 (high level of psychological propensity for DAB). Calculating this value for different universities and other subgroups of interest (age, gender, year, program of study, etc.).
8. Calculating the percentage of respondents whose results in stens are below 3.5 (low level of psychological propensity for DAB). Calculating this value for different universities and other subgroups of interest (age, gender, year, program of study, etc.).
9. Calculating the percentage of respondents whose results in stens are within the range from 3.5 to 7.5 (middle level of psychological propensity for DAB). Calculating this value for different universities and other subgroups of interest (age, gender, year, program of study, etc.).

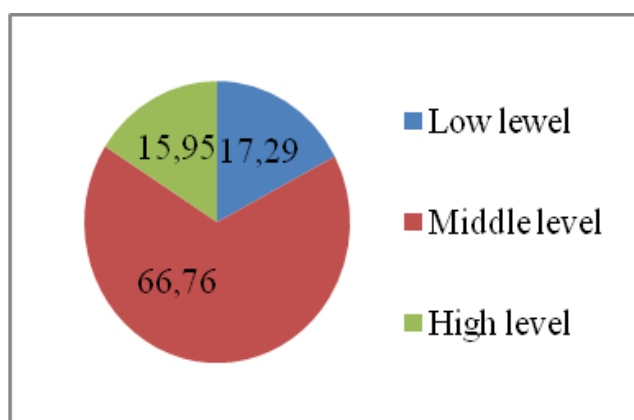
Figure 01 presents the distribution of these three levels of psychological propensity for DAB in different A and B universities. Labels A, ..., H of the horizontal axis stand for the universities, the percentage of the respondents with low, middle and high scores are shown in blue, red and green accordingly. One can see that the percentage of the high score respondents varies from 12.13% to 24.79%; the percentage of the middle score respondents varies from 64.23% to 69.7%; the percentage of the low score respondents varies from 9.6% to 21.42%, depending on the university. Therefore, the percentage of middle score respondents is much less changeable in comparison with high and low score respondents.



**Figure 01.** Distribution of three levels of psychological propensity for DAB in Tomsk and Seversk universities with respect to the total sample.

The main objective of the study was to estimate the total number of respondents with a high level of psychological propensity for drag-addict behaviour (Shelehov, Kornetov & Grebennikova, 2016), or high-risk population, as a percentage of the sample. Figure 02 demonstrates these results. Here the percentage of the respondents with low, middle and high scores are shown in blue, red and green

accordingly. One can see that the percentage of the high score respondents is 15.95%, the percentage of the middle score respondents is 66.76%, the percentage of the low score respondents is 17.29%.



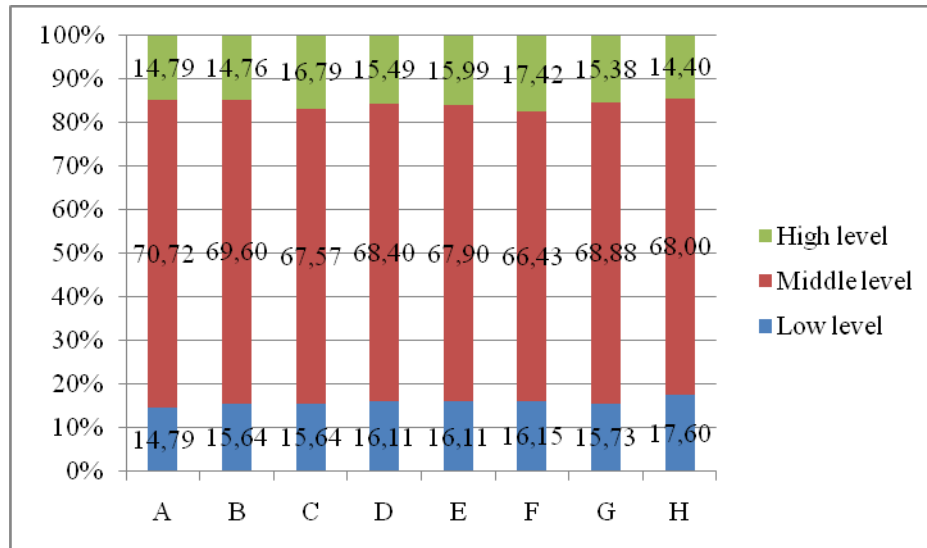
**Figure 02.** Distribution of three levels of psychological propensity for DAB for the total sample.

These results are very close to the results of similar studies conducted in Kazan, Birsk and Izhevsk (Cheverikina, 2012.). However, the authors can identify a number of problems arising when the tens scores for obtaining this objective are used.

First, the sten scores method needs information in numerical form, whereas in the test, respondents should answer ‘yes’, ‘rather yes’, ‘rather no’ and ‘no’, which means that the information is given in an order scale. Hence, at steps 1 and 2, answers are translated into numbers; consequently, the order scale is transformed into the numerical scale. From the statistical point of view, such transformations are unfounded and they result in the misrepresentation of information.

Then, the sten scores are in fact linearly transformed into normalised standard scores (see steps 4–6); after normalising, the authors investigate not absolute but relative numbers. Hence, the sten score indicates an individual's approximate position with respect to the population of values and, therefore, to other people in that population. So, the high-risk population can be estimated only in reference to the sample. In other words, the percentage does not depend significantly on the sample, if the authors investigate two groups of people, one with a high level of DAB, and another with a low level; in every group the authors will most probably obtain the same percentage of the high-risk population. This argument is supported by the investigation, the authors conducted using the same data. At step 5, the authors calculated the means  $M_k$  and standard deviations  $s_k$  for all universities  $k \in \{A, \dots, H\}$  and then the authors recalculated standard scores by using the formula  $z_i^k = (x_i^k - M_k) / s_k$ , where  $x_i^k$  is the raw score of the  $i$ -th respondent from the  $k$ -th university. After that, the authors performed steps 6–9. Figure 03 presents the results. One can see that the percentage of the high score respondents varies now from 14.4% to 17.42%; the percentage of the middle score respondents varies from 66.43% to 70.72%; the percentage of the low score respondents varies from 14.79% to 17.60%, depending on the university. So, the percentages of high and low score respondents are much less changeable in comparison with Figure 01. They are close to 16% in all cases, which can be explained as follows. The sten scores method is based on the hypothesis that the initial distribution of the raw scores is close to normal distribution (Kline, 2000); hence, the sten scores have normal distribution, too. The bounds of 3.5 stens and 7.5 stens

were calculated by using this hypothesis; in fact, these are 16% quantiles of normal distribution. Taking into account the large number of respondents and questions, the authors can expect that the distribution of the raw scores, according to the central limit theorem, tends to be normal, and the percentages of high-level score and low-level score respondents should be close to 16%.



**Figure 03.** Distribution of three levels of psychological propensity for DAB with respect to universities samples.

Consequently, the main objective of the stens method is to determine not the percentage of high-level score and low-level score respondents, but to recognize the respondents belonging to these 16% groups. It is the reason of the wide application of the method in the interpretation of IQ and other similar tests (Anastasi, 1988). In other studies, it can be applied in certain conditions (Kline, 2000): (1) where the original distribution resembles normal distribution; (2) where the authors have some theoretical basis to expect normal distribution, and (3) where the authors are confident that their normalization group is sufficiently large and representative to be a true reflection of the population.

Besides, sometimes the results are essentially different with respect to the group where the mean and the variance are calculated. For example, for university H, Figure 01 shows a relatively big percentage of high-level scores and relatively small percentage of low-level scores, but the results shown in Figure 03 are divisive, as the authors note a relatively small percentage of high-level scores and relatively big percentage of low-level scores. The authors think that the main reason for this is that the number of students at the university is rather small (only 136 respondents were questioned). So, the condition (3) set by Kline does not hold true. It stresses the fact that this method does not fit well for small-size samples.

## 5.2. Cluster analysis

Cluster analysis divides data into groups (clusters) based only on information that describes the objects and their relationships. The requirement is that the objects within a cluster must be similar to one another and significantly different from the objects in other groups. The greater the similarity within a

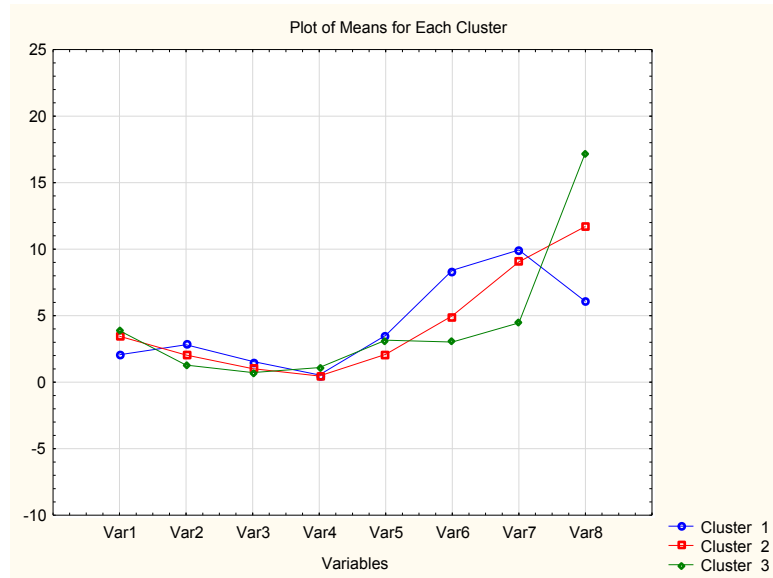
cluster, and the greater the difference between clusters, the better is the clustering. In cluster analysis, the characteristics of object (variables) can be of any type: nominal, order, quantitative. The objective is to define a distance between objects and between clusters (Everitt, Landau, Leese & Stahl, 2011).

In this case, the authors have 35 questions, which is a rather big number of variables. Besides, all direct markers are supposed to be equal, and all inverse markers too. Consequently, the authors chose as variables the numbers of every kind of responses for the direct and inverse markers; as a result, the authors obtained eight variables. After that, the authors divided all respondents within universities into three clusters, as in the experiment described above, by using software package “Statistica 10” (StatSoft, 2013). Here Figure 04 presents the results for university H (with a small number of respondents), and Figure 05 presents the results for university E (with a large number of respondents). Here the following variables are used:

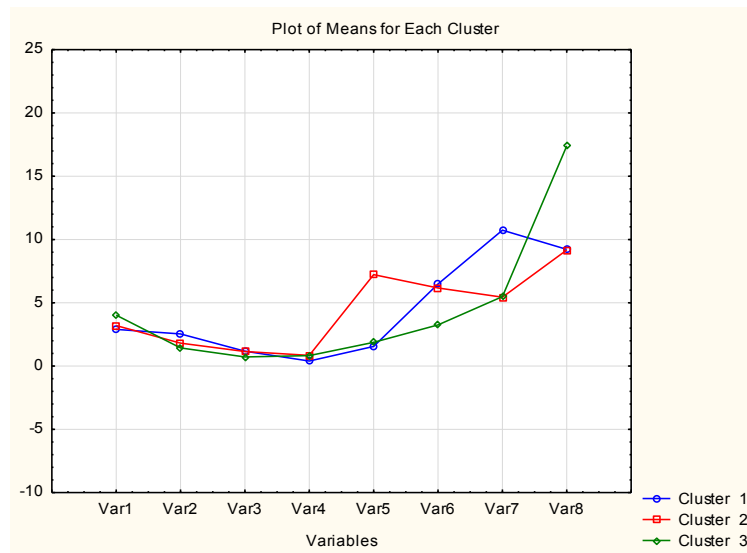
- Var1 – number of answers “yes” for the inverse markers;
- Var2 – number of answers “rather yes” for the inverse markers;
- Var3 – number of answers “rather no” for the inverse markers;
- Var4 – number of answers “no” for the inverse markers;
- Var5 – number of answers “yes” for the direct markers;
- Var6 – number of answers “rather yes” for the direct markers;
- Var7 – number of answers “rather no” for the direct markers;
- Var8 – number of answers “no” for the direct markers.

Points mark mean values of the variables for each cluster.

One can see that in both cases, Cluster 3 (green line) corresponds to the low-risk groups. The number of answers “yes” for the inverse markers (Var 1) and the number of answers “no” for the direct markers (Var 8) exceed such values in other clusters, and exceed the numbers of other answers within the clusters. For university H, Cluster 1 (blue line) corresponds to the high-risk group: it has the least (between the clusters) number of answers “yes” for the inverse markers (Var 1) and number of answers “no” for the direct markers (Var 8), and the greatest numbers of answers “yes” and “rather yes” for the direct markers (Var 5 and Var 6). For university E, Cluster 2 (red line) corresponds to the high-risk group: the number of answers “yes” for the direct markers (Var 6) essentially exceeds similar values for other clusters. Cluster 2 (red line) for university H and cluster 1 (blue line) for university E correspond to the middle-risk groups. The analysis of the variable variances shows that the probability of incorrect clustering is not more than 0.001.



**Figure 04.** Plot of means for each clusters for university H.



**Figure 05.** Plot of means for each clusters for university E.

Table 01 presents means of variables for clusters. One can see that the direct markers describe the situation better than the inverse markers. For example, the number of answers “no” for the direct markers is sufficiently greater for the low-risk groups than for other groups in both universities. The mean number is above 17; so, the low-risk respondents deny the direct markers in 17 cases from 27 in average. For both high-risk groups, the number of answers “yes” exceeds this numbers in other groups, and for university E, the difference is essential (7.22 against 1.88 and 1.55). The number of answers “yes” and “rather yes” is about 12-13; so, the high-risk respondents admit the direct markers in 12-13 cases from 27 in average, whereas in the middle-risk groups, this number is about 7; in the low-risk groups, it is about 5-6. So, the numbers of different answers for direct markers are suitable variables for clustering, both from mathematical and from psychological points of view.

As for the inverse markers, the number of answers “yes” is the biggest for the most cases (about 3–4), except for the high-risk group at university H. But in this case, the number of answers “rather yes”

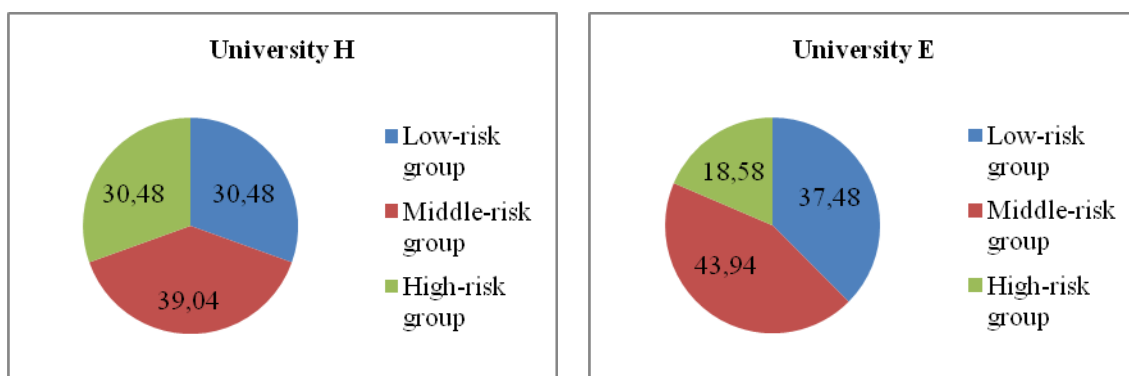


is the biggest. Therefore, the variables do not characterize well the groups with different DAB. Most probably, it is connected with a small number of markers.

**Table 01.** Means of variables within clusters.

| Markers         | Answers    | University H   |                   |                 | University E   |                   |                 |
|-----------------|------------|----------------|-------------------|-----------------|----------------|-------------------|-----------------|
|                 |            | Low-risk group | Middle-risk group | High-risk group | Low-risk group | Middle-risk group | High-risk group |
| Inverse markers | yes        | 3.88           | 3.45              | 2.06            | 4.02           | 2.9               | 3.19            |
|                 | rather yes | 1.29           | 2.04              | 2.82            | 1.43           | 2.54              | 1.82            |
|                 | rather no  | 0.73           | 1.02              | 1.56            | 0.71           | 1.17              | 1.15            |
|                 | no         | 1.09           | 0.46              | 0.54            | 0.82           | 0.4               | 0.85            |
| Direct markers  | yes        | 3.16           | 2.09              | 3.52            | 1.88           | 1.55              | 7.22            |
|                 | rather yes | 3.02           | 4.99              | 8.4             | 3.25           | 6.49              | 6.19            |
|                 | rather no  | 4.46           | 9.04              | 9.94            | 5.49           | 10.74             | 5.44            |
|                 | no         | 17.25          | 11.71             | 6.06            | 17.39          | 9.22              | 9.17            |

Figure 06 presents the distribution of respondents within universities H and E. One can see that the percentages of the low-risk and high-risk respondents differ from 16% usual for sten scores methodology. The reason is that in cluster analysis the number of cluster members is not prescribed; the goal is to divide the respondents into a given number of clusters. Consequently, cluster analysis can be used to estimate the percentage of groups. The obtained percentages of the high-risk groups and the low-risk groups are higher than ones obtained by sten scores method (see Figure 01 for universities H and E). They are 30.48 against 24 for university H, and 18.58 against 14.39 for university E. One should note that in cluster analysis, no extra information was used, whereas in the sten score method, the mean and the variance of the raw score was calculated by the total sample.



**Figure 06.** Distribution of respondents within universities H and E.

Remembering that the high-risk respondents of both universities admit the direct markers in 12-13 cases from 27 in average, it is rather a large number, and it is similar in both universities, in spite of the fact, that the number of respondents in these two cases differs a lot. Hence, cluster analysis can be used to estimate the number of high-risk respondents both for gross and for small samples.

## 6. Findings

The comparison of the two methods of psychological processing of the results of testing young people to determine the group of the greatest risk by propensity to use narcotic substances showed that the cluster analysis method is a mathematically and statistically correct method for this type of research. As advantages and disadvantages of the two methods for processing a large sample of data were compared, the authors suggest that the use of cluster analysis for the processing of statistical information makes it possible to work with both qualitative and qualitative data of various types.

## 7. Conclusion

A comparison of the two methods demonstrated the advantages of applying the cluster analysis method to the processing of statistical data (both qualitative and quantitative) to determine correctly the respondents' belonging to a certain class. The authors suggest that this method can be successfully applied in the processing of sociological data, and other data of socio-economic direction.

## Acknowledgments

The research was completed as part of the project State assignment №27.4344.2017/5.1 "Improving the mechanisms of mentoring under-aged citizens, including those from disadvantaged families, in work placements".

## References

- Anastasi, A. (1988). *Psychological testing*. New York: Macmillan; London: Collier Macmillan.
- Barrett, A. Turner, R. (2006) Family structure and substance use problems in adolescence and early adulthood: examining explanations for the relationship. *Addiction*. 101,109–120. doi: 10.1111/j.1360-0443.2005.01296.x
- Blum, K., Chen A., Giordano, J., Borsten, J., Chen, T., Hauser, M., Barh, D. (2012) The Addictive Brain: All Roads Lead to Dopamine. *Journal of Psychoactive Drugs*. 44, 2. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/02791072.2012.685407>
- Cheverikina, E.A. (2012) Sotsialno-psihologicheskie osobennosti studentov vuzov, sklonnyih k zavisimosti ot psihoaktivnyih veschestv. *Kazanskiy pedagogicheskiy zhurnal*. 5-6. URL: Retrieved from <http://cyberleninka.ru/article/n/sotsialno-psihologicheskie-osobennosti-studentov-vuzov-sklonnyih-k-zavisimosti-ot-psihoaktivnyih-veschestv>
- Dembo, R., Dertke, M., Borders, S., Washburn M, Schmeidler J.(1988). The relationship between physical and sexual abuse and tobacco, alcohol, and illicit drug use among youths in a juvenile detention center. *The International Journal of the Addictions*. 23, 351–378. Retrieved from <https://link.springer.com/article/10.1007/BF02888935>
- Drugs, Brains, and Behavior: The Science of Addiction. (2014). Retrieved from <https://www.drugabuse.gov/publications/drugs-brains-behavior-science-addiction/drug-abuse-addiction>
- Electronic Version: StatSoft, Inc. (2013). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft. Retrieved from: <http://www.statsoft.com/textbook/>.
- Everitt, B., Landau, S., Leese, M., Stahl, D.(2011) *Cluster Analysis*. John Wiley & Sons Ltd, United Kingdom.

- Hayaki, J., Stein, M.D., Lessor, J.A., Herman, D.S., Anderson, B.J. (2005) Adversity among drug users: relationship to impulsivity. *Drug Alcohol Depend.* 78, 65–71. doi: 10.1016/j.drugalcdep.2004.09.002
- Kline, P. (2000). *The handbook of psychological testing*. (2d ed.). London and New York. Routledge, Retrieved from [https://books.google.ru/books?id=lm2RxaKaok8C&printsec=frontcover&hl=ru&source=gbs\\_ViewAPI&redir\\_esc=y#v=onepage&q&f=false](https://books.google.ru/books?id=lm2RxaKaok8C&printsec=frontcover&hl=ru&source=gbs_ViewAPI&redir_esc=y#v=onepage&q&f=false)
- Lukianova, N., Fell E., Sibers J. (2015). Constructing images of the future: investigating the problem. *Vestnik nauki Sibiri.* 2(17).37-46.
- Newcomb, M.D., Bentler, P.M. (1988). Impact of adolescent drug use and social support on problems of young adults: A longitudinal study. *Journal of Abnormal Psychology.* 97, DOI: 10.1037/0021-843X.97.1.64
- Patrick, M. E., Patrick, W., Schoeni, R.F., & Schulenberg, J. E. (2012). Socioeconomic Status and Substance Use Among Young Adults: A Comparison Across Constructs and Drugs. *Journal of Studies on Alcohol and Drugs.* 73(5), 772–782. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410945/>
- Restivo, S., Loughlin, J. (2016) Critical Sociology of Science and Scientific Validity. *Science Communication* . 8(3), 486 – 508. Retrieved from 10.1177/107554708700800304 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410945/>
- Shelehov, I. L., Kornetov, A. N., Grebennikova, E. V. (2016). *Suitsidologiya: istoriya i sovremennyye predstavleniya*. Tomsk : Izdatelstvo Tomskogo gosudarstvennogo pedagogicheskogo universiteta.
- Sinha, R. (2008), Chronic Stress, Drug Use, and Vulnerability to Addiction. *Annals of the New York Academy of Sciences*, 1141, 105–130. doi:10.1196/annals.1441.030
- Spooner, C., Hetherington, K. (2004). *Social determinants of drug use* (Technical Report Number 228). Retrieved from National drug and alcohol research centre, university of new south Wales, Sydney <https://ndarc.med.unsw.edu.au/sites/default/files/ndarc/resources/TR.228.pdf>